B

b Parameter

The b parameter is an item response theory (IRT)–based index of item difficulty. As IRT models have become an increasingly common way of modeling item response data, the b parameter has become a popular way of characterizing the difficulty of an individual item, as well as comparing the relative difficulty levels of different items. This entry addresses the b parameter with regard to different IRT models. Further, it discusses interpreting, estimating, and studying the b parameter.

b Parameter Within Different Item Response Theory Models

The precise interpretation of the b parameter is dependent on the specific IRT model within which it is considered, the most common being the one-parameter logistic (1PL) or Rasch model, the two-parameter logistic (2PL) model, and three-parameter logistic (3PL) model. Under the 1PL model, the b parameter is the single item feature by which items are distinguished in characterizing the likelihood of a correct response. Specifically, the probability of correct response $(X_{ij} = 1)$ by examinee i to item j is given by

$$P(X_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

where θ_j represents an ability-level (or trait-level) parameter of the examinee. An interpretation of the b parameter follows from its being attached to the same metric as that assigned to θ.

Usually this metric is continuous and unbounded; the indeterminacy of the metric is often handled by

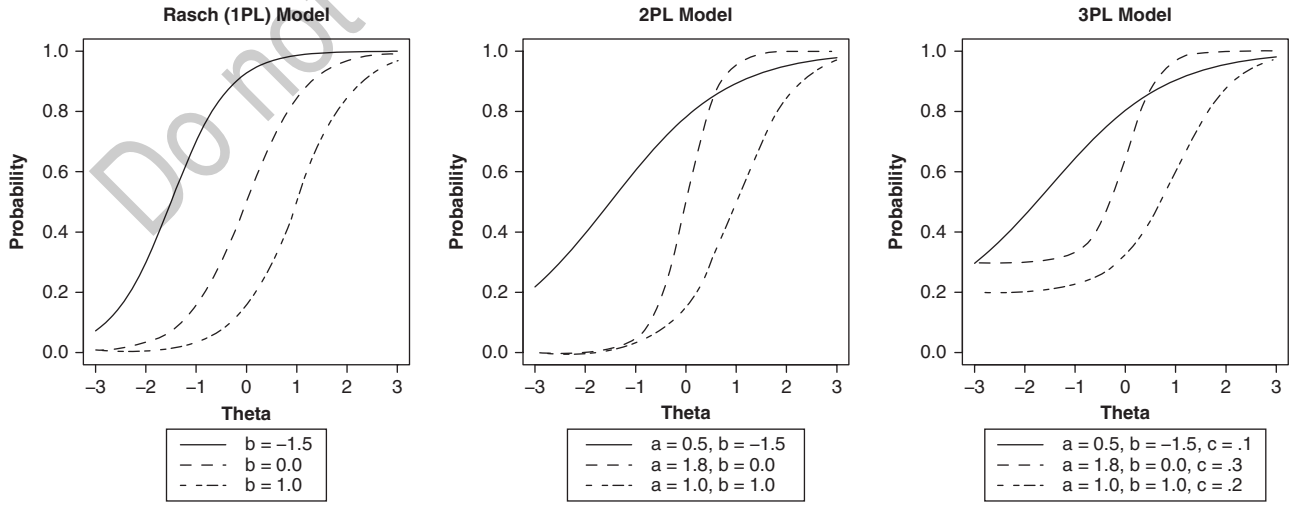


Figure 1 Item Characteristic Curves for Example Items, One-Parameter Logistic (1PL or Rasch), Two-Parameter Logistic (2PL), and Three-Parameter Logistic (3PL) Models

assigning either the mean of θ (across examinees) or b (across items) to 0. Commonly b parameters will assume values between -3 and 3, with more extreme positive values representing more difficult (or infrequently endorsed) items, and more extreme negative values representing easy (or frequently endorsed) items.

The 2PL and 3PL models include additional item parameters that interact with the b parameter in determining the probability of correct response. The 2PL model adds an item discrimination parameter (a_j), so the probability of correct response is

$$P(X_{ij}=1)=\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]},$$

and the 3PL model adds a lower asymptote ("pseudoguessing") parameter, resulting in

$$P(X_{ij}=1)=c_j+(1-c_j)\frac{\exp[a_j(\theta_i-b_j)]}{1+\exp[a_j(\theta_i-b_j)]}.$$

While the same general interpretation of the b parameter as a difficulty parameter still applies under the 2PL and 3PL models, the discrimination and lower asymptote parameters also contribute to the likelihood of a correct response at a given ability level.

Interpretation of the b Parameter

Figure 1 provides an illustration of the b parameter with respect to the 1PL, 2PL, and 3PL models. In this figure, item characteristic curves (ICCs) for three example items are shown with respect to each model. Each curve represents the probability of a correct response as a function of the latent ability level of the examinee. Across all three models, it can be generally seen that as the b parameter increases, the ICC tends to decrease, implying a lower probability of correct response.

In the 1PL and 2PL models, the b parameter has the interpretation of representing the level of the ability or trait at which the respondent has a .50 probability of answering correctly (endorsing the item). For each of the models, the b parameter also identifies the ability level that corresponds to the inflection point of the ICC, and thus the b parameter can be viewed as determining the ability level at which the item is maximally informative. Consequently, the b parameter is a critical element in determining where along the ability continuum an item provides its most effective estimation of ability, and thus the parameter has a strong influence on how items are selected when administered adaptively, such as in a computerized adaptive testing environment.

Under the 1PL model, the b parameter effectively orders all items from easiest to hardest, and this ordering is the same regardless of the examinee ability or trait level. This property is no longer present in the 2PL and 3PL models, as the ICCs of items may cross, implying a different ordering of item difficulties at different ability levels. This property can also be seen in the example items in Figure 1 in which the ICCs cross for the 2PL and 3PL models, but not for the 1PL model. Consequently, while the b parameter remains the key factor in influencing the difficulty of the item, it is not the sole determinant.

An appealing aspect of the b parameter for all IRT models is that its interpretation is invariant with respect to examinee ability or trait level. That is, its value provides a consistent indicator of item difficulty whether considered for a population of high, medium, or low ability. This property is not present in more classical measures of item difficulty (e.g., "proportion correct"), which are influenced not only by the difficulty of the item, but also by the distribution of ability in the population in which they are administered. This invariance property allows the b parameter to play a fundamental role in how important measurement applications, such as item bias (differential item functioning), test equating, and appropriateness measurement, are conducted and evaluated in an IRT framework.

Estimating the b Parameter

The b parameter is often characterized as a structural parameter within an IRT model and as such will generally be estimated in the process of fitting an IRT model to item response data. Various estimation strategies have been proposed and investigated, some being more appropriate for certain model types. Under the 1PL model, conditional maximum likelihood procedures are common. For all three model types, marginal maximum likelihood, joint maximum likelihood, and Bayesian estimation procedures have been developed and are also commonly used.

Studying the b Parameter

The b parameter can also be the focus of further analysis. Models such as the *linear logistic test model* and its variants attempt to relate the b parameter to task components within an item that account for its difficulty. Such models also provide a way in which the b parameter's estimates of items can ultimately be used to validate a test instrument. When the b parameter assumes the value expected given an item's known task components, the parameter provides evidence that the item is functioning as intended by the item writer.

Daniel Bolt

See also Differential Item Functioning; Item Analysis; Item Response Theory; Parameters; Validity of Measurement

Further Readings

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Balanced Incomplete Block Designs

Block designs are useful in experimental design when the experimental material is not uniform but can be divided into homogeneous blocks. Typically, blocks cannot contain all the treatments, so incomplete block designs are used. The condition of balance, if it can be achieved, guarantees that the designs are optimal in minimizing the variance of treatment differences.

This entry provides a survey of balanced incomplete block designs (BIBDs). It discusses necessary conditions on the parameters for the existence of the designs and optimality results. It also provides a brief introduction to constructions of BIBDs using difference families, finite fields, or recursive methods, and some generalizations, including pointers about what to do if no BIBD is available.

Overview

Balanced incomplete block designs, or BIBDs (known to mathematicians as 2-designs), were introduced into statistics by F. Yates at Rothamsted Experimental Station in 1936. So the introductory example provided in this entry will be a hypothetical agricultural experiment.

Suppose we have seven types of fertilizer, $A, B, \ldots, G$, to test. We have land available for the test on seven farms in different regions; each farm can provide three plots for the experiment. The designer's job is to allocate a fertilizer to each of the 21 plots. One possible solution is shown in Table 1.

The farms are represented by columns in the array; the three entries in a column are, in no particular order, the three fertilizers that will be applied to plots on that farm. We refer to the fertilizers as *treatments* and the farms as *blocks*.

Table 1 Possible Solution to Fertilizer Allocation

A	*A*	*A*	*B*	*B*	*C*	*C*
B	*D*	*F*	*D*	*E*	*D*	*E*
C	*E*	*G*	*F*	*G*	*G*	*F*

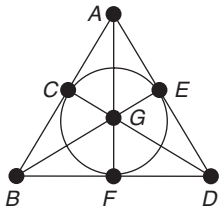


Figure 1 Fano Plane

This design has several good properties:

1. it is *binary*, that is, the same treatment (fertilizer) is not used more than once in a block;
2. it is *equireplicate*, that is, each treatment is used a constant number of times (3 times); and
3. it is *balanced*, that is, each pair of treatments occur together in a block the same number of times (once).

The design can be specified compactly by regarding each block as a set of three treatments (rather than a set of three plots). This is the way that mathematicians regard such a design, but they use the word *points* in place of *treatments*. A mathematician might represent the design by the diagram shown in Figure 1. This is known to finite geometers as the Fano plane.

The experimental units or plots are not easily visible in this picture; they are what geometers refer to as *flags* (incident point-block pairs). But from the list of blocks, as sets of treatments, we can draw a table similar to the one given earlier; the table entries correspond to the plots.

Definition and Properties

A BIBD consists of a set of v treatments, together with a collection of k-element subsets of the treatment set (blocks), with the property that any two treatments are contained in exactly λ blocks. We require that $1 < k < v$. (If $k = 1$, then no comparisons would be possible; if $k = v$, we would have a *complete block design*, with every treatment in every block.) The earlier introductory example has $v = 7$, $k = 3$, $\lambda = 1$.

Theorem 1. Suppose that a BIBD with parameters v, k, λ as above exists. Then

- it is equireplicate, that is, any treatment occurs in a constant number r of blocks;
- $r(k - 1) = (v - 1)\lambda$; and
- the number b of blocks satisfies $bk = vr$.

This result is proved by straightforward double counting. It follows that $r = (v - 1)\lambda/(k - 1)$ and $b = v$

(v – 1) λ/k (k – 1); so a necessary condition for the existence of such a design is that k – 1 divides (v – 1) λ and k divides v (v – 1) λ.

The next result is known as *Fisher's inequality*. It rules out the existence of BIBDs with certain parameter sets.

Theorem 2. The numbers v and b of treatments and blocks in a BIBD satisfy $b \geq v$.

If equality holds, then $k = r$, and any two blocks have exactly λ treatments in common.

A BIBD meeting Fisher's bound is referred to as a *symmetric* BIBD. We can form the dual by simply interchanging the labels "treatment" and "block"; the dual of a symmetric BIBD is again a symmetric BIBD.

Optimality

The job of the designer of an experiment like the one in the introductory example is to obtain the maximum possible information from the given experimental material. Typically, block designs are used when comparing v treatments, and bk experimental units are available, which are divided into b blocks of size k. The assumption is that plots within a block are relatively homogeneous but may differ in systematic but unknown ways from plots in a different block. (In the introductory example, different farms may have very different soil and climatic conditions.) In other words, the parameters v, b, and k are given. We wish to estimate differences between treatments as accurately as possible; this means that we want the variances of the estimators to be as small as possible. However, this is a multidimensional optimization problem; we cannot make all variances small simultaneously.

This has led to the introduction of a number of *optimality parameters* which give an overall summary of the variances. The most important are

1. A-optimality, the average variances of treatment differences;
2. D-optimality, the volume of a confidence ellipsoid; and
3. E-optimality, the largest variance of a treatment difference.

In general, different designs might optimize different parameters here. The importance of BIBDs, when they exist, is that they minimize these three parameters and a number of others. The following was shown by Anant M. Kshirgar (1958) and Jack Kiefer (1975).

Theorem 3. Suppose that a BIBD exists for v treatments in b blocks of size k. Then a design for these parameters minimizes any one of the parameters A, D, and E if and only if it is a BIBD.

So, if a BIBD exists, the designer should use it. If not, the situation is significantly more complicated.

Constructions

Many constructions of BIBDs are known. A few of important ones are mentioned here.

Finite Geometry

These designs are finite analogues of the usual projective and affine (Euclidean) geometry. We work over a finite field with q elements: This is a structure in which the arithmetic operations of addition, subtraction, multiplication, and division (except by zero) are defined and satisfy the usual rules. (It goes back to Évariste Galois, the 19th century mathematician who established that finite fields with q elements exist if and only if q is a prime power.)

The earlier introductory example is of this form. If we represent the treatments $A, \ldots, G$ by 3-dimensional vectors over the *binary field* {0,1}, specifically A = 001, B = 010, C = 011, D = 100, E = 101, F = 110, G = 111, then the blocks are precisely the triples of vectors with sum zero (recall that 1 + 1 = 0 in the binary field).

Geometries of dimension n over a field with q elements give rise to designs, where we take the blocks to be the m-dimensional subspaces with $0 < m < n$. The number of treatments is $(q^n - 1)/(q - 1)$ for projective geometry and q^n for affine geometry.

Difference Families

In the integers mod 7, represented as {0, 1, 2, . . . , 6}, the set {1, 2, 4} has the property that every nonzero element has a unique representation as the difference (mod 7) between two elements of the set:

1 = 2 – 1, 2 = 4 – 2, 3 = 4 – 1, 4 = 1 – 4, 5 = 2 – 4, 6 = 1 – 2.

(The set {1, 2, 4} is called a *difference set*.) It follows that the translates of {1, 2, 4}, namely

124, 235, 346, 450, 561, 602, 013

are the blocks of a symmetric BIBD. This matches the introductory example if we take 1 = A, 2 = B, 3 = D, 4 = C, 5 = F, 6 = G, and 0 = E.

The construction can be generalized: we can replace the integers mod 7 by an arbitrary group; we can replace the uniqueness property by the property that every nonidentity element has a constant number λ of representations, or we can take a difference family consisting of several sets of the same size, such that each nonidentity element has λ representations as the difference between two elements of the same set.

Recursive Constructions

There are many constructions that build larger BIBDs from smaller ones. These often require the existence of auxiliary structures such as Latin squares or transversal designs. These are complicated and so are not described here.

Existence Theorems

BIBDs were first considered in the 19th century, mostly as a branch of recreational mathematics. Designs with $k = 3$ and $\lambda = 1$ are called *Steiner triple systems*, since the question of their existence was posed by the Swiss geometer Jakob Steiner in 1853; unknown to him, the problem had been solved by Thomas P. Kirkman, the rector of a country parish in the north of England, in 1847 and published in the *Ladies' and Gentlemen's Diary*:

Theorem 4. *A BIBD with $k = 3$ and $\lambda = 1$ on v treatments exists if and only if v is congruent to 1 or 3 mod 6.*

The necessity of the condition follows from our necessary conditions: 2 divides $v - 1$, so v is odd; 3 divides $v(v - 1)$, so v is congruent to 0 or 1 mod 3. The sufficiency was proved by Kirkman by construction, partly direct and partly recursive.

In the 20th century, Haim Hanani showed that, for $k = 3$ and any value of λ, the necessary divisibility conditions are sufficient for the existence of a BIBD. Then in 1975, Richard M. Wilson solved the general existence question asymptotically:

Theorem 5. *There is a number $N(k, \lambda)$ such that, if $k - 1$ divides $(v - 1)\lambda$, k divides $v(v - 1)\lambda$, and $v \geq N(k, \lambda)$, then a BIBD with parameters v, k, λ exists.*

Tables of parameters of BIBDs can be found in a number of places. Marshall Hall's book *Combinatorial Theory* tabulates parameter sets by the replication number r for $r \leq 15$, giving either a construction or a reference to a nonexistence proof in each case.

Resolvability

A block design is resolvable if the set of blocks can be partitioned into *resolution classes*, so that the blocks in each class contain each treatment precisely once.

An example, with $v = 9$, $k = 3$ and $\lambda = 1$ (a Steiner triple system) consists of the following blocks:

ABC, DEF, GHI, ADG, BEH, CFI, AEI, BFG, CDH, AFH, BDI, CEG.

From the definition, we see that the blocks in a resolution class are pairwise disjoint and that the stronger necessary condition that k divides v holds. Kirkman, mentioned earlier, posed his celebrated *schoolgirls problem*, asking for a resolvable BIBD with $v = 15$, $k = 3$, and $\lambda = 1$. He solved this problem himself, but the general case with $k = 3$ and $\lambda = 1$ was not solved until the mid-20th century, when Dijen K. Ray-Chaudhuri and Richard M. Wilson showed that resolvable designs exist whenever v is an odd multiple of 3.

The designs from affine spaces described earlier are resolvable, the resolution classes being *parallel classes* of subspaces of the geometry. Indeed, in the example just given, the blocks are lines of the *affine plane* over the 3-element field.

Resolvable designs can be useful in managing an experiment. If we perform the experiment on the resolution classes in order, then losing (say) the last replication class leaves a block design which is still equireplicate, though no longer balanced.

Generalizations

The concept of BIBD has been generalized in various ways. However, the difference in outlook of statisticians and mathematicians means that many of the mathematical generalizations, though applicable in various fields such as information security, are not relevant to experimental design.

One case of interest to both groups is that of *partially balanced incomplete block designs*, or PBIBDs. There is an association scheme on the set of treatments, such that the concurrence of two treatments (the number of blocks containing both) depends only on the associate class containing the pair. This notion was introduced by Raj Chandra Bose and his students in the 1950s and was widely used since it simplified the computational problem of inverting the information matrix of the design; it also includes many examples of great interest to mathematicians, such as generalized polygons.

Our first condition asserted that BIBDs are binary, that is, a treatment occurs at most once in each block. Relaxing this is difficult for the mathematical approach,

since blocks would have to be multisets rather than sets of treatments, but it is quite natural in experimental design. Indeed, the (nonbinary) design for five treatments in seven blocks of three, with blocks *AAB*, *ACD*, *ACE*, *ADE*, *BCD*, *BCE*, *BDE*, is E-optimal for these parameters (it beats any binary design on this criterion). Since 1995, John. P. Morgan and his students have investigated what they call *variance-balanced designs*, which are E-optimal, and have given necessary and sufficient conditions for their existence when $k = 3$.

Peter J. Cameron

See also Block Design; Experimental Design; Factorial Design; Pairwise Comparisons; Randomized Block Design; Replication

Further Readings

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge, UK: Cambridge University Press.

Beth, T., Jungnickel, D., & Lenz, H. (1999). *Design theory* (2 vol.). Cambridge, UK: Cambridge University Press.

Hall, M. Jr. (1998). *Combinatorial theory* (2nd ed.). Hoboken, NJ: Wiley.

Bar Chart

The term *bar chart* refers to a category of diagrams in which values are represented by the height or length of bars, lines, or other symbolic representations. Bar charts are typically used to display variables on a nominal or ordinal scale. Bar charts are a very popular form of information graphics often used in research articles, scientific reports, textbooks, and popular media to visually display relationships and trends in data. However, for this display to be effective, the data must be presented accurately, and the reader must be able to analyze the presentation effectively. This entry provides information on the history of the bar charts, the types of bar charts, and the construction of a bar chart.

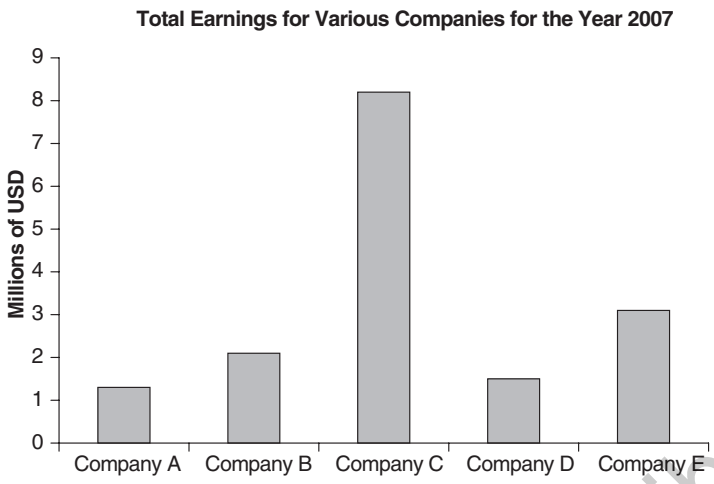
History

The creation of the first bar chart is attributed to William Playfair and appeared in *The Commercial and Political Atlas* in 1786. Playfair's bar graph was an adaptation of Joseph Priestley's time-line charts, which were popular at the time. Ironically, Playfair attributed his creation of the bar graph to a lack of data. In his *Atlas*, Playfair presented 34 plates containing line graphs or surface charts graphically representing the imports and exports from different countries over the years. Since he lacked the necessary time-series data for Scotland, he was forced to graph its trade data for a single year as a series of 34 bars, one for each of the imports and exports of Scotland's 17 trading partners. However, his innovation was largely ignored in Britain for a number of years. Playfair himself attributed little value to his invention, apologizing for what he saw as the limitations of the bar chart. It was not until 1801 and the publication of his *Statistical Breviary* that Playfair recognized the value of his invention. Playfair's invention fared better in Germany and France. In 1811 the German Alexander von Humboldt published adaptations of Playfair's bar graph and pie charts in *Essai Politique sur le Royaume de la Nouvelle Espagne*. In 1821, Jean Baptiste Joseph Fourier adapted the bar chart to create the first graph of cumulative frequency distribution, referred to as an ogive. In 1833, A. M. Guerry used the bar chart to plot crime data, creating the first histogram. Finally, in 1859 Playfair's work began to be accepted in Britain when Stanley Jevons published bar charts in his version of an economic atlas modeled on Playfair's earlier work. Jevons in turn influenced Karl Pearson, commonly considered the "father of modern statistics," who promoted the widespread acceptance of the bar chart and other forms of information graphics.

Types

Although the terms *bar chart* and *bar graph* are now used interchangeably, the term *bar chart* was reserved traditionally for corresponding displays that did not have scales, grid lines, or tick marks. The value each bar represented was instead shown on or adjacent to the data graphic.

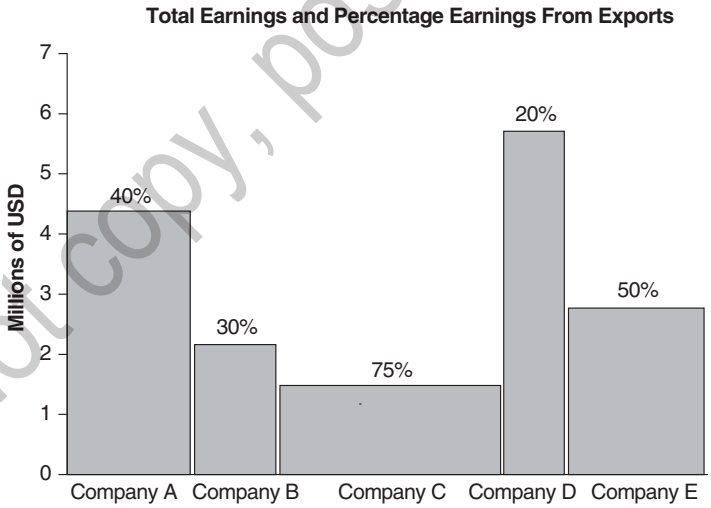
An example bar chart is presented in Figure 1. Bar charts can display data by the use of either horizontal or vertical bars; vertical bar charts are also referred to as *column graphs*. The bars are typically of a uniform width with a uniform space between bars. The end of the bar represents the value of the category being plotted. When there is no space between the bars, the graph is referred to as a *joined bar graph* and is used to emphasize the differences between conditions or discrete categories. When continuous quantitative scales are used on both axes of a joined bar chart, the chart is referred to as a *histogram* and is often used to display the distribution of variables that are of interval or ratio scale. If the widths of the bars are not uniform but are instead used to display some measure or characteristic of the data element represented by the bar, the graph is referred to as an *area bar graph* (see Figure 2). In this graph, the heights of the bars represent the total earnings in U.S. dollars, and the widths of the bars are used to represent the percentage of the earnings coming from exports. The information expressed by the bar width can be displayed by means of a scale on the horizontal axis or by a legend, or, as in this case, the values might be noted directly on



Company A	1.3
Company B	2.1
Company C	8.2
Company D	1.5
Company E	3.1

Figure 1 Simple Bar Chart and Associated Data

Note: USD = U.S. dollars.



	Earnings (USD)	*Percentage of Earnings from Exports*
Company A	4.2	40
Company B	2.1	30
Company C	1.5	75
Company D	5.7	20
Company E	2.9	50

Figure 2 Area Bar Graph and Associated Data

Note: USD = U.S. dollars.

the graph. If both positive and negative values are plotted on the quantitative axis, the graph is called a *deviation graph*. On occasion the bars are replaced with pictures or symbols to make the graph more attractive or to visually represent the data series; these graphs are referred to as *pictographs* or *pictorial bar graphs*.

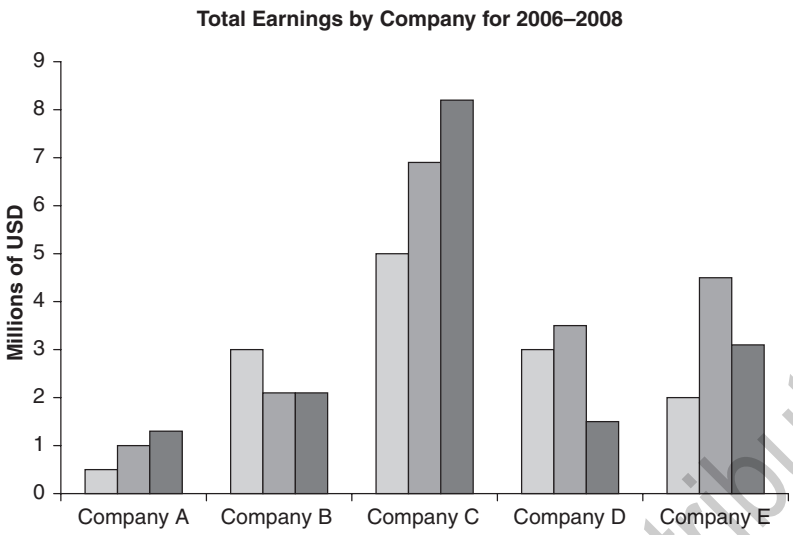
It may on occasion be desirable to display a confidence interval for the values plotted on the graph. In these cases the confidence intervals can be displayed by appending an error bar, or a shaded, striped, or tapered area to the end of the bar, representing the possible values covered in the confidence interval. If the bars are used to represent the range between the upper and lower values of a data series rather than one specific value, the graph is called a *range bar graph*. Typically the lower values are plotted on the left in a horizontal bar chart and on the bottom for a vertical bar chart. A line drawn across the bar can designate additional or inner values, such as a mean or median value. When the five-number summary (the minimum and maximum values, the upper and lower quartiles, and the median) is displayed, the graph is commonly referred to as a *box plot* or a *box-and-whisker diagram*.

A *simple bar graph* allows the display of a single data series, whereas a *grouped* or *clustered bar graph* displays two or more data series on one graph (see Figure 3). In clustered bar graphs, elements of the same category are plotted side by side; different colors, shades, or patterns, explained in a legend, may be used to differentiate the various data series, and the spaces between clusters distinguish the various categories.

While there is no limit to the number of series that can be plotted on the same graph, it is wise to limit the number of series plotted to no more than four in order to keep the graph from becoming confusing. To reduce the size of the graph and to improve readability, the bars for separate categories can be overlapped, but the overlap should be less than 75% to prevent the graph from being mistaken for a stacked bar graph. A *stacked bar graph*, also called a *divided* or *composite bar graph*, has multiple series stacked end to end instead of side by side. This graph displays the relative contribution of the components of a category; a different color, shade, or pattern differentiates each component, as described in a legend. The end of the bar represents the value of the whole category, and the heights of the various data series represent the relative contribution of the components of the category. If the graph represents the separate components' percentage of the whole value rather than the actual values, this graph is commonly referred to as a *100% stacked bar graph*. Lines can be drawn to connect the components of a stacked bar graph to more clearly delineate the relationship between the same components of different categories. A stacked bar graph can also use only one bar to demonstrate the contribution of the components of only one category, condition, or occasion, in which case it functions more like a pie chart. Two data series can also be plotted together in a *paired bar graph*, also referred to as a *sliding bar* or *bilateral bar graph*. This graph differs from a clustered bar graph because rather than being plotted side by side, the values for one data series are plotted with horizontal bars to the left and the values for the other data series are plotted with horizontal bars to the right. The units of measurement and scale intervals for the two data series need not be the same, allowing for a visual display of correlations and other meaningful relationships between the two data series. A paired bar graph can be a variation of either a simple, clustered, or stacked bar graph. A paired bar graph without spaces between the bars is often called a *pyramid graph* or a *two-way histogram*. Another method for comparing two data series is the *difference bar graph*. In this type of bar graph, the bars represent the difference in the values of two data series. For instance, one could compare the performance of two different classes on a series of tests or compare the different performance of males and females on a series of assessments. The direction of the difference can be noted at the ends of bars or by labeling the bars. When comparing multiple factors at two points in time or under two different conditions, one can use a *change bar graph*. The bars in this graph are used to represent the change between the two conditions or times. Since the direction of change is usually important with these types of graphs, a coding system is used to indicate the direction of the change.

Creating an Effective Bar Chart

A well-designed bar chart can effectively communicate a substantial amount of information relatively easily, but a poorly designed graph can create confusion and lead to inaccurate conclusions among readers. Choosing the correct graphing format or technique is the first step in creating an effective graphical presentation of data. Bar charts are best used for making discrete comparisons between several categorical variables because the eye can spot very small differences in relative height. However, a bar chart works best with four to six categories; attempting to display more than six categories on a bar graph can lead to a crowded and confusing graph. Once an appropriate graphing technique has been chosen, it is important to choose the direction and the measurement scale for the primary axes. The decision to present the data in a horizontal or vertical format is largely a matter of personal preference; a vertical presentation, however, is more intui-



	2006	*2007*	*2008*
Company A	0.5	1	1.3
Company B	3	2.1	2.1
Company C	5	6.9	8.2
Company D	3	3.5	1.5
Company E	2	4.5	3.1

Figure 3 Clustered Bar Chart and Associated Data

Note: USD =U.S. dollars.

tive for displaying amount or quantity, and a horizontal presentation makes more sense for displaying distance or time. A horizontal presentation also allows for more space for detailed labeling of the categorical axis. The choice of an appropriate scale is critical for accurate presentation of data in a bar graph. Simple changes in the starting point or the interval of a scale can make the graph look dramatically different and may possibly misrepresent the relationships within the data. The best method for avoiding this problem is to always begin the quantitative scale at 0 and to use a linear rather than a logarithmic scale. However, in cases in which the values to be represented are extremely large, a start value of 0 effectively hides any differences in the data because by necessity the intervals must be extremely wide. In these cases it is possible to maintain smaller intervals while still starting the scale at 0 by the use of a clearly marked scale break. Alternatively, one can highlight the true relationship between the data by starting the scale at 0 and adding an inset of a small section of the larger graph to demonstrate the true relationship. Finally, it is important to make sure the graph and its axes are clearly labeled so that the reader can understand what data are being presented. Modern technology allows the addition of many superfluous graphical elements to enhance the basic graph design. Although the addition of these elements is a matter of personal choice, it is important to remember that the primary aim of data graphics is to display data accurately and clearly. If the additional elements detract from this clarity of presentation, they should be avoided.

Teresa P. Clark and Sara E. Bolt

See also Box-and-Whisker Plot; Distribution; Graphical Display of Data; Histogram; Pie Chart

Further Readings

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229,* 828–833.

Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference.* New York: Oxford University Press.

Playfair, W. (1786). *The commercial and political atlas.* London: Corry.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14,* 47–69.

Spence, I. (2000). The invention and use of statistical charts. *Journal de la Société Française de Statistique, 141,* 77–81.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics.

Wainer, H. (1996). Depicting error. *American Statistician, 50,* 101–111.

Bartlett's Test

The assumption of equal variances across treatment groups may cause serious problems if violated in one-way analysis of variance models. A common test for homogeneity of variances is Bartlett's test. This statistical test checks whether the variances from different groups (or samples) are equal.

Suppose that there are r treatment groups and we want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$$

versus

$$H_1 : \sigma_m^2 \neq \sigma_k^2 \text{ for some } m \neq k.$$

In this context, we assume that we have independently chosen random samples of size $n_i, i = 1,\ldots,r$ from each of the r independent populations. Let $X_{ij} \sim N\left(\mu_i, \sigma_i^2\right)$ be independently distributed with a normal distribution having mean μ_i and variance σ_i^2 for each $j = 1,\ldots,n_i$ and each $i = 1,\ldots,r$. Let $\bar{X}_i$ be the sample mean and S_i^2 the sample variance of the sample taken from the ith group or population. The uniformly most powerful unbiased parametric test of size α for testing for equality of variances among r populations is known as Bartlett's test, and Bartlett's test statistic is given by

$$\ell_1^* = \frac{\prod_{i=1}^{r} (S_i^2)^{w_i}}{\sum_{i=1}^{r} w_i S_i^2},$$

where $w_i = (n_i - 1)/(N - r)$ is known as the weight for the ith group and $N = \sum_{i=1}^{r} n_i$ is the sum of the individual sample sizes. In the equireplicate case (i.e., $n_1 = \cdots = n_r = n$), the weights are equal, and $w_i = 1/r$ for each $i = 1,\ldots,r$. The test statistic is the ratio of the *weighted geometric mean* of the group sample variances to their *weighted arithmetic mean*. The values of the test statistic are bounded as $0 \leq \ell_1^* \leq 1$ by Jensen's inequality. Large values of $0 \leq \ell_1^* \leq 1$ (i.e., values near 1) indicate agreement with the null hypothesis, whereas small values indicate disagreement with the null. The terminology ℓ_1^* is used to indicate that Bartlett's test is based on M. S. Bartlett's modification of the likelihood ratio test, wherein he replaced the sample sizes n_i with their corresponding degrees of freedom, $n_i - 1$. Bartlett did so to make the test unbiased. In the equireplicate case, Bartlett's test and the likelihood ratio test result in the same test statistic and same critical region.

The distribution of ℓ_1^* is complex even when the null hypothesis is true. R. E. Glaser showed that the distribution of ℓ_1^* could be expressed as a product of independently distributed beta random variables. In doing so he renewed much interest in the exact distribution of Bartlett's test. We reject H_0 provided $\ell_1^* \leq b_\alpha(n_1,\ldots,n_r)$ where $\Pr\left(\ell_1^* < b_\alpha(n_1,\ldots,n_r)\right) = \alpha$ when H_0 is true. The Bartlett critical value $b_\alpha(n_1,\ldots,n_r)$ is indexed by level of significance and the individual sample sizes. The critical values were first tabled in the equireplicate case, and the critical value was simplified to $b_\alpha(n,\ldots,n) = b_\alpha(n)$. Tabulating critical values with unequal sample sizes becomes counterproductive because of possible combinations of groups, sample sizes, and levels of significance.

Example

Consider an experiment in which lead levels are measured at five different sites. The data in Table 1 come from Paul Berthouex and Linfield Brown:

Table 1 Ten Measurements of Lead Concentration (mG=L) Measured on Waste Water Specimens

	Measurement No.									
Lab	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	3.4	3.0	3.4	5.0	5.1	5.5	5.4	4.2	3.8	4.2
2	4.5	3.7	3.8	3.9	4.3	3.9	4.1	4.0	3.0	4.5
3	5.3	4.7	3.6	5.0	3.6	4.5	4.6	5.3	3.9	4.1
4	3.2	3.4	3.1	3.0	3.9	2.0	1.9	2.7	3.8	4.2
5	3.3	2.4	2.7	3.2	3.3	2.9	4.4	3.4	4.8	3.0

Source: Berthouex, P. M., & Brown, L. C. (2002). *Statistics for environmental engineers* (2nd ed., p. 170). Boca Raton, FL: Lewis.

From these data one can compute the sample variances and weights, which are:

Labs	*Weight*	*Variance*
1	0.2	0.81778
2	0.2	0.19344
3	0.2	0.41156
4	0.2	0.58400
5	0.2	0.54267

By substituting these values into the formula for ℓ_1^*, we obtain

$$\ell_1^* = \frac{0.46016}{0.509889} = 0.90248$$

Critical values, $b_\alpha(\alpha, n)$ of Bartlett's test are tabled for cases in which the sample sizes are equal and $\alpha = .05$.

Table 2 Table of Bartlett's Critical Values

	Number of Populations, r								
n	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
3	.3123	.3058	.3173	.3299	.	.	.	.	.
4	.4780	.4699	.4803	.4921	.5028	.5122	.5204	.5277	.5341
5	.5845	.5762	.5850	.5952	.6045	.6126	.6197	.6260	.6315
6	.6563	.6483	.6559	.6646	.6727	.6798	.6860	.6914	.6961
7	.7075	.7000	.7065	.7142	.7213	.7275	.7329	.7376	.7418
8	.7456	.7387	.7444	.7512	.7574	.7629	.7677	.7719	.7757
9	.7751	.7686	.7737	.7798	.7854	.7903	.7946	.7984	.8017
10	.7984	.7924	.7970	.8025	.8076	.8121	.8160	.8194	.8224
11	.8175	.8118	.8160	.8210	.8257	.8298	.8333	.8365	.8392
12	.8332	.8280	.8317	.8364	.8407	.8444	.8477	.8506	.8531
13	.8465	.8415	.8450	.8493	.8533	.8568	.8598	.8625	.8648
14	.8578	.8532	.8564	.8604	.8641	.8673	.8701	.8726	.8748
15	.8676	.8632	.8662	.8699	.8734	.8764	.8790	.8814	.8834
16	.8761	.8719	.8747	.8782	.8815	.8843	.8868	.8890	.8909
17	.8836	.8796	.8823	.8856	.8886	.8913	.8936	.8957	.8975
18	.8902	.8865	.8890	.8921	.8949	.8975	.8997	.9016	.9033
19	.8961	.8926	.8949	.8979	.9006	.9030	.9051	.9069	.9086
20	.9015	.8980	.9003	.9031	.9057	.9080	.9100	.9117	.9132
21	.9063	.9030	.9051	.9078	.9103	.9124	.9143	.9160	.9175
22	.9106	.9075	.9095	.9120	.9144	.9165	.9183	.9199	.9213
23	.9146	.9116	.9135	.9159	.9182	.9202	.9219	.9235	.9248
24	.9182	.9153	.9172	.9195	.9217	.9236	.9253	.9267	.9280
25	.9216	.9187	.9205	.9228	.9249	.9267	.9283	.9297	.9309
26	.9246	.9219	.9236	.9258	.9278	.9296	.9311	.9325	.9336
27	.9275	.9249	.9265	.9286	.9305	.9322	.9337	.9350	.9361
28	.9301	.9276	.9292	.9312	.9330	.9347	.9361	.9374	.9385
29	.9326	.9301	.9316	.9336	.9354	.9370	.9383	.9396	.9406
30	.9348	.9325	.9340	.9358	.9376	.9391	.9404	.9416	.9426
40	.9513	.9495	.9506	.9520	.9533	.9545	.9555	.9564	.9572
50	.9612	.9597	.9606	.9617	.9628	.9637	.9645	.9652	.9658
60	.9677	.9665	.9672	.9681	.9690	.9698	.9705	.9710	.9716
80	.9758	.9749	.9754	.9761	.9768	.9774	.9779	.9783	.9787
100	.9807	.9799	.9804	.9809	.9815	.9819	.9823	.9827	.9830

Source: Dyer, D., & Keating, J. P. (1980). On the determination of critical values for Bartlett's test. *Journal of the American Statistical Association*, *75*, 313–319. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 1980 by the American Statistical Association. All rights reserved.

Note: The table shows the critical values used in Bartlett's test of equal variance at the 5% level of significance.

These values are given in D. Dyer and Jerome Keating for various values of r, the number of groups, and n, the common sample size (see Table 2). Works by Glaser, M. T. Chao, and Glaser, and S. B. Nandi provide tables of exact critical values of Bartlett's test. The most extensive set is contained in Dyer and Keating. Extensions (for larger numbers of groups) to the table of critical values can be found in Keating, Glaser, and N. S. Ketchum.

Approximation

In the event that the sample sizes are not equal, one can use the *Dyer–Keating approximation* to the critical values:

$$b_\alpha(a;n_1,\ldots,n_a) \doteq \sum_{i=1}^{a} \frac{n_i}{N} \times b_\alpha(a,n_i).$$

So for the lead levels, we have the following values: $b_{0.05}(5;10) \doteq 0.8025$. At the 5% level of significance, there is not enough evidence to reject the null hypothesis of equal variances.

Because of the complexity of the distribution of ℓ_1^*, Bartlett's test originally employed an approximation. Bartlett proved that

$$\left[-\ln\left(\ell_1^*\right)\right] / c \sim \chi^2(r-1),$$

where

$$c = \frac{1 + \left[\frac{1}{3(r-1)}\right]\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{N-r}}{N-r}.$$

The approximation works poorly for small sample sizes. This approximation is more accurate as sample sizes increase, and it is recommended that $\min(n_i) \geq 3$ and that most $n_i > 5$.

Assumptions

Bartlett's test statistic is quite sensitive to nonnormality. In fact, W. J. Conover, M. E. Johnson, and M. M. Johnson echo the results of G. E. P. Box that Bartlett's test is very sensitive to samples that exhibit nonnormal kurtosis. They recommend that Bartlett's test be used only when the data conform to normality. Prior to using Bartlett's test, it is recommended that one test for normality using an appropriate test such as the Shapiro–Wilk W test. In the event that the normality assumption is violated, it is recommended that one test equality of variances using Howard Levene's test.

Mark T. Leung and Jerome P. Keating

See also Critical Value; Likelihood Ratio Statistic; Normality Assumption; Parametric Statistics; Variance

Further Readings

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A*, *160*, 268–282.

Berthouex, P. M., & Brown, L. C. (2002). *Statistics for environmental engineers* (2nd ed.). Boca Raton, FL: Lewis.

Box, G. E. P. (1953). Nonnormality and tests on variances. *Biometrika*, *40*, 318–335.

Chao, M. T., & Glaser, R. E. (1978). The exact distribution of Bartlett's test statistic for homogeneity of variances with unequal sample sizes. *Journal of the American Statistical Association*, *73*, 422–426.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests of homogeneity of variances with applications to the Outer Continental Shelf bidding data. *Technometrics, 23*, 351–361.

Dyer, D., & Keating, J. P. (1980). On the determination of critical values for Bartlett's test. *Journal of the American Statistical Association*, *75*, 313–319.

Glaser, R. E. (1976). Exact critical values for Bartlett's test for homogeneity of variances. *Journal of the American Statistical Association*, *71*, 488–490.

Keating, J. P., Glaser, R. E., & Ketchum, N. S. (1990). Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics*, *32*, 67–82.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Palo Alto, CA: Stanford University Press.

Madansky, A. (1989). *Prescriptions for working statisticians*. New York: Springer-Verlag.

Nandi, S. B. (1980). On the exact distribution of a normalized ratio of weighted geometric mean to the unweighted arithmetic mean in samples from gamma distributions. *Journal of the American Statistical Association*, *75*, 217–220.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Barycentric Discriminant Analysis

Barycentric discriminant analysis (BADA) generalizes discriminant analysis, and like discriminant analysis, it is performed when measurements made on some observations are combined to assign these observations, or *new* observations, to a priori defined categories. For example, BADA can be used (a) to assign people to a given diagnostic group (e.g., patients with Alzheimer's disease, patients with other dementia, or people aging without dementia) on the basis of brain imaging data or psychological tests (here the a priori categories are the clinical groups), (b) to assign wines to a region of production on the basis of

several physical and chemical measurements (here the a priori categories are the regions of production), (c) to use brain scans taken on a given participant to determine what type of object (e.g., a face, a cat, a chair) was watched by the participant when the scans were taken (here the a priori categories are the types of object), or (d) to use DNA measurements to predict whether a person is at risk for a given health problem (here the a priori categories are the types of health problem).

BADA is more general than standard discriminant analysis because it can be used in cases for which discriminant analysis cannot be used. This is the case, for example, when there are more variables than observations, when the predictors are colinear, or when the measurements are categorical.

BADA is a class of methods that all rely on the same principle: Each category of interest is represented by the *barycenter* of its observations (i.e., the weighted average of the observations of a given category; the barycenter is also called the *center of gravity* or *center of mass*), and a generalized principal components analysis (GPCA) is performed on the category by variable matrix. This analysis gives a set of discriminant factor scores for the categories and another set of factor scores for the variables. The original observations are then projected onto the category factor space, providing a set of factor scores for the observations. The distance of each observation to the set of categories is computed from the factor scores, and each observation is assigned to the closest category. The a priori and a posteriori category assignments are compared to assess the quality of the discriminant procedure. The prediction for the observations that were used to compute the barycenters is called the *fixed-effect prediction*. The fixed-effect performance is evaluated by counting the number of correct and incorrect assignments and storing these numbers in a confusion matrix. Another index of the performance of the fixed-effect model—equivalent to a squared coefficient of correlation—is the ratio of the category variance to the sum of the category variance and the variance of the observations within each category. This coefficient is denoted R^2 and is interpreted as the proportion of variance of the observations explained by the categories or as the proportion of the variance explained by the discriminant model. The performance of the fixed-effect model can also be represented graphically as a *tolerance ellipsoid* that encompasses a given proportion (say 95%) of the observations. The overlap between the tolerance ellipsoids of two categories is proportional to the number of misclassifications between these two categories.

New observations can also be projected onto the discriminant factor space, and they can be assigned to the closest category. When the actual assignment of these observations is not known, the model can be used to *predict* category membership. The model is then called a *random* model (as opposed to the fixed model). An obvious problem, then, is to evaluate the quality of the prediction for new observations. Ideally, the performance of the random-effect model is evaluated by counting the number of correct and incorrect classifications for new observations and computing a confusion matrix based on these new observations. However, it is not always practical or even feasible to obtain new observations, and therefore the random-effect performance is often evaluated using computational cross-validation techniques such as the *leave one out* (LOO) or the *bootstrap*. For example, an LOO approach can be used by which each observation is taken out of the set, in turn, and predicted from the model built on all the other observations. The predicted observations are then projected in the space of the fixed-effect discriminant scores. This can also be represented graphically as a *prediction ellipsoid*. A prediction ellipsoid encompasses a given proportion (say 95%) of the new observations. The overlap between the prediction ellipsoids of two categories is proportional to the number of misclassifications of new observations between these two categories.

The stability of the discriminant model can be assessed by a cross-validation model such as the bootstrap. In this procedure, multiple sets of observations are generated by sampling with replacement from the original set of observations, and the category barycenters are computed from each of these sets. These barycenters are then projected onto the discriminant factor scores. The variability of the barycenters can be represented graphically as a *confidence* ellipsoid that encompasses a given proportion (say 95%) of the barycenters. When the confidence intervals of two categories do not overlap, these two categories are significantly different.

In summary, BADA is a GPCA performed on the category barycenters. GPCA encompasses various techniques, such as correspondence analysis, biplot, Hellinger distance analysis, discriminant analysis, and canonical variate analysis. For each specific type of GPCA, there is a corresponding version of BADA. For example, when the GPCA is correspondence analysis, this gives the most well-known version of BADA: discriminant correspondence analysis. Because BADA is based on GPCA, it can also analyze data tables obtained by the concatenation of blocks (i.e., subtables). In this case, the importance (often called the *contribution*) of each block to the overall discrimination can also be evaluated and represented as a graph.

Hervé Abdi and Lynne J. Williams

See also Bootstrapping; Canonical Correlation Analysis; Correspondence Analysis; Jackknife; Matrix Algebra; Predictive Discriminant Analysis; Principal Components Analysis

Further Readings

Abdi, H. (2007). Discriminant correspondence analysis. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 270–275). Thousand Oaks, CA: Sage.

Abdi, H., Williams, L. J., Beaton, D., Posamentier, M., Harris, T. S., Krishnan, A., & Devous, M. D. (2012). Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, fronto-temporal dementia, and elderly controls: A multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer's Disease*, *31*, s189–s201. doi:10.3233/JAD-2012-112111.

Beaton, D., Dunlop, J., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Partial Least squares-correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, *21*, 621–651. doi:10.1037/met0000053.

Basket Trials Design

Basket trials test a therapeutic intervention for several different disease *indications* simultaneously in the same clinical trial. Indications are grouped together in a basket because they share a molecular marker or a disease mechanism believed to be predictive of clinical benefit from the therapeutic intervention under study. Basket trials are a type of master protocol, a class of clinical designs that investigate several hypotheses concurrently. Master protocols include basket trials, whereby one therapy is tested in multiple indications; umbrella trials, whereby multiple therapies are tested in one indication; and platform trials, whereby multiple therapies enter and exit an ongoing study in an assembly-line fashion.

This entry summarizes the motivation for using master protocols, metrics for basket trial performance evaluation, advantages and limitations of basket trials, and types of basket trials.

Motivation

Master protocols address the high cost and long time required for clinical drug development, the process of testing safety and effectiveness for therapies, to facilitate approval by national health authorities for patient use. The total cost of developing a new therapy, including basic research, medicinal chemistry, animal efficacy and toxicology, and clinical development, is estimated at approximately US$ 1 billion, including the cost of failed attempts. The resulting high cost of therapy means patients often cannot afford needed medications and must choose between medications and other basic necessities. The total time for development is a decade or longer. During this waiting period, patients facing debilitating and/or life-threatening diseases may have their medical needs unmet. By investigating several hypotheses at once, master protocols aim to reduce these severe cost and time issues. Alternatively, the savings may be utilized to more thoroughly investigate optimal dosing, scheduling, and matching of therapies to patient subpopulations, rather than to reduce cost and development time.

Performance Evaluation

The performance metrics discussed in this section provide a framework for comparing possible basket trials designs. Because basket trials are complex, it is usually not possible to write mathematical formulas for their performance. The performance is evaluated by computer simulation, in which the trial is simulated numerous times with chance variation, and the results observed.

False Positive Rate

For conventional clinical trials, the definition of the false positive rate is relatively simple. If investigators execute the clinical trial for a therapy that is actually ineffective in the proposed indication, the false positive rate is the proportion of times the trial will falsely reach the conclusion that the therapy is effective, due to the play of chance. However, for basket trials, the situation is more complicated, in that the therapy is being tested in multiple indications at once. A *positive basket* may be defined as a basket for which the drug is actually effective in one or more indications, and a *negative basket* as one in which the drug is actually ineffective in all indications. The *false positive rate by basket* is then the proportion of times among evaluations of a negative basket that the trial falsely concludes that it is positive. A more stringent approach is the *false positive rate by indication*. In this approach, each indication is defined as positive or negative by whether the therapy is actually effective or ineffective in that indication. The *false positive rate by indication* is the probability that a negative indication will be falsely considered positive by the study.

An important controversy in the field is whether (and when) control of false positive rates by indication is required in basket trials. The indications are subgroups of the overall basket. In a conventional clinical trial, control of the false positive rate by subgroups is not required. But basket trials feature nontraditional groupings, in which molecular characteristics that normally define subgroups define the overall group, and diseases which are traditionally considered different are grouped as one.

False Negative Rate

In analogy with the false positive rate, the *false negative rate by basket* is the proportion of times in

evaluating a positive basket that the trial design will, by the play of chance, falsely conclude it is negative. The more stringent *false negative rate by indication* is the proportion of times a positive individual indication will be falsely considered to be negative by the study. *Power* is one minus the false negative rate.

Efficiency

Efficiency is an important metric for any master protocol. It normalizes (divides) a results metric by a cost metric, such as financial cost or number of trial participants. Although power is a popular performance indicator, it is possible to increase the power of a study to any desired level by increasing the sample size. Yet, for each constant increment in power, a progressively higher cost must be paid in sample size, and the efficiency decreases due to the diminishing returns.

The choice of results and cost metrics is varied and depends on the situation. Moreover, the concept of efficiency applies not only to individual therapies and indications but to a portfolio of many therapies and indications. In some instances, with finite resources, it may be wise to invest less in one trial to preserve resources for another.

Advantages

Basket trials group multiple indications together as one. Under ideal conditions, a basket trial consisting of k indications can offer a nearly k-fold increase in efficiency. This increase in efficiency is potentially greater than other master protocols. Even if control of the false positive rate by indication is required, a 40% increase in efficiency may be achieved with optimization.

Basket trials typically study only one therapy, and therefore they can be conducted by a single sponsor, avoiding the complex negotiations between multiple sponsors required for other master protocols.

Limitations

Basket trials assume that the indications within the basket will have similar responses to the therapy, due to the shared marker or mechanism. However, this is not always true. The drug may work in melanoma with a given mutation but not in colon cancer with the same mutation. This heterogeneity can result in false positives and false negatives when data are pooled across indications.

Much of basket trial methodology is aimed at minimizing heterogeneity. In *adaptive basket designs*, interim study data are used to minimize heterogeneity. The final result is judged on a grouping that is more likely to be homogeneous; other indications may be discarded.

Types of Basket Trials

Basket trials may be randomized controlled or single arm. Randomized controlled trials are more conclusive in that they minimize biased selection of participants as well as biases due to improvements in supportive care and diagnostic technologies which affect comparisons of single-arm results to historical controls.

Basket trials also differ in the degree to which data from different indications are pooled. In early basket trials, the data were not pooled, and thus the studies were primarily a useful administrative strategy for rare diseases. Most current designs either fully pool the data or utilize partial pooling calibrated by the degree of similarity in the results (*information borrowing*).

Basket trials have been applied mostly to the *exploratory phase* of clinical development, which involves early safety and efficacy data and dose/schedule optimization. In contrast, applications in the *confirmatory phase* of development, where large studies are usually performed to obtain statistical proof of efficacy, have been limited to special cases supported by extraordinary scientific evidence, and/or unusual efficacy in conditions with limited or no other medical options. In these settings, single-arm designs, small sample sizes, and short-term binary end points with uncertain relationship to definitive end points like survival are appropriate even in the confirmatory phase, facilitating application of basket designs.

Cost and time savings would be greater if basket trials could be applied generally to the resource-intensive confirmatory space. A randomized controlled confirmatory basket trial design has been proposed, and the false positive rate can be strictly controlled as required in the confirmatory space. The design can be further modified to achieve control of the false positive rate by indication. Nonetheless, whether the design can be confirmatory despite possible differences in the control arms and the natural histories of the indications may depend on the strength of scientific and medical evidence for pooling.

Robert A. Beckman

See also Adaptive Designs in Clinical Trials; Multiplicity Problem; Type I Error; Type II Error; Umbrella Trials Designs

Further Readings

Antonijevic, Z., & Beckman, R. A. (Eds.). (2018). *Platform trials in drug development: Umbrella trials and basket trials*. Boca Raton, FL: CRC Press.

Beckman, R. A., Antonijevic, Z., Kalamegham, R., & Chen, C. (2016). Adaptive design for a confirmatory basket trial in multiple tumor types based on a putative predictive

biomarker. *Clinical Pharmacology and Therapeutics*, *100*, 617–625. doi:10.1002/cpt.446.

Collignon, O., Gartner C., Haidich, A. B., Hemmings, R. J., Hofner, B., Petavy, F., . . . Schiel, A. (2020). Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials. *Clinical Pharmacology and Therapeutics*, *107*, 1059–1067. doi:10.1002/cpt.1804.

Cunanan, K. M., Gönen, M., Shen, R., Hyman, D. M., Riely, D. J., Begg, C. B., & Iasonos, A. (2017). Basket trials in oncology: A trade-off between complexity and efficiency. *Journal of Clinical Oncology*, *35*, 271–273. doi:10.1200/JCO.2016.69.9751.

Woodcock, J., & LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, *377*, 62–70. doi:10.1056/NEJMra1510062.

Bayes's Theorem

Bayes's theorem is a simple mathematical formula used for calculating conditional probabilities. It figures prominently in subjectivist or Bayesian approaches to statistics, epistemology, and inductive logic. Subjectivists, who maintain that rational belief is governed by the laws of probability, lean heavily on conditional probabilities in their theories of evidence and their models of empirical learning. Bayes's theorem is central to these paradigms because it simplifies the calculation of conditional probabilities and clarifies significant features of the subjectivist position.

This entry begins with a brief history of Thomas Bayes and the publication of his theorem. Next, the entry focuses on probability and its role in Bayes's theorem. Last, the entry explores modern applications of Bayes's theorem.

History

Thomas Bayes was born in 1702, probably in London, England. Others have suggested the place of his birth to be Hertfordshire. He was the eldest of six children of Joshua and Ann Carpenter Bayes. His father was a nonconformist minister, one of the first seven in England. Information on Bayes's childhood is scarce. Some sources state that he was privately educated, and others state he received a liberal education to prepare for the ministry. After assisting his father for many years, he spent his adult life as a Presbyterian minister at the chapel in Tunbridge Wells. In 1742, Bayes was elected as a fellow by the Royal Society of London. He retired in 1752 and remained in Tunbridge Wells until his death in April 1761.

Throughout his life he wrote very little, and only two of his works are known to have been published. These two essays are *Divine Benevolence*, published in 1731, and *Introduction to the Doctrine of Fluxions*, published in 1736. He was known as a mathematician not for these essays but for two other papers he had written but never published. His studies focused in the areas of probability and statistics. His posthumously published article now known by the title "An Essay Towards Solving a Problem in the Doctrine of Chances" developed the idea of inverse probability, which later became associated with his name as Bayes's theorem. Inverse probability was so called because it involves inferring backward from the data to the parameter (i.e., from the effect to the cause). Initially, Bayes's ideas attracted little attention. It was not until after the French mathematician Pierre-Simon Laplace published his paper "Mémoire sur la Probabilité des Causes par les Évènements" in 1774 that Bayes's ideas gained wider attention. Laplace extended the use of inverse probability to a variety of distributions and introduced the notion of "indifference" as a means of specifying prior distributions in the absence of prior knowledge. Inverse probability became during the 19th century the most commonly used method for making statistical inferences. Some of the more famous examples of the use of inverse probability to draw inferences during this period include estimation of the mass of Saturn, the probability of the birth of a boy at different locations, the utility of antiseptics, and the accuracy of judicial decisions.

In the latter half of the 19th century, authorities such as Siméon-Denis Poisson, Bernard Bolzano, Robert Leslie Ellis, Jakob Friedrich Fries, John Stuart Mill, and A. A. Cournot began to make distinctions between probabilities about things and probabilities involving our beliefs about things. Some of these authors attached the terms *objective* and *subjective* to the two types of probability. Toward the end of the century, Karl Pearson, in his *Grammar of Science*, argued for using experience to determine prior distributions, an approach that eventually evolved into what is now known as *empirical Bayes*. The Bayesian idea of inverse probability was also being challenged toward the end of the 19th century, with the criticism focusing on the use of uniform or "indifference" prior distributions to express a lack of prior knowledge.

The criticism of Bayesian ideas spurred research into statistical methods that did not rely on prior knowledge and the choice of prior distributions. In 1922, Ronald Alymer Fisher's paper "On the Mathematical Foundations of Theoretical Statistics," which introduced the idea of likelihood and maximum likelihood estimates, revolutionized modern statistical thinking. Jerzy Neyman and Egon Pearson extended Fisher's work by adding the ideas of hypothesis testing and confidence intervals. Eventually the collective work of Fisher, Neyman, and Pearson became known as *frequentist* methods. From the 1920s to the 1950s, frequentist methods displaced inverse probability as the primary methods used by researchers to make statistical inferences.

Interest in using Bayesian methods for statistical inference revived in the 1950s, inspired by Leonard Jimmie Savage's 1954 book *The Foundations of Statistics*. Savage's work built on previous work of several earlier authors exploring the idea of subjective probability, in particular the work of Bruno de Finetti. It was during this time that the terms *Bayesian* and *frequentist* began to be used to refer to the two statistical inference camps. The number of papers and authors using Bayesian statistics continued to grow in the 1960s. Examples of Bayesian research from this period include an investigation by Frederick Mosteller and David Wallace into the authorship of several of the Federalist papers and the use of Bayesian methods to estimate the parameters of time-series models. The introduction of Monte Carlo Markov chain (MCMC) methods to the Bayesian world in the late 1980s made computations that were impractical or impossible earlier realistic and relatively easy. The result has been a resurgence of interest in the use of Bayesian methods to draw statistical inferences.

Publishing of Bayes's Theorem

Bayes never published his mathematical papers, and therein lies a mystery. Some suggest his theological concerns with modesty might have played a role in his decision. However, after Bayes's death, his family asked Richard Price to examine Bayes's work. Price was responsible for the communication of Bayes's essay on probability and chance to the Royal Society. Although Price was making Bayes's work known, he was occasionally mistaken for the author of the essays and for a time received credit for them. In fact, Price only added introductions and appendixes to works he had published for Bayes, although he would eventually write a follow-up paper to Bayes's work.

The present form of Bayes's theorem was actually derived not by Bayes but by Laplace. Laplace used the information provided by Bayes to construct the theorem in 1774. Only in later papers did Laplace acknowledge Bayes's work.

Inspiration of Bayes's Theorem

In "An Essay Towards Solving a Problem in the Doctrine of Chances," Bayes posed a problem to be solved: "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." Bayes's reasoning began with the idea of conditional probability:

If $P(B) > 0$, the conditional probability of A given B, denoted by $P(A \mid B)$, is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ or } \frac{P(AB)}{P(B)}.$$

Bayes's main focus then became defining $P(B \mid A)$ in terms of $P(A \mid B)$.

A key component that Bayes needed was the *law of total probability*. Sometimes it is not possible to calculate the probability of the occurrence of an event A. However, it is possible to find $P(A \mid B)$ and $P(A \mid B^c)$ for some event B where B^c is the complement of B. The weighted average, $P(A)$, of the probability of A given that B has occurred and the probability of A given that B has not occurred can be defined as follows:

Let B be an event with $P(B) > 0$ and $P(B^c) > 0$. Then for any event A,

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$

If there are k events, $B_1, \ldots, B_k$, that form a partition of the sample space, and A is another event in the sample space, then the events $B_1A, B_2A, \ldots, B_kA$ will form a partition of A. Thus, the law of total probability can be extended as follows:

Let B_j be an event with $P(Bj) > 0$ for $j = 1, \ldots, k$. Then for any event A,

$$P(A) = \sum_{j=1}^{k} P(B_j)P(A \mid B_j).$$

These basic rules of probability served as the inspiration for Bayes's theorem.

Bayes's Theorem

Bayes's theorem allows for a reduction in uncertainty by considering events that have occurred. The theorem is applicable as long as the probability of the more recent event (given an earlier event) is known. With this theorem, one can find the probability of the earlier event, given the more recent event that has occurred. The earlier event is known as the *prior* probability. The primary focus is on the probability of the earlier event given the more recent event that has occurred (called the *posterior* probability). The theorem can be described in the following manner:

Let B_j be an event with $P(B_j) > 0$ for $j = 1, \ldots, k$ and forming a partition of the sample space. Furthermore, let A be an event such that $P(A) > 0$. Then for $i = 1, \ldots, k$,

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j)P(A \mid B_j)}.$$

$P(B_i)$ is the *prior* probability and the probability of the earlier event. $P(A \mid B_i)$ is the probability of the more recent event given the prior has occurred and is referred

to as the *likelihood*. $P(B_i \mid A)$ is what one is solving for and is the probability of the earlier event given that the recent event has occurred (the posterior probability). There is also a version of Bayes's theorem based on a secondary event C:

$$P(B_i \mid AC) = \frac{P(B_i \mid C)P(A \mid B_iC)}{\sum_{j=1}^{k} P(B_j \mid C)P(A \mid B_jC)}.$$

Example

A box contains 7 red and 13 blue balls. Two balls are selected at random and are discarded without their colors being seen. If a third ball is drawn randomly and observed to be red, what is the probability that both of the discarded balls were blue?

Let *BB*, *BR*, and *RR* represent the events that the discarded balls are blue and blue, blue and red, and red and red, respectively. Let *R* represent the event that the third ball chosen is red. Solve for the posterior probability $P(BB|R)$.

$$P(BB \mid R) = \frac{P(R \mid BB)P(BB)}{P(R \mid BB)P(BB) + P(R \mid BR)P(BR) + P(R \mid RR)P(RR)}$$

The probability that the first two balls drawn were blue, red, or blue and red are 39/95, 21/190, and 91/190, in that order. Now,

$$P(BB \mid R) = \frac{\frac{7}{18} * \frac{39}{95}}{\frac{7}{18} * \frac{39}{95} + \frac{6}{18} * \frac{91}{190} + \frac{5}{18} * \frac{21}{190}} \approx 0.46$$

The probability that the first two balls chosen were blue given the third ball selected was red is approximately 46%.

Bayes's theorem can be applied to the real world to make appropriate estimations of probability in a given situation. Diagnostic testing is one example in which the theorem is a useful tool. Diagnostic testing identifies whether a person has a particular disease. However, these tests contain error. Thus, a person can test positive for the disease and in actuality not be carrying the disease. Bayes's theorem can be used to estimate the probability that a person truly has the disease given that the person tests positive. As an illustration of this, suppose that a particular cancer is found for every 1 person in 2,000. Furthermore, if a person has the disease, there is a 90% chance the diagnostic procedure will result in a positive identification. If a person does not have the disease, the test will give a false positive 1% of the time. Using Bayes's theorem, the probability that a person with a positive test result actually has the cancer (C), is

$$P(C \mid P) = \frac{\frac{1}{2000}(.90)}{\frac{1}{2000}(0.90) + \frac{1999}{2000}(0.01)} \approx 0.043.$$

If a person tests positive for the cancer test, there is only a 4% chance that the person has the cancer. Consequently, follow-up tests are almost always necessary to verify a positive finding with medical screening tests.

Bayes's theorem has also been used in psychometrics to make a classification scale rather than an ability scale in the classroom. A simple example of classification is dividing a population into two categories of mastery and nonmastery of a subject. A test would be devised to determine whether a person falls in the mastery or the nonmastery category. The posterior probabilities for different skills can be collected, and the results would show mastered skills and nonmastered skills that need attention. The test may even allow for new posterior probabilities to be computed after each question.

The two examples presented above are just a small sample of the applications in which Bayes's theorem has been useful. While certain academic fields concentrate on its use more than others do, the theorem has far-reaching influence in business, medicine, education, psychology, and so on.

Bayesian Statistical Inference

Bayes's theorem provides a foundation for Bayesian statistical inference. However, the approach to inference is different from that of a traditional (frequentist) point of view. With Bayes's theorem, inference is dynamic. That is, a Bayesian approach uses evidence about a phenomenon to update knowledge of prior beliefs.

There are two popular ways to approach inference. The traditional way is the frequentist approach, in which the probability P of an uncertain event A, written $P(A)$, is defined by the frequency of that event, based on previous observations. In general, population parameters are considered as fixed effects and do not have distributional form. The frequentist approach to defining the probability of an uncertain event is sufficient, provided that one has been able to record accurate information about many past instances of the event. However, if no such historical database exists, then a different approach must be considered.

Bayesian inference is an approach that allows one to reason about beliefs under conditions of uncertainty. Different people may have different beliefs about the probability of a prior event, depending on their specific knowledge of factors that might affect its likelihood. Thus, Bayesian inference has no one correct probability or approach. Bayesian inference is dependent on both prior and observed data.

In a traditional hypothesis test, there are two complementary hypotheses: H_0, the status quo hypothesis, and H_1, the hypothesis of change. Letting D stand for the observed data, Bayes's theorem applied to the hypotheses becomes

$$P(H_0 \mid D) = \frac{P(H_0)P(D \mid H_0)}{P(H_0)P(D \mid H_0) + P(H_1)P(D \mid H_1)}$$

and

$$P(H_1 \mid D) = \frac{P(H_1)P(D \mid H_1)}{P(H_0)P(D \mid H_0) + P(H_1)P(D \mid H_1)}.$$

The $P(H_0 \mid D)$ and $P(H_1 \mid D)$ are posterior probabilities (i.e., the probability that the null is true given the data and the probability that the alternative is true given the data, respectively). The $P(H_0)$ and $P(H_1)$ are prior probabilities (i.e., the probability that the null or the alternative is true prior to considering the new data, respectively).

In frequentist hypothesis testing, one considers only $P(D \mid H_0)$, which is called the *p value*. If the p value is smaller than a predetermined significance level, then one rejects the null hypothesis and asserts the alternative hypothesis. One common mistake is to interpret the p value as the probability that the null hypothesis is true, given the observed data. This interpretation is a Bayesian one. From a Bayesian perspective, one may obtain $P(H_0 \mid D)$, and if that probability is sufficiently small, then one rejects the null hypothesis in favor of the alternative hypothesis. In addition, with a Bayesian approach, several alternative hypotheses can be considered at one time.

As with traditional frequentist confidence intervals, a *credible interval* can be computed in Bayesian statistics. This credible interval is defined as the posterior probability interval and is used in ways similar to the uses of confidence intervals in frequentist statistics. For example, a 95% credible interval means that the posterior probability of the parameter lying in the given range is 0.95. A frequentist 95% confidence interval means that with a large number of repeated samples, 95% of the calculated confidence intervals would include the true value of the parameter; yet the probability that the parameter is inside the actual calculated confidence interval is either 0 or 1. In general, Bayesian credible intervals do not match a frequentist confidence interval, since the credible interval incorporates information from the prior distribution whereas confidence intervals are based only on the data.

Modern Applications

In Bayesian statistics, information about the data and a priori information are combined to estimate the posterior distribution of the parameters. This posterior distribution is used to infer the values of the parameters, along with the associated uncertainty. Multiple tests and predictions can be performed simultaneously and flexibly. Quantities of interest that are functions of the parameters are straightforward to estimate, again including the uncertainty. Posterior inferences can be updated as more data are obtained, so study design is more flexible than for frequentist methods.

Bayesian inference is possible in a number of contexts in which frequentist methods are deficient. For instance, Bayesian inference can be performed with small data sets. More broadly, Bayesian statistics is useful when the data set may be large but when few data points are associated with a particular treatment. In such situations standard frequentist estimators can be inappropriate because the likelihood may not be well approximated by a normal distribution. The use of Bayesian statistics also allows for the incorporation of prior information and for simultaneous inference using data from multiple studies. Inference is also possible for complex hierarchical models.

Lately, computation for Bayesian models is most often done via MCMC techniques, which obtain dependent samples from the posterior distribution of the parameters. In MCMC, a set of initial parameter values is chosen. These parameter values are then iteratively updated via a specially constructed Markovian transition. In the limit of the number of iterations, the parameter values are distributed according to the posterior distribution. In practice, after approximate convergence of the Markov chain, the time series of sets of parameter values can be stored and then used for inference via empirical averaging (i.e., Monte Carlo). The accuracy of this empirical averaging depends on the effective sample size of the stored parameter values, that is, the number of iterations of the chain after convergence, adjusted for the autocorrelation of the chain. One method of specifying the Markovian transition is via Metropolis–Hastings, which proposes a change in the parameters, often according to a random walk (the assumption that many unpredictable small fluctuations will occur in a chain of events), and then accepts or rejects that move with a probability that is dependent on the current and proposed state.

In order to perform valid inference, the Markov chain must have approximately converged to the posterior distribution before the samples are stored and used for inference. In addition, enough samples must be stored after convergence to have a large effective sample size; if the autocorrelation of the chain is high, then the number of samples needs to be large. Lack of convergence or high autocorrelation of the chain is detected via convergence diagnostics, which include autocorrelation and trace plots, as well as Geweke, Gelman–Rubin, and Heidelberger–Welch diagnostics. Software for MCMC can also be validated by a distinct set of techniques.

These techniques compare the posterior samples drawn by the software with samples from the prior and the data model, thereby validating the joint distribution of the data and parameters as estimated by the software.

Brandon K. Vaughn and Daniel L. Murphy

See also Estimation; Hypothesis; Inference: Deductive and Inductive; Parametric Statistics; Probability, Laws of

Further Readings

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

Dale, A. I. (1991). *A history of inverse probability from Thomas Bayes to Karl Pearson.* New York: Springer-Verlag.

Daston, L. (1994). How probabilities came to be objective and subjective. *Historia Mathematica*, *21*, 330–344.

Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science*, *7*, 208–225.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, *1*(1), 1–40.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, *222*, 309–368.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènements [Memoir on the probability of causes of events]. *Mémoires de la Mathématique et de Physique Presentés à l'Académie Royale des Sciences, Par Divers Savans, & Lûs dans ses Assemblées*, *6*, 621–656.

Mosteller, F., & Wallace, D. L. (1963). Inferences in an authorship problem. *Journal of the American Statistical Association*, *58*, 275–309.

Pearson, K. (1892). *The grammar of science.* London, UK: Walter Scott.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Bayesian Adaptive Randomization Design

The Bayesian adaptive randomization design is an extension of adaptive research designs, which modify the randomization ratio during a study in order to maximize individual participant outcomes. Bayesian adaptive randomization designs use Bayesian data analysis to estimate the probability that each intervention maximizes the outcome of interest, and the randomization ratio is adjusted according to these estimates. In adjusting the randomization ratio based on which intervention is estimated to produce the most favorable outcome, Bayesian adaptive randomization designs maximize the likelihood of participants experiencing positive outcomes and minimize the likelihood of participants experiencing poor outcomes.

Bayesian adaptive randomization designs follow a four-step procedure. In Step 1, participants are randomly allocated to one of the interventions with the probability of allocation into each intervention being specified by the randomization ratio. In Step 2, participants' outcome data are collected. In Step 3, the probability that each intervention produces the most favorable outcome is estimated with Bayesian data analysis. In Step 4, the randomization ratio is updated. Steps 1–4 are then repeated until either the desired number of participants has been recruited or one of the interventions has shown convincing evidence of producing the most favorable outcome.

This entry goes on to provide an examination of adaptive randomization ratios, two primary assumptions made by Bayesian adaptive randomization design, and logical considerations. Various types of data used in Bayesian data analysis are introduced, and the steps of early and late randomizations are then explained. The entry concludes by listing some advantages and limitations of Bayesian adaptive randomization design, followed by an example of a hypothetical study implementing a two-group Bayesian adaptive randomization design.

Adaptive Randomization Ratios

Most randomized study designs use a fixed randomization ratio, but Bayesian adaptive randomization designs implement an adaptive randomization ratio. A randomization ratio is defined as a ratio of the probabilities of being randomly allocated to each of the interventions. A fixed randomization ratio is held constant throughout the study, but an adaptive randomization ratio can change during the study.

To better understand randomization ratios—both fixed and adaptive—consider a two-group study. In this example and throughout the remainder of this entry, it is assumed that each group is implementing a different intervention. Researchers may consider a research design using a 1:1 fixed randomization ratio, which indicates a 50% chance a participant is allocated into the first group and a 50% chance a participant is allocated into the second group. Researchers may also consider a research design using an adaptive randomization ratio. The adaptive randomization ratio may be 1:1

during the early stages of the study, but the randomization ratio could change during the course of the study if collected data indicate one of the groups' interventions produces better outcomes.

Assumptions

Bayesian adaptive randomization designs make two primary assumptions. First, the randomization of participants must be dispersed throughout the course of the study, such that the outcome data collection for some participants occurs before the randomization of other participants. Because Bayesian adaptive randomization designs use participants' results to inform future randomizations, some participants must be randomized after collecting data from previous participants to maximize individual participant outcomes. If all the participants are randomized simultaneously, Bayesian adaptive randomization designs replicate fixed randomization designs, which would likely fail to maximize individual participant outcomes.

Second, the time between randomizing a participant into one of the groups and obtaining outcome data must be reasonably quick. Since participants' outcomes influence future participants' randomizations, long periods between randomization and outcome data collection force researchers to either slow recruitment so that previous participants' outcomes can be collected prior to randomizing new participants or randomizing new participants without incorporating data from some of the previously randomized participants.

Logistical Considerations

The randomization ratios in Bayesian adaptive randomization designs can be updated after each participant's outcome data are collected, although this may often be unrealistic. The problem with updating the randomization ratio after collecting each participant's data is that the amount of time, funding, and computational resources needed may not be feasible. For example, collecting a participant's outcome data in the morning prior to a participant's randomization in the afternoon would require dedicating time and effort to entering the outcome data immediately after collecting the data as well as having enough computational power to calculate the updated randomization ratio before the randomization. This also does not address concerns with multisite studies, which would need to maintain a centralized database with participant outcomes and the updated randomization ratio for the study teams at each site when there may be numerous data collections and randomizations happening each day.

To address these issues, block stratification can be used to improve the feasibility of Bayesian adaptive randomization designs. Block stratification is a strategy for updating the adaptive randomization ratio at prespecified intervals, which might be after every *n* participants have been randomized or after every *n* weeks. In practice, this might entail updating the randomization ratio after every 10 participants have been randomized or after every 4 weeks (i.e., monthly).

Data

Bayesian adaptive randomization designs utilize Bayesian data analysis to examine between-group differences in order to adjust the randomization ratio. Therefore, Bayesian adaptive randomization designs incorporate data from the prior distribution and the likelihood to estimate the posterior distribution, which models the probability that one of the treatment groups produces the most favorable outcome.

Prior Distribution

The prior distribution models the *a priori* belief regarding the probability that one of the groups produces the most favorable outcome. For two-group Bayesian adaptive randomization designs, a single prior distribution can fully reflect the probability that each group produces the most favorable outcome, since knowing the probability that one group produces the most favorable outcome allows for the probability the second group produces the most favorable outcome to be calculated. For more complex designs with three or more groups, a prior distribution for each group is needed to estimate the probability that each group produces the most favorable outcome. As is the case in Bayesian data analysis, the prior distribution can be based on the findings of previous studies or on theory. It should be noted, however, that the choice of the prior distribution can bias the findings if the chosen prior distribution is not reflective of the data collected in the study. It should also be noted that the confidence in the prior is reflected by the variance of the prior distribution, and the confidence in the prior impacts how much data in the likelihood are needed to produce findings contrary to the prior.

To give a brief example of how a prior distribution is incorporated into a Bayesian adaptive randomized design, consider a two-group study in which researchers begin with an assumption that the two groups will produce similar outcomes. Because the prior distribution is modeling probabilities, the prior distribution ranges from 0 to 1. Since the researchers believe the two groups will perform similarly, the mode of the prior distribution

is located at 0.50, and the prior distribution is symmetric. Put simply, the mode of the prior distribution implies there is a 50% chance that each group produces the most favorable outcome, and the symmetry of the prior distribution ensures both groups have the same probability of producing the most favorable outcome.

Likelihood

The likelihood models the probability that one of the treatment groups produces the most favorable outcome based on the collected data. As the collected data increase, the variance of the likelihood should decrease, resulting in a more precise estimation of the probability that one of the treatment groups produces the most favorable outcome based on the collected data.

Posterior Distribution

The posterior distribution synthesizes the prior distribution and the likelihood to model the probability that one of the treatment groups produces the most favorable outcome. The mode is often used to generate a point estimate for the probability that one of the treatment groups produces the most favorable outcome. A 95% highest density interval, which includes the 95% of the outcomes that are most likely, is often used to estimate the certainty of the posterior distribution. As more data are collected, the width of the 95% highest density interval often decreases, which reflects a more precise estimate of the probability that one of the treatment groups produces the most favorable outcome.

Early Randomizations

When planning a study using a Bayesian adaptive randomization design, researchers must take special care to plan how the first participants will be randomized. Bayesian adaptive randomization designs use Bayesian data analysis to estimate the probability that each of the groups is producing the most favorable outcome. Consequently, factors impacting the estimation of the posterior distribution may bias the randomization ratio used in a Bayesian adaptive randomization design. Specifically, unstable posterior distributions due to small sample sizes in the early stages of a study are perhaps the largest threat to a Bayesian adaptive randomization design.

There are two primary methods for a study using a Bayesian adaptive randomization design to stabilize the posterior distribution for early randomizations. First, researchers can utilize a burn-in stage, which is defined as a period where the study implements a fixed randomization ratio to allow for a sizable number of participants to be assigned to all the study groups prior to switching to an adaptive randomization ratio. In practice, this might mean randomizing the first 50 or 100 participants at a 1:1 fixed randomization ratio in a two-group study. Second, researchers can utilize a tuning factor, which weights the obtained outcome data so that studies using a Bayesian adaptive randomized design essentially use a fixed randomization ratio for early randomizations and gradually shift toward an adaptive randomization ratio as more data are collected.

Late Randomizations

In contrast to a fixed randomization design, Bayesian adaptive randomized designs have reduced statistical power because participant allocations and outcomes in Bayesian adaptive randomized designs are correlated. This correlation increases the variance of the outcomes, which makes it harder to distinguish whether there are statistical between-group differences in terms of intervention effectiveness. Hence, more participants are needed in a Bayesian adaptive randomized design to achieve the same level of statistical precision as a fixed randomization design.

To compensate for the reduced statistical power in a Bayesian adaptive randomization design, researchers will often implement early stopping rules. Early stopping rules are defined as a priori conditions for terminating the study based on the magnitude of the probability that one group produces superior outcomes as estimated in the posterior distribution. In practice, researchers may decide to conclude the study if one treatment group has a posterior probability greater than .95 that the treatment is superior to the other treatment(s). It should be noted, however, that there are no defined standards for what constitutes an appropriate early stopping rule. Consequently, the early stopping rules used in practice are variable.

Advantages

Maximizing Individual Participant Outcomes

Bayesian adaptive randomization designs seek to provide the best outcome to each of the individuals participating in the study. Bayesian adaptive randomized designs use empirical evidence regarding the effectiveness of the treatment interventions to adjust the randomization ratio. Consequently, Bayesian adaptive randomized designs are maximizing the likelihood of participants experiencing a positive outcome. At the same time, Bayesian adaptive randomized designs are

minimizing the likelihood of participants experiencing a poor outcome, which is critical in fields such as cancer research.

Incorporating Previous Findings

To better maximize the likelihood of participants experiencing a positive outcome, Bayesian adaptive randomized designs can also incorporate the findings from previous studies. By appropriately incorporating previous findings, the prior distribution used in a Bayesian adaptive randomization design should cause more participants to be allocated into the most effective intervention in the early stages of the study, which should maximize the likelihood of participants experiencing a positive outcome. While there are no guidelines of how similar the previous studies and the current study should be, substantial differences between previous studies and the current study may lead the prior distribution to be a poor reflection of the data collected in the study, which may decrease the likelihood of participants experiencing a positive outcome. Thus, care should be taken in choosing studies to inform the prior distribution.

Limitations

Threats to Internal Validity

Because Bayesian adaptive randomized designs rely on sequential randomization, participants may be differentially affected by threats to internal validity (e.g., history effect, maturation). For example, a study implementing a Bayesian adaptive randomization design for an intervention designed to improve reading comprehension for elementary school students may have reduced internal validity, since the elementary students are experiencing developmental changes during the intervention that could impact the effectiveness of the intervention.

Equally Effective Interventions

Bayesian adaptive randomization designs were created with the intention of maximizing the likelihood of positive outcomes for participants. However, this assumes the superiority of one of the implemented interventions. In studies where there is no superior intervention, the randomization ratio will replicate the fixed randomization ratio used in a traditional randomized design. In such cases, the reduced statistical power of Bayesian adaptive randomization designs dictates that more participants must be randomized to achieve the same level of statistical precision as a fixed randomization design.

Example

Having gone through the characteristics of Bayesian adaptive randomization designs, an example of a study implementing a two-group Bayesian adaptive randomization design will be provided. To preserve the generality of the example for a variety of readers, the interventions will not be described in detail. Rather, they will simply be described as Intervention A and Intervention B, which are associated with Group A and Group B, respectively.

In this hypothetical study, an uninformative prior distribution with a mode of 0.5 will be used, which implies neither intervention is superior. This study will recruit 300 participants, will set the burn-in stage to be 50 participants, will establish an early stopping rule of .95, and will not use block stratification. Of note, this study will use a fixed 1:1 randomization ratio for the 50 participants randomized during the burn-in stage.

When beginning this study, the first 50 participants will be randomized similar to any fixed randomization design, and the outcomes for these 50 participants will be collected. Because the randomization ratio for the first 50 participants is 1:1, there should be approximately 25 participants allocated into each group at this point, although the exact numbers may differ slightly. With approximately 25 participants per group, there should be enough data collected from both groups to obtain stable posterior distribution estimates.

After collecting the outcome data for the 50th randomized participant, the between-group difference in intervention effectiveness is estimated using Bayesian data analysis. For purposes of this example, the process for conducting this Bayesian data analysis will not be presented. Instead, consider the case where there is a 65% chance that Intervention A is more effective than Intervention B, and the randomization ratio adapts to 0.65. Thus, there is a 65% chance the next participant will be allocated into Group A and a 35% chance the next participant will be allocated into Group B.

Continuing the example, the 51st participant is subsequently allocated into Group B, and the outcome for the 51st participant is positive. Consequently, the probability that Group A produces superior outcomes decreases. For the purpose of demonstration, consider the case where there is now a 63% chance that Intervention A is more effective than Intervention B, and the randomization ratio adapts to 0.63. Thus, there is now a 63% chance that the 52nd participant will be randomized into Group A.

Assume the 52nd participant is allocated into Group A, and the outcome for the 52nd participant is positive.

The probability that Group A produces superior outcomes increases. Because more data have been accumulated, the changes in the probability of a group producing a superior outcome, and thus changes in the randomization ratio, will be less with each allocated participant. There is now a 64% chance that Intervention A is more effective than Intervention B, and the randomization ratio adapts to 0.64. Thus, there is now a 64% chance that the 53rd participant will be randomized into Group A.

Continuing the example, the 53rd participant is allocated into Group B, and the outcome for the 53rd participant is negative. Therefore, the probability that Group A produces superior outcomes increases. There is now a 65% chance that Intervention A is more effective than Intervention B, which causes the randomization ratio to adapt to 0.65. Thus, there is now a 65% chance that the 54th participant will be randomized into Group A.

This process will continue with the probability of one group producing superior outcomes being estimated after each participant's data are collected until either (1) the outcome data for the 300th participant is collected or (2) the posterior probability that one of the interventions produces superior outcomes exceeds the early stopping rule of .95.

Jeffrey C. Hoover

See also Adaptive Designs in Clinical Trials; Bayesian Data Analysis; Distribution; Posterior Distribution; Random Assignment; Research Design Principles; Risk (in Human Subjects Research)

Further Readings

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC press.

Hu, F., & Rosenberger, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, *98*(463), 671–678. doi:10.1198/016214503000000576.

Hu, F., & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. Hoboken, NJ: Wiley.

Korn, E. L., & Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology*, *29*(6), 771–776. doi:10.1200/JCO.2010.31.1423.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London, UK: Academic Press.

Lee, J. J., Chen, N., & Yin, G. (2012). Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*, *18*(17), 4498–4507. doi:10.1158/1078-0432.CCR-11-2555.

Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice* (2nd ed.). Hoboken, NJ: Wiley.

Thall, P. F., & Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, *43*(5), 859–866. doi:10.1016/j.ejca.2007.01.006.

Bayesian Data Analysis

Bayesian data analysis is an umbrella term that encompasses different data analysis approaches that have in common the use of Bayes's rule as a guiding principle, and the goal of quantifying evidence in favor of possible parameter values, models, or hypotheses, rather than making a decision on whether a parameter value is different than a specific value (e.g., zero) or not. This entry describes two approaches in detail—Bayesian parameter estimation and Bayesian hypothesis testing—and briefly describes two other approaches: Bayesian model comparison and hierarchical Bayesian models.

Bayesian Parameter Estimation

The goal of Bayesian parameter estimation is the same as the traditional frequentist parameter estimation: to make an estimation of the value of a population parameter such as a mean, a difference between two means, a correlation between two variables, or a regression coefficient. The difference between the two approaches is that in the traditional approach the estimation consists of providing a point estimate and a confidence interval, typically a 95% confidence interval, whereas in the Bayesian approach the estimation consists of providing a posterior distribution, which could be summarized, for instance, with the mean of the distribution and its 2.5th and the 97.5th percentiles (i.e., a 95% credible interval or the 95% high-density interval).

The advantage of the 95% credible (or high density) interval in Bayesian parameter estimation is that it provides the information that researchers are typically interested in. It indicates that the probability that the actual value of the parameter of interest (e.g., the population mean of a variable) is within the interval is 95%. In other words, the researcher can be 95% confident that the actual parameter value is in the interval. The traditional 95% confidence interval is more difficult to interpret: It is the interval generated by a procedure that provides confidence intervals that include the actual parameter value 95% of the times the procedure is used. The process to obtain a posterior distribution of a parameter of interest consists of four steps: choice of

distribution to create the likelihood function, choice of prior distribution, data collection and obtention of relevant statistics, and obtention of posterior distribution.

Choice of Distribution to Create the Likelihood Function

Let's consider the case in which a researcher aims to estimate the proportion of people who are currently experiencing anxiety in a specific city by administering an anxiety scale to 200 people in that city; that scale determines whether the person is experiencing anxiety (1) or not (0). The distribution to create the likelihood function has a resemblance with the sampling distribution in the traditional approach. In this case, the binomial distribution is the most appropriate distribution to construct the likelihood function. The binomial distribution provides the probability of obtaining a specific number of people with anxiety in the sample of 200 participants given that the proportion of people of anxiety in the population (denoted by π) is a determined value. For example, if $\pi = 0.50$, the most probable value in the sample is 100 people with anxiety, while 99 and 101 being the second most probable values, 98 and 102 following in probability, and so forth, with values close to 0 and 200 having an extremely low probability of occurrence. The binomial distribution is then used to construct a likelihood function over parameter values once the data are observed. The likelihood function will come into play again when this function is combined with the prior distribution to produce the posterior distribution.

Choice of Prior Distribution

In Bayesian parameter estimation, the prior distribution is the credibility for each possible parameter value given by a researcher before observing new data. The researcher declares what credibility they give to each possible parameter value by using a probability distribution. This distribution is known as the *prior distribution* (or just *the prior*). That is, the Bayesian researcher makes explicit their knowledge about the parameter values before collecting the data. There are two main approaches to choose the prior distribution. In the first approach, the researcher declares to be ignorant regarding the possible parameter values, and this ignorance is expressed by choosing a probability distribution in which all the possible parameter values are equally likely. Given that π can be any value between 0 and 1, a continuous uniform distribution with range 0 to 1 would be appropriate. However, for reasons that will be clear later, a beta distribution is typically used for proportions. A beta distribution with parameters $\alpha = 1$ and $\beta = 1$ is equivalent to the continuous uniform distribution over the range 0 to 1. This type of prior distribution is typically called *uninformed prior*.

In the second approach, the researcher uses their actual knowledge of the parameter values and expresses it in the prior distribution. Suppose the researcher knows of a previous study measuring anxiety in the same city with 50 participants, in which 20 of them had anxiety and 30 did not have anxiety. They can express this knowledge with a beta prior distribution with parameters $\alpha = 20$ and $\beta = 30$. In that distribution the most probable value is 0.40, and the lower bound and upper bound of the 95% credible interval of the distributions are 0.27 and 0.58, respectively. This type of prior distribution is typically referred to as *informed prior*.

Data Collection and Obtention of Relevant Statistic

Once the binomial distribution is chosen to construct the likelihood function and the beta distribution is chosen as prior distribution, it is time to collect the data (in this case, the administration of an anxiety scale to 200 participants) and obtain the value of the statistic of interest, that is, the number of people with anxiety. Assume that the results indicate that 40 participants have anxiety and 160 participants do not have anxiety.

Obtention of Posterior Distribution

The posterior distribution is obtained by using the Bayes's rule as guiding principle:

$$P(\theta|d) = \frac{P(d \mid \theta) \times P(\theta)}{P(d)},$$

which in English can be expressed as:

$$\text{Posterior distribution} = \frac{\text{Likelihood} \times \text{Prior distribution}}{\text{Marginal likelihood}},$$

where θ denotes any parameter value in a generic way. When referring to specific parameters, a different Greek letter takes the place of θ. For example, π is used for population proportions, μ is used for population means, σ is used for population standard deviations, ρ is used for population correlation, and β is used for population regression slopes. In this case, θ represents the proportion of persons with anxiety in the population; thus, in this explanation π and θ are used interchangeably. Note that we do not know the actual value of θ, so we must consider all the possible values between 0 and 1. $P(\theta)$ is the prior distribution, which in the example is either a beta distribution with parameters $\alpha = 1$ and $\beta = 1$ for the uninformed prior or a beta distribution with parameter $\alpha = 20$ and parameter $\beta = 30$

for the informed prior. $P(d \mid \theta)$ is the probability of the observed data (d) given each possible parameter value, also known as the *likelihood of the possible parameter values*. Some authors use the notation $L(\theta \mid d)$, which is very useful because it identifies that, as the prior and posterior distributions, we are dealing with a function that assigns a number to all the possible parameter values. It is not strictly a probability distribution because it does not add to 1 as discrete probability distributions do or it does not integrate to 1 as continuous probability distributions do. However, it works as a probability distribution in the sense that somehow indicates how likely each parameter value is.

The construction of the likelihood function is rather artificial; here is an example to explain the process. After observing that the data are 40 people with anxiety out of 200, we use the binomial distribution to determine the probability of obtaining 40 successful outcomes out of 200 trials for all possible values of π. For clarity of explanation, only four possible values of π are considered, but it should be considered that π can acquire infinite number of values in the range 0 to 1. (Note that we are now using as π instead of θ.)

$$L(\pi = 0.15 | d = 40) = P(d = 40 | \pi = 0.15) = 0.0115$$

$$L(\pi = 0.20 | d = 40) = P(d = 40 | \pi = 0.20) = 0.0704$$

$$L(\pi = 0.25 | d = 40) = P(d = 40 | \pi = 0.25) = 0.0173$$

$$L(\pi = 0.30 | d = 40) = P(d = 40 | \pi = 0.30) = 0.0041$$

The first line should be interpreted as the likelihood of the parameter π to be the value 0.15 in the population (i.e., the city of interest) given that we observed 40 out of 200 people with anxiety in the sample. This is calculated by using the binomial distribution and determining the probability of observing 40 out of 200 successes given that the probability of success in each trial (i.e., π) is 0.15. That probability is 0.0115. The second line indicates how likely is for π to be 0.20, the third how likely is for π to be 0.25, and the fourth line indicates how likely is for π to be 0.30. In this set of values, the most likely value is 0.20 with a probability of 0.0704. We can keep adding parameter values in the range 0 to 1 and obtaining the likelihood for each of them, and eventually we will obtain a function that resembles a probability distribution.

To calculate the posterior distribution, we need to follow three more steps. The first one is to follow the numerator of Bayes's rule and, for each parameter value, to obtain the product of its likelihood and its prior probability. This new set of values gives us a good idea of the plausibility of each parameter value after observing the data and taking into account previous knowledge; however, it is not a probability distribution because it does not add or integrate to 1. We can obtain a proper posterior distribution if there is a way of calculating the marginal likelihood [i.e., $P(d)$], which is a single number obtained with the following integral:

$$\int_{\theta} P(d \mid \theta) \times P(\theta)\ d\theta,$$

where the d in $d\theta$ does not refer to data, rather it indicates the integration is over all possible values of θ. Calculating this integral is not always feasible, and when this is the case, Monte Carlo simulation methods are used to approximate it, methods that are explained in this entry. In this example, the choice of the beta distribution as prior and the binomial distribution for constructing the likelihood was not coincidental. These two distributions have the property of conjunction, which means that there is a way of combining them to easily come up with a new probability distribution. When conjunction is possible, the posterior distribution can be obtained without calculating the marginal likelihood.

For the uninformed prior, the beta distribution with $\alpha = 1$ and $\beta = 1$ combined with the binomial with 40 success (and 160 failures), the posterior distribution is a beta distribution with $\alpha = 41$ (40 + 1) and $\beta = 161$ (160 + 1). In this case, the posterior distribution has exactly the same shape as the likelihood function, which makes sense because the prior represented the ignorance of the researcher, thus the information about the parameter value in the posterior distribution is entirely provided by the new observed data, which is captured by the likelihood function. For the informed prior, the posterior distribution becomes a beta distribution with $\alpha = 60$ (40 + 20) and $\beta = 190$ (160 + 30). In this posterior distribution, the mean is 0.24 and the 95% high-density interval is (0.189, 0.295). Note that the prior probability distribution was centered on 0.40 and that the likelihood was centered on 0.20 and the posterior distribution is centered on 0.24. The center of the posterior distribution is closer to that of the likelihood than that of the prior because the prior only incorporated knowledge of a sample of 50 participants, whereas the likelihood provided information over 200 participants; therefore, the latter has more weighting in the posterior distribution.

Practical Issues

Bayesian parameter estimation forms part of all the other approaches, but John Kruschke has been a strong proponent of Bayesian parameter estimation as a

stand-alone procedure. Kruschke's website (https://jkkweb.sitehost.iu.edu/index.html) provides a number of useful resources for conducting parameter estimation. The R packages rstan and rjags afford the possibility to conduct Bayesian parameter estimation with limited coding knowledge. The user friendly and free software JASP was built with the purpose of conducting Bayesian hypothesis testing, but it also provides posterior distributions with credible intervals.

Bayesian Hypothesis Testing

Bayesian hypothesis testing is an approach to compare a null hypothesis with an alternative hypothesis. There are a few differences between the Bayesian and the traditional hypothesis testing approaches. First, in the Bayesian approach, there is a model for the null hypothesis and a model for the alternative hypothesis, unlike the traditional approach in which only a model for the null hypothesis is specified. This allows the Bayesian approach to quantify the evidence in favor or against the alternative hypothesis relative to the null hypothesis and vice versa. Second, the Bayesian approach uses Bayes factors, not *p* values as in the traditional approach. Third, the Bayes factor quantifies the relative evidence in favor or against the hypotheses, it does not make a binary decision (reject or not reject) regarding the null hypothesis.

The process for conducting Bayesian hypothesis testing involves specifying two models and comparing them. Let's consider a case in which a *t*-test is used in the traditional approach. A sample of 80 participants was randomly allocated to two conditions to perform a memory task: fast presentation of stimulus (40 participants) or slow presentation of stimulus (40 participants). The mean percentage of correct answers was obtained for each group, and the goal of the study was to test the null hypothesis that there are no differences in performance in this memory task between the two groups. The mean percentage items correctly recalled in the group that was presented stimuli at a fast pace was 50.15 and *SD* = 16.4 and that of the group that was presented stimuli at a slow pace was 61.18, *SD* = 15.2. The mean difference is –11.03. The pooled *SD* is 15.8; therefore, the effect size in the sample is: –11.03/15.8 = –0.698.

Specification of Priors Over Parameter Values and Models

In Bayesian parameter estimation, one only considers one model at a time, whereas in Bayesian null hypothesis testing, two models are compared. One must both specify a prior distribution over parameter values for each of the models and also specify the prior distribution over the models.

Regarding the prior distribution of parameters, the parameter of interest is the difference between means in memory performance in a hypothetical population of people exposed to slow presentation of stimuli compared to one in which people are exposed to fast presentation of stimuli. For practical reasons, a standardized effect size is used for this parameter, and it is denoted by δ, which is calculated by δ = difference between population means/combined standard deviation in the population.

The prior distribution over the possible parameter values for the model of the null hypothesis is:

$$H0 : \delta = 0.$$

This means that for the model of the null hypothesis, the whole probability is given to the parameter value 0, values different than 0 are impossible, and are given a probability of 0. For the alternative hypothesis, there are many options for prior distribution. In this example, we will use a normal distribution centered on 0.

$$H1 : \delta \sim \text{Normal}\ (\pi = 0, \sigma = 1).$$

In this model of the alternative hypothesis, the prior distribution for the difference between means gives the value 0 the highest probability, given that the mean (π) of a normal distribution is its most probable value, and values higher and smaller than 0 that are close to 0 are highly probable whereas values far away from zero are less probable. The spread of the normal distribution is determined by its standard deviation (σ). There is an extensive literature on determining default values for parameters such as σ, which will not be discussed here. Suffice it to say that in experimental research in social sciences, large differences between means are unlikely, so the default value of σ should reflect this fact; in this case, $\sigma = 1$ was chosen. If the researcher has the knowledge or expectation that one of the groups will have higher performance than the other group, they may express that by using a half-normal distribution, that is, a normal distribution centered on 0, but with the left side of the distribution removed, making explicit the assumption that the value of δ in the population is higher than zero. This is equivalent to a directional hypothesis in traditional null hypothesis testing.

Regarding the prior distribution over models, the default distribution is the following:

$$P(\text{H0}) = 0.5,\ P(\text{H1}) = 0.5.$$

This uniform distribution is equivalent to the uninformed prior explained in the parameter estimation section. In this case, the researcher makes explicit that they do not have previous knowledge on which hypothesis is more plausible or they prefer not to use any previous knowledge for hypothesis testing. There are situations in which a researcher may decide to be conservative and assign H0 a higher prior probability than that for H1. For example, if H1 involves showing that telepathy exists, because it involves violating well established laws of physics, a researcher may assign a lower prior probability to H1 and only favor H1 over H0 if the data provide extraordinary evidence in favor of H1.

Specification of Likelihood

The specification of likelihood over parameter values is the same as in Bayesian parameter estimation. Again, there are a few distributions that can be used to construct the likelihood such as a *t* distribution centered on any possible parameter value. For simplicity, assume a normal distribution. The rationale is the following: If in the population $\delta = 0$, in the sample used in the study one would expect the difference in memory performance between the two groups to be 0 or close to 0, with values far away from 0 being very unlikely. Likewise, if $\delta = 8.5$, then one would expect the difference between means in the sample to be 8.5 or values close to 8.5, with values far from 8.5 being very implausible. Please refer to the likelihood in Bayesian parameter estimation to see how the likelihood is constructed after this step.

Bayes's Rule

In hypothesis testing, we need to use two versions of Bayes's rule. First, a version with a posterior distribution over parameter δ:

$$\mathrm{H0}: P\ (\delta|d,\ \mathrm{H0}) = \frac{P(d \mid \delta, \mathrm{H0}) \times P(\delta \mid \mathrm{H0})}{P(d \mid \mathrm{H0})} \times P(\mathrm{H0}),$$

$$\mathrm{H1}: P\ (\delta|d,\ \mathrm{H1}) = \frac{P(d \mid \delta, \mathrm{H1}) \times P(\delta \mid \mathrm{H1})}{P(d \mid \mathrm{H1})} \times P(\mathrm{H1}),$$

where $P(\mathrm{H0}) = \mathrm{P}(\mathrm{H1}) = 0.5$ are the priors for the hypotheses, $P(\delta \mid \mathrm{H0})$ is the prior of δ for the null hypothesis (i.e., $\delta = 0$), $P(\delta \mid \mathrm{H1})$ is the prior of δ for the alternative hypothesis (i.e., a normal distribution with $\mu = 0$ and $\sigma = 1$), and $P(d \mid \delta, \mathrm{H0})$ and $P(d|\delta, \mathrm{H1})$ denote the likelihoods over parameter values for the models of the null and the alternative hypotheses, which in both cases are obtained via a normal distribution.

Given that the observed standardized difference between means in our sample of 80 participants is –0.698, the likelihood in the model of the null hypothesis is the probability of observing that value given that $\delta = 0$, whereas in the model of the alternative hypothesis the likelihood includes the probability of observing –0.698, given all possible values of δ. As in parameter estimation, the marginal likelihoods $P(d \mid \mathrm{H0})$ and $P(d \mid \mathrm{H1})$ are calculated with the integrals

$$\int_{\delta} P(d \mid \delta, \mathrm{H0}) \times P(\delta \mid \mathrm{H0})\ \mathrm{d}\delta$$

and

$$\int_{\delta} P(d \mid \delta, \mathrm{H1}) \times P(\delta \mid \mathrm{H1})\ \mathrm{d}\delta,$$

respectively.

The second version of Bayes's rule contains a posterior over models, not parameter values, denoted by the following equations:

$$P\ (\mathrm{H0}|d) = \frac{P(d \mid \mathrm{H0})}{P(d)} \times P(\mathrm{H0}),$$

$$P\ (\mathrm{H1}|d) = \frac{P(d \mid \mathrm{H1})}{P(d)} \times P(\mathrm{H1}).$$

These versions of the Bayes's rule show the posterior of the models for the null and the alternative hypotheses, which are calculated with the likelihood of the hypothesis given the data [$L(\mathrm{H0} \mid d) = P(d \mid \mathrm{H0})$ and $L(\mathrm{H0} \mid d) = P(d \mid \mathrm{H0})$, respectively], the priors of the hypotheses [$P(\mathrm{H0})$ and $P(\mathrm{H1})$, respectively], and the probability of the data [$P(d)$]. Note that the likelihoods in this version of the Bayes's rule are equivalent to the marginal likelihoods in the previous version, and they denote how plausible is the observed effect size of –0.698 given the prior distribution assigned over δ in each model. In the current version, the probability of the data $P(d)$ represents how plausible is to observe –0.698, for a model that combines the model of H0 and the model of H1.

Bayes Factor

The Bayes factor is a ratio. In this instance, the ratio between the marginal likelihood under the model of the null hypothesis and the marginal likelihood under the model of the alternative hypothesis. Obtain the Bayes factor by obtaining the ratio of the two previous equations:

$$\frac{P(\mathrm{H0} \mid d)}{P(\mathrm{H1} \mid d)} = \frac{P(d|\mathrm{H0})\ \times P(\mathrm{H0}) / P(d)}{P(d|\mathrm{H1}) \times P(\mathrm{H1}) / P(d)},$$

The probability of the data $[P(d)]$ is the same value for both models; therefore, this value can be simplified from the equation and obtain:

$$\frac{P(\text{H0} \mid d)}{P(\text{H1} \mid d)} = \frac{P(d|\text{H0})}{P(d|\text{H1})} \times \frac{P(\text{H0})}{P(\text{H1})}$$

which can be expressed in English as:

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}$$

$$\text{That is, Bayes factor} = \frac{P(d|\text{H0})}{P(d|\text{H1})}.$$

The Bayes factor must be interpreted as how plausible the data are under a model (i.e., the model of the null hypothesis) relative to how plausible the data are under another model (i.e., the model of the alternative hypothesis). This ratio is named as Bayes factor 01. Using JASP, the Bayes factor 01 returns the value 0.0574. When the Bayes factor 01 is a number below 1, it is more useful to calculate the Bayes factor 10, that is, the ratio between the marginal likelihood of the model of the alternative hypothesis and that of the model of the null hypothesis. Alternatively, the same value can be achieved by Bayes factor 10 = 1/Bayes factor 01. The corresponding value is 17.4245, indicating that the data are more than 17 times more likely under the alternative hypothesis than under the null hypothesis. This quantification of the evidence in favor of the alternative hypothesis relative to the null hypothesis is the end of the analysis, unlike in the traditional approach in which a binary decision on whether to reject the null hypothesis is carried out.

Practical Issues

The use of Bayes factor to evaluate hypotheses was introduced by Robert E. Kass and Adrian E. Raftery in 1995. More recently, it has been popularized by psychologist Eric-Jan Wagenmakers, who has been developing a free and easy to use statistical software, JASP, with emphasis on Bayesian hypothesis testing, which can be downloaded for free from jasp-stats.org. This software builds on the work of other researchers, who developed default priors to test null hypotheses via Bayes factor to replace traditional tests such as *t* test, analysis of variance, regression, correlation, binomial test, and chi-squared test, among others.

Other Bayesian Approaches

Another important approach is Bayesian model comparison. The Bayesian model comparison approach also uses Bayes factor, but instead of comparing a null hypothesis with an alternative hypothesis, it compares any two models that aim at explaining the data. Model comparison can occur between nested models, such as a regression model with one intercept and one slope compared to a regression model with those parameters plus an additional slope parameter, and non-nested models such as one that explains the outcome variable with one predictor variable with another that does so with a completely different predictor variable. Models that have the same predictor variables but differ in the function that relates the predictor variable with the outcome variables can also be compared, such as a linear model with a quadratic model. Finally, an interesting feature of the Bayesian model comparison approach is that it can compare models that have the same predictor variables and functions linking them to the outcome variables but differ in their prior probability distribution for possible parameter values. Andrew Gelman has introduced these models to social sciences, and Richard McElreath's book, *Statistical Rethinking*, presents multiple examples of this approach in a didactic way.

The final approach is hierarchical Bayesian models. These models possess multiple prior distributions and at different levels; they have characteristics of traditional structural equation models because they have multiple variables and pathways connecting them, multilevel models because they have parameters at different levels (e.g., a parameter can vary among individuals or it is fixed for all the individuals, another may vary or be fixed among trials, and so forth), and directed acyclical graphs because they share the graphical representation with those models. An innovative feature of hierarchical Bayesian models is the use of plates, a graphical device to represent whether a parameter is fixed or varies among individuals, trials, or time, for example.

Final Thoughts

Bayesian data analysis approaches, as indicated earlier, share the use of Bayes's rule as a guiding principle for data analysis. They also share another unique feature: They explicitly manifest the knowledge the researcher has before observing new data, and that knowledge is updated after observing new data. This feature dovetails with the cumulative and collaborative nature of the scientific endeavor. Strong conclusions based on single studies are discouraged and the cumulative quantification of evidence is encouraged.

Guillermo Campitelli

See also Posterior Distribution

Further Readings

Campitelli, G., & Macbeth, G. (2014). Hierarchical graphical Bayesian models in psychology. *Revista Colombiana de Estadistica*, *37*, 319–339.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

JASP Team. (2020). JASP (Version 0.14.1) [Computer software]. Jamaica, NY: JASP Team.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London, UK: Academic Press.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Boca Raton, FL: CRC Press.

Bayesian Information Criterion

Bayesian (or Bayes) information criterion (BIC), also referred to as *Schwarz Bayes information criterion*, is one of the information criteria for selecting statistical models. BIC, as its name tells, is based on the asymptotic behavior of Bayes estimators under a class of priors. It is defined mathematically as L2 – $r\log(N)$, in which N is the total sample size and r is the number of degrees of freedom after model fitting.

History of BIC

BIC was first proposed by Gideon E. Schwarz in his 1978 paper, "Estimating the Dimension of a Model," published in the *Annals of Statistics*, based on a Bayesian approach to hypothesis testing developed by Harold Jeffreys in 1961. According to Schwarz, using the maximum likelihood principle to choose the appropriate model is problematic because this principle results in the highest possible dimension rather than the right dimension. The approach based on Bayes estimator assumes that observations are relative to some fixed measure on the sample space with some density. When the asymptotes of Bayes are fitted, researchers need not know the specification of a priori distribution on the parameters. Furthermore, it is assumed that there is a fixed penalty for estimating the wrong model, regardless of the sample size, because the Bayes solution selects a most probable a posteriori model based on the a priori distribution. Therefore, a model is considered true when it has the highest posterior probability, without specifying a prior distribution.

Although Schwartz was the first person who proposed BIC, his work was not paid attention to until Adrian E. Raftery used the BIC in a wide range of models and strongly advocated for its usefulness in his two papers published in 1986: one was "Choosing Models for Cross-Classifications" in *American Sociological Review* and the other was "A Note on Bayes Factors for Log-linear Contingency Table Models With Vague Prior Information" in *Journal of the Royal Statistical Society*. He argued that the Bayesian approach is better for hypothesis testing especially with a larger sample size and it can also account for model uncertainty with the use of priors.

Variants of BIC

Besides the original form of BIC, there are other revised forms or variants of BIC, including Kashyap Bayesian information criterion (KBIC) developed by Rangasami L. Kashyap in 1982, adjusted Bayesian information criterion (ABIC) developed by Stanley L. Sclove in 1987, the Haughton Bayesian information criterion (HBIC) developed by Dominique M. A. Haughton in 1988, the distributed Bayesian information criterion (DBIC) developed by David Draper in 1995, as well as the information matrix-based Bayesian information criterion (IBIC) and the scaled unit information proper Bayesian information criterion (SPBIC) developed by Kenneth A. Bollen and colleagues in 2012.

More specifically, KBIC selects a time-series model in BIC equation. ABIC focuses on the shortest description length in model selection and keeps a balance between model fit and complexity. It is also known as the *sample-size-adjusted BIC*. HBIC extends BIC's focus on linear models to curve models and reduces the penalty in BIC. DBIC adds a term to BIC equation that has better results with small to moderate sample sizes. IBIC adds two terms into the original equation of BIC, and SPBIC uses scaled unit information prior rather than unit information prior in the original form of BIC.

Use of BIC

BIC is used as a criterion of model selection and can be applied in many models, such as log linear, logits, covariance structure models, and linear regression. It includes an adjustment for sample size and has consistency property, which suggests that, when sample size increases and approaches infinity, the BIC selects the fitted mode with probability close to 1. This makes BIC useful

especially when researchers are building theories based on data because they assume that they can collect data, fit the model based on BIC, and select the true model as the sample size increases. Researchers need to compute the BIC for each model and select the model with the smallest criterion value. In this sense, BIC focuses on parsimony of the model.

Although BIC and other information criterion measures are not used as popularly as other fit indices in fields such as structural equation modeling (SEM), they outperform other fit indices in moderate to large samples and in models that have extra parameters. When the SEM model includes real-world criterion, or criterion that is measurable, BIC is probably the most effective for model accuracy and parsimony. Among variants of BIC, ABIC has a better performance than original BIC in selecting numbers of factors and latent classes. DBIC has an improved performance in model selection for real problems. HBIC improves BIC's performance particularly in confirmatory factor analysis models. It, along with SPBIC, performs the best in the selection of true models and has the highest accuracy ratios in model fitting.

Limitations of BIC

Despite its overall superior performance in model selection, BIC has its own limitations. For instance, because Bayes factors are sensitive to differences existing in prior beliefs, they are not always the same as the Bayes factor implied by prior beliefs. Furthermore, the advantages of including total sample size in the BIC makes it harder to discriminate between null and alternative hypotheses based only on sample size. When testing the difference between two group means, a sample of 1,000 participants in each group is more informative than a sample of 200 participants in one group and 1,800 in the other. Because the BIC uses the total sample size in the study, the difference between the samples in the two situations is not obvious and the use of prior beliefs in the BIC is difficult to justify. In addition, BIC seems too conservative in model selection.

Haiying Long

See also Bayes's Theorem; Model Fit

Further Readings

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 1–19. doi:10.1080/10705511.2014.856691.

Raftery, A. E. (1999). Bayes factors and BIC: Comment on "A critique of the Bayesian information criterion for model selection." *Sociological Methods and Research*, *27*(3), 411–427. doi:10.1177/0049124199027003005.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research*, *27*(3), 359–397. doi:10.1177/0049124199027003002.

Bayesian Networks

Bayesian networks are probabilistic graphical models depicting the relations within a potentially large set of variables. Commonly, Bayesian networks show the conditional dependencies, information that cannot easily be derived from, for example, a correlation table. Thus, these networks are a versatile tool for exploring and studying relations in large data sets and are used in many branches of science.

Graphical Model

In its essence, Bayesian network models are graphical visualizations of variables and their relations, whereby the values of their relations depend on random variation. Just as in structural equation models and factor models, for example, these visualizations show each variable as an ellipse, called a *node*, and the relations are depicted by arrows, called *edges* or *ties*, between variables.

The graph satisfies the formal requirements of Bayesian networks if it is a so-called directed acyclic graph (DAG). In a directed graph, all connections are arrows—usually indicating causal relations—and there are no cycles in the graph. No cycles implies that if there is a directed path from node A to node B, there cannot be a path from node B to node A.

Many studies focus on correlational relations rather than causal relations. A variant of the Bayesian networks, where directed arrows are replaced by undirected lines, is useful in this context. Although technically not a Bayesian network, these graphs are also commonly referred to as *Bayesian networks*.

Figure 1 displays the Bayesian network for a fictitious example. For 300 primary school children, three variables are recorded: age (X_1), their shoe size (X_2), and their numeracy level, based on some arithmetic test (X_3). These nodes are depicted as ellipses, with the relations as arrows. In this example, the direction of the

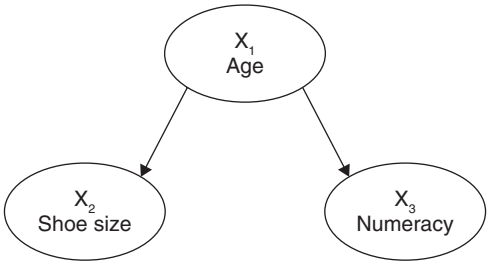


Figure 1 Example of a Bayesian Network

causal arrows is clear: one's feet grow with age, and so does one's arithmetic skills.

Perhaps the most interesting thing about Figure 1 is the lack of edge between X_2 and X_3. This does not imply that these two variables are unrelated. In fact, they are probably strongly correlated: kids with large feet tend to also be the kids with better arithmetic skills, as these children are the older ones in the sample. Variables X_2 and X_3 have no edge as these are *conditionally independent*: given the value for X_3, the two variables are unrelated. In other words, for children *of the same age*, there is no relation between shoe size and numeracy.

Formally, variables A and B are independent conditional on variables C_1, …, C_k if

$$\mathrm{P}A \cap BC_1, \ldots, C_k = \mathrm{P}AC_1, \ldots, C_k \times \mathrm{P}BC_1, \ldots, C_k.$$

As this notation relies on conditional probabilities, Bayes's theorem is required, hence the name Bayesian network. The notation for this is $A \perp\!\!\!\perp B | C_1, \ldots, C_k$. The conditional independence property of DAGs is what makes the Bayesian network so useful in practice. This is especially the case when working with a large number of variables. By conditioning on certain nodes, sets of other nodes can become independent, which is of great help in interpreting the results.

Gaussian Graphical Model

A common type of Bayesian network model is the Gaussian graphical model (GGM). This model assumes that the joint distribution of all variables is multivariate normal: $X \sim N\mu, \Sigma$. In this case, the inverse of the variance/covariance matrix Σ immediately provides the values for the edges of the network. Where Σ itself is used to compute the correlations between variables, $\Sigma - 1$ is used to compute the so-called partial correlations. The normality assumption of the GGM can be checked in a similar way as checking this assumption for standard linear regression models.

There are two methods for visualizing the edges of a graph: (1) lines and arrows are either present or absent, denoting the presence or absence of a relation, or (2) lines and arrows vary in thickness, denoting the strength of the relation, with another property (different line color or using dashed lines) to indicate when relations are negative. As partial correlations in practice never are exactly zero, raw visualizations of a GGM can yield visual overload: when all edges are drawn, it is difficult to assess where the interesting edges are. A straightforward solution to this is called *thresholding*: visualize only those partial correlations that, in absolute value, exceed some threshold (e.g., 0.2). Correlations smaller than this threshold are deemed to be of too little interest. More sophisticated solutions, such as the graphical lasso, might provide a more elegant graph.

Bayesian Networks in the Social Sciences

Within the social sciences, Bayesian networks occur in two classes: social networks and psychological networks. In social network analysis, the nodes represent entities, such as Facebook users, and the edges represent 0/1 variables, such as indicating whether or not two Facebook users are (mutual) friends. The interest does not lie in specific individuals in the network but on the behavior of the network as a whole. Relevant concepts in this field include density (how many edges are there, relative to the possible number of edges) and centrality (which nodes are most connected, who are the key influencers).

Psychological network analysis has a different focus. Here, the nodes concern psychological variables, such as symptoms of anxiety disorder, and the edges represent the partial correlations. Thus, the values of the edges are now realizations of a random process, rather than observed 0/1 scores. One of the main aims in this branch of network analysis is to find pathways connecting one symptom to another. By intervening on one of the symptoms on this path, and thus keeping this constant, one hopes to avoid having other symptoms worsen.

Especially thanks to recent innovations in software for network analysis, the use of network models has increased rapidly since the first decade of the 21st century.

Casper J. Albers

See also Bayes's Theorem; Network Analysis; Network Visualization

Further Readings

Bhushan, N., Mohnert, F., Sloot, D., Jans, L., Albers, C. J., & Steg, L. (2019). Using a Gaussian graphical model to explore relationships between items and variables in environmental

psychology research. *Frontiers in Psychology*, *10*, 1050. doi:10.3389/fpsyg.2019.01050.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. doi:10.1146/annurev-clinpsy-050212-185608.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*, 1–18. doi:10.18637/jss.v048.i04.

Jones, P. J., Mair, P., & McNally, R. J. (2018). Visualizing psychological networks: A tutorial in R. *Frontiers in Psychology*, 3389. doi:10.3389/fpsyg.2018.01742.

Lauritzen, S. L. (1996). *Graphical models*. Oxford, UK: Clarendon Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Van Duijn, M. A., & Vermunt, J. K. (2006). What is special about social network analysis? *Methodology*, *2*(1), 2–6.

Behavior Analysis Design

Behavior analysis is a specific scientific approach to studying behavior that evolved from John Watson's behaviorism and the operant research model popularized by B. F. Skinner during the middle of the 20th century. This approach stresses direct experimentation and measurement of observable behavior. A basic assumption of behavior analysis is that behavior is malleable and controlled primarily by consequences. B. F. Skinner described the basic unit of behavior as an *operant*, a behavior emitted to operate on the environment. Additionally, he proposed that the response rate of the operant serve as the basic datum of the scientific study of behavior. An operant is characterized by a response that occurs within a specific environment and produces a specific consequence. According to the principles of operant conditioning, behavior is a function of three interactive components, illustrated by the three-term contingency: context, response, and consequences of behavior. The relationship between these three variables forms the basis of all behavioral research. Within this framework, individual components of the three-term contingency can be studied by manipulating experimental context, response requirements, or consequences of behavior. A change in any one of these components often changes the overall function of behavior, resulting in a change in future behavior. If the consequence strengthens future behavior, the process is called *reinforcement*. If future behavior is weakened or eliminated as a result of changing the consequence of behavior, the process is called *punishment*.

Behavior analysis encompasses two types of research: the experimental analysis of behavior, consisting of research to discover basic underlying behavioral principles, and applied behavior analysis, involving research implementing basic principles in real-world situations. Researchers in this field are often referred to as behavior analysts, and their research can take place in both laboratory and naturalistic settings and with animals and humans. Basic behavioral processes can be studied in any species, and the findings may be applied to other species. Therefore, researchers can use animals for experimentation, which can increase experimental control by eliminating or reducing confounding variables. Since it is important to verify that findings generalize across species, experiments are often replicated with other animals and with humans. Applied behavior analysis strives to develop empirically based interventions rooted in principles discovered through basic research. Many empirically based treatments have been developed with participants ranging from children with autism to corporate executives and to students and substance abusers. Contributions have been made in developmental disabilities, intellectual developmental disorders, rehabilitation, delinquency, mental health, counseling, education and teaching, business and industry, and substance abuse and addiction, with potential in many other areas of social significance. Similar designs are employed in both basic and applied research, but they differ with regard to subjects studied, experimental settings, and degree of environmental control.

Regardless of the subject matter, a primary feature of behavior analytic research is that the behavior of individual organisms is examined under conditions that are rigorously controlled. One subject can provide a representative sample, and studying an individual subject thoroughly can sometimes provide more information than can studying many subjects because each subject's data are considered an independent replication. Behavior analysts demonstrate the reliable manipulation of behavior by changing the environment. Manipulating the environment allows researchers to discover the relationships between behavior and environment. This method is referred to as *single-subject* or *within-subject research* and requires unique designs, which have been outlined by James Johnston and Hank Pennypacker. Consequently, this method takes an approach to the collection, validity, analysis, and generality of data that is different from approaches that primarily use group designs and inferential statistics to study behavior.

Measurement Considerations

Defining Response Classes

Measurement in single-subject design is objective and restricted to observable phenomena. Measurement considerations can contribute to behavioral variability that can obscure experimental effects, so care must be taken to avoid potential confounding variables. Measurement focuses on targeting a *response class*, which is any set of responses that result in the same environmental change. Response classes are typically defined by function rather than topography. This means that the form of the responses may vary considerably but produce the same result. For example, a button can be pressed several ways, with one finger, with the palm, with the toe, or with several fingers. The exact method of action is unimportant, but any behavior resulting in button depression is part of a response class. Topographical definitions are likely to result in classes that include some or all of several functional response classes, which can produce unwanted variability. Researchers try to arrange the environment to minimize variability within a clearly defined response class.

There are many ways to quantify the occurrence of a response class member. The characteristics of the behavior captured in its definition must suit the needs of the experiment, be able to address the experimental question, and meet practical limits for observation. In animal studies, a response is typically defined as the closing of a circuit in an experimental chamber by depressing a lever or pushing a key or button. With this type of response, the frequency and duration that a circuit is closed can be recorded. Conditions can be arranged to measure the force used to push the button or lever, the amount of time that occurs between responses, and the latency and accuracy of responding in relation to some experimentally arranged stimulus. These measurements serve as dependent variables. In human studies, the response is typically more broadly defined and may be highly individualized. For example, self-injurious behavior in a child with autism may include many forms that meet a common definition of minimum force that leaves a mark. Just as in basic research, a variety of behavioral measurements can be used as dependent variables. The response class must be sensitive to the influence of the independent variable (IV) without being affected by extraneous variables so that effects can be detected. The response class must be defined in such a way that researchers can clearly observe and record behavior.

Observation and Recording

Once researchers define a response class, the methods of observation and recording are important in order to obtain a complete and accurate record of the subject's behavior. Measurement is direct when the focus of the experiment is the same as the phenomenon being measured. Indirect measurement is typically avoided in behavioral research because it undermines experimental control. Mechanical, electrical, or electronic devices can be used to record responses, or human observers can be selected and trained for data collection. Machine and human observations may be used together throughout an experiment. Behavior is continuous, so observational procedures must be designed to detect and record each response within the targeted response class.

Experimental Design and Demonstration of Experimental Effects

Experimental Arrangements

The most basic single-subject experimental design is the baseline–treatment sequence, the AB design. This procedure cannot account for certain confounds, such as maturation, environmental history, or unknown extraneous variables. Replicating components of the AB design provide additional evidence that the IV is the source of any change in the dependent measure. Replication designs consist of a *baseline* or *control* condition (A), followed by one or more *experimental* or *treatment* conditions (B), with additional conditions indicated by successive letters. Subjects experience both the control and the experimental conditions, often in sequence and perhaps more than once. An ABA design replicates the original baseline, while an ABAB design replicates the baseline and the experimental conditions, allowing researchers to infer causal relationships between variables. These designs can be compared with a light switch. The first time one moves the switch from the on position to the off position, one cannot be completely certain that one's behavior was responsible for the change in lighting conditions. One cannot be sure the light bulb did not burn out at that exact moment or the electricity did not shut off coincidentally. Confidence is bolstered when one pushes the switch back to the on position and the lights turn back on. With a replication of moving the switch to off again, one has total confidence that the switch is controlling the light.

Single-subject research determines the effectiveness of the IV by eliminating or holding constant any potential confounding sources of variability. One or more behavioral measures are used as dependent variables so

that data comparisons are made from one condition to another. Any change in behavior between the control and the experimental conditions is attributed to the effects of the IV. The outcome provides a detailed interpretation of the effects of an IV on the behavior of the subject.

Replication designs work only in cases in which effects are reversible. Sequence effects can occur when experience in one experimental condition affects a subject's behavior in subsequent conditions. The researcher must be careful to ensure consistent experimental conditions over replications. Multiple-baseline designs with multiple individuals, multiple behaviors, or multiple settings can be used in circumstances in which sequence effects occur, or as a variation on the AB design. Results are compared across control and experimental conditions, and factors such as irreversibility of effects, maturation of the subject, and sequence effect can be examined.

Behavioral Variability

Variability in single-subject design refers both to variations in features of responding within a single response class and to variations in summary measures of that class, which researchers may be examining across sessions or entire phases of the experiment. The causes of variability can often be identified and systematically evaluated. Behavior analysts have demonstrated that frequently changing the environment results in greater degrees of variability. Inversely, holding the environment constant for a time allows behavior to stabilize and minimizes variability. Murray Sidman has offered several suggestions for decreasing variability, including strengthening the variables that directly maintain the behavior of interest, such as increasing deprivation, increasing the intensity of the consequences, making stimuli more detectable, or providing feedback to the subject. If these changes do not immediately affect variability, it could be that behavior requires exposure to the condition for a longer duration. Employing these strategies to control variability increases the likelihood that results can be interpreted and replicated.

Reduction of Confounding Variables

Extraneous, or confounding, variables affect the detection of behavioral change due to the IV. Only by eliminating or minimizing external sources of variability can data be judged as accurately reflecting performance. Subjects should be selected that are similar along extra-experimental dimensions in order to reduce extraneous sources of variability. For example, it is common practice to use animals from the same litter or to select human participants on the basis of age, level of education, or socioeconomic status. Environmental history of an organism can also influence the target behavior; therefore, subject selection methods should attempt to minimize differences between subjects. Some types of confounding variables cannot be removed, and the researcher must design an experiment to minimize their effects.

Steady State Behavior

Single-subject designs rely on the collection of steady state baseline data prior to the administration of the IV. Steady states are obtained by exposing the subject to only one condition consistently until behavior stabilizes over time. Stabilization is determined by graphically examining the variability in behavior. Stability can be defined as a pattern of responding that exhibits relatively little variation in its measured dimensional quantities over time.

Stability criteria specify the standards for evaluating steady states. Dimensions of behavior such as duration, latency, rate, and intensity can be judged as stable or variable during the course of experimental study, with rate most commonly used to determine behavioral stability. Stability criteria must set limits on two types of variability over time. The first is systematic increases and decreases of behavior, or *trend*, and the second is unsystematic changes in behavior, or *bounce*. Only when behavior is stable, without trend or bounce, should the next condition be introduced. Specific stability criteria include time, visual inspection of graphical data, and simple statistics. Time criteria can designate the number of experimental sessions or discrete period in which behavior stabilizes. The time criterion chosen must encompass even the slowest subject. A time criterion allowing for longer exposure to the condition may needlessly lengthen the experiment if stability occurs rapidly; on the other hand, behavior might still be unstable, necessitating experience and good judgment when a time criterion is used. A comparison of steady state behavior under baseline and different experimental conditions allows researchers to examine the effects of the IV.

Scientific Discovery Through Data Analysis

Single-subject designs use visual comparison of steady state responding between conditions as the primary method of data analysis. Visual analysis usually involves the assessment of several variables evident in graphed data. These variables include upward or

downward trend, the amount of variability within and across conditions, and differences in means and stability both within and across conditions. Continuous data are displayed against the smallest unit of time that is likely to show systematic variability. Cumulative graphs provide the greatest level of detail by showing the distribution of individual responses over time and across various stimulus conditions. Data can be summarized with less precision by the use of descriptive statistics such as measures of central tendency (mean and median), variation (interquartile range and standard deviation), and association (correlation and linear regression). These methods obscure individual response variability but can highlight the effects of the experimental conditions on responding, thus promoting steady states. Responding summarized across individual sessions represents some combination of individual responses across a group of sessions, such as mean response rate during baseline conditions. This method should not be the only means of analysis but is useful when one is looking for differences among sets of sessions sharing common characteristics.

Single-subject design uses ongoing behavioral data to establish steady states and make decisions about the experimental conditions. Graphical analysis is completed throughout the experiment, so any problems with the design or measurement can be uncovered immediately and corrected. However, graphical analysis is not without criticism. Some have found that visual inspection can be insensitive to small but potentially important differences of graphic data. When evaluating the significance of data from this perspective, one must take into account the magnitude of the effect, variability in data, adequacy of experimental design, value of misses and false alarms, social significance, durability of behavior change, and number and kinds of subjects. The best approach to analysis of behavioral data probably uses some combination of both graphical and statistical methods because each approach has relative advantages and disadvantages.

Judging Significance

Changes in level, trend, variability, and serial dependency must be detected in order for one to evaluate behavioral data. *Level* refers to the general magnitude of behavior for some specific dimension. For example, 40 responses per minute is a lower level than 100 responses per minute. *Trend* refers to the increasing or decreasing nature of behavior change. *Variability* refers to changes in behavior from measurement to measurement. *Serial dependency* occurs when a measurement obtained during one time period is related to a value obtained earlier.

Several features of graphs are important, such as trend lines, axis units, number of data points, and condition demarcation. Trend lines are lines that fit the data best within a condition. These lines allow for discrimination of level and may assist in discrimination of behavioral trends. The axis serves as an anchor for data, and data points near the bottom of a graph are easier to interpret than data in the middle of a graph. The number of data points also seems to affect decisions, with fewer points per phase improving accuracy.

Generality

Generality, or how the results of an individual experiment apply in a broader context outside the laboratory, is essential to advancing science. The dimensions of generality include subjects, response classes, settings, species, variables, methods, and processes. Single-subject designs typically involve a small number of subjects that are evaluated numerous times, permitting in-depth analysis of these individuals and the phenomenon in question, while providing *systematic replication*. Systematic replication enhances generality of findings to other populations or conditions and increases internal validity. The *internal validity* of an experiment is demonstrated when additional subjects demonstrate similar behavior under similar conditions; although the absolute level of behavior may vary among subjects, the relationship between the IV and the relative effect on behavior has been reliably demonstrated, illustrating generalization.

Jennifer L. Bredthauer and Wendy D. Donlin-Washington

See also Animal Research; Applied Research; Experimental Design; Graphical Display of Data; Independent Variable; Research Design Principles; Single-Subject Design; Trend Analysis; Within-Subjects Design

Further Readings

Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of methods in human operant behavior* (pp. 3–14). New York: Plenum Press.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *Behavior Analyst*, *21*(1), 111–123.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Mazur, J. E. (2007). *Learning and behavior* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Poling, A., & Grossett, D. (1986). Basic research designs in applied behavior analysis. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 7–27). New York: Plenum Press.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* Boston: Authors Cooperative.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: D. Appleton-Century.

Skinner, B. F. (1965). *Science and human behavior.* New York: Free Press.

Watson, J. B. (1913). Psychology as behaviorist views it. *Psychological Review*, *20*, 158–177.

Behrens–Fisher *t'* Statistic

The Behrens–Fisher t' statistic can be employed when one seeks to make inferences about the means of two normal populations without assuming the variances are equal. The statistic was offered first by W. U. Behrens in 1929 and reformulated by Ronald A. Fisher in 1939:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = t_1 \sin\theta - t_2 \cos\theta,$$

where sample mean $\bar{x}_1$ and sample variance s_1^2 are obtained from the random sample of size n_1 from the normal distribution with mean μ_1 and variance σ_1^2, $t_1 = (\bar{x}_1 - \mu_1)/\sqrt{s_1^2\ n_1}$ has a t distribution with $\nu_1 = n_1 - 1$ degrees of freedom, the respective quantities with subscript 2 are defined similarly, and $\tan\theta = (s_1/\sqrt{n_1})(s_2/\sqrt{n_2})$ or $\theta = \tan^{-1}\left[(s_1/\sqrt{n_1})/(s_2/\sqrt{n_2})\right]$. The distribution of t' is the Behrens–Fisher distribution. It is, hence, a mixture of the two t distributions. The problem arising when one tries to test the normal population means without making any assumptions about their variances is referred to as the Behrens–Fisher problem or as the two means problem.

Under the usual null hypothesis of $H_0: \mu_1 = \mu_2$, the test statistic t' can be obtained and compared with the percentage points of the Behrens–Fisher distribution. Tables for the Behrens–Fisher distribution are available, and the table entries are prepared on the basis of the four numbers $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$, θ, and the Type I error rate α. For example, Ronald A. Fisher and Frank Yates in 1957 presented significance points of the Behrens–Fisher distribution in two tables, one for ν_1 and $\nu_2 = 6,8,12,24,\infty$; $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$; and $\alpha = .05, .01$, and the other for ν_1 that is greater than $\nu_2 = 1,2,3,4,5,6,7$; $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$ and $\alpha = .10, .05, .02, .01$. Seock-Ho Kim and Allan S. Cohen in 1998 presented significance points of the Behrens–Fisher distribution for ν_1 that is greater than $\nu_2 = 2,4,6,8,10,12$; $\theta = 0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ$; and $\alpha = .10, .05, .02, .01$, and also offered computer programs for obtaining tail areas and percentage values of the Behrens–Fisher distribution.

Using the Behrens–Fisher distribution, one can construct the $100(1-\alpha)\%$ interval that contains $\mu_1 - \mu_2$ with

$$\bar{x}_1 - \bar{x}_2 \pm t'_{\alpha/2}(\nu_1, \nu_2, \theta)\sqrt{s_1^2\ /\ n_1\ + s_2^2\ /\ n_2},$$

where the probability that $t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is $\alpha/2$ or, equivalently, $\Pr\left[t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)\right] = \alpha/2$.

This entry first illustrates the statistic with an example. Then related methods are presented, and the methods are compared.

Example

Driving times from a person's house to work were measured for two different routes with $n_1 = 5$ and $n_2 = 11$. The ordered data from the first route are 6.5, 6.8, 7.1, 7.3, 10.2, yielding $\bar{x}_1 = 7.580$ and $s_1^2 = 2.237$, and the data from the second route are 5.8, 5.8, 5.9, 6.0, 6.0, 6.0, 6.3, 6.3, 6.4, 6.5, 6.5, yielding $\bar{x}_2 = 6.136$ and $s_2^2 = 0.073$. It is assumed that the two independent samples were drawn from two normal distributions having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. A researcher wants to know whether the average driving times differed for the two routes.

The test statistic under the null hypothesis of equal population means is $t' = 2.143$ with $\nu_1 = 4$, $\nu_2 = 10$, and $\theta = 83.078$. From the computer program, $\Pr(t' > 2.143) = .049$, indicating the null hypothesis cannot be rejected at $\alpha = .05$ when the alternative hypothesis is nondirectional, $H_a: \mu_1 \neq \mu_2$, because $p = .098$. The corresponding 95% interval for the population mean difference is $[-0.421, 3.308]$.

Related Methods

The Student's t test for independent means can be used when the two population variances are assumed to be equal and $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\ /\ n_1\ + s_p^2\ /\ n_2}},$$

where the pooled variance that provides the estimate of the common population variance σ^2 is defined as $s_p^2 = \left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]\ /\ (n_1 + n_2 - 2)$. It has a t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. The example data yield the Student's $t = 3.220$, $\nu = 14$, the two-tailed $p = .006$, and the 95% confidence interval of [0.482, 2.405]. The null hypothesis of equal

population means is rejected at the nominal $\alpha = .05$, and the confidence interval does not contain 0.

When the two variances cannot be assumed to be the same, one of the solutions is to use the Behrens–Fisher t' statistic. There are several alternative solutions. One simple way to solve the two means problem, called the smaller degrees of freedom t test, is to use the same t' statistic that has a t distribution with different degrees of freedom

$$t' \sim t\left[\min(\nu_1, \nu_2)\right],$$

where the degrees of freedom is the smaller value of ν_1 or ν_2. Note that this method should be used only if no statistical software is available because it yields a conservative test result and a wider confidence interval. The example data yield $t' = 2.143$, $\nu = 4$, the two-tailed $p = .099$, and the 95% confidence interval of [–0.427, 3.314]. The null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

B. L. Welch in 1938 presented an approximate t test. It uses the same t' statistic that has a t distribution with the approximate degrees of freedom ν':

$$t' \sim t(\nu'),$$

Where $\nu' = 1 \Big/ \left[c^2 / \nu_1 + (1-c)^2 / \nu_2 \right]$ with $c = (s_1^2/n_1) \Big/ \left[(s_1^2/n_1) + (s_2^2/n_2) \right]$. The approximation is accurate when both sample sizes are 5 or larger. Although there are other solutions, Welch's approximate t test might be the best practical solution to the Behrens–Fisher problem because of its availability from the popular statistical software, including SPSS (an IBM company, formerly called PASW® Statistics) and SAS. The example data yield $t' = 2.143$, $\nu' = 4.118$, the two-tailed $p = .097$, and the 95% confidence interval of [–0.406, 3.293] The null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

In addition to the previous method, the Welch–Aspin t test employs an approximation of the distribution of t' by the method of moments. The example data yield $t' = 2.143$, and the critical value under the Welch–Aspin t test for the two-tailed test is 2.715 at $\alpha = .05$. The corresponding 95% confidence interval is [–0.386, 3.273]. Again, the null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

Comparison of Methods

The Behrens–Fisher t' statistic and the Behrens–Fisher distribution are based on Fisher's fiducial approach. The approach is to find a fiducial probability distribution that is a probability distribution of a parameter from observed data. Consequently, the interval that involves $t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is referred to as the $100(1-\alpha)\%$ fiducial interval.

The Bayesian solution to the Behrens–Fisher problem was offered by Harold Jeffreys in 1940. When uninformative uniform priors are used for the population parameters, the Bayesian solution to the Behrens–Fisher problem is identical to that of Fisher's in 1939. The Bayesian highest posterior density interval that contains the population mean difference with the probability of $1-\alpha$ is identical to the $100(1-\alpha)\%$ fiducial interval.

There are many solutions to the Behrens–Fisher problem based on the frequentist approach of Jerzy Neyman and Egon S. Pearson's sampling theory. Among the methods, Welch's approximate t test and the Welch–Aspin t test are the most important ones from the frequentist perspective. The critical values and the confidence intervals from various methods under the frequentist approach are in general different from those of either the fiducial or the Bayesian approach. For the one-sided alternative hypothesis, however, it is interesting to note that the generalized extreme region to obtain the generalized p developed by Kam-Wah Tsui and Samaradasa Weerahandi in 1989 is identical to the extreme area from the Behrens–Fisher t' statistic.

The critical values for the two-sided alternative hypothesis at $\alpha = .05$ for the example data are 2.776 for the smaller degrees of freedom t test, 2.767 for the Behrens–Fisher t' test, 2.745 for Welch's approximate t test, 2.715 for the Welch–Aspin t test, and 2.145 for the Student's t test. The respective 95% fiducial and confidence intervals are [–0.427, 3.314] for the smaller degrees of freedom test, [–0.421, 3.308] for the Behrens–Fisher t' test, [–0.406, 3.293] for Welch's approximate t test, [–0.386, 3.273] for the Welch–Aspin t test, and [0.482, 2.405] for the Student's t test. The smaller degrees of freedom t test yielded the most conservative result with the largest critical value and the widest confidence interval. The Student's t test yielded the smallest critical value and the shortest confidence interval. All other intervals lie between these two intervals. The differences between many solutions to the Behrens–Fisher problem might be less than their differences from the Student's t test when sample sizes are greater than 10.

The popular statistical software programs SPSS and SAS produce results from Welch's approximate t test and the Student's t test, as well as the respective confidence intervals. It is essential to have a table that contains the percentage points of the Behrens–Fisher distribution or computer programs that can calculate the tail areas and percentage values in order to use the

Behrens–Fisher t' test or to obtain the fiducial interval. Note that Welch's approximate t test may not be as effective as the Welch–Aspin t test. Note also that the sequential testing of the population means on the basis of the result from either *Levene's test* of the equal population variances from SPSS or the folded F test from SAS is not recommended in general because of the complicated nature of control of the Type I error (rejecting a true null hypothesis) in the sequential testing.

Seock-Ho Kim

See also Mean Comparisons; Student's *t* Test; *t* Test, Independent Samples

Further Readings

Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen [A contribution to error estimation with few observations]. *Landwirtschaftliche Jahrbücher*, *68*, 807–837.

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics*, *9*, 174–180.

Fisher, R. A., & Yates, F. (1957). *Statistical tables for biological, agricultural and medical research* (4th ed.). Edinburgh, UK: Oliver and Boyd.

Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Annals of Eugenics*, *10*, 48–51.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd ed.). New York: Wiley.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). New York: Oxford University Press.

Kim, S. H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, *23*, 356–377.

Tsui, K. H., & Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, *84*, 602–607; Correction, *86,* 256.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Belmont Report

The Belmont Report is an authoritative document governing the ethical conduct of research involving human subjects. While it was originally conceived for biomedical and behavioral research, its principles have been employed and expanded on in other disciplinary fields such as philosophy, law, political science, sociology, and computer science. This entry presents an overview of the Belmont Report's history, its three principles and their applications, and the issues that have been raised since its inception.

History

The Belmont Report was preceded by a number of other ethical codes developed since 1945 in order to protect human subjects against abuse. The Nuremberg Code, published in 1947 after the Nuremberg War Crime Trials revealed Nazi doctors' exploitative experiments with concentration camp prisoners, offers a set of guidelines to safeguard human subjects' physical and mental safety during clinical trials. The guidelines laid out by the Nuremberg Code were further expanded by the World Medical Association with the 1964 Declaration of Helsinki, focusing on physicians' ethical conduct when combining research and clinical practice.

A series of disreputable biomedical studies were also being conducted in the United States during the same period, yet not until 1978 did the public outcry become so overwhelming as to warrant congressional action. In the 1950s, a controversial hepatitis study was conducted among children with developmental disabilities attending the Willowbrook State School, deliberately infecting them with active hepatitis to examine the transmission of the disease. That same decade, pregnant women unknowingly participated in an experimental study of a new drug—thalidomide—to treat nausea, which was subsequently proven to cause birth defects. Other dubious biomedical studies followed suit, including an investigation into the spread of cancer by injecting live cancer cells into ailing elderly patients at the Jewish Chronic Disease Hospital in the 1960s and the study of contraceptive pills' efficacy involving unaware and poverty-stricken women in San Antonio, TX, in the 1970s.

The most infamous of the series of unethical medical studies in the United States was the Tuskegee syphilis study, whereby African American male farmers with syphilis were observed, tested, and left untreated between 1932 and 1972 to understand the natural progression of a disease that was known to affect a wider population. The human subjects were promised free medical care while being subjected to placebo treatments without their knowledge. Upon the revelation of the study's methods, the U.S. Congress enacted the National Research Act (Pub. L. 93–348) in 1974, standardizing institutional review boards' ethical oversight of studies with a research component involving human subjects. Additionally, it authorized the appointment of

the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research involving 11 members representing civil society and professionals from the fields of medicine, law, psychology, and ethics.

Following monthly public consultations over a period of 4 years, including intensive deliberations at the Belmont Conference Center in February 1976, the commission formally issued *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* on September 30, 1978. It formally appeared in the Federal Register on April 18, 1979. The document begins by making the distinction between medical practice and research. Whereas the former alludes to the application of widely accepted interventions with reasonable expectations of enhancing health and well-being, the latter is mainly geared toward hypothesis testing and the production of generalizable knowledge. In situations where clinical practice involves an element that can be considered as research, such as when new techniques need to be assessed for safety and efficacy, the Belmont Report requires ethical review and approval. The remaining sections of the report discuss the principles for ethical research involving human subjects and their applications.

Core Principles

The 5,500-word Belmont Report identifies three core principles for the conduct of ethically sound research involving human participants: respect for persons, beneficence, and justice.

Respect for Persons

Respect for persons in the Belmont Report is essentially framed as acknowledging the autonomy of human subjects. Under this principle, human subjects should be treated as autonomous agents capable of making self-directed choices when it comes to their participation in research. Consequently, when their capacity for self-determination is diminished due to immaturity, illness, mental or physical disability, or social disadvantage, additional measures should be taken for their protection.

The direct application of this principle involves securing informed consent. All relevant information regarding the study's purposes, procedures, and associated benefits and risks must be disclosed to human subjects. Respect entails that subjects be allowed to freely ask questions and withdraw from the study without reprisal, and their full understanding of the conditions surrounding their participation must be guaranteed by presenting information in a manner that is clear, organized, and undemanding. Consent must also be given voluntarily and not out of compliance, fear, or manipulation by offering inappropriate or disproportionate rewards. While autonomous decision-making must be assumed and granted to the fullest extent possible, this principle indicates that special measures may be warranted when subjects' capacities for comprehension are restricted or impaired by immaturity, mental or physical afflictions, or language limitations. Such a situation justifies the involvement of authorized third parties to exercise discretion and act in the human subject's best interest.

Beneficence

The Belmont Report's principle of beneficence involves two main dimensions: not doing harm and maximizing possible benefits while minimizing possible harms. This principle invokes the conduct of a risk–benefit analysis. This will allow researchers to make judgments on whether it is appropriate to renounce a potentially beneficial course of action when the harm it may generate outweighs its conceivable benefits. The nature and magnitude of risks and benefits will naturally differ when evaluated at the individual and social level in either the short term or the long term. For instance, study findings may benefit the wider society in the long term in the absence of direct advantages to the research subjects. In such situations, the Belmont Report suggests a systematic and rigorous examination of the nature, probability, and severity of harms and benefits arising from the study. To make such judgments as precise and accurate as possible, the analysis must include clear and explicit descriptions of the study's implications on the subject's psychological, physical, legal, social, and economic well-being. Additionally, there must be a conscious and thorough effort to consider possible alternatives offering comparable benefits while generating the least likelihood and magnitude of risks.

Justice

The Belmont Report conceptualizes justice as the fair distribution of the risks of research across society. It dictates that the benefits generated by publicly funded research must be shared equally among society regardless of wealth or financial status and that the involvement of subjects who are unable to benefit from the outcomes of research be avoided as much as possible.

As a matter of application, this principle manifests as fairness in the procedures and outcomes of selecting research subjects. Justice operates at two levels: individual and social. Individual justice requires researchers to act objectively when choosing subjects for high-risk research, avoiding social, racial, sexual, and cultural

biases. Consequently, this also implies the need for researchers to ensure that the recipients of beneficial studies are not limited to groups that are already privileged. Social justice, on the other hand, avoids the systematic recruitment of vulnerable populations in risk-laden research. The involvement of specific groups such as racial minorities or the underprivileged must be based on the relevance of their demographic characteristics to the problem or condition being studied—not due to their accessibility or likelihood of being coerced to participate. As such, the Belmont Report recommends an order of preference when recruiting subjects for research purposes (such as considering adults before children) and keeping the involvement of vulnerable populations to a minimum, only involving them when the nature of the condition or treatment being examined so requires.

Issues and Contemporary Reinterpretations

The Belmont Report is a critical document that has revolutionized the conduct of research both within and beyond the medical sciences. Its enduring quality can be attributed to its clarity and succinctness, although several issues have been raised since its inception regarding its capacity to comprehensively address the dilemmas arising from human research. Some authors argue that the principles of respect for persons, beneficience, and justice can be applied more broadly than what was stipulated in the Belmont Report. For instance, respect for persons does not only involve securing informed consent prior to research implementation but also ensuring that the participants are treated with dignity and fairness in all phases of the study and throughout the duration of their participation.

The biomedical and behavioral focus of the Belmont Report also limits its scope outside these disciplines. This was acknowledged by the Belmont Report's authors themselves in a footnote within the document and is a limitation that proves particularly salient in community-based research. Participatory methodologies, which are now widely employed in the health and social sciences, employ the term *participants* as opposed to the Belmont Report's *human subjects* to more appropriately capture the nature of the research relationship.

One of Belmont Report's more prominent criticisms is the protective stance it espouses and its negative implications. Categorical declarations of vulnerability are argued to generate unintended harms by promoting paternalism, perpetuating stereotypes, and conflating vulnerability with a lack of autonomy. They are also viewed to lead to undue and unnecessary exclusion of those who would otherwise be interested in participating in research. Current debates on ethical research have evolved from protecting disenfranchised individuals from exploitative biomedical experiments into avoiding subtle or hidden forms of oppression and systemic exclusion as well as enhancing the inclusion and participation of disadvantaged or marginalized populations. Discounting the participation of vulnerable groups, even when well-intended, can also lead to harmful deprivation of data and findings that could otherwise improve existing interventions and services.

While the Belmont Report reflects a necessary response to the social, cultural, and political climate of its time, recent developments have cast doubt on some of its premises in light of contemporary challenges. The processes of globalization and digitalization have brought forth a new set of ethical dilemmas and theoretical tools that challenge the concepts it laid out. The spread of feminist and non-Western bioethical perspectives have broadened the notions of harm and justice, revealing power hierarchies and various forms of bias (gender, ethnic, and cultural, among others) not addressed in the Belmont Report. These alternative paradigms have introduced additional dimensions to existing formulations of ethical and responsible research, including cultural sensitivity, respect for diversity, and epistemic justice. Contemporary ethicists also juxtapose the Belmont Report's individualism against a collectivist position that considers research's risks and benefits to relationships and communal life.

The practice of clinical medicine is also changing in ways that challenge the report's distinction between medical practice as standard treatment with a reasonable expectation of improving health on the one hand and medical research as the systematic process of generating generalizable knowledge on the other. Increasingly, learning, data collection, and research are seen as essential processes of quality service delivery. Given that the Belmont Report stipulates ethical review and oversight for all research-related activity, the question arises whether this should be applied in an all-encompassing way or whether exceptions should be made for learning activities presenting minimal risks. Additionally, some authors have pointed out how nonmaleficence is invoked implicitly in the Belmont Report, arguing for a deliberate inclusion of this principle.

More recently, the emergence of internet and data technologies as a research tool and space has generated possibilities and concerns not anticipated by the Belmont Report. Digital platforms now allow consent to be sought and secured by ticking a box in an online form and

adding a digital signature. While convenient, this raises concerns regarding the extent and depth of the respondent's understanding of the conditions surrounding his or her participation. Digital settings are also particularly susceptible to data breach and misuse, giving rise to harmful outcomes such as reputational damage and privacy infringement that can further result in psychological or social distress. The intense connectivity enabled by online platforms and internet technologies has amassed huge swaths of data that can feed beneficial research, yet it has also led to concerns about surveillance and a lack of transparency as to how these data are accessed and used. For this, additional measures have been recommended especially with regard to anonymization and the protection of participants' personal information during data collection, archiving, and dissemination. Further operationalizing the Belmont Report in the context of digital research, *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research* was published in 2012, adding respect for law and public interest as a fourth principle to specifically tackle issues of transparency and accountability.

Notwithstanding these limitations, the suite of ethical guidelines and frameworks that branched out from the Belmont Report is a testament to its legacy in ethical decision-making. It remains to be seen how these principles will continue to be applied and developed in the name of ethical and responsible research.

Icy F. Anabo, Iciar Elexpuru-Albizuri, and Lourdes Villardón-Gallego

See also Beneficence; Declaration of Helsinki; Ethics in the Research Process; Informed Consent; Justice and Social Science Research; Nuremberg Code; Respect for Persons; Risk in Human Subjects Research

Further Readings

Adashi, E., Walters, L. B., & Menikoff, J. A. (2018). The Belmont Report at 40: Reckoning with time. *American Journal of Public Health*, *108*(10), 1345–1348. doi:10.2105/AJPH.2018.304580.

Bailey, M., Dittrich, D., Kenneally, E., & Maughan, D. (2012). The Menlo Report. *IEEE Security and Privacy*, *10*(2), 71–75. doi:10.1109/MSP.2012.52.

Childress, J. F., Meslin, E. M., & Shapiro, H. T. (Eds.). *Belmont revisited: Ethical principles for research with human subjects*. Washington, DC: Georgetown University Press.

Friesen, P., Kearns, L., Redman, B., & Caplan, A. L. (2017). Rethinking the Belmont Report? *American Journal of Bioethics*, *17*(7), 15–21. doi:10.1080/15265161.2017.1329482.

Kass, E., Faden, R. R., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). The research-treatment distinction: A problematic approach for determining which activities should have ethical oversight. *Ethical Oversight of Learning Health Care Systems*, *43*(1), S4–S15. doi:10.1002/hast.133.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Rice, T. W. (2008). The historical, ethical, and legal background of human-subjects research. *Respiratory Care*, *53*(1), 1325–1329.

Shore, N. (2006). Re-conceptualizing the Belmont Report: A community-based participatory research perspective. *Journal of Community Practice*, *14*(4), 5–26. doi:10.1300/J125v14n04_02.

Beneficence

This entry presents the principle of beneficence in research involving humans. The concept is most concrete as used in medical research, which is the focus of this entry, but the concept is also central to social science research ethics. Beneficence is usefully detailed in the ethical theory of American ethicists Tom L. Beauchamp and James F. Childress. Their theory has been dominant worldwide for 40 years and has continued to develop in the eight editions of their book *Principles of Biomedical Ethics*.

Protection of Human Subjects in Biomedical Research—the Belmont Report

Experimental research on human beings attracted notoriety during the 1946–1947 Doctors' Trial in Nuremberg, Germany, on war crimes in concentration camps. The resulting Nuremberg Code (1947) influenced the Helsinki Declaration (1964), and together these documents offer a basis for the ethical obligations in biomedical research of free and informed consent, risk–benefit analysis, and review by independent committees (Briggle & Mitcham, 2018, pp. 134–138).

In the United States, however, it was the Tuskegee syphilis study in Alabama from 1932 to 1974 that did most to promote public awareness of ethical perspectives in biomedical experiments on human subjects. In that study, poor African American men suffering from syphilis received free medical examinations and food in exchange for participation but were not informed that they were enrolled in an experiment or even that they had syphilis, and they were deprived of effective treatment when penicillin became available in the 1940s (Briggle & Mitcham, 2018, pp. 140–143). Their low socioeconomic status made them vulnerable to manipulation and exploitation. Ethical obligations regarding free and informed consent, not causing harm, and promoting welfare were not met.

After public outrage over the Tuskegee syphilis study, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research published the Belmont Report (1979), providing a general framework of three basic ethical principles for biomedical research: respect for persons, beneficence, and justice (Beauchamp, 2010a, pp. 18–21). Beauchamp joined the employees of the commission in 1976 with the task to draft the report, expressing the views of the commissioners. The report presents a universal perspective on basic ethical principles (Beauchamp, 2010a, pp. 3, 6–9). Respect for persons relates to informed consent and means "that individuals should be treated as autonomous agents" and "that persons with diminished autonomy are entitled to protection" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. Part B, 1). Beneficence is an obligation to "do not harm" and "maximize possible benefits and minimize possible harms" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. Part B, 2, Part C, 2), which means it relates to risk–benefit assessment (Beauchamp, 2010a, pp. 21–22). Justice is an obligation of fairness in distribution of the benefits and burdens of research (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. Part B, 3, Part C, 3), relates to the selection of research subjects, and "requires special levels of protection for vulnerable and disadvantaged parties" (Beauchamp, 2010a, p. 22).

The Belmont Report says that research involving human subjects is often justified by the principle of beneficence but states that considerations of informed consent, fairness in recruitment of research subjects, and risk assessment should set limits for social utility (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. Part B, 2, Part C, 2), which means that the interests of research subjects outweigh those of the benefit to science and society. In his essay "Codes, Declarations, and Other Ethical Guidance for Human Subjects Research: The *Belmont Report*," Beauchamp writes "it is doubtful that the question of how best to control utilitarian balancing was ever resolved by the National Commission" (2010a, pp. 25–26). Furthermore, he and Childress both argued that there were clear boundaries that the commission did not make between the principles of beneficence, nonmaleficence, and respect for autonomy. According to Beauchamp, many writers wrongly presume that the Belmont Report forms the foundation of his and Childress's book *Principles of Biomedical Ethics*, first published in 1979. These two works were written at the same time and affected each other to the benefit of both (Beauchamp, 2010b, pp. 6–7). Beauchamp states that "*Principles of Biomedical Ethics* became the sole work expressing my deepest philosophical convictions about principles" (Beauchamp, 2010b, p. 7).

The Four Principles of Biomedical Ethics

Unlike the Belmont Report, Beauchamp and Childress treat the principle of nonmaleficence as a separate principle not included in the principle of beneficence. They present a general moral framework of four principles for biomedical ethics: beneficence, nonmaleficence, justice, and respect for autonomy. As in the Belmont Report, these principles are not limited to the domain of biomedical ethics but are generally acknowledged as part of a common universal morality (Beauchamp & Childress, 2019, pp. 3–5). The four principles are equally important and presented as prima facie binding (Beauchamp & Childress, 2019, pp. ix, 15). As Beauchamp and Childress (2019, p. 15) write, "A *prima facie* obligation must be fulfilled unless it conflicts with an equal or stronger obligation."

When competing moral considerations conflict in biomedical practice, the principles are specified, weighted, and balanced depending on the particular context in which they are applied (Beauchamp & Childress, 2019, pp. 15–24). Respect for autonomy is an obligation to respect and support autonomous decisions. Insufficiently autonomous persons are protected by the principles of nonmaleficence, beneficence, and justice. Nonmaleficence is an obligation to avoid causing physical and mental harm. Beneficence is an obligation to promote the good, hinder and remove harm and pain, and balance benefits against risks and costs. Justice is an obligation of fairness in the distribution of benefits, risks, and costs (Beauchamp & Childress, 2019, pp. 13, 156–159).

The Principle of Beneficence

The Belmont Report and Beauchamp and Childress's theory both stress that the utilitarian social beneficence of biomedical experimentation should be limited by considerations of informed consent, fairness in recruitment of research subjects, and risk assessment. Beauchamp and Childress analyze two principles of beneficence: utility and positive beneficence (Beauchamp & Childress, 2019, p. 217).

Beauchamp and Childress' principle of utility differs from the classical utilitarian principle of utility, which is an absolute principle. Instead, Beauchamp and Childress defend a principle of utility "as one among a number of equally important prima facie principles" (Beauchamp & Childress, 2019, p. 218). The principle of utility as a prima facie binding principle "can be applied to health policies through tools that analyze and assess benefits relative to costs and risks" (Beauchamp & Childress, 2019, p. 243). These tools are commonly referred to as

cost-effectiveness analysis and *cost-benefit analysis* (Beauchamp & Childress, 2019, p. 251).

Beauchamp and Childress state that morality requires that we take positive steps to contribute to people's welfare (positive beneficence) and not merely abstain from harming them (nonmaleficence; Beauchamp & Childress, 2019, p. 217). They write, "In our view, conflating nonmaleficence and beneficence into a single principle obscures critical moral distinctions as well as different types of moral theory" (Beauchamp & Childress, 2019, p. 156). Positive beneficence requires preventing or removing evil or harm by doing or promoting good, whereas nonmaleficence only requires avoiding intentionally inflicting evil or harm (Beauchamp & Childress, 2019, p. 157). The principle of beneficence therefore requires taking the positive step of performing risk–benefit analysis on biomedical experiments, with the risk–benefit relationship viewed "in terms of a ratio between the probability and magnitude of an anticipated benefit and the probability and magnitude of an anticipated harm" (Beauchamp & Childress, 2019, p. 244).

Mette Ebbesen

See also Belmont Report; Declaration of Helsinki; Ethics in the Research Process

Further Readings

Beauchamp, T. L. (2010a). Codes, declarations, and other ethical guidance for human subjects research: The Belmont Report. In T. L. Beauchamp (Ed.), *Standing on principles. Collected essays* (pp. 18–32). New York, NY: Oxford University Press.

Beauchamp, T. L. (2010b). The origins and evolution of the Belmont report. In T. L. Beauchamp (Ed.), *Standing on principles. Collected essays* (pp. 3–17). New York, NY: Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). New York, NY: Oxford University Press.

Briggle, A., & Mitcham, C. (2018). *Ethics and science. An introduction* (8th ed.). Cambridge, UK: Cambridge University Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Retrieved from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution for a random variable that takes only two possible values, 0 and 1. Examples of events that lead to such a random variable include coin tossing (heads or tails), answers to a test item (correct or incorrect), outcomes of a medical treatment (recovered or not recovered), and so on. Although it is the simplest probability distribution, it provides a basis for other important probability distributions, such as the binomial distribution and the negative binomial distribution.

Definition and Properties

An experiment of chance whose result has only two possibilities is called a *Bernoulli trial* (or *Bernoulli experiment*). Let p denote the probability of success in a Bernoulli trial $(0 < p < 1)$. Then, a random variable X that assigns value 1 for a success with probability p and value 0 for a failure with probability $1-p$ is called a *Bernoulli random variable*, and it follows the Bernoulli distribution with probability p, which is denoted by $X \sim \text{Ber}(p)$. The probability mass function of Ber(p) is given by

$$P(X = x) = p^x(1-p)^{1-x},\ x = 0,1.$$

The mean of X is p, and the variance is $p(1-p)$. Figure 1 shows the probability mass function of Ber(.7). The horizontal axis represents values of X, and the vertical axis represents the corresponding probabilities. Thus, the height is .7 at $X = 1$, and .3 for $X = 0$. The mean of Ber(0.7) is 0.7, and the variance is .21.

Suppose that a Bernoulli trial with probability p is independently repeated for n times, and we obtain a random sample $X_1, X_2, \ldots, X_n$. Then, the number of successes $Y = X_1 + X_2 + \cdots + X_n$ follows the *binomial distribution* with probability p and the number of trials n, which is denoted by $Y \sim \text{Bin}(n, p)$. Stated in the opposite way, the Bernoulli distribution is a special case of the binomial distribution in which the number of trials n is 1. The probability mass function of Bin(n,p) is given by

$$P(Y = y) = \frac{n!}{y!(n-y)!}p^y(1-p)^{n-y},$$
$$y = 0,1,\ldots,n,$$

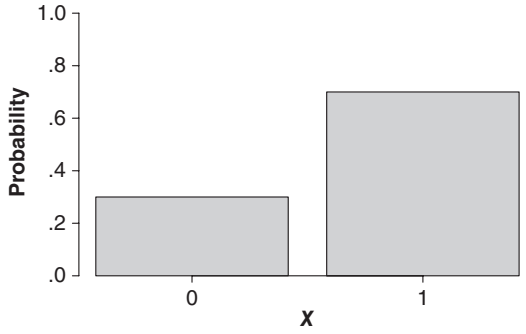


Figure 1 Probability Mass Function of the Bernoulli Distribution With $p = .7$

where $n!$ is the factorial of n, which equals the product $n(n-1)\cdots 2\cdot 1$. The mean of Y is np, and the variance is $np(1-p)$. Figure 2 shows the probability mass function of Bin(10, .7), which is obtained as the distribution of the sum of 10 independent random variables, each of which follows Ber(.7). The height of each bar represents the probability that Y takes the corresponding value; for example, the probability of $Y = 7$ is about .27. The mean is 7 and the variance is 2.1. In general, the distribution is skewed to the right when $p < .5$, skewed to the left when $p > .5$, and symmetric when $p = .5$.

Relationship to Other Probability Distributions

The Bernoulli distribution is a basis for many probability distributions, as well as for the binomial distribution. The number of failures before observing a success t times in independent Bernoulli trials follows the *negative binomial distribution* with probability p and the number of successes t. The *geometric distribution* is a special case of the negative binomial distribution in which the number of failures is counted before observing the first success (i.e., $t = 1$).

Assume a finite Bernoulli population in which individual members are denoted by either 0 or 1. If sampling is done by randomly selecting one member at each time *with replacement* (i.e., each selected member is returned to the population before the next selection is made), then the resulting sequence constitutes independent Bernoulli trials, and the number of successes follows the binomial distribution. If sampling is done at random but *without* replacement, then each of the individual selections is still a Bernoulli trial, but they are no longer independent of each other. In this case, the number of successes follows the *hypergeometric distribution,* which is specified by the population probability p, the number of trials n, and the population size m.

Various approximations are available for the binomial distribution. These approximations are extremely useful when n is large because in that case the factorials in the binomial probability mass function become prohibitively large and make probability calculations tedious. For example, by the central limit theorem, $Z = (Y - np) / \sqrt{np(1-p)}$ approximately follows the standard normal distribution $N(0, 1)$ when $Y \sim \text{Bin}(n,p)$. The constant 0.5 is often added to the denominator to improve the approximation (called *continuity correction*). As a rule of thumb, the normal approximation works well when either (a) $np(1 - p) > 9$ or (b) $np > 9$ for $0 < p \le .5$. The Poisson distribution with parameter np also well approximates Bin(n,p) when n is large and p is small. The Poisson approximation works well if $n^{0.31}p > .47$; for example, $p > .19$, .14, and .11 when $n = 20$, 50, and 100, respectively. If $n^{0.31}p \ge .47$, then the normal distribution gives better approximations.

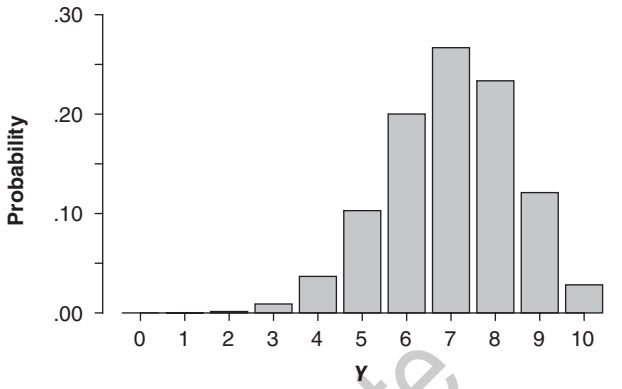


Figure 2 Probability Mass Function of the Binomial Distribution With $p = 7$ and $n = 10$

Estimation

Inferences regarding the population proportion p can be made from a random sample $X_1, X_2, ..., X_n$ from Ber(p), whose sum follows Bin(n,p). The population proportion p can be estimated by the sample mean (or the sample proportion) $\hat{p} = \bar{X} = \sum_{i=1}^{n} X_i / n$, which is an unbiased estimator of p.

Interval estimation is usually made by the normal approximation. If n is large enough (e.g., $n > 100$), a $100(1 - \alpha)\%$ confidence interval is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ is the sample proportion and $z_{\alpha/2}$ is the value of the standard normal variable that gives the probability $\alpha/2$ in the right tail. For smaller ns, the quadratic approximation gives better results:

$$\frac{1}{1+z_{\alpha/2}/n}\left(\hat{p}+\frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{\alpha/2}^2}{4n^2}}\right).$$

The quadratic approximation works well if $.1 < p < .9$ and n is as large as 25.

There are often cases in which one is interested in comparing two population proportions. Suppose that we obtained sample proportions $\hat{p}_1$ and $\hat{p}_2$ with sample sizes n_1 and n_2, respectively. Then, the difference between the population proportions is estimated by the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$. Its standard error is given by

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

from which one can construct a 100(1 – α) confidence interval as

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} SE(\hat{p}_1 - \hat{p}_2).$$

Applications

Logistic Regression

Logistic regression is a regression model about the Bernoulli probability and used when the dependent variable takes only two possible values. Logistic regression models are formulated as *generalized linear models* in which the canonical link function is the logit link and the Bernoulli distribution is assumed for the dependent variable.

In the standard case in which there are K linear predictors $x_1, x_2, ..., x_K$ and the dependent variable Y, which represents a Bernoulli random variable (i.e., $Y = 0, 1$), the logistic regression model is expressed by the equation

$$\ln \frac{p(x)}{1-p(x)} = b_0 + b_1 x_1 + \cdots + b_K x_K,$$

where ln is the natural logarithm, $p(x)$ is the probability of $Y = 1$ (or the expected value of Y) given $x_1, x_2, ..., x_K$, and $b_0, b_1, ..., b_K$ are the regression coefficients. The left-hand side of the above equation is called the *logit*, or the *log-odds ratio,* of proportion p. The logit is symmetric about zero; it is positive (negative) if $p > .5$ ($p < .5$), and zero if $p = .5$. It approaches positive (negative) infinity as p approaches 1 (0). Another representation equivalent to the above is

$$p(x) = \frac{\exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}.$$

The right-hand side is called the *logistic* regression function. In either case, the model states that the distribution of Y given predictors $x_1, x_2, ..., x_K$ is Ber[$p(x)$], where the logit of $p(x)$ is determined by a linear combination of predictors $x_1, x_2, ..., x_K$. The regression coefficients are estimated from N sets of observed data $(Y_i, x_{i1}, x_{i2}, ..., x_{iK})$, $i = 1, 2, ..., N$.

The Binomial Error Model

The binomial error model is one of the measurement models in the classical test theory. Suppose that there are n test items, each of which is scored either 1 (correct) or 0 (incorrect). The binomial error model assumes that the distribution of person i's total score X_i given his or her "proportion-corrected" true score $\zeta_i (0 < \zeta_i < 1)$ is Bin(n, ζ_i):

$$P(X_i = x \mid \zeta_i) = \frac{n!}{x!(n-x)!} \zeta_i^{\,x} (1 - \zeta_i)^{n-x},$$
$$x = 0, 1, ..., n.$$

This model builds on a simple assumption that for all items, the probability of a correct response for a person with true score ζ_i is equal to ζ_i, but the error variance, $n\zeta_i (1 - \zeta_i)$, varies as a function of ζ_i unlike the standard classical test model.

The observed total score $X_i = x_i$ serves as an estimate of $n\zeta_i$, and the associated error variance can also be estimated as $\hat{\sigma}_i^2 = x_i(n - x_i)/(n-1)$. Averaging this error variance over N persons gives the overall error variance $\hat{\sigma}^2 = |\bar{x}(n - \bar{x}) - s^2|/(n-1)$, where $\bar{x}$ is the sample mean of observed total scores over the N persons and s^2 is the sample variance. It turns out that by substituting $\hat{\sigma}^2$ and s^2 in the definition of reliability, the reliability of the n-item test equals the Kuder–Richardson formula 21 under the binomial error model.

History

The name *Bernoulli* was taken from Jakob Bernoulli, a Swiss mathematician in the 17th century. He made many contributions to mathematics, especially in calculus and probability theory. He is the first person who expressed the idea of the law of large numbers, along with its mathematical proof (thus, the law is also called *Bernoulli's theorem*). Bernoulli derived the binomial distribution in the case in which the probability p is a rational number, and his result was published in 1713. Later in the 18th century, Thomas Bayes generalized Bernoulli's binomial distribution by removing its rational restriction on p in his formulation of a statistical theory that is now known as Bayesian statistics.

Kentaro Kato and William M. Bart

See also Logistic Regression; Normal Distribution; Odds Ratio; Poisson Distribution; Probability, Laws of

Further Readings

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). Hoboken, NJ: Wiley.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Beta

Beta (β) refers to the probability of Type II error in a statistical hypothesis test. Frequently, the power of a test, equal to $1-\beta$ rather than β itself, is referred to as a measure of quality for a hypothesis test. This entry discusses the role of β in hypothesis testing and its relationship with significance (α).

Hypothesis Testing and Beta

Hypothesis testing is a very important part of statistical inference: the formal process of deciding whether a particular contention (called the *null hypothesis*) is supported by the data, or whether a second contention (called the *alternative hypothesis*) is preferred. In this context, one can represent the situation in a simple 2×2 decision table in which the columns reflect the true (unobservable) situation and the rows reflect the inference made based on a set of data:

Decision	*Null Hypothesis Is True/Preferred*	*Alternative Hypothesis Is True/Preferred*
Fail to reject null hypothesis	Correct decision	Type II error
Reject null hypothesis in favor of alternative hypothesis	Type I error	Correct decision

The language used in the decision table is subtle but deliberate. Although people commonly speak of accepting hypotheses, under the maxim that scientific theories are not so much proven as *supported* by evidence, we might more properly speak of failing to reject a hypothesis rather than of accepting it. Note also that it may be the case that neither the null nor the alternative hypothesis is, in fact, true, but generally we might think of one as preferable over the other on the basis of evidence. Semantics notwithstanding, the decision table makes clear that there exist two distinct possible types of error: that in which the null hypothesis is rejected when it is, in fact, true; and that in which the null hypothesis is not rejected when it is, in fact, false. A simple example that helps one in thinking about the difference between these two types of error is a criminal trial in the U.S. judicial system. In that system, there is an initial presumption of innocence (null hypothesis), and evidence is presented in order to reach a decision to convict (reject the null hypothesis) or acquit (fail to reject the null). In this context, a Type I error is committed if an innocent person is convicted, while a Type II error is committed if a guilty person is acquitted. Clearly, both types of error cannot occur in a single trial; after all, a person cannot be both innocent and guilty of a particular crime. However, *a priori* we can conceive of the probability of each type of error, with the probability of a Type I error called the significance level of a test and denoted by α, and the probability of a Type II error denoted by β, with $1-\beta$, the probability of not committing a Type II error, called the *power of the test*.

Relationship With Significance

Just as it is impossible to realize both types of error in a single test, it is also not possible to minimize both α and β in a particular experiment with fixed sample size. In this sense, in a given experiment, there is a trade-off between α and β, meaning that both cannot be specified or guaranteed to be low. For example, a simple way to guarantee no chance of a Type I error would be to never reject the null hypothesis regardless of the data, but such a strategy would typically result in a very large β. Hence, it is common practice in statistical inference to fix the significance level at some nominal, low value (usually .05) and to compute and report β in communicating the result of the test. Note the implied asymmetry between the two types of error possible from a hypothesis test: α is held at some prespecified value, while β is not constrained. The preference for controlling α rather than β also has an analogue in the judicial example above, in which the concept of "beyond reasonable doubt" captures the idea of setting α at some low level, and where there is an oft-stated preference for setting a guilty person free over convicting an innocent person, thereby preferring to commit a Type II error over a Type I error. The common choice of .05 for α most likely stems from Sir Ronald Fisher's 1926 statement that he "prefers to set a low standard of significance at the 5% point, and ignore entirely all results that fail to reach that level." He went on to say that "a scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" (Fisher, 1926, p. 504).

Although it is not generally possible to control both α and β for a test with a fixed sample size, it is typically possible to decrease β while holding α constant if the sample size is increased. As a result, a simple way to conduct tests with high power (low β) is to select a sample size sufficiently large to guarantee a specified power for the test. Of course, such a sample size may be prohibitively large or even impossible, depending on the

nature and cost of the experiment. From a research design perspective, sample size is the most critical aspect of ensuring that a test has sufficient power, and *a priori* sample size calculations designed to produce a specified power level are common when designing an experiment or survey. For example, if one wished to test the null hypothesis that a mean μ was equal to μ_0 versus the alternative that μ was equal to $\mu_1 > \mu_0$, the sample size required to ensure a Type II error of β if $\alpha = .05$ is $n = \{\sigma(1.645 - \Phi^{-1}(\beta)) / (\mu_1 - \mu_0)\}^2$, where Φ is the standard normal cumulative distribution function and σ is the underlying standard deviation, an estimate of which (usually the sample standard deviation) is used to compute the required sample size.

The value of β for a test is also dependent on the effect size—that is, the measure of how different the null and alternative hypotheses are, or the size of the effect that the test is designed to detect. The larger the effect size, the lower β will typically be at fixed sample size, or, in other words, the more easily the effect will be detected.

Michael A. Martin and Steven Roberts

See also Hypothesis; Power; *p* Value; Type I Error; Type II Error

Further Readings

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *23*, 503–513.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. New York: Wiley.

Moore, D. (1979). *Statistics: Concepts and controversies*. San Francisco: W. H. Freeman.

Beta Distribution

The beta distribution describes a probability distribution for a continuous random variable, say, *X*, that has the property $0 < X < 1$. Thus, this distribution can be used for modeling proportions, percentages, and other doubly bounded continuous random variables that can be linearly transformed to the (0,1) interval.

Modeling doubly bounded variables requires a distribution that respects the variable's bounds and is able to deal with the fact that central tendency and dispersion in such variables are not independent of each other. Moreover, the distribution shapes should include extreme skew and even *bathtub* shapes. The beta distribution has only two parameters, but it is capable of modeling a considerable variety of distribution shapes. Its flexibility and straightforward interpretability have made the beta distribution the most widely used distribution for modeling variables in (0,1).

This entry begins with definitions of the beta distribution's density and cumulative density functions, and a brief overview of its genesis and its connections with other well-known *probability distributions*. Properties of the beta distribution are then described, with an emphasis on its strengths and limitations for modeling doubly bounded variables. Thereafter, methods of parameter estimation are discussed, including model diagnostics such as residuals and influence statistics. The entry concludes with a brief overview of multivariate distributions whose marginals are beta distributions.

Properties of the Beta Distribution

This section introduces the properties of the beta distribution, including an alternative parameterization that is used in beta regression models. It also describes relevant relationships between the beta distribution and other distributions.

Distribution Function

The probability density function (PDF) of the beta (α, β) distribution is

$$f(x,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad (1)$$

where $0 < x < 1$, $\alpha > 0$, $\beta > 0$, and Γ denotes the gamma function. Equation 1 implies that a beta distribution of $1 - X$ mirror images the distribution of X, that is, $f(1-x,\alpha,\beta) = f(x,\beta,\alpha)$.

The mean of a beta(α, β) distribution is

$$\mu = \alpha/(\alpha+\beta) \quad (2)$$

The variance of a beta distribution is

$$\sigma^2 = \left[\alpha\beta/(\alpha+\beta)^2\right]/(\alpha+\beta+1) = \mu(1-\mu)/(\alpha+\beta+1), \quad (3)$$

which suggests a popular alternative parameterization of the beta distribution, namely a mean, μ, and a precision, $\phi = \alpha + \beta$. From Equation 3, we can see that φ is in the denominator of the variance and so larger φ implies lower σ^2. This parameterization is the one most often used in generalized linear models (GLMs) for beta-distributed dependent variables.

The beta(α, β) distribution can assume a wide variety of shapes. When it has a mode or antimode, this occurs at $x = (\alpha-1)/(\alpha+\beta-2)$. When $\alpha < 1$ and $\beta < 1$, it has a U-shape with modes at 0 and 1 and an antimode. When $\alpha = 1$ and $\beta = 1$, it is flat (the uniform distribution).

When $(\alpha - 1)(\beta - 1) < 0$, the distribution is J- or reverse-J-shaped with no mode or antimode. When $\alpha > 1$ and $\beta > 1$, it has a unimodal shape; and if $\alpha = \beta$, it is symmetric.

The variance, σ^2, of a beta(α, β) variable is less than 1/4 for any α and β. To see this, recall that in Equation 3, the largest possible value of $\mu(1 - \mu)$ is 1/4 (when $\mu = 1/2$), and the denominator in Equation 3 must be >1. The variance has additional constraints. If $\alpha > 1$ and $\beta > 1$, then $\sigma^2 < 1/12$; and if $\alpha < 1$ and $\beta < 1$, then $1/12 < \sigma^2 < 1/4$.

Relations With Other Distributions

While there is no general agreement about the processes generating continuous random variables in (0,1), the beta distribution has direct links to other kinds of random variables whose geneses have widely agreed-on accounts. Two of these that are relevant to the human sciences are as follows:

(1) If Y_1 and Y_2 are gamma-distributed random variables with PDFs $g_1(\alpha, \delta)$ and $g_2(\beta, \delta)$, for $\delta > 0$, then $X = Y_1/(Y_1 + Y_2)$ has a beta(α, β) distribution.

(2) If W_1 and W_2 are χ^2-distributed random variables with PDFs $q_1(\gamma)$ and $q_2(\eta)$, for $\gamma > 0$ and $\eta > 0$, then $X = W_1/(W_1 + W_2)$ has a beta($\gamma/2$, $\eta/2$) distribution. If one or both of W_1 and W_2 are noncentral χ^2 variables, then X follows a noncentral beta distribution whose parameters include the noncentrality parameter(s).

Several other relationships between the beta and other distributions also are relevant for applications in the human sciences:

(1) If X has a beta(1, β) distribution, then it has a Kumaraswamy(1, β) distribution; and if X has a beta(α, 1) distribution, then it has a Kumaraswamy(α, 1) distribution. The Kumaraswamy distribution often is applied to modeling quantiles of X.

(2) If X has a beta(α, β) distribution, then $\alpha X/(\beta(1 - X))$ has a F(2α, 2β) distribution.

(3) If X has a beta(α, 1) distribution, then $-\ln(X)$ has an exponential(α) distribution. There also is an extensive literature on $X = \ln(Y)$ when X has a beta(α, β) distribution.

Parameter Estimation and Applications

This section begins by describing methods for estimating the parameters of the beta distribution and examples of its application. It then moves on to considerations about estimation bias, model checking, and limitations of the beta distribution.

Parameter Estimation

The beta distribution's parameters may be estimated by method of moments or by maximum likelihood techniques. The latter is favored for GLMs with covariates predicting either or both of the parameters. Several authors such as Paolino (2001), Ferrari and Cribari-Neto (2004), and Smithson and Verkuilen (2006) present beta GLMs using the mean-precision parameterization described in Equations 2 and 3. Taking into account that the precision is negatively associated with dispersion, the beta GLM is a location–dispersion model in the sense of Smyth (1989). There are two submodels, typically using the following link functions:

$$\begin{aligned} \log\left(\mu_i/\left(1-\mu_i\right)\right) &= \mathbf{x}_i^{\top}\beta \\ \log\left(\phi_i\right) &= \mathbf{w}_i^{\top}\delta \end{aligned}, \tag{4}$$

where $\mathbf{x}$ and $\mathbf{w}$ are vectors of covariates and β and δ are vectors of coefficients. The two sets of covariates may or may not overlap. Other link functions could be used (e.g., the cauchit or probit for the mean), but the logit and log link functions are the ones implemented in available software at the time this is written.

Applications

Early use of the beta distribution was primarily either as a means to form conjugate priors for binomial random variables or for theoretical purposes, such as its relationship with the F distribution. Only since 1993 has it been used for statistical modeling, as in GLMs.

Doubly bounded random variables occur throughout psychology and cognate areas such as political science, economics, and biology. The most commonplace examples in psychology include proportions and percentages, such as probability judgments, the proportion of the brain's volume occupied by a specific part of the brain, and the proportion of a period of time spent on an activity. Examples from economics include rates such as fractional repayments on debts, market shares, and capital structure. Many psychological scales are doubly bounded, and in some applications, it is sensible to treat the bounds as true scores (rather than as censored scores). Examples of this kind include analyses of Likert-type scale data and some kinds of summative scales (e.g., a quality of life index).

Model Checking and Extensions

Maximum likelihood estimation seems to work well for GLMs using the beta distribution, despite the fact that standard GLM regularity conditions do not apply to the beta distribution because it is not a member of the exponential family. There is some evidence that this may also hold for random-effects as well as fixed-effects models. Daniel Zimprich (2010) fitted a mixed-effects beta GLM, and Jay Verkuilen and Michael Smithson (2012) examined such models in some depth, along with Bayesian Markov chain Monte Carlo estimation approaches. Likewise, Yvonnick Nöel and Bruno Dauvier (2007) developed item-response models for doubly bounded continuous scale items using the beta distribution, which Nöel (2014) extended to unfolding models.

That said, there is some evidence of estimation bias, especially in the precision parameter submodel when sample sizes are small. Bias correction methods suggested by Ioannis Kosmidis and David Firth (2010) have been implemented in some software packages for beta GLMs such as the betareg package in R (Grün, Kosmidis, & Zeileis, 2012).

Model checking in beta GLMs has proved somewhat less straightforward, although dfbetas as influence statistics seem to work reasonably well and conventional residuals such as the Pearson also can be helpful. Espinheira, Ferrari, and Cribari-Neto (2008a, 2008b) discuss issues regarding the uses of influence statistics and residuals for model checking in beta GLMs. At the time of this writing, there is no agreed-on residual for beta GLMs, although several alternatives have been investigated.

Flexible though it is, the beta distribution is limited in its ability to capture some kinds of distribution shapes such as heavy-tailed or bimodal distributions. Some attempts have been made to extend beta GLMs for greater flexibility. Most of these have focused on mixture distribution models. Smithson and Segale (2009) and Smithson, Merkle, and Verkuilen (2011) employed mixture models for analyzing experimental data where they could assume that for at least one component distribution, the location parameter is known *a priori*. More generally, Hahn (2008) introduced the beta rectangular distribution, a mixture of a uniform and a beta distribution; and Migliorati, Di Brisco, and Ongaro (2018) presented a mixture of two beta distributions with arbitrary means but common variance. These mixture distributions yield identifiable GLMs with bounded likelihoods that have a greater variety of density shapes than the beta, especially in tail behavior and bimodality.

A major limitation of the beta distribution in applications to real data is the fact that its density is not defined at 0 or at 1. In applications where the data contain 0s and 1s, this limitation has been problematic. One solution is a so-called *0-1-inflated model* (Ospina & Ferrari, 2012). Technically, this is a hurdle model, a mixture distribution with degenerate probability masses at 0 and 1 and the beta distribution for modeling data in the (0,1) interval.

Multivariate Extensions

There are two types of multidimensional extension of the beta distribution (i.e., multivariate distributions with beta marginals): compositional, where the variables must sum to 1 across dimensions, and noncompositional. The classic compositional distribution with beta marginals is the Dirichlet:

$$f(x,\eta)=\frac{1}{B(\eta)}\prod_{k=1}^{K}x_i^{\eta_k-1}, \tag{5}$$

where

$$B(\eta)=\frac{\prod_{k=1}^{K}\Gamma(\eta_k)}{\Gamma\left(\sum_{k=1}^{K}\eta_k\right)},\eta_k>0 \tag{6}$$

and $\sum_{k=1}^{K}x_k=1$. The covariances in the Dirichlet PDF all are negative:

$$\sigma_{jk}=-\frac{\eta_j\eta_k}{\eta_0^2(\eta_0+1)}. \tag{7}$$

The marginal PDFs for the Dirichlet distribution are $\text{beta}(\eta_k,\eta_0-\eta_k)$, where $\eta_0=\sum_{k=1}^{K}\eta_k$. Dirichlet regression models have been proposed and implemented in software.

Noncompositional multivariate extensions of the beta distribution have been provided in two forms. One is the standard multilevel (or mixed) model approach, as described by Verkuilen and Smithson (2012). The other is to use copulas to model the dependency structure separately from the marginal PDFs. Copulas are multivariate cumulative distribution functions with uniform marginal distributions, so a copula model often is estimated in two stages. The marginal PDFs are modeled and their parameter estimates are fed to their respective quantile functions, whose output then yields a multivariate distribution with uniform marginals. A copula model then is estimated using this multivariate distribution as input.

Michael Smithson

See also Distribution; Logistic Regression; Loglinear Models; Normal Distribution

Further Readings

Balakrishnan, N., & Lai, C. D. (2009). *Continuous bivariate distributions*. Berlin, Germany: Springer Science & Business Media.

Espinheira, P. L., Ferrari, S. L., & Cribari-Neto, F. (2008a). Influence diagnostics in beta regression. *Computational Statistics & Data Analysis*, *52*(9), 4417–4431.

Espinheira, P. L., Ferrari, S. L., & Cribari-Neto, F. (2008b). On beta regression residuals. *Journal of Applied Statistics*, *35*(4), 407–419.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Grün, B., Kosmidis, I., & Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, *48*(11), 1–25.

Gupta, A. K., & Nadarajah, S. (Eds.). (2004). *Handbook of beta distribution and its applications*. Boca Raton, FL: CRC Press.

Hahn, E. D. (2008). Mixture densities for project management activity times: A robust approach to PERT. *European Journal of Operational Research*, *188*(2), 450–459.

Kosmidis, I., & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, *4*, 1097–1112.

Maier, M. J. (2014). *DirichletReg: Dirichlet regression for compositional data in R*. Research Report Series. Vienna, Austria: Department of Statistics and Mathematics, Vienna University of Economics and Business. Retrieved from http://epub.wu.ac.at/4077/

Migliorati, S., Di Brisco, A. M., & Ongaro, A. (2018). A new regression model for bounded responses. *Bayesian Analysis*, *13*(3), 845–872.

Nöel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika*, *79*(4), 647–674.

Nöel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47–73. doi:10.1177/0146621605287691.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, *9*(4), 325–346.

Smithson, M., Merkle, E. C., & Verkuilen, J. (2011). Beta regression finite mixture models of polarization and priming. *Journal of Educational and Behavioral Statistics*, *36*(6), 804–831. doi:10.3102/1076998610396893.

Smithson, M., & Segale, C. (2009). Partition priming in judgments of imprecise probabilities. *Journal of Statistical Theory and Practice*, *3*(1), 169–181. doi:10.1080/15598608.2009.10411918.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54. doi:10.1037/1082-989X.11.1.54.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *51*(1), 47–60.

Verkuilen, J., & Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, *37*(1), 82–113.

Zimprich, D. (2010). Modeling change in skewed variables using mixed beta regression models. *Research in Human Development*, *7*(1), 9–26. doi:10.1080/15427600903578136.

Between-Subjects Design

See Single-Case Research Design; Within-Subjects Design

Bias

Bias is systematic error that occurs in the research enterprise. All research includes some forms of bias, and investigators are responsible for recognizing such potential when sharing their findings. Naming how different types of bias may or may not affect the quality of research conclusions is a central feature of the validation process. Consider, for example, what happens when comparing the behaviors of people who are classified into different groups and studied over time. If the groups are unequal at the beginning of a study, any subsequent variance between the groups cannot be explained using only the new research evidence. Uncontrolled and unnamed biases may be the cause of any differences between groups or changes that might occur over time. To minimize distortion in any research, it is incumbent on all investigators to define and explore the qualities and influence of bias.

When bias is not considered in the interpretation process, the credibility of conclusions is called into question. Thus, investigators within and across disciplines have collaborated to name the most common forms of biases they encounter. Studies of bias can be classified into one of three nested categories, each of which is explored by considering particular disciplinary assumptions and norms. Bias can be the result of distortions in theoretical logic, the design of a study, and/or the measurement and evaluation of variables.

Theoretical Bias

Bias in the theoretical logic of a research program or study involves flaws in critical thinking and the interpretation of evidence. This type of bias is commonly

detected by reviewing existing literature and looking for systematic flaws in the explicit and implicit assumptions generated as part of the research process. Advocates of a particular theory commonly evaluate the qualities of the arguments generated and the evidence used to support those arguments. Across theories, it is also possible to identify common argument forms and fallacies in how premises are generated and tested.

Generally speaking, arguments can take valid and invalid forms. Philosophers of science endeavor to distinguish between these types of arguments when naming bias. Pennock (2019), for example, described how the quest for truth simultaneously guides curiosity and generates moral norms that investigators are bound to imagine when conducting research. Vaughn (2019) named some of the common fallacies that are found in research and described valid and invalid argument forms.

Valid research, in Vaughn's model, systematically tests whether antecedent assumptions can be affirmed, whether consequences are explored well enough to disprove them should they be untrue, and whether hypothetical syllogisms comprised of two or more premises can be verified or refuted. Invalid research instigates bias when consequences are assumed to support an initial, unsubstantiated premise. Likewise, bias would be likely when consequences are assumed to be true even though initial antecedents are not supported. Starting a study with unequal groups, for example, could yield either of these two forms of argumentative bias if investigators did not find a way to control for such inequalities.

A second form of theoretical bias is apparent when the content of operating premises or assumptions is laden with fallacies. Fallacies in critical thinking are easy to overlook when investigators become too wedded to a particular set of beliefs or premises. Some fallacies are grounded in irrelevant premises (Vaughn, 2019). For example, appeals to ignorance about a topic or to tradition introduce bias in any truth claims because they are grounded in a lack of evidence. Likewise, raising irrelevant issues or distorting, weakening, or oversimplifying a theoretical premise serve as distractions because supporting evidence is unavailable.

Other fallacies are grounded in unacceptable premises. Creating a false dilemma or using the conclusion of an argument as a premise are two ways in which investigators can add theoretical inaccuracies into a quest for truth. Similarly, slippery slope claims about undesirable consequences or hasty generalizations using an inadequate sample to support claims about an entire group are common forms of theoretical bias. Taken as a whole, these forms of bias are sometimes called *experimenter expectancy effects* or *generalizability threats*. Biases related to the era in which some studies are conducted and/or to changes that may not be attributable to the research are additional threats to theoretical validity, so much so that entire disciplines have been formulated to better understand the consequences of these biases.

Bias in Research Design

Research designs, the relations between variables under investigation, can be biased to such an extent that they cannot yield valid results even when theoretical claims are sound. Two classic explorations of how to address biases in experimental and quasi-experimental research design illustrate how investigators control distortion (Campbell & Stanley, 1963; Cook & Campbell, 1979). Since then, there has been a proliferation of studies on the biases associated with a broad range of research designs. Naming and controlling for internal and external biases that potentially distort research evidence strengthens the truth value of conclusions. Biases in research design, therefore, are commonly depicted as threats to internal and external validity.

Some of the most common sources of bias in research design proliferate from misunderstandings about the concept of randomization. Randomization, the reliance on chance-driven procedures when making decisions, is often used as a practical solution to bias that stems from uncontrollable error. In principle, when simple random sampling is repeated across multiple implementations of the same biased procedures, the repetition will result in a normal distribution of uncontrollable error. This procedure controls for bias when the same study is repeated often enough for normal distributions to emerge in the evidence. Yet, few studies are actually replicated often enough for simple random sampling to eliminate uncontrollable error from a research design.

Despite its usefulness in some situations, randomization often results in unequal groups when used in isolation, so much so that a number of design biases may be inferred. External to a study, for example, the participants selected for inclusion in one sample may not fully represent the distribution of members in the targeted population: Subsets of a population may be excluded or oversampled when participants are randomly selected. Names such as selection–treatment interaction as well as diffusion, compensatory equalization, or compensatory rivalry of treatments have been used to depict some of these external biases.

Internal to a study, a second source of bias is likely when the members of any sample are randomly assigned to one group or another: Subsets of the sample may be missing from one group and overrepresented in another. When left uncontrolled, these decisions results in exponential forms of error that can yield biased conclusions. Names such as selection–maturation interaction,

resentful demoralization, and experimental mortality have been used to depict some of these internal biases.

Investigators address both random selection and random assignment biases by identifying those forms of distortion that might undermine their theoretical claims and adding controls for such distortion into their research designs. Studies of bias in research design differ in the extent to which they focus on interventions or descriptive depictions of the concepts and constructs and the settings in which such observations occur (Larzelere, Kuhn, & Johnson, 2004; Rosenthal, 2002). Ideally, choices for addressing biases in research design become progressively more sophisticated as solutions to research problems become more truthful, and as truthful premises are accurately distinguished from false premises.

Measurement and Evaluation Bias

A third category of bias focuses on the use of data to answer research questions. First, the instrumentation used to track, measure, and interpret key concepts, constructs, or variables includes bias. Second, the methods used to aggregate data, compare variables, and answer research questions include sources of bias that warrant consideration.

The detection of measurement and evaluation biases hinge on how a research program balances questions of *objectivity*, *individuality*, and *solidarity* when making decisions. Criteria for objectivity lead investigators to rely on standardized tools that will yield predictable outcomes on repeated use, representing bias as any deviation from such standards. Criteria for individuality lead investigators to look for the relative uniqueness of each measurement encounter, expressing bias as distortion in the description of such uniqueness. Criteria for solidarity lead investigators to look for commonalities across measurement opportunities, depicting bias as inaccuracies in how such measurements are recorded and aggregated.

Investigators who endeavor to generate strong predictions or control outcomes place the strongest emphasis on objectivity when they construct measurement plans. They tend to rely on random sampling theory and address its concomitant forms of bias when determining what to measure and how to measure the ideas under investigation. *Psychometric measurement* approaches, for example, depend heavily on probability theory and randomization, drawing comparisons between hypothetical true scores and biased observed scores. In such work, bias is defined as a combination of known error and unknown error. Biases that are salient in individuals' reports of their own and others' experience is perhaps the most frequently studied form of psychometric bias.

Developmental measurement and its corresponding bias account for change as it emerges in the structure and function of what is to be measured and/or in how people evolve over time. Individuality and solidarity are sometimes emphasized more than objectivity, but an ideal research plan balances all three concerns. Specialized applications of random sampling theory and its concomitant biases are used when the constructs can be defended and measured using standardized tools to track changes over time. Additional measurement approaches allow for the detection of structural changes in what is to be measured or to detect particular functions and change mechanisms, but each includes detectable biases. Detection of the physical and social-emotional changes that occur as adolescents progress through puberty, for example, requires respect for the highly individualized nature of hormonal change as well as aggregated evidence depicting age-related commonalities. Bias proliferates when measures designed for one purpose are used for another but is addressed when the limits of each tool are considered when interpreting results.

Interpretive measurement occurs when investigators celebrate individuality and narrow forms of solidarity by investigating bias (Thorkildsen, 2005). Interpretive research focuses on rule-governed descriptions of everyday events such as describing the world, challenging assumptions, discovering contrasting evidence, clarifying the dimensions of particular constructs, and resisting attempts to reify truth claims. Seeking dependable, credible evidence, investigators distinguish the respectable bias needed to place reasonable parameters on the scope of a research project and distorting bias that should be minimized. Recording observations with precision and relying on multiple methods of interpreting results help to restrict problematic bias.

The final source of bias emerges when investigators use statistical analyses and descriptive methods to report their findings. When tools of analysis do not align well with the instrumentation decisions, bias can yield invalid conclusions. When analytical tools align with the research questions and accommodate the types of biases embedded in the instrumentation process, truthful conclusions are likely.

Research as a Study of Bias

Depicting bias as systematic error places pressure on investigators to define truth as often as they offer claims about distortion. Disciplinary assumptions and the nature of the theories under consideration help investigators identify the parameters for determining bias by evaluating the qualities of arguments for their relative truth-value. Thoughtful exploration of design limitations and how well evidence yields informative truth places the study of bias at the heart of conducting strong research. Accepting that bias is a feature of all

research constrains curiosity and pressures investigators to consider the moral consequences of their claims.

Theresa A. Thorkildsen

See also Cluster Sampling; Experimenter Expectancy Effect; Response Bias; Sampling; Systematic Error; Validity of Measurement

Further Readings

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An unrecognized confound in intervention research. *Psychological Bulletin*, *130*, 289–303. doi:10.1037/0033-2909.130.2.289.

Pennock, R. T. (2019). *An instinct for truth: Curiosity and the moral character of science*. Cambridge, MA: MIT Press.

Thorkildsen, T. A. (2005). *Fundamentals of measurement in applied research*. New York, NY: Allyn & Bacon.

Vaughn, L. (2019). *The power of critical thinking: Effective reasoning about ordinary and extraordinary claims* (6th ed.). New York, NY: Oxford University Press.

Biased Estimator

In many scientific research fields, statistical models are used to describe a system or a population, to interpret a phenomenon, or to investigate the relationship among various measurements. These statistical models often contain one or multiple components, called *parameters*, that are unknown and thus need to be estimated from the data (sometimes also called the *sample*). An estimator, which is essentially a function of the observable data, is biased if its expectation does not equal the parameter to be estimated.

To formalize this concept, suppose θ is the parameter of interest in a statistical model. Let $\hat{\theta}$ be its estimator based on an observed sample. Then $\hat{\theta}$ is a biased estimator if $E\left(\hat{\theta}\right) \neq \theta$, where E denotes the expectation operator. Similarly, one may say that $\hat{\theta}$ is an unbiased estimator if $E\left(\hat{\theta}\right) = \theta$. Some examples follow.

Example 1

Suppose an investigator wants to know the average amount of credit card debt of undergraduate students from a certain university. Then the population would be all undergraduate students currently enrolled in this university, and the population mean of the amount of credit card debt of these undergraduate students, denoted by θ, is the parameter of interest. To estimate θ, a random sample is collected from the university, and the sample mean of the amount of credit card debt is calculated. Denote this sample mean by $\hat{\theta}_1$. Then $E\left(\hat{\theta}_1\right) = \theta$; that is, $\hat{\theta}_1$ is an unbiased estimator. If the largest amount of credit card debt from the sample, call it $\hat{\theta}_2$, is used to estimate θ, then obviously $\hat{\theta}_2$ is biased. In other words, $E\left(\hat{\theta}_2\right) \neq \theta$.

Example 2

In this example a more abstract scenario is examined. Consider a statistical model in which a random variable X follows a normal distribution with mean μ and variance σ^2, and suppose a random sample $X_1,\ldots,X_n$ is observed. Let the parameter θ be μ. It is seen in Example 1 that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean of $X_1,\ldots,X_n$, is an unbiased estimator for θ. But $\bar{X}^2$ is a biased estimator for μ^2 (or θ^2). This is because $\bar{X}$ follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. Therefore, $E\left(\bar{X}^2\right) = \mu^2 + \frac{\sigma^2}{n} \neq \mu^2$.

Example 2 indicates that one should be careful about determining whether an estimator is biased. Specifically, although $\hat{\theta}$ is an unbiased estimator for θ, $g\left(\hat{\theta}\right)$ may be a biased estimator for $g(\theta)$ if g is a nonlinear function. In Example 2, $g(\theta) = \theta^2$ is such a function. However, when g is a linear function, that is, $g(\theta) = a\theta + b$ where a and b are two constants, then $g\left(\hat{\theta}\right)$ is always an unbiased estimator for $g(\theta)$.

Example 3

Let $X_1,\ldots,X_n$ be an observed sample from some distribution (not necessarily normal) with mean μ and variance σ^2. The sample variance S^2, which is defined as $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, is an unbiased estimator for σ^2, while the intuitive guess $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ would yield a biased estimator. A heuristic argument is given here. If μ were known, $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ could be calculated, which would be an unbiased estimator for σ^2. But since μ is not known, it has to be replaced by $\bar{X}$. This replacement actually makes the numerator smaller. That is, $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \leq \sum_{i=1}^{n}\left(X_i - \mu\right)^2$ regardless of the value of μ. Therefore, the denominator has to be reduced a little bit (from n to $n-1$) accordingly.

A closely related concept is the bias of an estimator, which is defined as $E(\hat{\theta})-\theta$. Therefore, an unbiased estimator can also be defined as an estimator whose bias is zero, while a biased estimator is one whose bias is nonzero. A biased estimator is said to underestimate the parameter if the bias is negative or overestimate the parameter if the bias is positive.

Biased estimators are usually not preferred in estimation problems, because in the long run, they do not provide an accurate "guess" of the parameter. Sometimes, however, cleverly constructed biased estimators are useful because although their expectation does not equal the parameter under estimation, they may have a small variance. To this end, a criterion that is quite commonly used in statistical science for judging the quality of an estimator needs to be introduced. The mean square error (MSE) of an estimator $\hat{\theta}$ for the parameter θ is defined as $E\left[\left(\hat{\theta}-\theta\right)^2\right]$. Apparently, one should seek estimators that make the MSE small, which means that $\hat{\theta}$ is "close" to θ. Notice that

$$E\left[\left(\hat{\theta}-\theta\right)^2\right]=E\left[\left(\hat{\theta}-E\left\{\hat{\theta}\right\}\right)^2\right]+\left(E\left\{\hat{\theta}\right\}-\theta\right)^2$$
$$=Var\left(\hat{\theta}\right)+Bias^2,$$

meaning that the magnitude of the MSE, which is always nonnegative, is determined by two components: the variance and the bias of the estimator. Therefore, an unbiased estimator (for which the bias would be zero), if possessing a large variance, may be inferior to a biased estimator whose variance and bias are both small. One of the most prominent examples is the shrinkage estimator, in which a small amount of bias for the estimator gains a great reduction of variance. Example 4 is a more straightforward example of the usage of a biased estimator.

Example 4

Let X be a Poisson random variable, that is, $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$, for $x=0,s1,s2,\ldots$. Suppose the parameter $\theta=e^{-2\lambda}$, which is essentially $\left[P(X=0)\right]^2$, is of interest and needs to be estimated. If an unbiased estimator, say $\hat{\theta}_1(X)$, for θ is desired, then by the definition of unbiasedness, it must satisfy $\sum_{x=0}^{\infty}\hat{\theta}_1(x)\frac{e^{-\lambda}\lambda^x}{x!}=e^{-2\lambda}$ or, equivalently, $\sum_{x=0}^{\infty}\frac{\hat{\theta}_1(x)\lambda^x}{x!}=e^{-\lambda}$ for all positive values of λ. Clearly, the only solution is that $\hat{\theta}_1(x)=(-1)^x$. But this unbiased estimator is rather absurd. For example, if $X=10$, then the estimator $\hat{\theta}_1$ takes the value of 1, whereas if $X=11$, then $\hat{\theta}_1$ is 1. As a matter of fact, a much more reasonable estimator would be $\hat{\theta}_2(X)=e^{-2X}$, based on the maximum likelihood approach. This estimator is biased but always has a smaller MSE than $\hat{\theta}_1(X)$.

Zhigang Zhang and Qianxing Mo

See also Distribution; Estimation; Expected Value

Further Readings

Kay, S., & Eldar, Y. C. (2008). Rethinking biased estimation [lecture notes]. *IEEE Signal Processing Magazine*, *25*(3), 133–136. doi:10.1109/MSP.2008.918027.

Mayer, L. S., & Willke, T. A. (1973). On biased estimation in linear models. *Technometrics*, *15*(3), 497–508.

Özkale, M. R., & Arıcan, E. (2016). A new biased estimator in logistic regression model. *Statistics*, *50*(2), 233–253. doi:10.1080/02331888.2015.1123711.

So, S. (2008). *Why is the sample variance a biased estimator?* (p. 9). Queensland, Australia: Griffith University, Tech. Rep.

Binomial Distribution

The binomial distribution describes the results of repeated independent trials of an event for which the outcome space has two possible values (e.g., yes or no, true or false, heads or tails, success or failure), and each trial shares the same probability of success. The binomial distribution represents the probability of different combinations of successes or failures when the experiment is repeated n number of times, and X represents the number of successes in n trials. The probability that a single trial succeeds can be represented with parameter p. The probability that a single trial fails can be represented with parameter q. The sum of p and q must always equal 1. The mean of a binomial distribution is always $\mu = n \times p$, and the variance of X can be approximated by $\sigma^2 = n \times p \times q$. Some statistical inference can be made using the binomial distribution in specific examples. This entry presents a description of the history of the binomial distribution as well as the applications for statistical inference and some specific examples where the binomial distribution can be used.

History

The desire to calculate probabilities in games of chance led to the first studies of binomial distribution. Essentially, mathematically inclined gamblers wanted to calculate their probability of winning on a certain number of dice rolls. In 1713, Jakob Bernoulli, a Swiss

mathematician, published a proof that determined that the probability of X equaling a specified number, x, in n trials was equal to the xth term in the binomial expansion of the expression $(p + q)^n$, thus creating the binomial distribution.

The binomial distribution was used in 1936 to publish evidence of possible scientific chicanery by Gregor Mendel in the famous 1866 pea genetics experiments. Ronald Fisher noted that the reported laws of inheritance in peas would dictate that the number of certain colors of peas would have a binomial distribution, and the results reported by Mendel should have a probability of only about .1.

Applications for Statistical Inference

Any experiment using the binomial distribution has two assumptions: identical trials and independent trials. Identical trials is the assumption that p and q take on the same probability value across n number of trials. The compound binomial distribution does not assume identical trials. Independent trials is the assumption that subsequent trials are not affected by previous outcomes, so in order for the binomial distribution to be used, sampling with replacement must occur.

When the random variable X has parameters n and p in the binomial distribution, it is mathematically represented as X~$B(n, p)$. The probability that X is equal to a certain number, x, where $x = 0, 1, 2, \ldots, n$, is given by the probability mass function:

$$f(x) = P(C = x) = C_x^n p^x q^{n-x},$$

where C_x^n is a mathematical combination, called the *binomial coefficient*, given by the equation:

$$C_x^n = \frac{n!}{x!(n-x)!}.$$

The cumulative distribution function for the binomial distribution is simply the sum of the probability mass function results for all applicable values. For example, the cumulative distribution function for the probability that X is less than or equal to a certain number, x, where $x = 0, 1, 2, \ldots, n$, is given by

$$P(C \le x) = \sum_{i=0}^{x} C_i^n \; p^i \; q^{n-i},$$

while the cumulative distribution function for the probability that X is greater than or equal to a certain number, x, where $x = 0, 1, 2, \ldots, n$, is given by

$$P(C \ge x) = \sum_{i=x}^{n} C_i^n \; p^i \; q^{n-i}.$$

Properties

When p represents the probability of a success in one trial and n represents the number of trials, the expected value or mean of the binomial distribution is $\mu = n \times p$ and the variance is $\sigma^2 = n \times p \times q$, where q represents the probability of failure on one trial, or $1 - p$. If p is unknown, $\hat{p}$ can be estimated to represent p as an unbiased estimator such that $\hat{p} = \frac{x}{n}$, where x is the number of observed successes in n trials.

The mode of the binomial distribution is dependent on different cases of the distribution and can be calculated as

$$\text{Mode} = \begin{cases} \text{floor}[(n+1)p] & \text{if } (n+1)p \text{ is } 0 \text{ or a noninteger,} \\ (n+1)p \text{ and } (n+1)p - 1 & \text{if } (n+1)p \in \{1, \ldots, n\}, \\ n & \text{if } (n+1)p = n+1, \end{cases}$$

where floor() is the floor function indicating the lowest previous integer in a series. For example, if a fair six-sided die is rolled six times, $n = 6$, $p = \frac{1}{6}$, so the mode would be $\text{floor}\left[(6+1)\frac{1}{6}\right]$, which is equal to $\text{floor}\left(\frac{7}{6}\right)$. When taking the lowest previous integer of $\frac{7}{6}$, the result is that the mode equals 1.

The median of the binomial distribution does not have a single formula; however, the median conforms to several statements, including

(1) The median, m, must lie within the interval $\text{floor}(np) \le m \le \text{ceiling}(np)$ where floor() is the lowest previous integer in a series and ceiling() is the highest previous integer in a series.

(2) The median must not be far from the mean such that

$$|m - np| \le \min\{\ln 2, \max(p, 1-p)\}.$$

(3) When $p = .5$ and n is odd, the median is a number in the interval:

$$\frac{1}{2}(n-1) \le m \le \frac{1}{2}(n+1).$$

(4) When $p = .5$ and n is even, the median equals $\frac{n}{2}$.

Confidence Intervals

A number of methods exist for determining a confidence level for the probability of success in a binomial

distribution. The Wald method is the most commonly recommended method; other methods include the Clopper–Pearson interval, the Agresti–Coull method, and the Arcsine method.

The Clopper–Pearson method, sometimes known as the *exact method*, calculates a confidence interval based on the binomial distribution's cumulative properties. When x is the number of successes observed in a binomial sample with n trials, the Clopper–Pearson method interval is given by

$$\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),\,2x,\,\frac{\alpha}{2}}} \le p \le \frac{\frac{x+1}{n-x}F_{2(x+1),\,2(n-x),\frac{\alpha}{2}}}{1+\frac{x+1}{n-x}F_{2(x+1),\,2(n-x),\frac{\alpha}{2}}}.$$

For example, if in 10 trials there are three successes, the Clopper–Pearson method gives a 95% confidence interval of

$$\frac{1}{1+\frac{10-3+1}{3}F_{16,6,0.025}} \le p \le \frac{\frac{3+1}{10-3}F_{8,\,14,\,0.025}}{1+\frac{3+1}{10-3}F_{8,\,14,\,0.025}},$$

thus, $.087 \le p \le .607$.

Given that $\hat{p}$ is the proportion of successes representing the estimate of p, and z is the quantile from the standard normal distribution that corresponds to the desired error rate, α, the Wald method for calculating a confidence interval is given by $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. In an attempt to reduce some of the bias as a result of the Wald method, the Agresti–Coull method modifies the estimate of p to be $\tilde{p} = \frac{x + \frac{1}{2}z^2}{n + z^2}$ and modifies the calculation of the confidence interval such that it is given by $\tilde{p} \pm z\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+z^2}}$.

Finally, the Arcsine method calculates the confidence interval through the use of the equation:

$$\mathrm{Sin}^2\left(\arcsin\left(\sqrt{\hat{p}}\right) \pm \frac{z}{2\sqrt{n}}\right).$$

Visualization

When visualizing the binomial distribution, a histogram like that shown in Figure 1 is constructed for each specific example. Number of successes, X, is plotted along the x-axis, while the probability of X is plotted along the y-axis according to the probability mass function results. As n increases, the binomial distribution will begin to look smoother. However, the binomial distribution does not have a set shape, as p and q change the mean of the distribution, and n, p, and q change the variance of the distribution. When n is large and p approaches .5, the binomial distribution approaches a bell-shaped curve. When p is smaller than .5, the visualization of the binomial distribution is right skewed. As p becomes smaller, the distribution becomes even more right skewed. When p is larger than .5, the visualization of the binomial distribution is left skewed. Again, as p becomes larger, the distribution is even more left skewed.

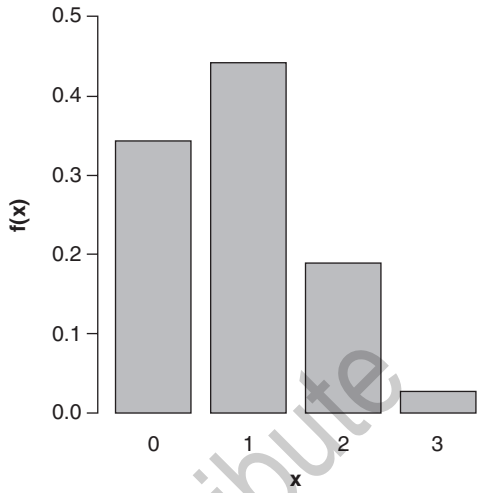


Figure 1 Binomial Distribution of Baseball Outcomes

Examples

Example 1 shows an application of the binomial distribution using the probability mass function. Example 2 shows an application of the binomial distribution using the cumulative distribution function.

Example 1: Baseball Batting Percentage

One example of applying the binomial distribution can be seen by predicting the number of hits a baseball player will get in three at bats. A batting average is calculated by taking the number of times a player has gotten a hit divided by the number of times they have been at bat, minus the number of times they took a base on balls. So a batter's probability of succeeding in getting a hit, p, can be represented by their batting average, usually somewhere around $p = .3$, and their probability of failing and getting an out, q, can be 1 minus batting average: $q = .7$. In three at bats, there are eight different patterns of outcomes. With a hit represented as "H" and an out repre-

Table 1 Probability for All Possible Patterns of at Bat Outcomes

Outcome	*Probability*	*X*
HHH	$.3^3$	3
HHO	$.3^2 \cdot .7$	2
HOH	$.3^2 \cdot .7$	2
HOO	$.3 \cdot .7^2$	1
OHH	$.3^2 \cdot .7$	2
OHO	$.3 \cdot .7^2$	1
OOH	$.3 \cdot .7^2$	1
OOO	$.7^3$	0

sented at "O," the possible outcomes include HHH, HHO, HOH, HOO, OHH, OHO, OOH, and OOO. However, each of these outcomes has its own probability of happening. To calculate the probability of a pattern, the probability of each individual event is multiplied. So, for example, the probability of HOH would be .3 × .7 × .3, which can be simplified to $.3^2 \times .7$. Table 1 provides the probabilities for all possible patterns of outcomes.

As Table 1 shows, one combination results in three hits, X = 3; three combinations result in two hits and one out, X = 2; three combinations result in one hit and two outs, X = 1; and one combination results in three outs, X = 0. Notice that each pattern that results in the same X value has the same probability value, regardless of what order the hits or outs happen in. This is precisely why the binomial coefficient is included in the probability mass function. Most of the time, a researcher would not ask what the probability of a certain pattern is; instead, a researcher would ask something like "what is the probability that the player gets two hits in three at bats." In this example, it was fairly easy to calculate the probability of every possible pattern. However, that is not possible in most models, so the binomial coefficient calculates how many patterns result in the same number of successes or the same X value. Table 2 illustrates the frequency table for each X number of successes. The histogram shown in Figure 1 represents this example.

Example 2: Marathon Entry Percentage

Other applications of the binomial distribution need to use the cumulative distribution function rather than the probability mass function. For example, 70% of people who apply to run a local marathon are randomly drawn to receive entry into the race. If five friends all apply, what is the probability that at least four of them

Table 2 Frequency Table of Number of Successes in Three at Bats

X	*f(x)*		
0	$.7^3$	=	.343
1	$3(.3 \cdot .7^2)$	=	.441
2	$3(.3^2 \cdot .7)$	=	.189
3	$.3^3$	=	.027

are accepted? In this case, at least four runners means that the probabilities of having 4 of 5 and 5 of 5 admitted need to be summed. So the cumulative distribution function would be

$$P(C \geq 4) = \sum_{i=4}^{5} C_i^5 \, .7^i \, .3^{5-i},$$

$$P(C \geq 4) = C_4^5 \, .7^4 \, .3^{5-4} + C_5^5 \, .7^5 \, .3^{5-5},$$

$$P(C \geq 4) = \left(\frac{5!}{4!1!}(.2401)(.3)\right) + \left(\frac{5!}{5!0!}(.16807)(1)\right),$$

$$P(C \geq 4) = .52822.$$

Thus, the probability of at least four of the runners being admitted is approximately .52822 or 52.82%.

Jessica Hess and Neal M. Kingston

See also Bernoulli Distribution; Beta Distribution; Mode; Normal Distribution; Poisson Distribution; Stochastic Processes; *z* Distribution

Further Readings

Agresti, A., & Coull, B. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. doi:10.1080/00031305.1998.10480550.

Čekanavičius, V., & Roos, B. (2006). Compound binomial approximations. *Annals of the Institutes of Statistical Mathematics*, *58*, 187–210. doi:10.1007/s10463-005-0018-4.

Edwards, A. W. (1960). The meaning of binomial distribution. *Nature*, *186*, 1074. doi:10.1038/1861074a0.

Jowett, G. H. (1963). The relationship between the binomial and *F* distributions. *Journal of the Royal Statistical Society*, *13*(1), 55–57.

Kaas, R., & Buhrman, J. (1980). Mean, median, and mode in binomial distributions. *Statistica Neerlandica*, *34*(1), 13–18. doi:10.1111/j.1467-9574.1980.tb00681.x.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University.

Wallis, S. (2013). Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, *20*(3), 178–208. doi:10.1080/09296174.2013.799918.

Biological and Technical Replicates

Biological and technical replicates are two different approaches to making repeated measurements of an underlying biological phenomenon in biomedical research. It is generally a good practice to include both types of replicates in each experiment.

Broadly speaking, technical replicates are independently repeated measurements of the same sample using the same procedure. As such, these replicates represent independent measures of the noise (typically random) associated with protocols or equipment: They help measure the reproducibility of an assay and not the reproducibility of the underlying biological phenomenon. Biological replicates, on the other hand, are parallel measurements of biologically distinct samples. These replicates help capture the variation (random or otherwise) of the biological phenomenon under study and help measure its reproducibility. The distinction between technical and biological replicates is a *functional* one, in that it depends on which type of data variability—procedural or biological—they capture and not necessarily on how the replicates are obtained.

There is no one-size-fits-all formula for designing replicates that are optimal for a given experiment. The optimal design, including the optimal mix of technical and biological replicates in a given experiment, depends on the potential sources and magnitudes of variability in a given experiment and the questions that the experiment seeks to answer.

Historical Origins of the Replicate Nomenclature

Until the 1990s, much of the reproducibility testing in biomedical research, especially in the *wet laboratory* experimental sciences such as molecular and cellular biology, employed what would be considered technical replicates today. The push to systematically include biological replicates in experiments originated primarily in these fields in the 2000s with the widespread realization that biomedical research findings were not sufficiently reproducible, in large part because technical replicates by themselves did not properly account for the variability of the underlying biological phenomena. Major funding agencies in these fields, such as the U.S. National Institutes of Health, spurred the widespread adoption of the current replicate nomenclature and practices of replicate design by incentivizing researchers to employ both biological and technical replicates in their research as a way of enhancing the reproducibility of the research findings.

While the practice of quantitative measurements and statistical testing was much better established in many other fields of biomedical research, such as epidemiology, psychophysics, ecology and evolutionary biology, reproducibility of results was not necessarily commensurately better in these fields, arguably also because of poor replicate design.

Reproducibility Requires Representative Replicates

Research is primarily about learning general truths about the phenomenon under study. A set of findings is useful only to the extent that the same findings are obtained when the given experiment is independently but precisely repeated. The term *reproducibility* typically means this type of across-experiment reproducibility of the *findings* or *conclusions* and not of the *measurements* on which the conclusions are based. One accepts the mathematical reality that the measurements themselves will not be exactly reproducible from one instance to the next, be it within or across experiments, even as one expects the conclusions to be reproducible.

The only way one can draw reproducible conclusions based on inherently variable measurements is to use sound practices of statistical sampling and, where necessary, statistical testing. To the extent that the empirical measurements are truly representative of the underlying phenomenon, one can have a quantifiable degree of confidence, say 95%, that the conclusions will be reproduced when the experiment is exactly repeated. Thus, the key to obtaining reproducible results is to ensure that replicates as a group adequately represent the relevant statistical properties of the phenomenon of interest. This, in a nutshell, is the goal of replicate design: to ensure that the replicates are representative and that they adequately capture the study-relevant statistical properties, including the variability, of the phenomenon.

In biology, the substrates of phenomena of interest tend to be highly variable. Therefore, ensuring that this biological variability is properly represented in the empirical measurements of a given phenomenon is a necessary and proper way of improving the reproducibility of the findings about the phenomenon. In other

words, it is usually a good idea to include biological replicates in an experiment because the underlying biological substrates are usually variable. It follows from the elementary principles of statistics that the greater the variability of the relevant biological substrates, the larger the number of replicates needed to adequately capture this variability.

Basic Principles of Replicate Design

Consider a simple hypothetical experiment to determine the levels of a blood component called *albumin* in adult Spraque Dawley laboratory rats 24 hours after skin injury. We induce injury in a designated spot, say a hind thigh, using standard procedures in one rat. We draw a vial of blood from this rat 24 hours after the procedure and measure the albumin levels using standard procedures. We repeat this measurement twice more independently using the same vial of blood. These constitute three technical replicates because they measure the reproducibility of the albumin assay (Figure 1A). From the three replicates, we determine the mean and the standard deviation of albumin levels.

Note, however, that these findings apply only to the particular vial of blood. We have no way of evaluating whether the results are likely to be reproducible across additional blood draws from the same mouse because we have not tested any additional blood draws. Obviously, this is not a useful outcome and reflects poor replicate design. To make the results more generalizable, we make three mutually independent blood draws from this rat and measure the albumin levels in each (Figure 1B). Although the sample size remains the same at three, these technical replicates are better designed because the albumin level estimate is likely to be better reproducible for this mouse.

It is desirable to have our findings apply to all Spraque Dawley rats and not just to the one rat we tested. We therefore repeat the experiment using three different rats, making one measurement each (Figure 1C). These three biological replicates allow us to draw conclusions about the three rats in question. To the extent that these three rats are representative of all Spraque Dawley rats, the results should be reproducible across all rats of this strain when the experiment is exactly repeated.

In general, it is a good practice to include both technical and biological replicates (Figure 1D), since the two types of replicates measure different types of variability in the measurements, as noted earlier.

Additional Observations About Replicate Design

There are a few additional things we always wanted to know about replicate design but our mothers never told us. First, it ultimately does not matter whether a given replicate is designated a technical replicate or a biological replicate, as long as the sources of the replicates are faithfully kept track of. For instance, if we arbitrarily shuffle the replicate designations of one or

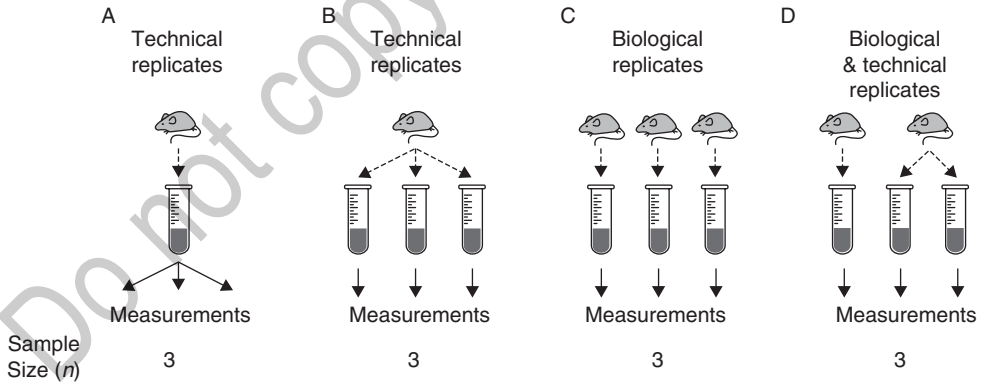


Figure 1 A Hypothetical Example That Illustrates the Distinction Between Technical and Biological Replicates

Note: In each case, rat blood is drawn and the concentration of albumin in the blood is measured. In each panel, dotted arrows at the top denote blood draws, and solid arrows at the bottom denote albumin measurements. (A) Repeated measurements from the same blood draw nominally represent technical replicates but reflect poor replicate design. (B) A better design for technical replicates is to make multiple independent blood draws and measure albumin level in each draw. (C) Biological replicates are those in which each blood draw is made from a different selected rat. (D) It is generally a good design practice to include both technical and biological replicates in an experiment and to adjust the relative proportions of the two types of replicates so that the replicates are representative of the phenomenon under study. Of the four replicate designs shown, the one in panel D is likely to yield the most reproducible results, even though the sample size ($n = 3$), and therefore the nominal statistical power, is the same in all cases. Note that statistically desirable values of n tend to be higher than 3 in actual experiments.

more replicates in Figure 1, it will not in any way affect the experimental findings, our observations as to which aspects of the experiments the findings apply to, or the reproducibility of the findings. Indeed, the technical replicates in the earlier example meet some of the criteria of biological replicates, in that they capture some of the *procedural* variability with a *biological* basis, such as the variability in the induced injury across rats. In some cases, it can be difficult to decide whether a given replicate qualifies as either type of replicate or both. For instance, what constitutes a technical replicate in a study that seeks to determine whether home value appraisers estimate lower values for homes owned by African Americans than those owned by Caucasians? Moreover, there are vast areas of research where the phenomena of interest are not biological at all. For instance, what would constitute biological replicates in a study about the effect of ethanol on catalytic converters in cars? Yet, no one would dispute the importance of these research questions or of ensuring the reproducibility of the findings by designing the replicates properly. In thinking about these issues, it helps to keep in mind the aforementioned historical origins of the replicate nomenclature and that while reproducibility is desirable in all research, not all research involves biology or even experimentation.

Second, proper replicate design requires that the goals of the study and the planned data analyses be precisely specified beforehand. This is especially important when the underlying phenomena are complex, multivariate, and/or subject to dynamic change. For instance, in the case of the aforementioned rat experiment, we need to decide which aspects of the underlying phenomenon we want to draw conclusions about and how broadly we want to draw them: injury to which body regions, which types of the injury (e.g., abrasions, cuts, chemicals, or burns), which ages, and so forth. Specifying the research question has the effect of specifying which statistical properties are relevant to the study and which are not, thus making replicate design more tractable. Without specifying the study parameters in this fashion, we would risk either having to obtain an unmanageably large number of replicates to try and capture all potential statistical variability of the underlying substrates or designing poor replicates that fail to capture the study-relevant variations of the phenomenon, and thereby reducing the reproducibility of the research findings. Specifying the planned tests is necessary for, among other things, planning the number of various replicates. For instance, in the aforementioned rat experiment, we planned no statistical tests because the goal of this simple experiment was to simply estimate the albumin levels, not to test any hypotheses. For a more complex experiment, in which we test the hypothesis that albumin levels in rats with skin injury are higher than in control rats that underwent a sham procedure, we would need to obtain replicates from both the treatment group and control group of rats. The numbers of the replicates do not necessarily have to be the same between the two groups. For instance, sham injury may be quite consistent from one rat to the next, so that fewer technical replicates might suffice for the control group.

Third, replicate design is closely related to, but not the same as, sample size calculation. A common, recommended practice is to first perform power analyses based on the expected strength of the effect under study (estimated based on the best available information from published results or pilot data), planned statistical tests, and the desired level of statistical significance and statistical power. This will yield the required total number of replicates (or sample size n). The n value can then be broken down into the desired numbers of technical versus biological replicates based on the estimated variability from various sources.

A fourth, related principle is that it is the responsibility of the researcher to report the replicate design, along with the rest of the study methods in sufficient detail as to enable other researchers to replicate the findings independently. Reporting a study poorly is tantamount to designing it poorly.

Finally, there are many cases in which a statistically optimal replicate design is not possible, not desirable, or both. For instance, in invasive studies of neural activity in monkey brains, it is typical to use only two monkeys because using additional monkeys is inadvisable without a compelling reason. Instead, researchers typically study a large number of individual neurons in either monkey, which results in a statistically suboptimal nested replicate design. One can nonetheless draw the best possible reproducible conclusions from such nested data using commonly available statistical tools. In cases such as this, the imperatives of sound statistical design must be balanced against other principles of sound research, and reproducibility must be maximized using the best available alternative methods.

Jay Hegdé

See also Animal Research; Nested Sampling; Power Analysis; Replication; Sample Size

Further Readings

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. doi: 10.1038/nn.3648.

Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: Replication. *Nature Methods*, *11*(9), 879–880. doi: 10.1038/nmeth.3091.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*(7485), 612–613. doi: 10.1038/505612a.

Maddox, J. (1992). Is molecular biology yet a science? *Nature*, *355*, 201. doi:10.1038/355201a0.

Naegle, K., Gough, N. R., & Yaffe, M. B. (2015). Criteria for biological reproducibility: What does "n" mean? *Science Signaling*, *8*(371), fs7. doi: 10.1126/scisignal.aab1125.

Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—What is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Reports*, *13*(4), 291–296. doi: 10.1038/embor.2012.36.

Bivariate Regression

Regression is a statistical technique used to help investigate how variation in one or more variables predicts or explains variation in another variable. This popular statistical technique is flexible in that it can be used to analyze experimental or nonexperimental data with multiple categorical and continuous independent variables. If only one variable is used to predict or explain the variation in another variable, the technique is referred to as *bivariate regression*. When more than one variable is used to predict or explain variation in another variable, the technique is referred to as *multiple regression*. Bivariate regression is the focus of this entry.

Various terms are used to describe the independent variable in regression, namely, *predictor variable*, *explanatory variable*, or *presumed cause*. The dependent variable is often referred to as an *outcome variable*, *criterion variable*, or *presumed effect*. The choice of independent variable term will likely depend on the preference of the researcher or the purpose of the research. Bivariate regression may be used solely for predictive purposes. For example, do scores on a college entrance exam predict college grade point average? Or it may be used for explanation. Do differences in IQ scores explain differences in achievement scores? It is often the case that although the term *predictor* is used by researchers, the purpose of the research is, in fact, explanatory.

Suppose a researcher is interested in how well reading in first grade predicts or explains fifth-grade science achievement scores. The researcher hypothesizes that those who read well in first grade will also have high science achievement in fifth grade. An example bivariate regression will be performed to test this hypothesis. The data used in this example are a random sample of students (10%) with first-grade reading and fifth-grade science scores and are taken from the Early Childhood Longitudinal Study public database. Variation in reading scores will be used to explain variation in science achievement scores, so first-grade reading achievement is the explanatory variable and fifth-grade science achievement is the outcome variable. Before the analysis is conducted, however, it should be noted that bivariate regression is rarely used in published research. For example, intelligence is likely an important common cause of both reading and science achievement. If a researcher was interested in explaining fifth-grade science achievement, then potential important common causes, such as intelligence, would need to be included in the research.

Regression Equation

The simple equation for bivariate linear regression is $Y = a + bX + e$. The science achievement score, Y, for a student equals the intercept or constant (a), plus the slope (b) times the reading score (X) for that student, plus error (e). Error, or the residual component (e), represents the error in prediction, or what is not explained in the outcome variable. The error term is not necessary and may be dropped so that the following equation is used: $Y' = a + bX$. Y' is the expected (or predicted) score. The intercept is the predicted fifth-grade science score for someone whose first-grade reading score is zero. The slope (b, also referred to as the *unstandardized regression coefficient*) represents the predicted unit increase in science scores associated with a one-unit increase in reading scores. X is the observed score for that person. The two parameters (a and b) that describe the linear relation between the predictor and outcome are thus the intercept and the regression coefficient. These parameters are often referred to as least squares estimators and will be estimated using the two sets of scores. That is, they represent the optimal estimates that will provide the least error in prediction.

Returning to the example, the data used in the analysis were first-grade reading scores and fifth-grade science scores obtained from a sample of 1,027 school-age children. *T*-scores, which have a mean of 50 and standard deviation of 10, were used. The means for the scores in the sample were 51.31 for reading and 51.83 for science.

Because science scores are the outcome, the science scores are regressed on first-grade reading scores. The easiest way to conduct such analysis is to use a statistical program. The estimates from the output may then be plugged into the equation. For these data, the prediction equation is $Y' = 21.99 + (.58)X$. Therefore, if a student's first-grade reading score was 60, the predicted fifth-grade science achievement score for that student would be 21.99 + (.58)60, which equals 56.79. One might ask, why even conduct a regression analysis to obtain a predicted science score when Johnny's science score was already available? There are a few possible reasons.

First, perhaps a researcher wants to use the information to predict later science performance, either for a new group of students or for an individual student, based on current first-grade reading scores. Second, a researcher may want to know the relation between the two variables, and a regression provides a nice summary of the relation between the scores for all the students. For example, do those students who tend to do well in reading in first grade also do well in science in fifth grade? Last, a researcher might be interested in different outcomes related to early reading ability when considering the possibility of implementing an early reading intervention program. Of course a bivariate relation is not very informative. A much more thoughtfully developed causal model would need to be developed if a researcher was serious about this type of research.

Scatterplot and Regression Line

The regression equation describes the linear relation between variables; more specifically, it describes science scores as a function of reading scores. A scatterplot could be used to represent the relation between these two variables, and the use of a scatterplot may assist one in understanding regression. In a scatterplot, the science scores (outcome variable) are on the *y*-axis, and the reading scores (explanatory variable) are on the *x*-axis.

A scatterplot is shown in Figure 1. Each person's reading and science scores in the sample are plotted. The scores are clustered fairly closely together, and the general direction looks to be positive. Higher scores in reading are generally associated with higher scores in science. The next step is to fit a regression line. The regression line is plotted so that it minimizes errors in prediction, or simply, the regression line is the line that is closest to all the data points. The line is fitted automatically in many computer programs, but information obtained in the regression analysis output can also be used to plot two data points that the line should be drawn through. For example, the intercept (where the line crosses the *y*-axis) represents the predicted science score when reading equals zero. Because the value of the intercept was 21.99, the first data point would be found at 0 on the *x*-axis and at 21.99 on the *y*-axis. The second point on the line may be located at the mean reading score (51.31) and mean science score (51.83). A line can then be drawn through those two points. The line is shown in Figure 1. Points that are found along this regression line represent the predicted science achievement score for Person A with a reading score of X.

Unstandardized and Standardized Coefficients

For a more thorough understanding of bivariate regression, it is useful to examine in more detail the output obtained after running the regression. First, the intercept has no important substantive meaning. It is unlikely that anyone would score a zero on the reading test, so it does not make much sense. It is useful in the unstandardized solution in that it is used to obtain predicted scores (it is

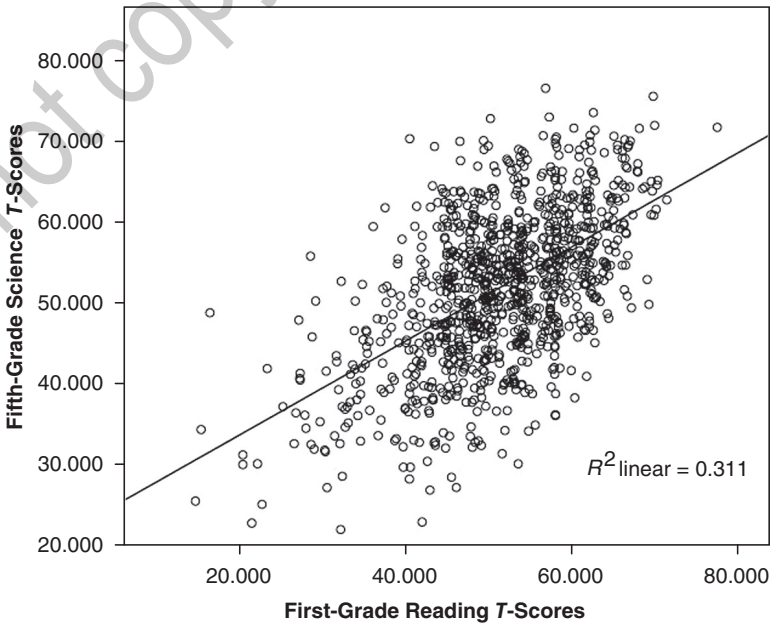


Figure 1 Scatterplot and Regression Line

a constant value added to everyone's score), and as demonstrated above, it is useful in plotting a regression line. The slope ($b = .58$) is the unstandardized coefficient. It was statistically significant, indicating that reading has a statistically significant influence on fifth-grade science. A 1-point *T*-score increase in reading is associated with a .58 *T*-score point increase in science scores. The *b*s are interpreted in the metric of the original variable. In the example, all the scores were *T*-scores. Unstandardized coefficients are especially useful for interpretation when the metric of the variables is meaningful. Sometimes, however, the metric of the variables is not meaningful.

Two equations were generated in the regression analysis. The first, as discussed in the example above, is referred to as the *unstandardized solution*. In addition to the unstandardized solution, there is a *standardized solution*. In this equation, the constant is dropped, and *z* scores (mean = 0, standard deviation = 1), rather than the *T*-scores (or raw scores), are used. The standardized regression coefficient is referred to as a *beta weight* (β). In the example, the beta weight was .56. Therefore, a one-standard-deviation increase in reading was associated with a .56-standard-deviation increase in science achievement. The unstandardized and standardized coefficients were similar in this example because *T*-scores are standardized scores, and the sample statistics for the *T*-scores were fairly close to the population mean of 50 and standard deviation of 10.

It is easy to convert back and forth from standardized to unstandardized regression coefficients:

$$\beta = b\left(\frac{\text{standard deviation of reading scores}}{\text{standard deviation of science scores}}\right)$$

or

$$b = \beta\left(\frac{\text{standard deviation of science scores}}{\text{standard deviation of reading scores}}\right).$$

From an interpretative standpoint, should someone interpret the unstandardized or the standardized coefficient? There is some debate over which one to use for interpretative statements, but in a bivariate regression, the easiest answer is that if both variables are in metrics that are easily interpretable, then it would make sense to use the unstandardized coefficients. If the metrics are not meaningful, then it may make more sense to use the standardized coefficient. Take, for example, number of books read per week. If number of books read per week was represented by the actual number of books read per week, the variable is in a meaningful metric. If the number of books read per week variable were coded so that 0 = no books read per week, 1 = one to three books read per week, and 2 = four or more books read per week, then the variable is not coded in a meaningful metric, and the standardized coefficient would be the better one to use for interpretation.

R and R^2

In bivariate regression, typically the regression coefficient is of greatest interest. Additional information is provided in the output, however. *R* is used in multiple regression output and represents a multiple correlation. Because there is only one explanatory variable, *R* (.56) is equal to the correlation coefficient ($r = .56$) between reading and science scores. Note that this value is also identical to the β. Although the values of β and *r* are the same, the interpretation differs. The researcher is not proposing an agnostic relation between reading scores and science scores. Rather the researcher is positing that early reading explains later science achievement. Hence, there is a clear direction in the relation, and this direction is not specified in a correlation.

R^2 is the variance in science scores explained by reading scores. In the current example, $R^2 = 31$. First-grade reading scores explained 31% of the variance in fifth-grade science achievement scores.

Statistical Significance

R and R^2 are typically used to evaluate the statistical significance of the overall regression equation (the tests for the two will result in the same answer). The null hypothesis is that R^2 equals zero in the population. One way of calculating the statistical significance of the overall regression is to use an *F* test associated with the value obtained with the formula

$$\frac{R^2 / k}{(1 - R^2) / (N - k - 1)}.$$

In this formula, R^2 equals the variance explained, $1 - R^2$ is the variance unexplained, and *k* equals the degrees of freedom (*df*) for the regression (which is 1 because one explanatory variable was used). With the numbers plugged in, the formula would look like

$$\frac{.31 / 1}{.69 / (1027 - 1 - 1)}$$

and results in $F = 462.17$. An *F* table indicates that reading did have a statistically significant effect on science achievement, $R^2 = .31$, $F(1,1025) = 462.17$, $p < .01$.

In standard multiple regression, a researcher typically interprets the statistical significance of R^2 (the statistical significance of the overall equation) and the statistical significance of the unique effects of each individual explanatory variable. Because this is bivariate regression, however, the statistical significance test of

the overall regression and the regression coefficient (b) will yield the same results, and typically the statistical significance tests for each are not reported.

The statistical significance of the regression coefficient (b) is evaluated with a t test. The null hypothesis is that the slope equals zero, that is, the regression line is parallel with the x-axis. The t-value is obtained by

$$\frac{b}{\text{standard error of } b}.$$

In this example, $b = .58$, and its associated standard error was .027. The t-value was 21.50. A t-table could be consulted to determine whether 21.50 is statistically significant. Or a rule of thumb may be used that given the large sample size and with a two-tailed significance test, a t-value greater than 2 will be statistically significant at the $p < .05$ level. Clearly, the regression coefficient was statistically significant. Earlier it was mentioned that because this is a bivariate regression, the significance of the overall regression and b provide redundant information. The use of F and t tests may thus be confusing, but note that F (462.17) equals $t^2(21.50^2)$ in this bivariate case. A word of caution: This finding does not generalize to multiple regression. In fact, in a multiple regression, the overall regression might be significant, and some of the bs may or may not be significant. In a multiple regression, both the overall regression equation and the individual coefficients are examined for statistical significance.

Residuals

Before completing this explanation of bivariate regression, it will be instructive to discuss a topic that has been for the most part avoided until now: the residuals. Earlier it was mentioned that e (residual) was also included in the regression equation. Remember that regression parameter estimates minimize the prediction errors, but the prediction is unlikely to be perfect. The residuals represent the error in prediction. Or the residual variance represents the variance that is left unexplained by the explanatory variable. Returning to the example, if reading scores were used to predict science scores for those 1,026 students, each student would have a prediction equation in which his or her reading score would be used to calculate a predicted science score. Because the actual score for each person is also known, the residual for each person would represent the observed fifth-grade science score minus the predicted score obtained from the regression equation. Residuals are thus observed scores minus predicted scores, or conceptually they may be thought of as the fifth-grade science scores with effects of first-grade reading removed.

Another way to think of the residuals is to revert back to the scatterplot in Figure 1. The x-axis represents the observed scores, and the y-axis represents the science scores. Both predicted and actual scores are already plotted on this scatterplot. That is, the predicted scores are found on the regression line. If a person's reading score was 40, the predicted science score may be obtained by first finding 40 on the x-axis, and then moving up in a straight line until reaching the regression line. The observed science scores for this sample are also shown on the plot, represented by the dots scattered about the regression line. Some are very close to the line whereas others are farther away. Each person's residual is thus represented by the distance between the observed score and the regression line. Because the regression line represents the predicted scores, the residuals are the difference between predicted and observed scores. Again, the regression line minimizes the distance of these residuals from the regression line. Much as residuals are thought of as science scores with the effects of reading scores removed, the residual variance is the proportion of variance in science scores left unexplained by reading scores. In the example, the residual variance was .69, or $1 - R^2$.

Regression Interpretation

An example interpretation for the reading and science example concludes this entry on bivariate regression. The purpose of this study was to determine how well first-grade science scores explained fifth-grade science achievement scores. The regression of fifth-grade science scores on first-grade reading scores was statistically significant, $R^2 = .31$, $F(1,1025) = 462.17$, $p < .01$. Reading accounted for 31% of the variance in science achievement. The unstandardized regression coefficient was .58, meaning that for each T-score point increase in reading, there was a .58 T-score increase in science achievement. Children who are better readers in first grade also tend to be higher achievers in fifth-grade science.

Matthew R. Reynolds

See also Correlation; Multiple Regression; Path Analysis; Scatterplot; Variance

Further Readings

Bobko, P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Keith, T. Z. (2006). *Multiple regression and beyond*. Boston: Pearson.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Thousand Oaks, CA: Sage.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Thousand Oaks, CA: Sage.

Weisburg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, NJ: Wiley.

Block Design

Sir Ronald Fisher, the father of modern experimental design, extolled the advantages of block designs in his classic book, *The Design of Experiments*. He observed that block designs enable researchers to reduce error variation and thereby obtain more powerful tests of false null hypotheses. In the behavioral sciences, a significant source of error variation is the nuisance variable of individual differences. This nuisance variable can be isolated by assigning participants or experimental units to blocks so that at the beginning of an experiment, the participants within a block are more homogeneous with respect to the dependent variable than are participants in different blocks. Three procedures are used to form homogeneous blocks.

1. Match participants on a variable that is correlated with the dependent variable. Each block consists of a set of matched participants.
2. Observe each participant under all or a portion of the treatment levels or treatment combinations. Each block consists of a single participant who is observed two or more times. Depending on the nature of the treatment, a period of time between treatment level administrations may be necessary in order for the effects of one treatment level to dissipate before the participant is observed under other levels.
3. Use identical twins or litter mates. Each block consists of participants who have identical or similar genetic characteristics.

Block designs also can be used to isolate other nuisance variables, such as the effects of administering treatments at different times of day, on different days of the week, or in different testing facilities. The salient features of the five most often used block designs are described next.

Block Designs With One Treatment

Dependent Samples t-Statistic Design

The simplest block design is the randomization and analysis plan that is used with a *t* statistic for dependent samples. Consider an experiment to compare two ways of memorizing Spanish vocabulary. The dependent variable is the number of trials required to learn the vocabulary list to the criterion of three correct recitations. The null and alternative hypotheses for the experiment are, respectively,

$$H_0: \mu_1 - \mu_2 = 0$$

and

$$H_1: \mu_1 - \mu_2 \neq 0,$$

where μ_1 and μ_2 denote the population means for the two memorization approaches. It is reasonable to believe that IQ is negatively correlated with the number of trials required to memorize Spanish vocabulary. To isolate this nuisance variable, n blocks of participants can be formed so that the two participants in each block have similar IQs. A simple way to form blocks of matched participants is to rank the participants in terms of IQ. The participants ranked 1 and 2 are assigned to Block 1, those ranked 3 and 4 are assigned to Block 2, and so on. Suppose that 20 participants have volunteered for the memorization experiment. In this case, $n = 10$ blocks of dependent samples can be formed. The two participants in each block are randomly assigned to the memorization approaches. The layout for the experiment is shown in Figure 1.

The null hypothesis is tested using a *t* statistic for dependent samples. If the researcher's hunch is correct—that IQ is correlated with the number of trials to learn—the design should result in a more powerful test of a false null hypothesis than would a *t*-statistic design for

	Treat. Level	*Dep. Var.*	*Treat. Level*	*Dep. Var.*
Block 1	a_1	Y_{11}	a_2	Y_{12}
Block 2	a_1	Y_{21}	a_2	Y_{22}
Block 3	a_1	Y_{31}	a_2	Y_{32}
⋮	⋮	⋮	⋮	⋮
Block 10	a_1	$Y_{10,1}$	a_2	$Y_{10,2}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$

Figure 1 Layout for a Dependent Samples *t*-Statistic Design

Note: a*j* denotes a treatment level (*Treat. Level*); Y*ij* denotes a measure of the dependent variable (Dep. Var.). Each block in the memorization experiment contains two matched participants. The participants in each block are randomly assigned to the treatment levels. The means of the treatments levels are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$.

independent samples. The increased power results from isolating the nuisance variable of IQ so that it does not appear in the estimates of the error effects.

Randomized Block Design

The randomized block analysis of variance design can be thought of as an extension of a dependent samples *t*-statistic design for the case in which the treatment has two or more levels. The layout for a randomized block design with $p = 3$ levels of Treatment A and $n = 10$ blocks is shown in Figure 2. A comparison of this layout with that in Figure 1 for the dependent samples *t*-statistic design reveals that the layouts are the same except that the randomized block design has three treatment levels.

In a randomized block design, a block might contain a single participant who is observed under all p treatment levels or p participants who are similar with respect to a variable that is correlated with the dependent variable. If each block contains one participant, the order in which the treatment levels are administered is randomized independently for each block, assuming that the nature of the research hypothesis permits this. If a block contains p matched participants, the participants in each block are randomly assigned to the treatment levels.

The statistical analysis of the data is the same whether repeated measures or matched participants are used. However, the procedure used to form homogeneous blocks does affect the interpretation of the results. The results of an experiment with repeated measures generalize to a population of participants who have been exposed to all the treatment levels. The results of an experiment with matched participants generalize to a population of participants who have been exposed to only one treatment level.

The total sum of squares (*SS*) and total degrees of freedom for a randomized block design are partitioned as follows:

$$SSTOTAL = SSA + SSBLOCKS + SSRESIDUAL$$
$$np - 1 = (p - 1) + (n - 1) + (n - 1)(p - 1),$$

where *SSA* denotes the Treatment *A SS* and *SSBLOCKS* denotes the blocks *SS*. The *SSRESIDUAL* is the interaction between Treatment *A* and blocks; it is used to estimate error effects. Many test statistics can be thought of as a ratio of error effects and treatment effects as follows:

$$\text{Test statistic} = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})},$$

where $f()$ denotes a function of the effects in parentheses. The use of a block design enables a researcher to isolate variation attributable to the blocks variable so that it does not appear in estimates of error effects. By removing this nuisance variable from the numerator and denominator of the test statistic, a researcher is rewarded with a more powerful test of a false null hypothesis.

Two null hypotheses can be tested in a randomized block design. One hypothesis concerns the equality of the Treatment *A* population means; the other hypothesis concerns the equality of the blocks population means. For this design and those described later, assume that the treatment represents a fixed effect and the nuisance variable, blocks, represents a random effect. For this mixed model, the null hypotheses are

	Treat. Level	*Dep. Var.*	*Treat. Level*	*Dep. Var.*	*Treat. Level*	*Dep. Var.*	
Block 1	a_1	Y_{11}	a_2	Y_{12}	a_3	Y_{13}	$\overline{Y}_{1\cdot}$
Block 2	a_1	Y_{21}	a_2	Y_{22}	a_3	Y_{23}	$\overline{Y}_{2\cdot}$
Block 3	a_1	Y_{31}	a_2	Y_{32}	a_3	Y_{33}	$\overline{Y}_{3\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Block 10	a_1	$Y_{10,1}$	a_2	$Y_{10,2}$	a_3	$Y_{10,3}$	$\overline{Y}_{10\cdot}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$		$\overline{Y}_{\cdot 3}$	

Figure 2 Layout for a Randomized Block Design With p = 3 Treatment Levels and n = 10 Blocks

Note: a_j denotes a treatment level (Treat. Level); Y_{ij} denotes a measure of the dependent variable (Dep. Var.). Each block contains three matched participants. The participants in each block are randomly assigned to the treatment levels. The means of Treatment *A* are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$ and $\overline{Y}_{\cdot 3}$ and the means of the blocks are denoted by $\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{10\cdot}$.

$$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot p}$$

(treatment A population means are equal)

$$H_0: \sigma^2_{BL} = 0$$

(variance of the blocks, BL, population means is equal to zero)

where μ_{ij} denotes the population mean for the ith block and the jth level of treatment A.

The F statistics for testing these hypotheses are

$$F = \frac{SSA/(p-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSA}{MSRESIDUAL}$$

and

$$F = \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSBL}{MSRESIDUAL}.$$

The test of the blocks null hypothesis is generally of little interest because the population means of the nuisance variable are expected to differ.

The advantages of the design are simplicity in the statistical analysis and the ability to isolate a nuisance variable so as to obtain greater power to reject a false null hypothesis. The disadvantages of the design include the difficulty of forming homogeneous blocks and of observing participants p times when p is large and the restrictive sphericity assumption of the design. This assumption states that in order for F statistics to be distributed as central F when the null hypothesis is true, the variances of the differences for all pairs of treatment levels must be homogeneous; that is,

$$\sigma^2_{Y_j - Y_{j'}} = \sigma^2_j + \sigma^2_{j'} - 2\sigma_{jj'}, \text{ for all } j \text{ and } j'.$$

Generalized Randomized Block Design

A generalized randomized block design is a variation of a randomized block design. Instead of having n blocks of p homogeneous participants, the generalized randomized block design has w groups of np homogeneous participants. The w groups, like the n blocks in a randomized

		Treat. Level	Dep. Var.		Treat. Level	Dep. Var.		Treat. Level	Dep. Var.
Group 1	1	a_1	Y_{111}	3	a_2	Y_{321}	5	a_3	Y_{531}
	2	a_1	Y_{211}	4	a_2	Y_{421}	6	a_3	Y_{631}
			$\overline{Y}_{\cdot 11}$			$\overline{Y}_{\cdot 21}$			$\overline{Y}_{\cdot 31}$
Group 2	7	a_1	Y_{712}	9	a_2	Y_{922}	11	a_3	$Y_{11,32}$
	8	a_1	Y_{812}	10	a_2	$Y_{10,22}$	12	a_3	$Y_{12,32}$
			$\overline{Y}_{\cdot 12}$			$\overline{Y}_{\cdot 22}$			$\overline{Y}_{\cdot 32}$
Group 3	13	a_1	$Y_{13,13}$	15	a_2	$Y_{15,23}$	17	a_3	$Y_{17,33}$
	14	a_1	$Y_{14,13}$	16	a_2	$Y_{16,23}$	18	a_3	$Y_{18,33}$
			$\overline{Y}_{\cdot 13}$			$\overline{Y}_{\cdot 23}$			$\overline{Y}_{\cdot 33}$
Group 4	19	a_1	$Y_{19,14}$	21	a_2	$Y_{21,24}$	23	a_3	$Y_{23,34}$
	20	a_1	$Y_{20,14}$	22	a_2	$Y_{22,24}$	24	a_3	$Y_{24,34}$
			$\overline{Y}_{\cdot 14}$			$\overline{Y}_{\cdot 24}$			$\overline{Y}_{\cdot 34}$
Group 5	25	a_1	$Y_{25,15}$	27	a_2	$Y_{27,25}$	29	a_3	$Y_{29,35}$
	26	a_1	$Y_{26,15}$	28	a_2	$Y_{28,25}$	30	a_3	$Y_{30,35}$
			$\overline{Y}_{\cdot 15}$			$\overline{Y}_{\cdot 25}$			$\overline{Y}_{\cdot 35}$

Figure 3 Generalized Randomized Block Design With $N = 30$ Participants, $p = 3$ Treatment Levels, and $w = 5$ Groups of $np = (2)(3) = 6$ Homogeneous Participants

	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	
$Block_1$	a_1b_1	Y_{111}	a_1b_2	Y_{112}	a_2b_1	Y_{121}	a_2b_2	Y_{122}	$\overline{Y}_{1..}$
$Block_2$	a_1b_1	Y_{211}	a_1b_2	Y_{212}	a_2b_1	Y_{221}	a_2b_2	Y_{222}	$\overline{Y}_{2..}$
$Block_3$	a_1b_1	Y_{311}	a_1b_2	Y_{312}	a_2b_1	Y_{321}	a_2b_2	Y_{322}	$\overline{Y}_{3..}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Block_{10}$	a_1b_1	$Y_{10,\,11}$	a_1b_2	$Y_{10,\,12}$	a_2b_1	$Y_{10,\,21}$	a_2b_2	$Y_{10,\,22}$	$\overline{Y}_{10..}$
		$\overline{Y}_{\cdot 11}$		$\overline{Y}_{\cdot 12}$		$\overline{Y}_{\cdot 21}$		$\overline{Y}_{\cdot 22}$	

Figure 4 Layout for a Two-Treatment, Randomized Block Factorial Design in Which Four Homogeneous Participants Are Randomly Assigned to the $pq = 2 \times 2 = 4$ Treatment Combinations in Each Block

Note: a_jb_k denotes a treatment combination (Treat. Comb.); Y_{ijk} denotes a measure of the dependent variable (Dep. Var.).

block design, represent a nuisance variable that a researcher wants to remove from the error effects. The generalized randomized block design can be used when a researcher is interested in one treatment with $p \geq 2$ treatment levels and the researcher has sufficient participants to form w groups, each containing np homogeneous participants. The total number of participants in the design is $N = npw$. The layout for the design is shown in Figure 3.

In the memorization experiment described earlier, suppose that 30 volunteers are available. The 30 participants are ranked with respect to IQ. The $np = (2)(3) = 6$ participants with the highest IQs are assigned to Group 1, the next 6 participants are assigned to Group 2, and so on. The $np = 6$ participants in each group are then randomly assigned to the $p = 3$ treatment levels with the restriction that $n = 2$ participants are assigned to each level.

The total *SS* and total degrees of freedom are partitioned as follows:

$$SSTOTAL = SSA + SSG + SSA \times G + SSWCELL$$
$$npw - 1 = (p-1) + (w-1) + (p-1)(w-1) + pw(n-1),$$

where *SSG* denotes the groups *SS* and $SSA \times G$ denotes the interaction of Treatment *A* and groups. The within-cells *SS*, *SSWCELL*, is used to estimate error effects. Three null hypotheses can be tested:

1. $H_o : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{p\cdot}$

   (Treatment *A* population means are equal),

2. $H_0 : \sigma^2_{\mathrm{G}} = 0$

   (variance of the groups, *G*, population means is equal to zero),

3. $H_0 : \sigma^2_{\mathrm{A\times G}} = 0$

   (variance of the $A \times G$ interaction is equal to zero),

where μ_{jz} denotes a population mean for the *i*th participant in the *j*th treatment level and *z*th group. The three null hypotheses are tested using the following *F* statistics:

1. $$F = \frac{SSA/(p-1)}{SSWCELL/[pw(n-1)]} = \frac{MSA}{MSWCELL},$$

2. $$F = \frac{SSG/(w-1)}{SSWCELL/[pw(n-1)]} = \frac{MSG}{MSWCELL},$$

3. $$F = \frac{SSA \times G/(p-1)(w-1)}{SSWCELL/[pw(n-1)]} = \frac{MSA \times G}{MSWCELL}$$

The generalized randomized block design enables a researcher to isolate one nuisance variable—an advantage that it shares with the randomized block design. Furthermore, the design uses the within-cell variation in the $pw = (3)(5) = 15$ cells to estimate error effects rather than an interaction, as in the randomized block design. Hence, the restrictive sphericity assumption of the randomized block design is replaced with the assumption of homogeneity of within-cell population variances.

Block Designs With Two or More Treatments

The blocking procedure that is used with a randomized block design can be extended to experiments that have two or more treatments, denoted by the letters *A*, *B*, *C*, and so on.

Randomized Block Factorial Design

A randomized block factorial design with two treatments, denoted by *A* and *B*, is constructed by crossing the p levels of Treatment *A* with the q levels of

Treatment *B*. The design's n blocks each contain $p \times q$ treatment combinations: $a_1b_1, a_1b_2 \ldots a_pb_q$. The design enables a researcher to isolate variation attributable to one nuisance variable while simultaneously evaluating two treatments and associated interaction.

The layout for the design with $p = 2$ levels of Treatment *A* and $q = 2$ levels of Treatment *B* is shown in Figure 4. It is apparent from Figure 4 that all the participants are used in simultaneously evaluating the effects of each treatment. Hence, the design permits efficient use of resources because each treatment is evaluated with the same precision as if the entire experiment had been devoted to that treatment alone.

The total *SS* and total degrees of freedom for a two-treatment randomized block factorial design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSBL + SSA + SSB \\ npq - 1 &= (n-1) + (p-1) + (q-1) \\ &+ SSA \times B + SSRESIDUAL \\ &+ (p-1)(q-1) + (n-1)(pq-1). \end{aligned}$$

Four null hypotheses can be tested:

1. H_0: $\sigma^2_{BL} = 0$ (variance of the blocks, *BL*, population means is equal to zero),
2. H_0: $\mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot} = \cdots = \mu_{\cdot p\cdot}$ (Treatment *A* population means are equal),
3. H_0: $\mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot q}$ (Treatment *B* population means are equal),
4. H_0: $A \times B$ interaction = 0 (Treatments *A* and *B* do not interact),

where μ_{ijk} denotes a population mean for the ith block, jth level of Treatment *A*, and kth level of treatment *B*. The F statistics for testing the null hypotheses are as follows:

$$\begin{aligned} F &= \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(pq-1)]} \\ &= \frac{MSBL}{MSRESIDUAL}, \end{aligned}$$

$$\begin{aligned} F &= \frac{SSA/(p-1)}{SSRESIDUAL/[(n-1)(pq-1)]} \\ &= \frac{MSA}{MSRESIDUAL}, \end{aligned}$$

$$\begin{aligned} F &= \frac{SSB/(q-1)}{SSRESIDUAL/[(n-1)(pq-1)]} \\ &= \frac{MSB}{MSRESIDUAL}, \end{aligned}$$

$$\begin{aligned} F &= \frac{SSA \times B/(p-1)(q-1)}{SSRESIDUAL/[(n-1)(pq-1)]} \\ &= \frac{MSA \times B}{MSRESIDUAL}. \end{aligned}$$

The design shares the advantages and disadvantages of the randomized block design. Furthermore, the design enables a researcher to efficiently evaluate two or more treatments and associated interactions in the same experiment. Unfortunately, the design lacks simplicity in the interpretation of the results if interaction effects are present. The design has another disadvantage: If Treatment *A* or *B* has numerous levels, say four or five, the block size becomes prohibitively large. For example, if $p = 4$ and $q = 3$, the design has blocks of size $4 \times 3 = 12$. Obtaining n blocks with 12 matched participants or observing n participants on 12 occasions is often not feasible. A design that reduces the size of the blocks is described next.

Split-Plot Factorial Design

In the late 1920s, Fisher and Frank Yates addressed the problem of prohibitively large block sizes by developing confounding schemes in which only a portion of the treatment combinations in an experiment are assigned to each block. The split-plot factorial design achieves a reduction in the block size by confounding one or more treatments with groups of blocks. *Group–treatment confounding* occurs when the effects of, say, Treatment *A* with p levels are indistinguishable from the effects of p groups of blocks.

The layout for a two-treatment split-plot factorial design is shown in Figure 5. The block size in the split-plot factorial design is half as large as the block size of the randomized block factorial design in Figure 4 although the designs contain the same treatment combinations. Consider the sample means $\overline{Y}_{\cdot 1\cdot}$ and $\overline{Y}_{\cdot 2\cdot}$ in Figure 5. Because of confounding, the difference between $\overline{Y}_{\cdot 1\cdot}$ and $\overline{Y}_{\cdot 2\cdot}$ reflects both group effects and Treatment *A* effects.

The total *SS* and total degrees of freedom for a split-plot factorial design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSBL(A) + SSB \\ &+ SSA \times B + SSRESIDUAL \\ npq - 1 &= (p-1) + p(n-1) + (q-1) \\ &+ (p-1)(q-1) + p(n-1)(q-1), \end{aligned}$$

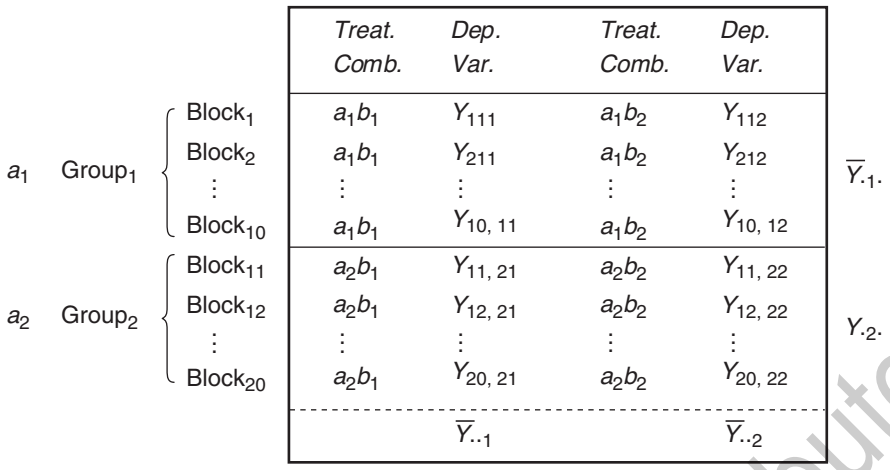


Figure 5 Layout for a Two-Treatment, Split-Plot Factorial Design in Which 10 + 10 = 20 Homogeneous Blocks Are Randomly Assigned to the Two Groups

Note: a_jb_k denotes a treatment combination (Treat. Comb.); Y_{ijk} denotes a measure of the dependent variable (Dep. Var.). Treatment A is confounded with groups. Treatment B and the $A \times B$ are not confounded.

where $SSBL(A)$ denotes the SS for blocks within Treatment A. Three null hypotheses can be tested:

1. H_0: $\mu_{.1.} = \mu_{.2.} = \cdots = \mu_{.p}$ (Treatment A population means are equal),
2. H_0: $\mu_{..1} = \mu_{..2} = \cdots = \mu_{..q}$ (Treatment B population means are equal),
3. H_0: $A \times B$ interaction = 0 (Treatments A and B do not interact),

where μ_{jk} denotes the jth block, jth level of treatment A, and kth level of treatment B. The F statistics are

$$F = \frac{SSA/(p-1)}{SSBL(A)/[p(n-1)]} = \frac{MSA}{MSBL(A)},$$

$$F = \frac{SSB/(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]} = \frac{MSB}{MSRESIDUAL},$$

$$F = \frac{SSA \times B/(p-1)(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]} = \frac{MSA \times B}{MSRESIDUAL}.$$

The split-plot factorial design uses two error terms: $MSBL(A)$ is used to test Treatment A; a different and usually much smaller error term, $MSRESIDUAL$, is used to test Treatment B and the $A \times B$ interaction. Because $MSRESIDUAL$ is generally smaller than $MSBL(A)$, the power of the tests of Treatment B and the $A \times B$ interaction is greater than that for Treatment A.

Roger E. Kirk

See also Analysis of Variance (ANOVA); Confounding; F Test; Nuisance Variable; Null Hypothesis; Sphericity; Sums of Squares

Further Readings

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer-Verlag.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (2002). Experimental design. In I. B. Weiner (Series Ed.) & J. Schinka & W. F. Velicer (Vol. Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 3–32). New York: Wiley.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

BLOCKMODELING

Blockmodeling is an approach for partitioning or clustering units (e.g., nodes, vertices, actors) of a network based on patterns (i.e., structure) of their ties to each other. By shrinking the groups that are then obtained, a new network called a *blockmodel* is

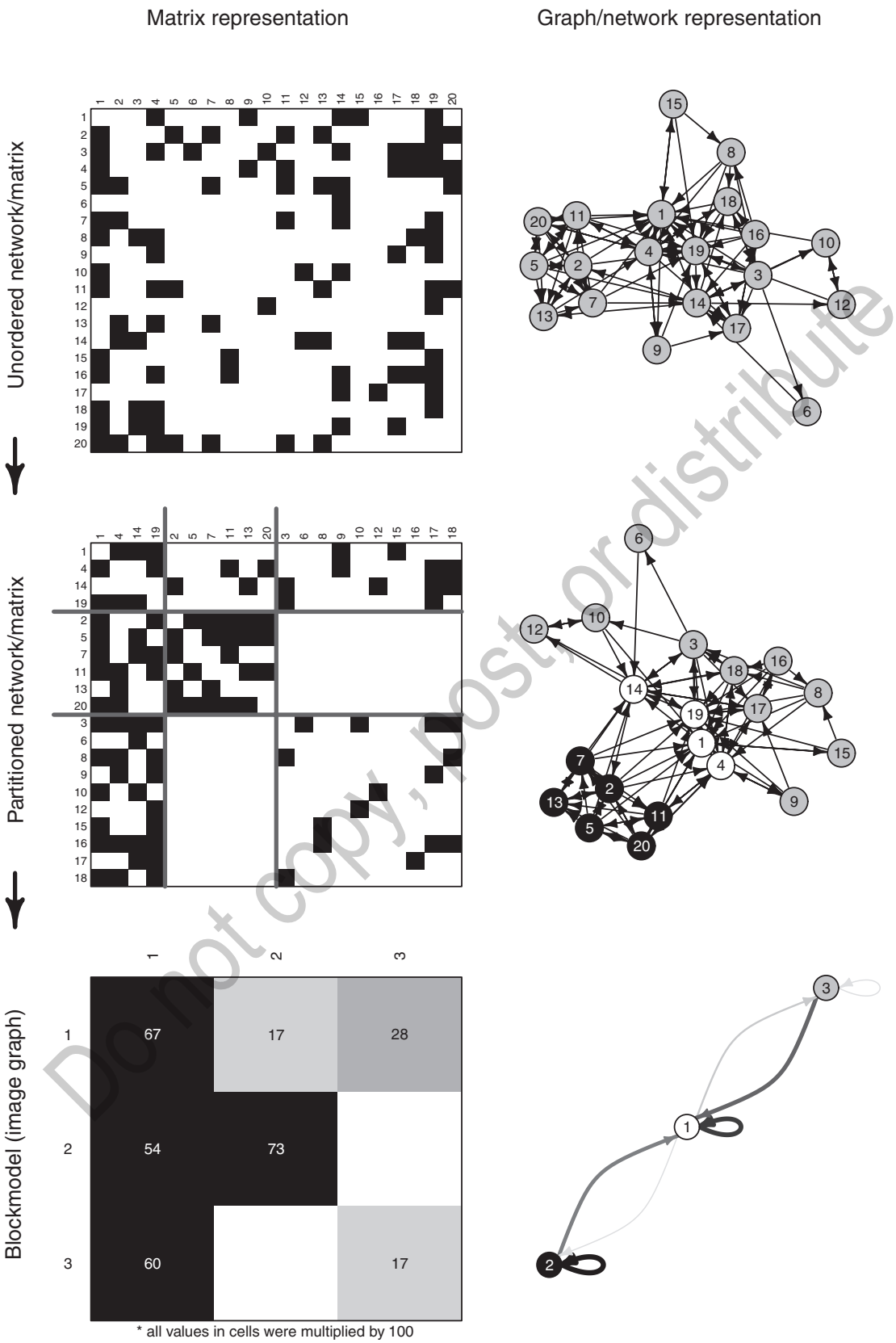


Figure 1 From an Unordered Network/Matrix to a Blockmodel

created whereby units represent groups and ties represent ties among these groups. This process is depicted in Figure 1: The right column shows graph representation and the left column shows the matrix representation of the network. Blockmodeling is therefore used to transform a large and complex network into a smaller and more comprehensible one. Blockmodeling has also been used to operationalize social roles.

A More Detailed Description

The term *blockmodeling* arises from the fact that when a network is represented by a matrix, and this matrix is rearranged according to a partition (so that the units that are part of the same groups are located together), blocks that represent ties within and between groups appear if one separates the groups by lines. These blocks are seen in the second *row* of the left column in Figure 1, where nine blocks are visible.

More formally, blockmodeling is an approach for clustering units in a network based on some form of equivalence. Equivalence is something that tells which units can be considered equal in some respect. As units typically are not perfectly equivalent to each other, blockmodeling methods usually attempt to cluster together units that are more equivalent to each other than to other units.

Equivalences

The form of equivalence class that is most commonly used is structural equivalence. Units are structurally equivalent if they are connected in the same way to the same units. If a partition is compatible with structural equivalence—that is, if the units within all the groups are perfectly structurally equivalent—then all ties within each block that are induced by the partition into such groups are all equal. With binary networks, this means that either the induced blocks are null (empty, meaning that no ties are present or, equivalently, that all ties have a value of 0) or they are complete (all possible ties are present or, equivalently, all of the ties have a value of 1).

Similarly, units are regularly equivalent if they are connected in the same way to equivalent others. The definition is circular by design. For binary networks, if a partition is compatible with regular equivalence, then all the blocks induced by that partition are either null or regular. A regular block is a block that has at least one tie in each row and each column. Namely, each unit from a *row* group must be connected to at least one unit from the *column* group.

Generalized equivalence is defined by block types and possibly their position in the blockmodel. Block type defines the allowed pattern of ties within a block. While describing structural and regular equivalence, these patterns are described earlier for binary networks for null, complete, and regular block types, yet numerous block types are defined for both binary and valued networks. Therefore, generalized equivalence is not a single equivalence class, but a way of specifying custom equivalences.

Another frequently used form of equivalence is stochastic equivalence. Units are stochastically equivalent if they have the same probability of having a tie to all other units.

Types of Blockmodeling

Blockmodeling approaches are characterized by several features. One such feature is the kind of networks the approaches are designed to analyze. For example, there are approaches for one-mode networks (all units can connect to any other unit and are of the same type) as well as for two-mode networks (ties are only possible between units of different types). Similarly, there are approaches that can handle either only single-relational networks or also multirelational ones. Certain approaches can also handle temporal or multilevel networks. Still pertaining to the type of networks, some approaches can handle only binary networks, while others can handle signed (where negative ties are also possible) or values networks (where ties can also have other values beyond simply 0 or 1).

Other divisions are based on the underlying characteristics of the algorithm being used. An important distinction is between stochastic and deterministic blockmodeling. Stochastic blockmodeling approaches are essentially those based on stochastic equivalence. They rely on some probabilistic model or generative model of network formation. Similar models also go by the name of mixture models for block clustering.

In contrast, deterministic approaches are not based on such a model and can be split into direct and indirect methods. Indirect methods involve two steps. First, a dissimilarity matrix among all units compatible with the selected equivalence must be computed. Second, building on that, a partition is obtained via some classical clustering technique, usually hierarchal clustering.

In contrast with the indirect approaches, the direct approaches partition units by directly optimizing some criterion function (which again must be compatible with the selected equivalence). Direct approaches vary according to both the criterion function they optimize and the optimization method. For instance, the criterion function for binary networks could be the number of

errors in blocks or, for both binary and valued networks, the sum of squares of deviations from the block mean (analogous to Ward's criterion function or *k*-means criterion from classical cluster analysis).

Most deterministic approaches are only able to find partitions compatible with structural equivalence. A notable exception is Patrick Doreian, Vladimir Batagelj, and Anuška Ferligoj's generalized blockmodeling. Generalized blockmodeling is also able to find partitions compatible with regular and generalized equivalence; therefore, it can be used to find partitions and blockmodels with specific properties. When using generalized blockmodeling, some or all parts of the blockmodel can be prespecified.

Aleš Žiberna

Author's Note: The author acknowledges the financial support from the Slovenian Research Agency (Research Core Funding No. P5-0168 and the project "Blockmodeling Multilevel and Temporal Networks" No. J7-8279).

See also Cluster Analysis; Mixture Models; Network Analysis; Social Network Analysis

Further Readings

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized blockmodeling*. New York: Cambridge University Press.

Doreian, P., Batagelj, V., & Ferligoj, A. (Eds.). (2020). *Advances in network clustering and blockmodeling*. Hoboken, NJ: Wiley-Blackwell.

Funke, T., & Becker, T. (2019). Stochastic block models: A comparison of variants and inference methods. *PLoS One*, *14*(4), e0215296. doi:10.1371/journal.pone.0215296.

Bonferroni Procedure

The Bonferroni procedure is a statistical adjustment to the significance level of hypothesis tests when multiple tests are being performed. The purpose of an adjustment such as the Bonferroni procedure is to reduce the probability of identifying significant results that do not exist, that is, to guard against making Type I errors (rejecting null hypotheses when they are true) in the testing process. This potential for error increases with an increase in the number of tests being performed in a given study and is due to the multiplication of probabilities across the multiple tests. The Bonferroni procedure is often used as an adjustment in multiple comparisons after a significant finding in an analysis of variance (ANOVA) or when constructing simultaneous confidence intervals for several population parameters, but more broadly, it can be used in any situation that involves multiple tests. The Bonferroni procedure is one of the more commonly used procedures in multiple testing situations, primarily because it is an easy adjustment to make. A strength of the Bonferroni procedure is its ability to maintain Type I error rates at or below a nominal value. A weakness of the Bonferroni procedure is that it often overcorrects, making testing results too conservative because of a decrease in statistical power.

A variety of other procedures have been developed to control the overall Type I error level when multiple tests are performed. Some of these other multiple comparison and multiple testing procedures, including the Student–Newman–Keuls procedure, are derivatives of the Bonferroni procedure, modified to make the procedure less conservative without sacrificing Type I error control. Other multiple comparison and multiple testing procedures are simulation based and are not directly related to the Bonferroni procedure.

This entry describes the procedure's background, explains the procedure, and provides an example. This entry also presents applications for the procedure and examines recent research.

Background

The Bonferroni procedure is named after the Italian mathematician Carlo Emilio Bonferroni. Although his work was in mathematical probability, researchers have since applied his work to statistical inference. Bonferroni's principal contribution to statistical inference was the identification of the probability inequality that bears his name.

Explanation

The Bonferroni procedure is an application of the *Bonferroni inequality* to the probabilities associated with multiple testing. It prescribes using an adjustment to the significance level for individual tests when simultaneous statistical inference for several tests is being performed. The adjustment can be used for bounding simultaneous confidence intervals, as well as for simultaneous testing of hypotheses.

The Bonferroni inequality states the following:

1. Let $A_i, i = 1$ to k, represent k events. Then, $P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P\left(\overline{A}_i\right)$, where $\overline{A}_i$ is the complement of the event *Ai*.

2. Consider the mechanics of the Bonferroni inequality, $P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P\left(\overline{A}_i\right)$,

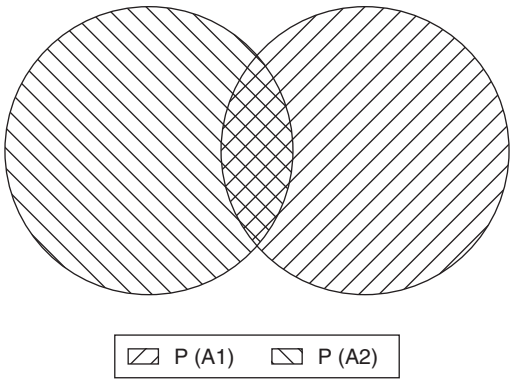


Figure 1 Illustration of Bonferroni's Inequality

Note: A1 = event 1; A2 = event 2; P(A1) = probability of A1; P(A2) = probability of A2. Interaction between events results in redundance in probability.

3. and rewrite the inequality as follows:

$$1-P\left(\bigcap_{i=1}^{k}A_i\right)\leq\sum_{i=1}^{k}P\left(\overline{A}_i\right).$$

Now, consider $\overline{A}_i$ as a Type I error in the ith test in a collection of k hypothesis tests. Then $P\left(\bigcap_{i=1}^{k}A_i\right)$ represents the probability that no Type I errors occur in the k hypothesis tests, and $1-P\left(\bigcap_{i=1}^{k}A_i\right)$ represents the probability that at least one Type I error occurs in the k hypothesis tests. $P\left(\overline{A}_i\right)$ represents the probability of a Type I error in the ith test, and we can label this probability as $\alpha_i=P\left(\overline{A}_i\right)$. So Bonferroni's inequality implies that the probability of at least one Type I error occurring in k hypothesis tests is $\leq\sum_{i=1}^{k}\alpha_i$.

If, as is often assumed, all k tests have the same probability of a Type I error, α, then we can conclude that the probability of at least one Type I error occurring in k hypothesis tests is $\leq k\alpha$.

Consider an illustration of Bonferroni's inequality in the simple case in which $k=2$: Let the two events A1 and A2 have probabilities P(A1) and P(A2), respectively. The sum of the probabilities of the two events is clearly greater than the probability of the union of the two events because the former counts the probability of the intersection of the two events twice, as shown in Figure 1.

The Bonferroni procedure is simple in the sense that a researcher need only know the number of tests to be performed and the probability of a Type I error for those tests in order to construct this upper bound on the experiment-wise error rate. However, as mentioned earlier, the Bonferroni procedure is often criticized for being too conservative. Consider that the researcher does not typically know what the actual Type I error rate is for a given test. Rather, the researcher constructs the test so that the maximum allowable Type I error rate is α. Then the actual Type I error rate may be considerably less than α for any given test.

For example, suppose a test is constructed with a nominal $\alpha=.05$. Suppose the researcher conducts $k=10$ such tests on a given set of data, and the actual Type I error rate for each of the tests is .04. Using the Bonferroni procedure, the researcher concludes that the experiment-wise error rate is at most 10*.05, or .50. The error rate in this scenario is in fact .40, which is considerably less than .50.

As another example, consider the extreme case in which all $k=10$ hypothesis tests are exactly dependent on each other—the same test is conducted 10 times on the same data. In this scenario, the experiment-wise error rate does not increase because of the multiple tests. In fact, if the Type I error rate for one of the tests is $\alpha=.05$, the experiment-wise error rate is the same, .05, for all 10 tests simultaneously. We can see this result from the Bonferroni inequality: $P\left(\bigcap_{i=1}^{k}A_i\right)=P\left(A_i\right)$ when the events, A_i, are all the same because $\bigcap_{i=1}^{k}A_i=A_i$. The Bonferroni procedure would suggest an upper bound on this experiment-wise probability as .50—overly conservative by 10-fold! It would be unusual for a researcher to conduct k equivalent tests on the same data. However, it would not be unusual for a researcher to conduct k tests and for many of those tests, if not all, to be partially interdependent. The more interdependent the tests are, the smaller the experiment-wise error rate and the more overly conservative the Bonferroni procedure is.

Other procedures have sought to correct for inflation in experiment-wise error rates without being as conservative as the Bonferroni procedure. However, none are as simple to use. These other procedures include the Student–Newman–Keuls, Tukey, and Scheffé procedures, to name a few. Descriptions of these other procedures and their uses can be found in many basic statistical methods textbooks, as well as this encyclopedia.

Example

Consider the case of a researcher studying the effect of three different teaching methods on the average words per minute (μ_1,μ_2,μ_3) at which a student can read. The researcher tests three hypotheses: $\mu_1=\mu_2(\text{vs. }\mu_1\neq\mu_2),\mu_1=\mu_3(\text{vs. }\mu_1\neq\mu_3)$, and $\mu_2=\mu_3(\text{vs. }\mu_2\neq\mu_3)$. Each test is conducted at a nominal level, $\alpha_0=.05$, resulting in a comparison-wise error rate of $\alpha_c=.05$ for

each test. Denote A1, A2, and A3 as the event of falsely rejecting the null hypotheses 1, 2, and 3, respectively, and denote $p1$, $p2$, and $p3$ the probability of events A1, A2, and A3, respectively. These would be the individual p values for these tests. It may be assumed that some dependence exists among the three events, A1, A2, and A3, principally because the events are all based on data collected from a single study. Consequently, the experiment-wise error rate, the probability of falsely rejecting any of the three null hypotheses, is at least equal to $\alpha_e = .05$ but potentially as large as $.05 * 3 = .15$. For this reason, we may apply the Bonferroni procedure by dividing our nominal level of $\alpha_0 = .05$ by $k = 3$ to obtain $\alpha_0^* = .0167$. Then, rather than comparing the p values $p1$, $p2$, and $p3$ to $\alpha_0 = 05$, we compare them to $\alpha_0^* = .0167$. The experiment-wise error rate is therefore adjusted down so that it is less than or equal to the original intended nominal level of $\alpha_0 = .05$.

It should be noted that although the Bonferroni procedure is often used in the comparison of multiple means, because the adjustment is made to the nominal level, α_0, or to the test's resulting p value, the multiple tests could be hypothesis tests of any population parameters based on any probability distributions. So, for example, one experiment could involve a hypothesis test regarding a mean and another hypothesis test regarding a variance, and an adjustment based on $k = 2$ could be made to the two tests to maintain the experiment-wise error rate at the nominal level.

Applications

As noted above, the Bonferroni procedure is used primarily to control the overall α level (i.e., the experiment-wise level) when multiple tests are being performed. Many statistical procedures have been developed at least partially for this purpose; however, most of those procedures have applications exclusively in the context of making multiple comparisons of group means after finding a significant ANOVA result. While the Bonferroni procedure can also be used in this context, one of its advantages over other such procedures is that it can also be used in other multiple testing situations that do not initially entail an omnibus test such as ANOVA.

For example, although most statistical tests do not advocate using a Bonferroni adjustment when testing beta coefficients in a multiple regression analysis, it has been shown that the overall Type I error rate in such an analysis involving as few as eight regression coefficients can exceed .30, resulting in almost a 1 in 3 chance of falsely rejecting a null hypothesis. Using a Bonferroni adjustment when one is conducting these tests would control that overall Type I error rate. Similar adjustments can be used to test for main effects and interactions in ANOVA and multivariate ANOVA designs because all that is required to make the adjustment is that the researcher knows the number of tests being performed. The Bonferroni adjustment has been used to adjust the experiment-wise Type I error rate for multiple tests in a variety of disciplines, such as medical, educational, and psychological research, to name a few.

Recent Research

One of the main criticisms of the Bonferroni procedure is the fact that it overcorrects the overall Type I error rate, which results in lower statistical power. Many modifications to this procedure have been proposed over the years to try to alleviate this problem. Most of these proposed alternatives can be classified either as *step-down procedures* (e.g., the Holm method), which test the most significant (and, therefore, smallest) p value first, or *step-up procedures* (e.g., the Hochberg method), which begin testing with the least significant (and largest) p value. With each of these procedures, although the tests are all being conducted concurrently, each hypothesis is not tested at the same time or at the same level of significance.

More recent research has attempted to find a divisor between 1 and k that would protect the overall Type I error rate at or below the nominal .05 level but closer to that nominal level so as to have a lesser effect on the power to detect actual differences. This attempt was based on the premise that making no adjustment to the α level is too liberal an approach (inflating the experiment-wise error rate), and dividing by the number of tests, k, is too conservative (overadjusting that error rate). It was shown that the optimal divisor is directly determined by the proportion of nonsignificant differences or relationships in the multiple tests being performed. Based on this result, a divisor of $k(1-q)$, where q = the proportion of nonsignificant tests, did the best job of protecting against Type I errors without sacrificing as much power. Unfortunately, researchers often do not know, a priori, the number of nonsignificant tests that will occur in the collection of tests being performed. Consequently, research has also shown that a practical choice of the divisor is $k/1.5$ (rounded to the nearest integer) when the number of tests is greater than three. This modified Bonferroni adjustment will outperform alternatives in keeping the experiment-wise error rate at or below the nominal .05 level and will have higher power than other commonly used adjustments.

Jamis J. Perrett and Daniel J. Mundfrom

See also Analysis of Variance (ANOVA); Hypothesis; Multiple Comparison Tests; Newman–Keuls Test and Tukey Test; Scheffé Test

Further Reqdings

Bain, L. J., & Engelhandt, M. (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Boston: PWS-Kent.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, *43*, 417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386.

Mundfrom, D. J., Perrett, J., Schaffer, J., Piccone, A., & Roozeboom, M. A. (2006). Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints*, *32*(1), 1–6.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*(3), 663–665.

Roozeboom, M. A., Mundfrom, D. J., & Perrett, J. (2008, August). *A single-step modified Bonferroni procedure for multiple tests*. Paper presented at the Joint Statistical Meetings, Denver, CO.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*(395), 826–831.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests using SAS*. Cary, NC: SAS Institute.

BOOTSTRAPPING

Bootstrapping is a computerized simulation operation that involves random resampling from the data set one is using to produce perhaps thousands of new data sets that have similar participant compositions to the original data set. For example, if a researcher had a data set containing 500 participants, the bootstrapping procedure would sample from the original data set to create new data sets, each of 500 participants. It is used to determine the statistical significance of parameter estimates whose significance cannot be tested using established statistical distributions (e.g., *t*, *F*, and chi-square distributions). One might use bootstrapping when data are nonnormal or the distribution of a parameter is difficult to anticipate (e.g., indirect effects in path analysis). Bootstrapping estimates the standard error (standard deviation of a set of statistical values such as a set of means from multiple samples) for a given statistical analysis, as well as confidence intervals for a parameter. Standard errors and confidence intervals are key to determining statistical significance.

This entry describes how and why bootstrapping samples are created, and how bootstrap samples indicate statistical significance of results. Other resampling techniques that are alternatives to bootstrapping are also introduced. The entry concludes with a review of software that can be used for bootstrapping.

How Bootstrap Samples Are Drawn

The key to bootstrapping is that the new bootstrap simulation samples are drawn *with replacement*. Imagine that a researcher places five ping-pong balls into a hat, the number "1" having been painted onto one ball, a "2" onto another, and the same for "3," "4," and "5." Suppose further that the researcher plans to draw a random sample of three balls. The term *with replacement* refers to the fact that the ball that is drawn first is put back into the hat to possibly be drawn again. The same is done with the ball drawn second. Hence, the final three-ball sample conceivably could consist of balls 2, 2, and 4, or balls 1, 3, and 3, for example. If sampling is done *without* replacement, a ball selected into the sample is not returned to the hat; hence, the same number cannot appear more than once in the sample. Possible samples might, therefore, include balls 1, 2, and 3, or balls 2, 4, and 5.

Without replacement, drawing 500 people to create new data sets of 500 (a more realistic example for bootstrapping) would simply keep reproducing the original sample of the same 500 people. With replacement, however, the same participant in the original data set could be selected into a new sample, put back into the original sampling pool, then selected again into the same new sample. Other participants in the original sample may not be selected at all into one of the new samples. Suppose a researcher who has a sample of 500 participants wishes to draw 1,000 new bootstrap samples, each with the same sample size of 500. One of these bootstrap samples might include participants with the identification (ID) number 1, ID number 2, ID number 2 again (so that his or her data are used twice), ID number 5, ID number 6, and so forth, all the way to a total of 500 participants. Another bootstrap sample might include participants with ID number 3, ID number 3 again, ID number 4, ID number 7, ID number 7 again, ID number 7 a third time, ID number 10, and so forth.

The Rationale for Bootstrap Samples

Amassing perhaps 1,000 or 5,000 synthetic samples from an original source sample is thought to approximate the real samples that would appear if a researcher drew thousands of random samples from a population of actual people (e.g., residents of Chicago or London). Bootstrapping is used in place of actual population sampling because it would be virtually impossible logistically and financially to draw and interview people comprising thousands of random samples of conventional sizes (e.g., 500–1,000 respondents) from a given geographic entity. For this reason, the unseen source of the bootstrap samples is sometimes referred to as a *surrogate population*, with the resulting samples sometimes called *phantom samples*.

How Bootstrap Samples Indicate Statistical Significance of a Result

Once the bootstrap samples are obtained, the computer program runs the desired analysis (e.g., multiple regression) in each sample and compiles a histogram of the focal statistic (e.g., a beta coefficient from one predictor to the outcome) in each of the bootstrap samples. This step, thus, creates a sampling distribution of the beta coefficient. As a consequence of the Central Limit Theorem, the histogram of regression coefficients will tend toward a normal distribution. Christoph Hanck and colleagues (2020) provide an interactive online animated demonstration of the process. Statistics blogger Jim Frost (2020) notes that the bootstrap method can analyze a large variety of sample statistics, including the mean, median, mode, standard deviation, analysis of variance, correlations, regression coefficients, proportions, odds ratios, variance in binary data, and multivariate statistics.

One way to obtain a bootstrap confidence interval (95% CI, in this example) is to remove the highest 2.5% and lowest 2.5% of parameter estimates (here, regression coefficients) from the histogram of results from the bootstrap samples. This is known as a *percentile* confidence interval and is not necessarily symmetric to the left and right of the mean of the focal parameter. The percentile CI is considered a nonparametric technique. After removal of the upper and lower 2.5% of the distribution, the remaining range from the lowest to highest parameter estimate constitutes the CI. With a large sample and a roughly normal bootstrap sampling distribution, a CI can be determined via equations involving the bootstrap standard error. The focal parameter (e.g., regression coefficient) is significantly different from zero if the CI end points are either both greater than zero (i.e., significantly positive coefficient) or both less than zero (i.e., significantly negative coefficient). Software packages typically display the lower and upper end points of confidence intervals in their output.

Refinements to Bootstrap Solutions

More complex alternatives, including bias-corrected and accelerated (BCa) CIs, can be used to adjust percentile intervals to make them more accurate. Two flaws for which the BCa attempts to correct are *bias* and *skewness* in the distribution. According to David Moore and colleagues (2016, pp. 16–10), "Bias is the difference between the mean of the resample means and the original [sample] mean." Skewness is used here in the ordinary sense, in terms of asymmetry between the two sides of the distribution. Moore and colleagues note that calculation of such CIs is highly technical but urge the use of BCa or similar methods when available in the software package one is using.

Illustration With Path Analysis Indirect Effects

One common use of bootstrapping in psychology and other social sciences is to determine the significance of indirect or mediational effects in path analysis modeling from an antecedent variable to an outcome variable. The magnitude of indirect effects is calculated by multiplying the two respective path coefficients, from an antecedent to a proposed mediator variable and from the mediator to an outcome (e.g., exposure to stressful life circumstances leading to increases in cortisol and cortisol leading to physical symptoms). Bootstrap tests of the significance of indirect effects would be implemented in line with the aforementioned general steps, by generating large numbers of random resamplings of the original data set, running path analysis models in each of the new synthetic samples, and plotting a histogram of the resulting indirect effects. Andrew Hayes has created a macro (add-on) to perform bootstrap mediation (and other) analyses with indirect effects in the SPSS and SAS packages (with an R macro in development at this writing).

Alternative Resampling Techniques

To provide context on bootstrapping's capabilities, limitations, and scope, two alternatives to it that also use resampling to determine statistical significance—jackknifing and permutation tests—are discussed briefly. Jackknifing creates repeated synthetic samples by deleting one observation from the original sample.

Permutation tests provide a way to assess significance by comparing a test statistic from one's sample with a critical value (akin to how, before computers automatically generated *p* levels for *t*- or chi-square tests, researchers used to consult tables in the back of a statistics textbook to see if their obtained statistic exceeded a critical value for significance). Permutation tests create synthetic distributions (centered at zero, akin to a *z*-distribution, representing the null hypothesis) from one's data, providing a way to compare an obtained test statistic to a critical value.

Bootstrapping Versus Other Resampling Techniques

Because bootstrapping and jackknifing are the most similar to each other and often are discussed in conjunction with each other, direct comparison of the two techniques is warranted. First, jackknifing is a simpler approximation to bootstrapping, suggesting the latter should be preferred unless prohibitively difficult (e.g., limited computer resources). Second, properties of a statistical parameter including *linearity* (i.e., a function involving only basic mathematical operations such as adding and multiplying, as opposed to raising numbers to higher powers such as squaring) and *smoothness* affect the usefulness of bootstrapping and jackknifing. Jackknife analyses of nonsmooth parameters such as the median and of nonlinear functions perform poorly, suggesting that bootstrapping can be used with a broader range of statistics than can jackknifing. Third, as discussed by Rodgers, the sampling frame of possible combinations of cases is much larger for bootstrapping than for other resampling techniques, making bootstrapping advantageous. On the other hand, bootstrapping is prone to difficulties with small samples.

Software Packages for Bootstrapping

The major statistical packages have general bootstrapping routines (i.e., applicable to a wide variety of techniques), sometimes at extra cost beyond the basic package. SAS provides macros known as *%BOOT* for a normally distributed sampling distribution and *%BOOTCI* for confidence intervals from nonnormal distributions. SPSS provides bootstrap options in its menu-driven modules for several techniques (e.g., mean, correlation). Stata offers syntax-based bootstrap routines. Finally, many different bootstrap routines in R can easily be found through internet searches.

Alan Reifman and Sylvia Niehuis

See also Confidence Intervals; Jackknife; Randomization Tests; Standard Error of the Mean

Further Readings

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London, UK: Chapman & Hall. doi:10.1007/978-1-4899-4541-9.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed., chap. 21). Thousand Oaks, CA: Sage.

Frost, J. (2020). Introduction to bootstrapping in statistics with an example. *Statistics by Jim*. Retrieved from https://statisticsbyjim.com/hypothesis-testing/bootstrapping/

Giles, D. (2019). What is a permutation test? *R-bloggers*. Retrieved from https://www.r-bloggers.com/2019/04/what-is-a-permutation-test/

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2020). *Introduction to econometrics with R* [section 4.5]. Retrieved from https://www.econometrics-with-r.org/ (animated demonstration at https://www.econometrics-with-r.org/4-5-tsdotoe.html).

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York: Guilford Press.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *American Statistician*, *69*, 371–386. doi:10.1080/00031305.2015.1089789.

McIntosh, A. I. (2016). The jackknife estimation method. *Statistics ArXiv*. Retrieved from https://arxiv.org/abs/1606.00497

Moore, D. S., McCabe, G. P., & Craig, B. A. (2016). *Introduction to the practice of statistics* (9th ed., chap. 16). Equitable Building. Basingstoke, UK: MacMillan.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441–456. doi:10.1207/S15327906MBR3404_2.

Singh, K., & Xie, M. (2010). Bootstrap method. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 46–51). Amsterdam, Netherlands: Elsevier Science. doi:10.1016/B978-0-08-044894-7.01309-9.

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, *115*, 929–943. doi:10.1037/pspa0000132.

Box-And-Whisker Plot

A box-and-whisker plot, or box plot, is a tool used to visually display the range, distribution symmetry, and central tendency of a distribution in order to illustrate

the variability and the concentration of values within a distribution. The box plot is a graphical representation of the five-number summary, or a quick way of summarizing the center and dispersion of data for a variable. The five-number summary includes the minimum value, 1st (lower) quartile (Q^1), median, 3rd (upper) quartile (Q^3), and the maximum value. Outliers are also indicated on a box plot. Box plots are especially useful in research methodology and data analysis as one of the many ways to visually represent data. From this visual representation, researchers glean several pieces of information that may aid in drawing conclusions, exploring unexpected patterns in the data, or prompting the researcher to develop future research questions and hypotheses. This entry provides an overview of the history of the box plot, key components and construction of the box plot, and a discussion of the appropriate uses of a box plot.

History

A box plot is one example of a graphical technique used within exploratory data analysis (EDA). EDA is a statistical method used to explore and understand data from several angles in social science research. EDA grew out of work by John Tukey and his associates in the 1960s and was developed to broadly understand the data, graphically represent data, generate hypotheses and build models to guide research, add robust measures to an analysis, and aid the researcher in finding the most appropriate method for analysis. EDA is especially helpful when the researcher is interested in identifying any unexpected or misleading patterns in the data. Although there are many forms of EDA, researchers must employ the most appropriate form given the specific procedure's purpose and use.

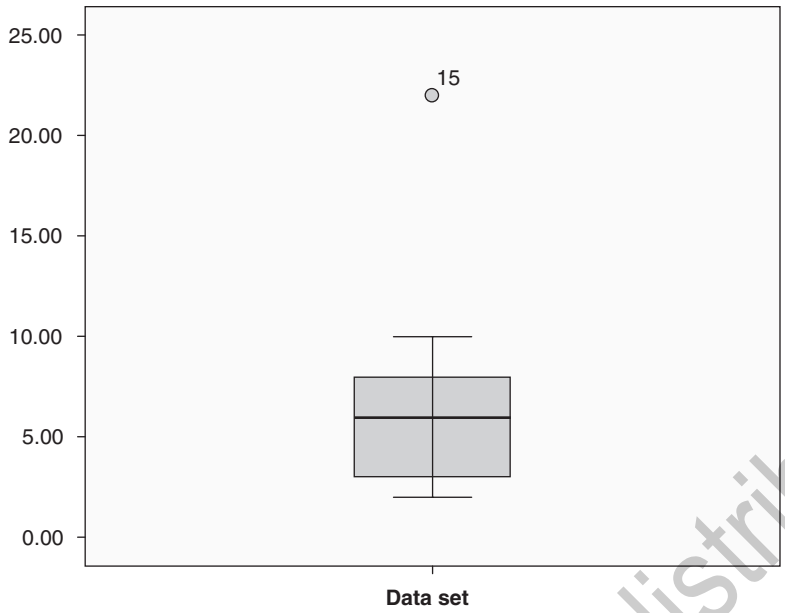
Definition and Construction

One of the first steps in any statistical analysis is to describe the central tendency and the variability of the values for each variable included in the analysis. The researcher seeks to understand the center of the distribution of values for a given variable (*central tendency*) and how the rest of the values fall in relation to the center (*variability*). Box plots are used to visually display variable distributions through the display of robust statistics, or statistics that are more resistant to the presence of outliers in the data set. Although there are somewhat different ways to construct box plots depending on the way in which the researcher wants to display outliers, a box plot always provides a visual display of the five-number summary. The median is defined as the value that falls in the middle after the values for the selected variable are ordered from lowest to highest value, and it is represented as a line in the middle of the rectangle within a box plot. As it is the central value, 50% of the data lie above the median and 50% lie below the median. When the distribution contains an odd number of values, the median represents an actual value in the distribution. When the distribution contains an even number of values, the median represents an average of the two middle values.

To create the rectangle (or box) associated with a box plot, one must determine the 1st and 3rd quartiles, which represent values (along with the median) that divide all the values into four sections, each including approximately 25% of the values. The 1st (lower) quartile (Q_1) represents a value that divides the lower 50% of the values (those below the median) into two equal sections, and the 3rd (upper) quartile (Q_3) represents a value that divides the upper 50% of the values (those above the median) into two equal sections. As with calculating the median, quartiles may represent the average of two values when the number of values below and above the median is even. The rectangle of a box plot is drawn such that it extends from the 1st quartile through the 3rd quartile and thereby represents the *interquartile range* (IQR; the distance between the 1st and 3rd quartiles). The rectangle includes the median.

In order to draw the "whiskers" (i.e., lines extending from the box), one must identify *fences*, or values that represent minimum and maximum values that would not be considered outliers. Typically, fences are calculated to be $Q - 1.5$ IQR (lower fence) and $Q_3 + 1.5$ IQR (upper fence). Whiskers are lines drawn by connecting the most extreme values that fall within the fence to the lines representing Q_1 and Q_3. Any value that is greater than the upper fence or lower than the lower fence is considered an outlier and is displayed as a special symbol beyond the whiskers. Outliers that extend beyond the fences are typically considered *mild outliers* on the box plot. An *extreme outlier* (i.e., one that is located beyond 3 times the length of the IQR from the 1st quartile (if a low outlier) or 3rd quartile (if a high outlier) may be indicated by a different symbol. Figure 1 provides an illustration of a box plot.

Box plots can be created in either a vertical or a horizontal direction. (In this entry, a vertical box plot is generally assumed for consistency.) They can often be very helpful when one is attempting to compare the distributions of two or more data sets or variables on the same scale, in which case they can be constructed side by side to facilitate comparison.



Data set values	*Defining features of this box plot*
2.0	Median = 6.0
2.0	First (Lower) Quartile = 3.0
2.0	Third (Upper) Quartile = 8.0
3.0	Interquartile Range (IQR) = 5.0
3.0	Lower Inner Fence = -4.5
5.0	Upper Inner Fence = 15.5
6.0	Range = 20.0
6.0	Mild Outlier = 22.0
7.0	
7.0	
8.0	
8.0	
9.0	
10.0	
22.0	

Figure 1 Box Plot Created With a Data Set and SPSS (an IBM company, formerly called PASW® Statistics)

Note: Data set values: 2.0, 2.0, 2.0, 3.0, 3.0, 5.0, 6.0, 6.0, 7.0, 7.0, 8.0, 8.0, 9.0, 10.0, 22.0. Defining features of this box plot: median = 6.0; first (lower) quartile = 3.0; third (upper) quartile = 8.0; interquartile range (IQR) = 5.0; lower inner fence = 4.5; upper inner fence = 15.5; range = 20.0; mild outlier = 22.0.

Steps to Creating a Box Plot

The following six steps are used to create a vertical box plot:

1. Order the values within the data set from smallest to largest and calculate the median, lower quartile (Q_1), upper quartile (Q_3), and minimum and maximum values.
2. Calculate the IQR.
3. Determine the lower and upper fences.
4. Using a number line or graph, draw a box to mark the location of the 1st and 3rd quartiles. Draw a line across the box to mark the median.
5. Make a short horizontal line below and above the box to locate the minimum and maximum values that fall within the lower and upper fences. Draw a line connecting each short horizontal line to the box. These are the box plot whiskers.
6. Mark each outlier with an asterisk or an "o."

Making Inferences

R. Lyman Ott and Michael Longnecker described five inferences that one can make from a box plot. First, the researcher can easily identify the median of the data by locating the line drawn in the middle of the box. Second, the researcher can easily identify the variability of the data by looking at the length of the box. Longer boxes illustrate greater variability whereas shorter box lengths illustrate a tighter distribution of the data around the median. Third, the researcher can easily examine the symmetry of the middle 50% of the data distribution by looking at where the median line falls in the box. If the median is in the middle of the box, then the data are evenly distributed on either side of the median, and the distribution can be considered symmetrical. Fourth, the researcher can easily identify outliers in the data by the asterisks outside the whiskers. Fifth, the researcher can easily identify the skewness of the distribution. On a distribution curve, data skewed to the right show more of the data to the left with a long "tail" trailing to the right. The opposite is shown when the data are skewed to the left. To identify skewness on a box plot, the researcher looks at the length of each half of the box plot. If the lower or left half of the box plot appears longer than the upper or right half, then the data are skewed in the lower direction or skewed to the left. If the upper half of the box plot appears longer than the lower half, then the data are skewed in the upper direction or skewed to the right. If a researcher suspects the data are skewed, it is recommended that the researcher investigate further by means of a histogram.

Variations

Over the past few decades, the availability of several statistical software packages has made EDA easier for social science researchers. However, these statistical packages may not calculate parts of a box plot in the same way, and hence some caution is warranted in their use. One study conducted by Michael Frigge, David C. Hoaglin, and Boris Iglewicz found that statistical packages calculate aspects of the box plot in different ways. In one example, the authors used three different statistical packages to create a box plot with the same distribution. Though the median looked approximately the same across the three box plots, the differences appeared in the length of the whiskers. The reason for the differences was the way the statistical packages used the interquartile range to calculate the whiskers. In general, to calculate the whiskers, one multiplies the interquartile range by a constant and then adds the result to Q_3 and subtracts it from Q_1. Each package used a different constant, ranging from 1.0 to 3.0. Though packages typically allow the user to adjust the constant, a package typically sets a default, which may not be the same as another package's default. This issue, identified by Frigge and colleagues, is important to consider because it guides the identification of outliers in the data. In addition, such variations in calculation lead to the lack of a standardized process and possibly to consumer confusion. Therefore, the authors provided three suggestions to guide the researcher in using statistical packages to create box plots. First, they suggested using a constant of 1.5 when the number of observations is between 5 and 20. Second, they suggested using a constant of 2.0 for outlier detection and rejection. Finally, they suggested using a constant of 3.0 for extreme cases. In the absence of standardization across statistical packages, researchers should understand how a package calculates whiskers and follow the suggested constant values.

Applications

As with all forms of data analysis, there are many advantages and disadvantages, appropriate uses, and certain precautions researchers should consider when using a box plot to display distributions. Box plots provide a good visualization of the range and potential skewness of the data. A box plot may provide the first step in exploring unexpected patterns in the distribution because box plots provide a good indication of how the data are distributed around the median. Box plots also clearly mark the location of mild and extreme outliers in the distribution. Other forms of graphical representation that graph individual values, such as dot plots, may not make this clear distinction. When used appropriately, box plots are useful in comparing more than one sample distribution side by side. In other forms of data analysis, a researcher may choose to compare data sets using a t test to compare means or an F test to compare variances. However, these methods are more vulnerable to skewness in the presence of extreme values. These methods must also meet normality and equal variance assumptions. Alternatively, box plots can compare the differences between variable distributions without the need to meet certain statistical assumptions.

However, unlike other forms of EDA, box plots show less detail than a researcher may need. For one, box plots may display only the five-number summary. They do not provide frequency measures or the quantitative measure of variance and standard deviation. Second, box plots are not used in a way that allows the researcher to compare the data with a normal distribution, which stem plots and histograms do allow. Finally, box plots would not be appropriate to use with a small

sample size because of the difficulty in detecting outliers and finding patterns in the distribution.

Besides taking into account the advantages and disadvantages of using a box plot, one should consider a few precautions. In a 1990 study conducted by John T. Behrens and colleagues, participants frequently made judgment errors in determining the length of the box or whiskers of a box plot. In part of the study, participants were asked to judge the length of the box by using the whisker as a judgment standard. When the whisker length was longer than the box length, the participants tended to overestimate the length of the box. When the whisker length was shorter than the box length, the participants tended to underestimate the length of the box. The same result was found when the participants judged the length of the whisker by using the box length as a judgment standard. The study also found that compared with vertical box plots, box plots positioned horizontally were associated with fewer judgment errors.

Sara C. Lewandowski and Sara E. Bolt

See also Exploratory Data Analysis; Histogram; Outlier

Further Readings

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*, 131–160.

Behrens, J. T., Stock, W. A., & Sedgwick, C. E. (1990). Judgment errors in elementary box-plot displays. *Communications in Statistics B: Simulation and Computation*, *19*, 245–262.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, *43,* 50–54.

Massart, D. L., Smeyers-Verbeke, J., Capron, X., & Schlesier, K. (2005). Visual presentation of data by means of box plots. *LCGC Europe*, *18*, 215–218.

Moore, D. S. (2001). *Statistics: Concepts and controversies* (5th ed.). New York: W. H. Freeman.

Moore, D. S., & McCabe, P. G. (1998). *Introduction to the practice of statistics* (3rd ed.). New York: W. H. Freeman.

Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis* (5th ed.). Pacific Grove, CA: Wadsworth.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.