



Haldane estimator– In a 2×2 contingency table, an estimator of the odds ratio obtained by adding $\frac{1}{2}$ to each cell frequency in order to avoid the possibility of division by zero. It is calculated by the formula:

$$\frac{(a + \frac{1}{2})(d + \frac{1}{2})}{(b + \frac{1}{2})(c + \frac{1}{2})}$$

where a , b , c , and d are the four cell counts. See also *Jewell's estimator*.

half-normal distribution– The probability distribution of a random variable $Z = |X|$ where X has a normal distribution with mean zero and variance σ^2 . Its probability density function is given by

$$f(z) = \frac{1}{\sigma} \sqrt{\frac{2}{\pi}} e^{-z^2/2\sigma^2}$$

The half-normal distribution has its probability mass distributed to the positive half of the real line.

half-normal plot– A graphical method for assessing the adequacy of a specified model and/or detecting the presence of outliers. The method involves plotting the residuals against the quantiles of the standard normal distribution.

half-normal probability paper– A normal probability paper where the negative abscissa is omitted, leaving only the positive half of the x axis.

haphazard selection– A method of selecting a sample of individuals by taking whoever is available or happens to be first on a list. It should not be confused with a true random selection.

hardware– The physical components or units making up a computer system. The term is used in contrast to programs and software which make up the operating instructions.

harmonic analysis– In time-series analysis, a procedure for calculating the period of the cyclic component.

harmonic mean– An **average** calculated by using the reciprocals of a set of numbers. It is obtained as the reciprocal of the **arithmetic mean** of the reciprocals. Given x_1, x_2, \dots, x_n , a set of n numbers, it is defined by the formula

$$HM = n / \sum_{i=1}^n 1/x_i$$

It is generally used to average **data sets** involving unequal **sample sizes**. It is useful in the averaging of certain **ratios**, such as miles per hour or miles per gallon of fuel. In many economic applications, it is used in averaging such **data** as time rates and rate-per-dollar prices. The harmonic mean is either smaller than or equal to the arithmetic mean.

Hartley's test– A **test procedure** for testing three or more **independent samples** for **homogeneity of variances** before using an **analysis of variance** procedure. It is based on the **ratio** between the largest and smallest **sample variances** and was proposed by Hartley in 1950. Like **Bartlett's test**, however, it is found to be sensitive to any departures from **normality**. See also *Box's test*, *Cochran's test*.

hazard– The instantaneous **risk** of failure or death.

hazard function– The **probability** that an individual dies in a certain time interval, given that the individual has survived until the beginning of the interval. Its reciprocal is equal to the **mean** survival time. The hazard function at time t , known as hazard rate, is determined as the limit of the probability of nearly immediate death for an individual known to be alive at time t . See also *survival function*.

hazard rate– See *hazard function*.

hazard ratio– In **survival analysis**, a measure of the **relative risk**, calculated as

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

where O_i and E_i ($i = 1, 2$) denote the observed and expected number of subjects experiencing the **event** of interest in the i th group. An HR of 1 suggests that the two groups being compared have the same **hazard** or **risk** of experiencing the event. An HR of greater than 1 suggests that the group 1 is more likely to experience the event while an HR of less than 1 indicates just the contrary. The **clinical significance** of a high hazard ratio depends on other information including the **absolute risk**, the **significance level**, and the clinical context.

heterogeneity of effects– In **meta-analysis**, the term is used to indicate that the individual studies being combined have effects of different magnitude. In the presence of substantial heterogeneity, it is not advisable to synthesize the individual results of different studies with a view to produce a single summary index. There are formal **statistical tests** to test for heterogeneity of effects; however, they lack sufficient **power** and their use can be misleading.

heterogeneity of effect size– Same as *heterogeneity of effects*.

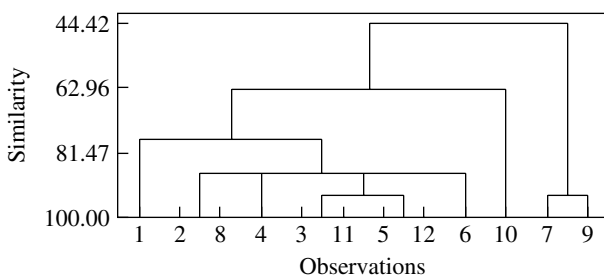
heterogeneity of variances– When **samples** differ markedly in terms of magnitude of their **variances**, they are said to exhibit heterogeneity of variances. This property of **data sets** is known as heteroscedasticity. Compare *homogeneity of variances*.

heterogeneous– A term used to describe the **variability** in the composition of different groups or within the elements of the same group.

heteroscedasticity– Compare *homoscedasticity*. Same as *heterogeneity of variances*.

hierarchical cluster analysis– Same as *hierarchical clustering*.

hierarchical clustering– An **algorithm** used for implementing one of the techniques of **cluster analysis**. The algorithm proceeds by either combining or dividing clusters.



Schematic illustration of hierarchical clustering

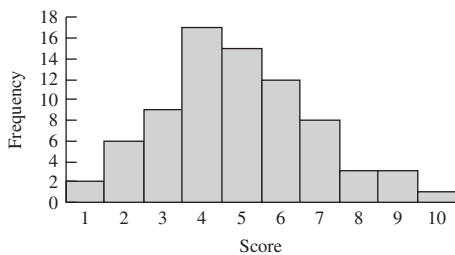
hierarchical design– Same as *nested design*.

hierarchical models– A series of **models** where each model is nested within the preceding one or the one immediately following it.

hierarchical regression– Same as *multilevel regression*.

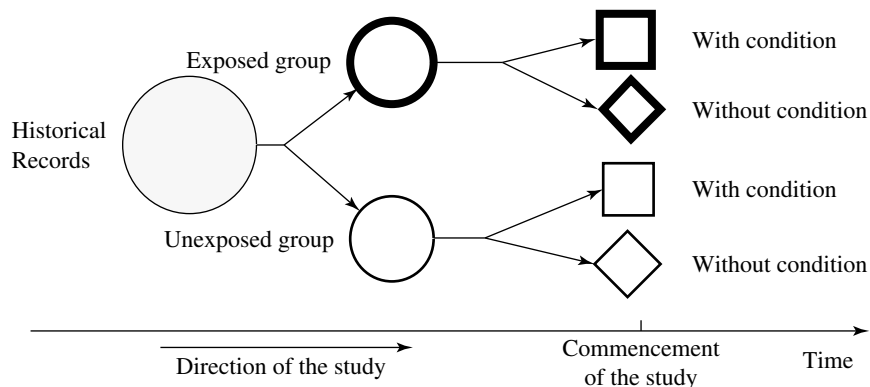
hinge– See *five-number summary*.

histogram– A **graphical presentation** of **frequency distribution** of a **quantitative variable** constructed by placing the **class intervals** on the **horizontal axis** of a graph and the **frequencies** on the **vertical axis**. Each class corresponds to a rectangle whose base is the real class interval and whose height is the **class frequency**. It differs from a **bar chart** in that bars are continuous and no spaces are left between the rectangles, indicating that the scoring categories represent a continuum of values that have been categorized into class intervals. A histogram can be viewed as a **bar diagram** for quantitative variables. In a histogram, the areas of the rectangles correspond to the frequencies being displayed.



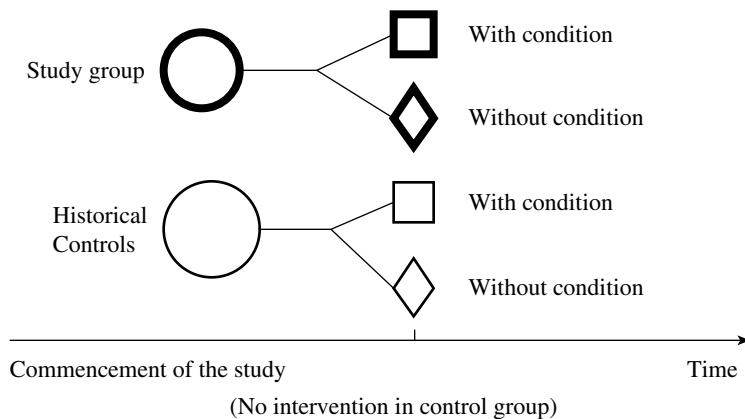
Histogram of some hypothetical data

historical cohort study– A **cohort study** based on **data** about persons at a time, or times, in the past. This method uses existing records or historical data about the health to determine the effect of a **risk factor** or **exposure** on a group of patients. Exposure to different levels of risk factors is then identified for subgroups of the population.



Schematic diagram of a historical cohort study

historical controls– In **clinical trials**, historical controls are **control subjects** for whom **data** were collected at a time previous to that at which the data are gathered on the **treatment group** being studied. Historical controls are generally obtained from clinical records or from the literature. Because of differences in exposures in the treatment group and historical controls, use of historical controls can lead to biased results.



Schematic diagram of a trial with historical controls

historical prospective study– Same as *historical cohort study*.

homogeneity– The extent to which the members of the group tend to be the same on the variables being investigated. The term is also used as a clipped form of **homogeneity of variances**.

homogeneity analysis– A **multivariate statistical technique** used to describe the relationships between two or more **variables** measured on a categorical or **nominal scale**. It is similar to **correspondence analysis**, but is not limited to two variables. Like correspondence analysis, it uses a set of coordinate values to display the relationship graphically. Objects within the same category are plotted close to each other whereas objects in different categories are plotted far apart. Homogeneity analysis is also known as multiple correspondence analysis; it can also be viewed as **principal components analysis** for **nominal data**.

homogeneity of regression– In **analysis of covariance**, the assumption that the **regression lines** within each group are equal.

homogeneity of variances– In an **analysis of variance**, when **samples** are assumed to have been drawn from **populations** with equal **variances**, they are said to exhibit homogeneity of variances. Many of the parametric **tests of significance** require that the variances of the underlying populations, from which the samples are drawn, should be homogeneous. In **regression analysis**, the condition in which the variance of the **dependent variable** (Y) is the same for all the values of the **independent variable** (X). Compare *heterogeneity of variances*.

homogeneous variance– Same as *homogeneity of variances*.

homoscedasticity– Compare *heteroscedasticity*. Same as *homogeneity of variances*.

honestly significant difference (HSD) test– Same as *Tukey's test*.

horizontal axis– The **abscissa** or baseline in a two-dimensional graph. It is also called the x axis.

Hosmer–Lemeshow statistic– A **statistic** used to assess the **goodness of fit** or predictive ability of a **logistic regression**. The procedure consists of computing the **probability** of a particular **event** for each **observation** by using the **model** being fitted. Subsequently, the **data** are grouped into “risk of event” categories (e.g., 0 to 10%, 10 to 20%, 20 to 30%, . . . , 90 to 100%) leading to an $r \times 2$ **contingency table** with the columns representing yes/no outcome and the rows representing risk-of-event categories as indicated above. The tabular entries in each **cell** contain the **observed** and **expected frequencies** for each **cross-tabulation**. The **chi-square statistic** is computed from the differences between observed and expected frequencies in each cell and is based on $r - 2$ **degrees of freedom**.

hospital controls– In **case-control studies**, the selection of **controls** from the same clinical source (hospital) from which **cases** are taken so that they represent the same catchment population and are subject to the same type of **selection biases**. See also *community controls*.

hot deck– A widely used and popular method of imputing **missing values** in **survey data**. See also *imputation*.

Hotelling–Lawley trace– See *multivariate analysis of variance*.

Hotelling's T^2 – A generalization of **Student's t distribution** to the case of **multivariate observations**. Like Student's t , T^2 can be used to test **hypotheses** involving a broad class of multivariate **statistics**, including **means** and differences of means, **regression coefficients** and their differences. **Tests of significance** involving T^2 can be carried out by using **variance ratio distribution**.

household survey– A **sample survey** conducted by interviewing people in their own homes. These surveys generally employ complex **sampling** methodology involving several stages of sampling. For each geographical unit sampled, there are additional levels of successive subsampling of smaller geographic areas; for example, census tracts, blocks within census tracts, and households within blocks. Finally, the individuals within a household may also be sampled.

HSD test– Acronym for *honestly significant difference test*.

hybrid series– A statistical series consisting of mixture of **time series** and **cross-section series**.

hypergeometric distribution– The **probability distribution** of a set of n elements randomly selected without replacement from a set of N elements, with D elements of one type and $N - D$ elements of a second type, such that the **sample** selected contains x elements of the first type and $n - x$ elements of the second type. The hypergeometric probability distribution is given by the formula

$$p(x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}} \quad x = 0, 1, \dots, \min(n, D)$$

When N is large and n is small compared to N , the hypergeometric distribution can be approximated by the **binomial distribution**. A hypergeometric distribution is frequently used in **quality control**, **sample surveys**, and in estimating the size of a wildlife population.

hypergeometric function– The hypergeometric function denoted by $F(\alpha, \beta, \gamma, x)$ is defined as

$$F(\alpha, \beta, \gamma, x) = 1 + \frac{\alpha \cdot \beta}{1 \cdot \gamma} x + \frac{\alpha(\alpha + 1) \cdot \beta(\beta + 1)}{1 \cdot 2 \cdot \gamma(\gamma + 1)} x^2 + \frac{\alpha(\alpha + 1)(\alpha + 2) \cdot \beta(\beta + 1)(\beta + 2)}{1 \cdot 2 \cdot 3 \cdot \gamma(\gamma + 1)(\gamma + 2)} x^3 + \dots$$

Hypergeometric functions have been found useful in the derivation of **characteristic functions** of **probability distributions**.

hyper-Graeco–Latin square– An **experimental design** that is an extension of **Latin** and **Graeco–Latin squares** to control for four sources of **variation**. It can also be used to investigate simultaneous **effects** of five **factors**: rows, columns, Latin letters, Greek letters, and Hebrew letters. It is obtained by juxtaposing or superimposing three Latin squares, one with treatments denoted by Greek letters, the second with treatments denoted by Latin letters, and the third with treatments denoted by Hebrew letters, such that each Hebrew letter appears once and only once with each Greek and Latin letter.

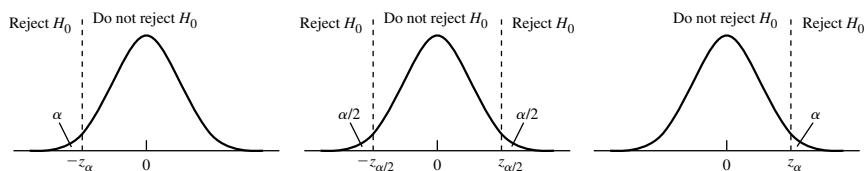
hyper square– A design obtained by superimposing three or more orthogonal **Latin squares**. In general a $p \times p$ hyper square is a design in which three or more orthogonal $p \times p$ Latin squares are superimposed. In using such a design, the researcher must assume that there would be no **interactions** between different **factors**. See also *Graeco–Latin square*, *hyper-Graeco–Latin square*.

hypothesis— A proposition or conjecture, tentatively advanced as being possibly true, that a researcher intends to test from **observations**. It is a working theory that forms the basis of a scientific investigation. Experience shows that a carefully and well-prepared hypothesis may ultimately save a great deal of time, effort, and money.

hypothesis test— See *hypothesis testing*.

hypothesis testing— In **inferential statistics**, a procedure for testing **hypotheses** about a **population parameter** of interest. The process begins with the choice of the so-called **null hypothesis** and an **alternative hypothesis**. A null hypothesis is usually tested and either rejected in favor of an alternative hypothesis or not rejected, in which case the alternative hypothesis cannot be sustained. Hypothesis testing is a scientific approach to assessing beliefs about a reality or phenomenon under investigation. The following are general steps in hypothesis testing:

1. State a null hypothesis (H_0) based on the specific question or phenomenon to be investigated.
 2. State an alternative hypothesis. This may be one-sided or two-sided depending on the problem being investigated as defined in the null hypothesis.
 3. Specify the **level of significance** (α). This is commonly taken as 0.05 and represents the maximum acceptable **probability** of incorrectly rejecting the null hypothesis.
 4. Determine an appropriate **sampling distribution** of the **sample statistic** of interest. Select a one-tailed or two-tailed test, depending on the alternative hypothesis.
 5. Evaluate the **standard error** or, more generally, an **estimate** of the standard error of the sample statistic; the formula for the standard error depends on the sample statistic in question.
 6. Compute the true value of the **test statistic** and locate its value on the sampling distribution.
 7. Reject or do not reject H_0 , depending on whether or not the sample statistic is located on the sampling distribution at or beyond the value of the test statistic at a given α .
- It is now a standard convention to report a **p-value** as justification for rejecting H_0 , which is the probability of obtaining a result equal to or more extreme than the observed value of the test statistic if the null hypothesis were true.



Graphical illustration of hypothesis testing based on the z statistic

See also *composite hypothesis, simple hypothesis, statistical test, type I error, type II error*.