## Hájek, Jaroslav

Hájek, Jaroslav was born in 1926 in Podebrady, Bohemia. A statistical engineer by profession, he obtained his doctorate in 1954. From 1954 to 1964, he worked as a researcher at the Institute of Mathematics of the Czechoslovakian Academy of Sciences. He then joined Charles University in Prague, where he was a professor from 1966 until his premature death in 1974. The principal works of Hájek, J. concern sampling probability theory and the rank test theory. In particular, he developed an asymptotic theory of the statistics of linear ranks. He was the first to apply the concept of invariance to the theory of rank testing.

*Some principal works and articles of Hájek, Jaroslav:*

**1955** Some rank distributions and their applications. Cas. Pest. Mat. 80, 17–31 (in Czech); translation in (1960) Select. Transl. Math. Stat. Probab., 2, 41–61.

**1958** Some contributions to the theory of probability sampling. Bull. Inst. Int. Stat., 36, 127–134.

**1961** Some extensions of the Wald–Wolfowitz–Noether theorem. Ann. Math. Stat. 32, 506–523.

**1964** Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann. Math. Stat. 35, 1419–1523.

**1965** Extension of the Kolmogorov–Smirnov test to regression alternatives. In: Neyman, J. and LeCam, L. (eds) Bernoulli–Bayes–Laplace: Proceedings of an International Seminar, 1963. Springer, Berlin Heidelberg New York, pp. 45–60.

**1981** Dupac, V. (ed) Sampling from a Finite Population. Marcel Dekker, New York.

## Harmonic Mean

The harmonic **mean** of $n$ observations is defined as $n$ divided by the sum of the inverses of all of the observations.

### HISTORY

See **geometric mean**.

The relationship between the harmonic mean, the geometric mean and the **arithmetic mean** is described by Mitrinovic, D.S. (1970).

### MATHEMATICAL ASPECTS

Let $x_1, x_2, \ldots, x_n$ be $n$ nonzero quantities, or $n$ observations related to a **quantitative**

**variable** $X$. The harmonic mean $H$ of these $n$ quantities is calculated as follows:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} .$$

If $\{x_i\}_{i=1,\ldots,n}$ represents a finite series of positive numbers, we state that:

$$\min_i x_i \leq H \leq G \leq \bar{x} \leq \max_i x_i .$$

## DOMAINS AND LIMITATIONS

The harmonic mean is rarely used in **statistics**. However, it can sometimes be useful, such as in the following cases:

- If a set of investments are invested at different interest rates, and they all give the same income, the unique rate at which all of the capital tied up in those investments must be invested to produce the same revenue as given by the set of investments is equal to the harmonic mean of the individual rates.
- Say we have a group of different materials, and each material can be purchased at a given price per amount of material (where the price per amount can be different for each material). We then buy a certain amount of each material, spending the same amount of money on each. In this case, the mean price per amount across all materials is given by the harmonic mean of the prices per amount for all of the materials.
- One property of the harmonic mean is that it is largely insensitive to outliers that have much larger values than the other data. For example, consider the following values: 1, 2, 3, 4, 5 and 100. Here the harmonic mean equals 2.62 and the **arithmetic mean** equals 19.17. However, the harmonic mean is much more sensitive to outliers when they have much smaller values than the rest of the data. So, for the observations 1, 6, 6, 6, 6, 6, we get $H = 3.27$ whereas the arithmetic mean equals 5.17.

## EXAMPLES

Three investments that each yield the same income have the following interest rates: 5%, 10% and 15%.

The harmonic mean gives the interest rate at which all of the capital would need to be invested in order to produce the same total income as the three separate investments:

$$H = \frac{3}{\left[\frac{1}{5} + \frac{1}{10} + \frac{1}{15}\right]} = \frac{3}{\frac{11}{30}} = 8.18\% .$$

We note that this result is different from the **arithmetic mean** of 10% $(5 + 10 + 15)/3$.

A representative buys three lots of coffee, each of a different grade (quality), at 3, 2 and 1.5 euros per kg respectively. He buys 200 euros of each grade.

The mean price per kg of coffee is then obtained by dividing the total cost by the total quantity bought:

$$\text{mean price} = \frac{\text{total cost}}{\text{total quantity}} = \frac{3 \cdot 200}{66.66 + 100 + 133.33} = 2 .$$

This corresponds to the harmonic mean of the prices of the three different grades of coffee:

$$\text{mean price} = \frac{3}{\left[\frac{1}{3} + \frac{1}{2} + \frac{1}{1.5}\right]} = \frac{6}{3} = 2 .$$

## FURTHER READING

- ▸ **Arithmetic mean**
- ▸ **Geometric mean**

▸ Mean
▸ Measure of central tendency

## REFERENCES

Mitrinovic, D.S. (with Vasic, P.M.): Analytic Inequalities. Springer, Berlin Heidelberg New York (1970)

# Hat Matrix

The hat matrix is a **matrix** used in **regression analysis** and **analysis of variance**. It is defined as the matrix that converts values from the observed **variable** into estimations obtained with the **least squares** method. Therefore, when performing linear regression in the matrix form, if $\hat{\mathbf{Y}}$ is the **vector** formed from estimations calculated from the least squares parameters, and $\mathbf{Y}$ is a vector of observations related to the **dependent variable**, then $\hat{\mathbf{Y}}$ is given by vector $\mathbf{Y}$ multiplied by $\mathbf{H}$, that is, $\hat{\mathbf{Y}} = \mathbf{HY}$ converts to Y's into $\hat{\mathrm{Y}}$'s.

## HISTORY

The hat matrix $\mathbf{H}$ was introduced by **Tukey, John Wilder** in 1972. An article by Hoaglin, D.C. and Welsch, R.E. (1978) gives the properties of the matrix $\mathbf{H}$ and also many examples of its application.

## MATHEMATICAL ASPECTS

Consider the following linear regression **model**:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \, .$$

where

- $\mathbf{Y}$ is an $(n \times 1)$ **vector** of observations on the **dependent variable**;
- $\mathbf{X}$ is the $(n \times p)$ **matrix** of independent variables (there are $p$ independent variables);
- $\boldsymbol{\varepsilon}$ is the $(n \times 1)$ vector of **errors**, and;
- $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of parameters to be estimated.

The estimation $\hat{\boldsymbol{\beta}}$ of the vector $\boldsymbol{\beta}$ is given by

$$\hat{\beta} = \left(\mathbf{X}' \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}' \cdot \mathbf{Y} \, .$$

and we can calculate the estimated values $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ if we know $\hat{\boldsymbol{\beta}}$:

$$\hat{\mathbf{Y}} = \mathbf{X} \cdot \hat{\boldsymbol{\beta}} = \mathbf{X} \cdot \left(\mathbf{X}' \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}' \cdot \mathbf{Y} \, .$$

The matrix $\mathbf{H}$ is then defined by:

$$\mathbf{H} = \mathbf{X} \cdot \left(\mathbf{X}' \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}' \, .$$

In particular, the diagonal element $h_{ii}$ will be defined by:

$$h_{ii} = x_i \cdot \left(\mathbf{X}' \cdot \mathbf{X}\right)^{-1} \cdot x_i' \, .$$

where $x_i$ is the $i$th line of $\mathbf{X}$.

## DOMAINS AND LIMITATIONS

The matrix $\mathbf{H}$, which allows us to obtain $n$ estimations of the **dependent variable** from $n$ observations, is an idempotent symmetric square matrix of order $n$. The element $(i, j)$ of this **matrix** measures the influence of the $j$th observation on the $i$th predicted **value**. In particular, the diagonal elements evaluate the effects of the observations on the corresponding estimations of the dependent variables. The value of each diagonal element of the matrix $\mathbf{H}$ ranges between 0 and 1.
Writing $\mathbf{H} = (h_{ij})$ for $i, j = 1, \ldots, n$, we have the relation:

$$h_{ii} = h_{ii}^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} h_{ij}^2 \, .$$

which is obtained based on the idempotent nature of $\mathbf{H}$; in other words $\mathbf{H} = \mathbf{H}^2$.

H

Therefore $\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p =$ number of parameters to estimate.
The matrix **H** is used to determine leverage points in regression analysis.

## EXAMPLES

Consider the following table where $Y$ is a **dependent variable** related to the **independent variable** $X$:

| X | Y |
|---|---|
| 50 | 6 |
| 52 | 8 |
| 55 | 9 |
| 75 | 7 |
| 57 | 8 |
| 58 | 10 |

The **model** of **simple linear regression** is written in the following manner in the matrix form:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} .$$

where

$$\mathbf{Y} = \begin{bmatrix} 6 \\ 8 \\ 9 \\ 7 \\ 8 \\ 10 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 50 \\ 1 & 52 \\ 1 & 55 \\ 1 & 75 \\ 1 & 57 \\ 1 & 58 \end{bmatrix} .$$

$\boldsymbol{\varepsilon}$ is the $(6 \times 1)$ **vector** of errors, and $\boldsymbol{\beta}$ is the $(2 \times 1)$ **vector** of parameters.
We find the **matrix H** using the result:

$$\mathbf{H} = \mathbf{X} \cdot \left(\mathbf{X}' \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}' .$$

By stepwise matrix calculations, we obtain:

$$(\mathbf{X}' \cdot \mathbf{X})^{-1} = \frac{1}{2393} \cdot \begin{bmatrix} 20467 & -347 \\ -347 & 6 \end{bmatrix} = \begin{bmatrix} 8.5528 & -0.1450 \\ -0.1450 & 0.0025 \end{bmatrix} .$$

and finally:

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}' = \begin{bmatrix} 0.32 & 0.28 & 0.22 & -0.17 & 0.18 & 0.16 \\ 0.28 & 0.25 & 0.21 & -0.08 & 0.18 & 0.16 \\ 0.22 & 0.21 & 0.19 & 0.04 & 0.17 & 0.17 \\ -0.17 & -0.08 & 0.04 & 0.90 & 0.13 & 0.17 \\ 0.18 & 0.18 & 0.17 & 0.13 & 0.17 & 0.17 \\ 0.16 & 0.16 & 0.17 & 0.17 & 0.17 & 0.17 \end{bmatrix} .$$

We remark, for example, that the weight of $y_1$ used during the **estimation** of $\hat{y}_1$ is 0.32. We then verify that the **trace** of **H** equals 2; in other words, it equals the number of parameters of the **model**.

$$\begin{aligned} \text{tr}(\mathbf{H}) &= 0.32 + 0.25 + 0.19 + 0.90 \\ &\quad + 0.17 + 0.17 \\ &= 2 . \end{aligned}$$

## FURTHER READING

- **Leverage point**
- **Matrix**
- **Multiple linear regression**
- **Regression analysis**
- **Simple linear regression**

## REFERENCES

Belsley, D.A., Kuh, E., Welsch, R.E.: Regression diagnostics. Wiley, New York pp. 16–19 (1980)

Hoaglin, D.C., Welsch, R.E.: The hat matrix in regression and ANOVA. Am. Stat. **32**, 17–22 (and correction at **32**, 146) (1978)

Tukey, J.W.: Some graphical and semigraphical displays. In: Bancroft, T.A. (ed.) Statistical Papers in Honor of George W. Snedecor. Iowa State University Press, Ames, IA , pp. 293–316 (1972)

# Histogram

The histogram is a **graphical representation** of the distribution of **data** that has been

grouped into classes. It consists of a series of rectangles, and is a type of frequency chart. Each data value is sorted and placed in an appropriate class **interval**. The number of data values within each class interval dictates the **frequency** (or relative frequency) of that class interval.

Each rectangle in the histogram represents a class of data. The width of the rectangle corresponds to the width of the class interval, and the surface of the rectangle represents the weight of the class.

## HISTORY

The term histogram was used for the first time by **Pearson, Karl** in 1895.

Also see **graphical representation**.

## MATHEMATICAL ASPECTS

The first step in the construction of a histogram consists of presenting the **data** in the form of a **frequency table**.

This requires that the class **intervals** are established and the data values are sorted and placed in the classes, which makes it possible to calculate the **frequencies** of the classes. The class intervals and frequencies are then added to the **frequency table**.

We then make use of the frequency table in order to construct the histogram. We divide the horizontal axis of the histogram into intervals, where the widths of these intervals correspond to those of the class intervals. We then draw the rectangles on the histogram. The width of each rectangle is the same as the width of the class that it corresponds to. The height of the rectangle is such that the surface area of the rectangle is equal to the relative frequency of the corresponding class. The sum of the surface areas of the rectangles must be equal to 1.

## DOMAINS AND LIMITATIONS

Histograms are used to present a **data** set in a visual form that is easy to understand. They allow certain general characteristics (such as typical **values**, the range or the shape of the data) to be visualized and extracted.

Reviewing the shape of a histogram can allow us to detect the **probability model** followed by the data (**normal distribution**, **log-normal distribution**, . . . ).

It is also possible to detect unexpected behavior or abnormal **values** with a histogram.

This type of **graphical representation** is most frequently used in economics, but since it is an extremely simple way of visualizing a **data** set, it is used in many other fields too.

We can also illustrate relative **frequency** in a histogram. In this case, the height of each rectangle equals the relative frequency of the corresponding class divided by the length of the class **interval**. In this case, if we sum the surface areas of all of the rectangles in the histogram, we obtain unity.

## EXAMPLES

The following table gives raw **data** on the average annual precipitation in 69 cities of the USA. We will use these data to establish a **frequency table** and then a corresponding histogram.

Annual average precipitations in 69 cities of the USA (in inches)

| | | | | | |
|---|---|---|---|---|---|
| Mob. | 67.0 | Chic. | 34.4 | St.L. | 35.9 |
| Sacr. | 17.2 | Loui. | 43.1 | At.C. | 45.5 |
| Wash. | 38.9 | Detr. | 31.0 | Char. | 42.7 |
| Boise | 11.5 | K.C. | 37.0 | Col. | 37.0 |
| Wich. | 30.6 | Conc. | 36.2 | Prov. | 42.8 |
| Bost. | 42.5 | N.Y. | 40.2 | Dall. | 35.9 |
| Jack. | 49.2 | Clev. | 35.0 | Norf. | 44.7 |
| Reno | 7.2 | Pitt. | 36.2 | Chey. | 14.6 |

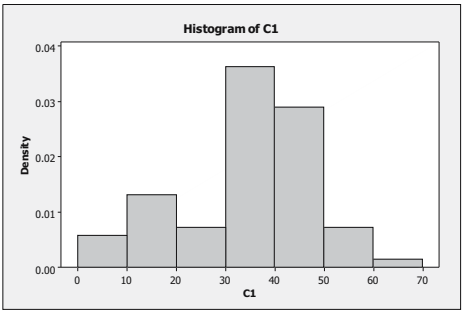
| | | | | | |
|---|---|---|---|---|---|
| Buff. | 36.1 | Nash. | 46.0 | L.R. | 48.5 |
| Cinc. | 39.0 | Burl. | 32.5 | Hart. | 43.4 |
| Phil. | 39.9 | Milw. | 29.1 | Atl. | 48.3 |
| Memp. | 49.1 | Phoen. | 7.0 | Indi. | 38.7 |
| S.L. | 15.2 | Denv. | 13.0 | Port. | 40.8 |
| Char. | 40.8 | Miami | 59.8 | Dul. | 30.2 |
| Jun. | 54.7 | Peor. | 35.1 | G.Fls | 15.0 |
| S.F. | 20.7 | N.Or. | 56.8 | Albq | 7.8 |
| Jack. | 54.5 | S.S.M. | 31.7 | Ral. | 42.5 |
| Spok. | 17.4 | L.A. | 14.0 | Omaha | 30.2 |
| Ok.C. | 31.4 | Wilm. | 40.2 | Alb. | 33.4 |
| Col. | 46.4 | Hon. | 22.9 | Bism. | 16.2 |
| El Paso | 7.8 | D.Mn. | 30.8 | Port. | 37.6 |
| S-Tac | 38.8 | Balt. | 41.8 | S.Fls | 24.7 |
| S.J. | 59.2 | Mn-SP | 25.9 | Hstn. | 48.2 |

Source: U.S. Census Bureau (1975) Statistical Abstract of the United States. U.S. Census Bureau, Washington, DC.

These **data** can be represented by the following **frequency table**:

| Class | Frequency | Relative Frequency |
|---|---|---|
| 0–10 | 4 | 0.058 |
| 10–20 | 9 | 0.130 |
| 20–30 | 5 | 0.072 |
| 30–40 | 24 | 0.348 |
| 40–50 | 21 | 0.304 |
| 50–60 | 4 | 0.058 |
| 60–70 | 2 | 0.030 |
| Total | 69 | 1.000 |

We can now construct the histogram:




The horizontal axis is divided up into classes, and in this case the relative **frequencies** are given by the heights of the rectangles because the classes all have the same width.

## FURTHER READING

- ▸ Frequency distribution
- ▸ Frequency polygon
- ▸ Frequency table
- ▸ Graphical representation
- ▸ Ogive
- ▸ Stem-and-leaf diagram

## REFERENCES

Dodge, Y.: Some difficulties involving nonparametric estimation of a density function. J. Offic. Stat. **2**(2), 193–202 (1986)

Freedman, D., Pisani, R., Purves, R.: Statistics. Norton, New York (1978)

Pearson, K.: Contributions to the mathematical theory of evolution. II: Skew variation in homogeneous material. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, Cambridge, pp. 41–112 (1948). First published in 1895 in Philos. Trans. Roy. Soc. Lond. Ser. A **186**, 343–414

# Homogeneity Test

One issue that often needs to be considered when analyzing **categorical data** obtained for many groups is that of the homogeneity of the groups. In other words, we need to find out whether there are significant differences between these groups in relation to one or many qualitative categorical variables. A homogeneity test can show whether the differences are significant or not.

## MATHEMATICAL ASPECTS

We consider the chi-square homogeneity test here, which is a specific type of **chi-square test**.

Let $I$ be the number of groups considered and $J$ the number of categories considered.

Let:

- $n_{i.} = \sum_{j=1}^{J} n_{ij}$ correspond to the size of group $i$;
- $n_{.j} = \sum_{i=1}^{I} n_{ij}$;
- $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ be the total number of observations
- $n_{ij}$ be the empirical **frequency** (that is, the number of occurrences observed) corresponding to group $i$ and category $j$;
- $m_{ij}$ is the theoretical frequency corresponding to group $i$ and category $j$, which, assuming homogeneity among the groups, equals:

$$m_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}.$$

If we represent the data in the form of a **contingency table** with $I$ lines and $J$ columns, we can calculate the $n_{i.}$ that contribute to the sum of the elements of line $i$ and the $n_{.j}$ that contribute to the sum of all of the elements of column $j$.

We calculate

$$\chi_c^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}.$$

and the chi-square homogeneity test is expressed in the following way: we reject the homogeneity hypothesis (at a significance level of 5%) if the value $\chi_c^2$ is greater then the value of the $\chi^2$ (chi-square) distribution with $(J-1)\cdot(I-1)$ degrees of freedom.

*Note*: We have assumed here that the same number of units are tested for each combination of group and category. However, we may want to test different numbers of units for different combinations. In this case, if we have a proportion $p_{ij}$ of units for group $i$ and category $j$, it is enough to replace $m_{ij}$ by $n_{i.} \cdot p_{ij}$.

## EXAMPLES

We consider a study performed in the pharmaceutical domain that concerns 100 people suffering from a particular illness. In order to examine the effect of a medical treatment, 100 people were chosen at random. Half of them were placed in a control group and received a placebo. The other patients received the medical treatment. Then the number of healthy people in each group was monitored for 24 hours following administration the treatment. The results are provided in the following table.

| Observed frequency | Healthy for 24 hours | Not healthy | Total |
|---|---|---|---|
| Placebo | 2 | 48 | 50 |
| Treatment | 9 | 41 | 50 |

The theoretical frequencies are obtained by assuming that the general state of health would have been the same for both groups if no treatment had been applied. In this case we obtain $m_{11} = m_{21} = \frac{50\cdot 11}{100} = 5.5$ and $m_{12} = m_{22} = \frac{50\cdot 89}{100} = 44.5$.

We calculate the value of $\chi^2$ by comparing the theoretical frequencies with the observed frequencies:

$$\chi_c^2 = \frac{(2-5.5)^2}{5.5} + \frac{(48-44.5)^2}{44.5} + \frac{(9-5.5)^2}{5.5} + \frac{(41-44.5)^2}{44.5} = 5.005.$$

If we then refer to the $\chi^2$ distribution table for one degree of freedom, we obtain the value $\chi^2_{0.05} = 3.84$ for a significance level of 5%, which is smaller then the value we calculated, $\chi^2_c = 5.005$. We conclude that the groups were not homogeneous and the treatment is efficient.

## FURTHER READING

- ▸ **Analysis of categorical data**
- ▸ **Analysis of variance**
- ▸ **Categorical data**
- ▸ **Chi-square distribution**
- ▸ **Chi-square test**
- ▸ **Frequency**

## REFERENCE

See **analysis of categorical data**.

# Hotelling, Harold

Hotelling, Harold (1895–1973) is considered to be one of the pioneers in the field of economical mathematics over the period 1920–1930. He introduced the $T^2$ multivariate test, principal components analysis and canonical correlation analysis.

He studied at University of Washington, where he obtained a B.A. in journalism in 1919. He then moved to Princeton University, obtaining a doctorate in mathematics from there in 1924. The began teaching at Stanford University that same year. His applications of mathematics to the social sciences initially concerned journalism and political science, and then he moved his focus to population and predictian.

In 1931, he moved to Colombia University, where he actively participated in the creation of its statistical department. During the Second World War, he performed statistical research for the military.

In 1946, he was hired by North Carolina University at Chapel Hill to create a statistics department there.

*Some principal works and articles of Hotelling, Harold:*

**1933** Analysis of a complex of statistical variables with principal components. J. Educ. Psychol., 24, 417–441 and 498–520.

**1936** Relation between two sets of variates. Biometrika 28, 321–377.

# Huber, Peter J.

Huber, Peter J. was born at Wohlen (Switzerland) in 1934. He performed brilliantly during his studies and his doctorate in mathematics at the Federal Polytechnic School of Zurich, where he received the Silver Medal for the scientific quality of his thesis. He worked as Professor of Mathematical Statistics at the Federal Polytechnic School of Zurich. He then moved to the USA and worked at the most prestigious universities (Princeton, Yale, Berkeley) as an invited professor. In 1977 he was named Professor of Applied Mathematics at the Massachusetts Institute of Technology. He is member of the prestigious American Academy of Arts and Sciences, the Bernoulli Society and the National Science Foundation in the USA, in which foreign members are extremely rare. Since the publication of his article "Robust estimation of a location parameter" in 1964, he has been considered to be the founder of robust statistics.

Huber, Peter J. received the title of Docteur Honoris Causa from Neuchâtel University in 1994.

*Some principal works and publications of Huber, Peter J.:*

**1964** Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.

**1968** Robust statistical procedures. SIAMCBMS-NSF Reg. Conf. Ser. Appl. Math.,

**1981** Robust Statistics. Wiley, New York.

**1995** Robustness: Where are we now? Student, Vol.1, 75–86.

## Hypergeometric Distribution

The hypergeometric distribution describes the **probability** of success if a series of objects are drawn from a population (which contains some objects that represent failure while the others represent success), without replacement.

It is therefore used to describe a **random experiment** where there are only two possible results: "success" and "failure."

Consider a set of $N$ **events** in which there are $M$ "successes" and $N - M$ "failures." The **random variable** $X$, corresponding to the number of successes obtained if we draw $n$ events without replacement follows a hypergeometric distribution with **parameters** $N$, $M$ and $n$, denoted by $H(N, M, n)$.

The hypergeometric distribution is a **discrete probability distribution**.

The number of ways that $n$ events can be drawn from $N$ events is equal to:

$$C_N^n = \binom{N}{n} = \frac{n!}{N! \cdot (n-N)!} .$$

The number of elementary **events** depends on $X$ and is:

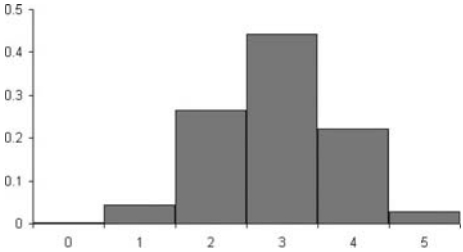
$$C_M^x \cdot C_{N-M}^{n-x} .$$

which gives the following **probability function**:

$$P(X = x) = \frac{C_M^x \cdot C_{N-M}^{n-x}}{C_N^n} ,$$

$$\text{for } x = 0, 1, \ldots, n$$

$$(\text{where } C_v^u = 0 \text{ if } u < v \text{by convention}) .$$




Hypergeometric distribution, $N = 12$, $M = 7$, $n = 5$

### MATHEMATICAL ASPECTS

Consider the **random variable** $X = X_1 + X_2 + \ldots + X_n$, where:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th drawing is a success} \\ 0 & \text{if the } i\text{th drawing is a failure} \end{cases} .$$

In this case, the **probability distribution** for $X_i$ is:

| $X_i$ | 0 | 1 |
|---|---|---|
| $P(X_i)$ | $\frac{N-M}{N}$ | $\frac{M}{N}$ |

The **expected value** of $X_i$ is therefore given by:

$$\begin{aligned} E[X_i] &= \sum_{j=1}^{2} x_j P(x_i = x_j) \\ &= 0 \cdot \frac{N-M}{N} + 1 \cdot \frac{M}{N} \\ &= \frac{M}{N} . \end{aligned}$$

Utilizing the fact that $X = X_1 + X_2 + \ldots + X_n$, we have:

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{M}{N} = n\frac{M}{N}\,.$$

The **variance** of $X_i$ is, by definition:

$$\begin{aligned}\mathrm{Var}(X_i) &= E\left[X_i^2\right] - (E[X_i])^2 \\ &= \frac{M}{N} - \left(\frac{M}{N}\right)^2 \\ &= \frac{M(N-M)}{N^2}\,.\end{aligned}$$

Since the $X_i$, $i = 1, 2, \ldots, n$ are dependent **random variables**, the **covariance** should be taken into account when calculating the **variance** of $X$.

The **probability** that $X_i$ and $X_j$ $(i \neq j)$ are both successes is equal to:

$$P(X_i = 1, X_j = 1) = \frac{M(M-1)}{N(N-1)}\,.$$

If we put $V = X_i \cdot X_j$, the **values** of $V$ and the associated **probabilities** are:

| $V$ | 0 | 1 |
|---|---|---|
| $P(V)$ | $1 - \frac{M(M-1)}{N(N-1)}$ | $\frac{M(M-1)}{N(N-1)}$ |

The **expected value** of $V$ is therefore:

$$\begin{aligned}E[V] &= 0 \cdot \left(1 - \frac{M(M-1)}{N(N-1)}\right) \\ &\quad + 1 \cdot \frac{M(M-1)}{N(N-1)}\,.\end{aligned}$$

The **covariance** of $X_i$ and $X_j$ is, by definition, equal to:

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= E[X_i \cdot X_j] - E[X_i] \cdot E[X_j] \\ &= \frac{M(M-1)}{N(N-1)} - \left(\frac{M}{N}\right)^2 \\ &= -\frac{M(N-M)}{N^2(N-1)} \\ &= -\frac{1}{N-1}\mathrm{Var}(X_i)\,.\end{aligned}$$

We can now calculate the **variance** of $X$:

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &\quad + 2\sum_{j=1}^{n}\sum_{i<j} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &\quad + n(n-1)\mathrm{Cov}(X_i, X_j) \\ &= n\left[\mathrm{Var}(X_i) - \frac{n-1}{N-1}\mathrm{Var}(X_i)\right] \\ &= n\mathrm{Var}(X_i)\frac{N-n}{N-1} \\ &= n\frac{M(N-M)}{N^2}\frac{N-n}{N-1} \\ &= \frac{N-n}{N-1}n\frac{M}{N}\left(1 - \frac{M}{N}\right)\,.\end{aligned}$$

## DOMAINS AND LIMITATIONS

The hypergeometric distribution is often used in quality control.

Suppose that a production line produces $N$ products, which are then submitted to verification. A **sample** of size $n$ is taken from this batch of products, and the number of defective products in this sample is noted. It is possible to to use this to obtain (by **inference**) information on the probable total number of defective products in the whole batch.

## EXAMPLES

A box contains 30 fuses, and 12 of these are defective. If we take five fuses at random, the **probability** that none of them is defective is equal to:

$$\begin{aligned}P(X = 0) &= \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} = \frac{C_{12}^0 C_{18}^5}{C_{30}^5} \\ &= 0.0601\,.\end{aligned}$$

### FURTHER READING

- ▸ Bernoulli distribution
- ▸ Binomial distribution
- ▸ Discrete probability distribution

## Hypothesis

A statistical hypothesis is an assertion regarding the distribution(s) of one or several **random variables**. It may concern the **parameters** of a given distribution or the **probability distribution** of a **population** under study.

The validity of the hypothesis is examined by performing **hypothesis testing** on **observations** collected for a **sample** of the studied **population**.

When performing **hypothesis testing** on the **probability distribution** of the **population** being studied, the hypothesis that the studied population follows a given probability distribution is called the **null hypothesis**. The hypothesis that affirms that the population does not follow a given probability distribution is called the **alternative hypothesis** (or opposite hypothesis).

If we perform **hypothesis testing** on the **parameters** of a **distribution**, the hypothesis that the studied parameter is equal to a given **value** is called the **null hypothesis**. The hypothesis that states that the value of the parameter is different to this given value is called the **alternative hypothesis**.

The **null hypothesis** is usually denoted by $H_0$ and the **alternative hypothesis** by $H_1$.

### HISTORY

In **hypothesis testing**, the hypothesis that is to be tested is called the **null hypothesis**. We owe the term "null" to **Fisher, R.A.** (1935). Introducing this concept, he mentioned the well-known tea tasting problem, where a lady claimed to be able to recognize by taste whether the milk or the tea was poured into her cup first. The hypothesis to be tested was that the taste was absolutely not influenced by the order in which the tea was made.

Originally, the null hypothesis was usually taken to mean that a particular **treatment** has no effect, or that there was no difference between the effects of different treatments.

Nowadays the null hypothesis is mostly used to indicate the hypothesis has that to be tested, in contrast to the **alternative hypothesis**. Also see **hypothesis testing**.

### EXAMPLES

Many problems involve repeating an **experiment** that has two possible results.

One example of this is the gender of a newborn child. In this case we are interested in the proportion of boys and girls in a given **population**. Consider $p$, the proportion of girls, which we would like to estimate from an observed **sample**. To determine whether the proportions of newborn boys and girls are the same, we make the statistical hypothesis that $p = \frac{1}{2}$.

### FURTHER READING

- ▸ Alternative hypothesis
- ▸ Analysis of variance
- ▸ Hypothesis testing
- ▸ Null hypothesis

### REFERENCES

Fisher, R.A.: The Design of Experiments. Oliver & Boyd, Edinburgh (1935)

## Hypothesis Testing

Hypothesis testing is a procedure that allow, us to (depending on certain decision rules) confirm a starting **hypothesis**, called the **null hypothesis**, or to reject this null hypothesis in favor of the **alternative hypothesis**.

### HISTORY

The theory behind hypothesis testing developed under study. The first steps were taken when works began to appear that discussed the significance (or insignificance) of a group of observations. Some examples of such works date from the eighteenth century, including those by Arbuthnott, J. (1710), Bernoulli, Daniel (1734) and **Laplace, Pierre Simon de** (1773). These works were seen more frequently in the nineteenth century, such as those by Gavarett (1840) and **Edgeworth, Francis Y.** (1885). The development of hypothesis testing occurred in parallel with the theory of **estimation**. Hypothesis testing seems to have been first elaborated by workers in the experimental sciences and the management domain. For example, the **Student test** was developed by **Gosset, William Sealy** during his time working for Guinness.

**Neyman, Jerzy** and **Pearson, Egon Sharpe** developed the mathematical theory of hypothesis testing, which they presented in an article published in 1928 in the review Biometrika. They were the first to recognize that the rational choice to be made during hypothesis testing had to be between the **null hypothesis** that we want to test and an **alternative hypothesis**. A second fundamental article on the theory of hypothesis testing was published in 1933 by the same mathematicians, where they also distinguished between a **type I error** and a **type II error**. The works resulting from the collaboration between Neyman, J. and Pearson, E.S. are described in Pearson (1966) and in the biography of Neyman, published by Reid (1982).

### MATHEMATICAL ASPECTS

Hypothesis testing of a **sample** generally involves the following steps:

1. Formulate the hypotheses:
   - The **null hypothesis** $H_0$,
   - The **alternative hypothesis** $H_1$.
2. Determine the **significance level** $\alpha$ of the test.
3. Determine the **probability distribution** that corresponds to the **sampling distribution**.
4. Calculate the **critical value** of the null hypothesis and deduce the **rejection region** or the **acceptance region**.
5. Establish the decision rules:
   - If the **statistics** observed in the sample are located in the acceptance region, we do not reject the null hypothesis $H_0$;
   - If the statistics observed on the sample are located in the rejection region, we reject the null hypothesis $H_0$ for the alternative hypothesis $H_1$.
6. Take the decision to accept or to reject the null hypothesis on the basis of the observed sample.

### DOMAINS AND LIMITATIONS

The most frequent types of hypothesis testing are described below.

1. *Hypothesis testing of a sample*: We want to test whether the **value** of a **parameter** $\theta$ of the **population** is identical to a presumed value. The hypotheses will be as follows:

$$H_0: \quad \theta = \theta_0\,,$$
$$H_1: \quad \theta \neq \theta_0\,.$$

where $\theta_0$ is the presumed value of the unknown parameter $\theta$.

2. *Hypothesis testing on two samples*: The goal in this case is to find out whether two populations that are both described by a particular parameter are different. Let $\theta_1$ and $\theta_2$ be parameters that describe populations 1 and 2 respectively. We can then formulate the following hypotheses:

$$H_0: \quad \theta_1 = \theta_2 ,$$
$$H_1: \quad \theta_1 \neq \theta_2 .$$

or

$$H_0: \quad \theta_1 - \theta_2 = 0 ,$$
$$H_1: \quad \theta_1 - \theta_2 \neq 0 .$$

3. *Hypothesis testing of more than two samples*: As for a test performed on two samples, hypothesis testing is performed on more than two samples to determine whether these populations are different, based on comparing the same parameter from all of the populations being tested. In this case, we test the following hypotheses:

$$H_0: \quad \theta_1 = \theta_2 = \ldots = \theta_k ,$$
$$H_1: \quad \text{The values of } \theta_i \ (i = 1, 2, \ldots, k) \text{ are not all identical.}$$

Here $\theta_1, \ldots, \theta_k$ are the unknown parameters of the populations and $k$ is the number of populations to be compared.

In hypothesis testing theory, and in practice, we can distinguish between two types of tests: a parametric test and a nonparametric test.

## Parametric Tests

A parametric test is a hypothesis test that presupposes a particular form for each of the distributions related to the underlying populations. This case applies, for example, when these populations follow a **normal distribution**.

The **Student test** is an example of a parametric test. This test compares the means of two normally distributed populations.

## Nonparametric Test

A nonparametric test is a hypothesis test where it is not necessary to specify the parametric form of the distribution of the underlying population.

There are many examples of this type of test, including the **sign test**, the **Wilcoxon test**, the **signed Wilcoxon test**, the **Mann–Whitney test**, the **Kruskal–Wallis test**, and the **Kolmogorov–Smirnov test**.

## EXAMPLES

For examples of parametric hypothesis testing, see **binomial test**, **Fisher test** or **Student test**. For examples of nonparametric hypothesis testing, see **Kolmogorov–Smirnov test**, **Kruskal–Wallis test**, **Mann–Whitney test**, **Wilcoxon test**, **signed Wilcoxon test** and **sign test**.

## FURTHER READING

- ▸ Acceptance region
- ▸ Alternative hypothesis
- ▸ Nonparametric test
- ▸ Null hypothesis
- ▸ One-sided test
- ▸ Parametric test
- ▸ Rejection region
- ▸ Sampling distribution
- ▸ Significance level
- ▸ Two-sided test

## REFERENCES

Arbuthnott, J.: An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. Philos. Trans. **27**, 186–190 (1710)

Bernoulli, D.: Quelle est la cause physique de l'inclinaison des planètes (...). Rec. Pièces Remport. Prix Acad. Roy. Sci. **3**, 95–122 (1734)

Edgeworth, F.Y.: Methods of Statistics. Jubilee Volume of the Royal Statistical Society, London (1885)

Gavarret, J.: Principes généraux de statistique médicale. Beché Jeune & Labé, Paris (1840)

Laplace, P.S. de: Mémoire sur l'inclinaison moyenne des orbites des comètes. Mém. Acad. Roy. Sci. Paris **7**, 503–524 (1773)

Lehmann, E.L.: Testing Statistical Hypotheses, 2nd edn. Wiley, New York (1986)

Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference, Parts I and II. Biometrika **20**A, 175–240, 263–294 (1928)

Neyman, J., Pearson, E.S.: On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. Lond. Ser. A **231**, 289–337 (1933)

Pearson, E.S.: The Neyman-Pearson story: 1926–34. In: David, F.N. (ed.) Research Papers in Statistics: Festschrift for J. Neyman. Wiley, New York (1966)

Reid, C.: Neyman—From Life. Springer, Berlin Heidelberg New York (1982)