The Concise Encyclopedia of Statistics

Yadolah Dodge

The Concise Encyclopedia of Statistics

With 247 Tables

Springer

Author
Yadolah Dodge
Honorary Professor
University of Neuchâtel
Switzerland
yadolah.dodge@uninc.ch

A C.i.P. Catalog record for this book is available from the Library of Congress Control

ISBN: 978-0-387-32833-1

This publication is available also as:
Print publication under ISBN 978-0-387-31742-7and
Print and electronic bundle under ISBN 978-0-387-33828-6

Springer is part of Springer Science+Business Media

springer.com

Printed on acid free paper SPIN: 10944523 2109 – 5 4 3 2 1 0

To the memory of my beloved wife K,
my caring mother,
my hard working father
and
to my two kind and warm-hearted sons, Ali and Arash

Preface

With this concise volume we hope to satisfy the needs of a large scientific community previously served mainly by huge encyclopedic references. Rather than aiming at a comprehensive coverage of our subject, we have concentrated on the most important topics, but explained those as deeply as space has allowed. The result is a compact work which we trust leaves no central topics out.

Entries have a rigid structure to facilitate the finding of information. Each term introduced here includes a definition, history, mathematical details, limitations in using the terms followed by examples, references and relevant literature for further reading. The reference is arranged alphabetically to provide quick access to the fundamental tools of statistical methodology and biographies of famous statisticians, including some currents ones who continue to contribute to the science of statistics, such as Sir David Cox, Bradley Efron and T.W. Anderson just to mention a few. The critera for selecting these statisticians, whether living or absent, is of course rather personal and it is very possible that some of those famous persons deserving of an entry are absent. I apologize sincerely for any such unintentional omissions.

In addition, an attempt has been made to present the essential information about statistical tests, concepts, and analytical methods in language that is accessible to practitioners and students and the vast community using statistics in medicine, engineering, physical science, life science, social science, and business/economics.

The primary steps of writing this book were taken in 1983. In 1993 the first French language version was published by Dunod publishing company in Paris. Later, in 2004, the updated and longer version in French was published by Springer France and in 2007 a student edition of the French edition was published at Springer.

In this encyclopedia, just as with the *Oxford Dictionary of Statistical Terms*, published for the International Statistical Institute in 2003, for each term one or more references are given, in some cases to an early source, and in others to a more recent publication. While some care has been taken in the choice of references, the establishment of historical priorities is notoriously difficult and the historical assignments are not to be regarded as authoritative.

For more information on terms not found in this encyclopedia short articles can be found in the following encyclopedias and dictionaries:

International Encyclopedia of Statistics, eds. William Kruskal and Judith M. Tanur (The Free Press, 1978).

Encyclopedia of Statistical Sciences, eds. Samuel Kotz, Norman L. Johnson and Cambell Reed (John Wiley and Sons, 1982).

The Encyclopedia of Biostatistics, eds. Peter Armitage and Ted Colton (Chichester: John Wiley and Sons, 1998).

The Encyclopedia of Environmetrics, eds. A.H. El-Sharaawi and W.W. Paregoric (John Wiley and Sons, 2001).

The Encyclopedia of Statistics in Quality and Reliability, eds. F. Ruggeri, R.S. Kenett and F.W. Faltin (John Wiley and Sons, 2008).

Dictionnaire- Encylopédique en Statistique, Yadolah Dodge, Springer 2004

In between the publication of the first version of the current book in French in 1993 and the later edition in 2004 to the current one, the manuscript has undergone many corrections. Special care has been made in choosing suitable translations for terms in order to achieve sound meaning in both the English and French languages. If in some cases this has not happen, I apologize. I would be very grateful to readers for any comments regarding inaccuracies, corrections, and suggestions for the inclusion of new terms, or any matter that could improve the next edition. Please send your comments to Springer-Verlag.

I wish to thank many people who helped me throughout these many years to bring this manuscript to its current form. Starting with my former assistants from 1983 to 2004, Nicole Rebetez, Sylvie Gonano-Weber, Maria Zegami, Jurg Schmid, Severine Pfaff, Jimmy Brignony Elisabeth Pasteur, Valentine Rousson, Alexandra Fragnieire, and Theiry Murrier. To my colleagues Joe Whittaker of University of Lancaster, Ludevic Lebart of France Telecom, and Bernard Fisher, University of Marseille, for reading parts of the manuscript. Special thanks go to Gonna Serbinenko and Thanos Kondylis for their remarkable cooperation in translating some of terms from the French version to English. Working with Thanos, my former Ph.D. student, was a wonderful experience. To my colleague Shahriar Huda whose helpful comments, criticisms, and corrections contributed greatly to this book. Finally, I thank the Springer-Verlag, especially John Kimmel, Andrew Spencer, and Oona Schmid for their meticulous care in the production of this encyclopedia.

January 2008

Yadolah Dodge
Honorary Professor
University of Neuchâtel
Switzerland

About the Author

Founder of the Master in Statistics program in 1989 for the University of Neuchâtel in Switzerland, Professor Yadolah Dodge earned his Master in Applied Statistics from the Utah State University in 1970 and his Ph.D in Statistics with a minor in Biometry from the Oregon State University in 1973. He has published numerous articles and authored, co-authored, and edited several books in the English and French languages, including *Mathematical Programming in Statistics* (John Wiley 1981, Classic Edition 1993), *Analysis of Experiments with Missing Data* (John Wiley 1985), *Alternative Methods of Regression* (John Wiley 1993), *Premier Pas en Statistique* (Springer 1999), *Adaptive Regression* (Springer 2000), *The Oxford Dictionary of Statistical Terms* (2003), *Statistique: Dictionnaire encyclopédique* (Springer 2004), and *Optimisation appliquée* (Springer 2005). Professor Dodge is an elected member of the International Statistical Institute (1976) and a Fellow of the Royal Statistical Society.

Acceptance Region

The acceptance region is the **interval** within the **sampling distribution** of the test **statistic** that is consistent with the **null hypothesis** H_0 from **hypothesis testing**.
It is the complementary region to the **rejection region**.
The acceptance region is associated with a **probability** $1 - \alpha$, where α is the **significance level** of the test.

MATHEMATICAL ASPECTS
See **rejection region**.

EXAMPLES
See **rejection region**.

FURTHER READING
- Critical value
- Hypothesis testing
- Rejection region
- Significance level

Accuracy

The general meaning of accuracy is the proximity of a **value** or a **statistic** to a reference value. More specifically, it measures the proximity of the **estimator** T of the unknown **parameter** θ to the true value of θ.

The accuracy of an **estimator** can be measured by the **expected value** of the squared deviation between T and θ, in other words:

$$E\left[(T-\theta)^2\right] .$$

Accuracy should not be confused with the term **precision**, which indicates the degree of exactness of a measure and is usually indicated by the number of decimals after the comma.

FURTHER READING
- Bias
- Estimator
- Parameter
- Statistics

Algorithm

An algorithm is a process that consists of a sequence of well-defined steps that lead to the solution of a particular type of problem. This process can be iterative, meaning that it is repeated several times. It is generally a numerical process.

HISTORY
The term algorithm comes from the Latin pronunciation of the name of the ninth century mathematician al-Khwarizmi, who lived in Baghdad and was the father of algebra.

DOMAINS AND LIMITATIONS

The word algorithm has taken on a different meaning in recent years due to the advent of computers. In the field of computing, it refers to a process that is described in a way that can be used in a computer program.

The principal goal of statistical software is to develop a programming language capable of incorporating statistical algorithms, so that these algorithms can then be presented in a form that is comprehensible to the user. The advantage of this approach is that the user understands the results produced by the algorithm and trusts the precision of the solutions. Among various statistical reviews that discuss algorithms, the *Journal of Algorithms* from the *Academic Press* (New York), the part of the *Journal of the Royal Statistical Society Series C (Applied Statistics)* that focuses on algorithms, *Computational Statistics* from *Physica-Verlag* (Heidelberg) and *Random Structures and Algorithms* edited by *Wiley* (New York) are all worthy of special mention.

EXAMPLES

We present here an algorithm that calculates the absolute value of a nonzero number; in other words $|x|$.

Process:

Step 1. Identify the algebraic sign of the given number.

Step 2. If the sign is negative, go to step 3. If the sign is positive, specify the absolute value of the number as the number itself:

$$|x| = x$$

and stop the process.

Step 3. Specify the absolute value of the given number as its opposite number:

$$|x| = -x$$

and stop the process.

FURTHER READING

- ▸ **Statistical software**
- ▸ **Yates' algorithm**

REFERENCES

Chambers, J.M.: Computational Methods for Data Analysis. Wiley, New York (1977)

Khwarizmi, Musa ibn Meusba (9th cent.). Jabr wa-al-muqeabalah. The algebra of Mohammed ben Musa, Rosen, F. (ed. and transl.). Georg Olms Verlag, Hildesheim (1986)

Rashed, R.: La naissance de l'algèbre. In: Noël, E. (ed.) Le Matin des Mathématiciens. Belin-Radio France, Paris (1985)

Alternative Hypothesis

An alternative hypothesis is the hypothesis which differs from the hypothesis being tested.

The alternative hypothesis is usually denoted by H_1.

HISTORY

See **hypothesis** and **hypothesis testing**.

MATHEMATICAL ASPECTS

During the **hypothesis testing** of a **parameter** of a **population**, the **null hypothesis** is presented in the following way:

$$H_0: \quad \theta = \theta_0 ,$$

where θ is the parameter of the population that is to be estimated, and θ_0 is the presumed **value** of this parameter. The alternative hypothesis can then take three different forms:

1. $H_1 : \theta > \theta_0$
2. $H_1 : \theta < \theta_0$
3. $H_1 : \theta \neq \theta_0$

In the first two cases, the **hypothesis test** is called the **one-sided**, whereas in the third case it is called the **two-sided**.

The alternative hypothesis can also take three different forms during the **hypothesis testing** of **parameters** of two **populations**. If the **null hypothesis** treats the two parameters θ_1 and θ_2 equally, then:

$$H_0 \colon \theta_1 = \theta_2 \quad \text{or}$$
$$H_0 \colon \theta_1 - \theta_2 = 0\,.$$

The alternative hypothesis could then be

- $H_1 : \theta_1 > \theta_2$ or $H_1 : \theta_1 - \theta_2 > 0$
- $H_1 : \theta_1 < \theta_2$ or $H_1 : \theta_1 - \theta_2 < 0$
- $H_1 : \theta_1 \neq \theta_2$ or $H_1 : \theta_1 - \theta_2 \neq 0$

During the comparison of more than two **populations**, the **null hypothesis** supposes that the **values** of all of the **parameters** are identical. If we want to compare k populations, the null hypothesis is the following:

$$H_0 \colon \theta_1 = \theta_2 = \ldots = \theta_k\,.$$

The alternative hypothesis will then be formulated as follows:

***H*₁:** the values of $\theta_i (i = 1, \ldots, k)$ are not all identical.

This means that only one **parameter** needs to have a different value to those of the other parameters in order to reject the **null hypothesis** and accept the alternative hypothesis.

EXAMPLES

We are going to examine the alternative hypotheses for three examples of **hypothesis testing**:

1. *Hypothesis testing on the percentage of a population*

   An election candidate wants to know if he will receive more than 50% of the votes.

   The **null hypothesis** for this problem can be written as follows:

   $$H_0 \colon \pi = 0.5\,,$$

   where π is the **percentage** of the **population** to be estimated.

   We carry out a **one-sided test** on the right-hand side that allows us to answer the candidate's question. The alternative hypothesis will therefore be:

   $$H_1 \colon \pi > 0.5\,.$$

2. *Hypothesis testing on the mean of a population*

   A bolt maker wants to test the precision of a new machine that should make bolts 8 mm in diameter.

   We can use the following **null hypothesis**:

   $$H_0 \colon \mu = 8\,,$$

   where μ is the **mean** of the **population** that is to be estimated.

   We carry out a **two-sided test** to check whether the bolt diameter is too small or too big.

   The alternative hypothesis can be formulated in the following way:

   $$H_1 \colon \mu \neq 8\,.$$

3. *Hypothesis testing on a comparison of the means of two populations*

   An insurance company decided to equip its offices with microcomputers. It wants to buy these computers from two different companies so long as there is no significant difference in durability between the two brands. It therefore tests the time that passes before the first breakdown on a **sample** of microcomputers from each brand.

According to the **null hypothesis**, the **mean** of the elapsed time before the first breakdown is the same for each brand:

$$H_0 \colon \mu_1 - \mu_2 = 0 \,.$$

Here μ_1 and μ_2 are the respective means of the two **populations**.

Since we do not know which mean will be the highest, we carry out a **two-sided test**. Therefore the alternative hypothesis will be:

$$H_1 \colon \mu_1 - \mu_2 \neq 0 \,.$$

FURTHER READING

- ► Analysis of variance
- ► Hypothesis
- ► Hypothesis testing
- ► Null hypothesis

REFERENCE

Lehmann, E.I., Romann, S.P.: Testing Statistical Hypothesis, 3rd edn. Springer, New York (2005)

Analysis of Binary Data

The study of how the probability of success depends on expanatory variables and grouping of materials.

The analysis of binary data also involves **goodness-of-fit** tests of a sample of binary variables to a theoretical distribution, as well as the study of 2×2 **contingency tables** and their subsequent analysis. In the latter case we note especially **independence tests** between attributes, and **homogeneity tests**.

HISTORY

See **data analysis**.

MATHEMATICAL ASPECTS

Let Y be a binary **random variable** and $X_1, X_2, \ldots, X_k$ be supplementary binary variables. So the **dependence** of Y on the variables $X_1, X_2, \ldots, X_k$ is represented by the following models (the coefficients of which are estimated via the **maximum likelihood**):

1. *Linear model:* $P(Y = 1)$ is expressed as a linear function (in the parameters) of X_i.
2. *Log-linear model:* $\log P(Y = 1)$ is expressed as a linear function (in the parameters) of X_i.
3. *Logistic model:* $\log\left(\frac{P(Y=1)}{P(Y=0)}\right)$ is expressed as a linear function (in the parameters) of X_i.

Models 1 and 2 are easier to interpret. Yet the last one has the advantage that the quantity to be explained takes all possible values of the linear models. It is also important to pay attention to the extrapolation of the model outside of the domain in which it is applied.

It is possible that among the independent variables $(X_1, X_2, \ldots, X_k)$, there are categorical variables (eg. binary ones). In this case, it is necessary to treat the nonbinary categorical variables in the following way: let Z be a random variable with m categories. We enumerate the categories from 1 to m and we define $m - 1$ random variables $Z_1, Z_2, \ldots, Z_{m-1}$. So Z_i takes the value 1 if Z belongs to the category represented by this index. The variable Z is therefore replaced by these $m - 1$ variables, the coefficients of which express the influence of the considered category. The reference (used in order to avoid the situation of **collinearity**) will have (for the purposes of comparison with other categories) a parameter of zero.

FURTHER READING

- ► **Binary data**
- ► **Data analysis**

REFERENCES

Cox, D.R., Snell, E.J.: The Analysis of Binary Data. Chapman & Hall (1989)

Analysis of Categorical Data

The analysis of **categorical data** involves the following methods:

(a) A study of the **goodness-of-fit test**;

(b) The study of a **contingency table** and its subsequent analysis, which consists of discovering and studying relationships between the attributes (if they exist);

(c) An **homogeneity test** of some populations, related to the distribution of a binary qualitative categorical variable;

(d) An examination of the **independence** hypothesis.

HISTORY

The term "contingency", used in the relation to cross tables of **categorical data** was probably first used by **Pearson, Karl** (1904). The **chi-square test**, was proposed by Barlett, M.S. in 1937.

MATHEMATICAL ASPECTS

See **goodness-of-fit** and **contingency table**.

FURTHER READING

- ► **Data**
- ► **Data analysis**
- ► **Categorical data**
- ► **Chi-square goodness of fit test**
- ► **Contingency table**
- ► **Correspondence analysis**
- ► **Goodness of fit test**
- ► **Homogeneity test**
- ► **Test of independence**

REFERENCES

Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)

Bartlett, M.S.: Properties of sufficiency and statistical tests. Proc. Roy. Soc. Lond. Ser. A **160**, 268–282 (1937)

Cox, D.R., Snell, E.J.: Analysis of Binary Data, 2nd edn. Chapman & Hall, London (1990)

Haberman, S.J.: Analysis of Qualitative Data. Vol. I: Introductory Topics. Academic, New York (1978)

Pearson, K.: On the theory of contingency and its relation to association and normal correlation. Drapers' Company Research Memoirs, Biometric Ser. I., pp. 1–35 (1904)

Analysis of Residuals

An analysis of **residuals** is used to test the validity of the statistical model and to control the assumptions made on the error term. It may be used also for outlier detection.

HISTORY

The analysis of residuals dates back to Euler (1749) and Mayer (1750) in the middle of the eighteenth century, who were confronted with the problem of the **estimation** of **parameters** from **observations** in the field of astronomy. Most of the methods used to analyze residuals are based on the works of Anscombe (1961) and Anscombe and Tukey (1963). In 1973, Anscombe also presented an interesting discussion on the reasons for using graphical methods of analysis. Cook and Weisberg (1982) dedicated a complete book to the analysis of residuals. Draper and Smith (1981) also addressed this problem in a chapter of their work *Applied Regression Analysis*.

MATHEMATICAL ASPECTS

Consider a general model of **multiple linear regression**:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i\,, \quad i = 1, \ldots, n\,,$$

where ε_i is the nonobservable random **error** term.

The **hypotheses** for the **errors** ε_i are generally as follows:

- The errors are independent;
- They are normally distributed (they follow a **normal distribution**);
- Their **mean** is equal to zero;
- Their **variance** is constant and equal to σ^2.

Regression analysis gives an **estimation** for Y_i,denoted $\hat{Y}_i$. If the chosen **model** is adequate, the distribution of the **residuals** or "observed **errors**" $e_i = Y_i - \hat{Y}_i$ should confirm these hypotheses.

Methods used to analyze residuals are mainly graphical. Such methods include:

1. Representing the residuals by a frequency chart (for example a **scatter plot**).
2. Plotting the residuals as a function of time (if the chronological order is known).
3. Plotting the residuals as a function of the estimated **values** $\hat{Y}_i$.
4. Plotting the residuals as a function of the **independent variables** X_{ij}.
5. Creating a **Q–Q plot** of the residuals.

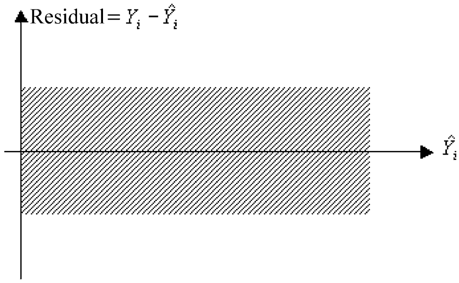
DOMAINS AND LIMITATIONS

To validate the analysis, some of the hypotheses need to hold (like for example the normality of the residuals in estimations based on the **mean square**).

Consider a plot of the **residuals** as a function of the estimated **values** $\hat{Y}_i$. This is one of the most commonly used graphical approaches to verifying the validity of a **model**. It consists of placing:

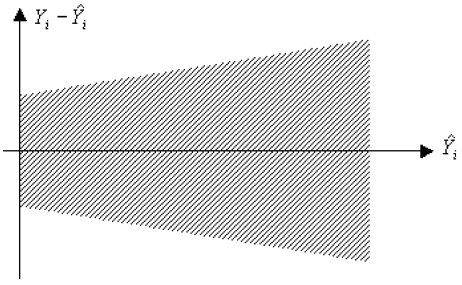
- The **residuals** $e_i = Y_i - \hat{Y}_i$ in increasing **order**;
- The estimated values $\hat{Y}_i$ on the abscissa.

If the chosen **model** is adequate, the **residuals** are uniformly distributed on a horizontal band of points.

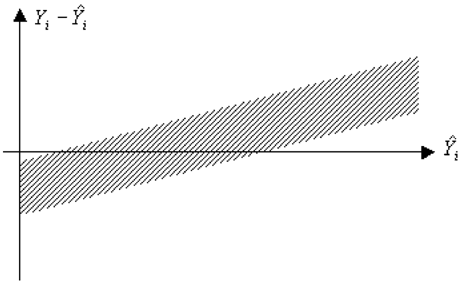


However, if the **hypotheses** for the **residuals** are not verified, the shape of the plot can be different to this. The three figures below show the shapes obtained when:

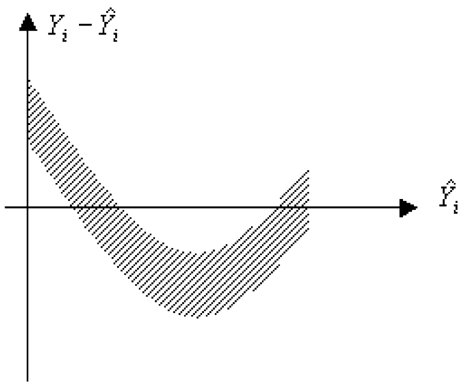
1. The **variance** σ^2 is not constant. In this case, it is necessary to perform a **transformation** on the **data** Y_i before tackling the **regression analysis**.



2. The chosen **model** is inadequate (for example, the model is linear but the constant term was omitted when it was necessary).



3. The chosen **model** is inadequate (a parabolic tendency is observed).



Different **statistics** have been proposed in order to permit numerical measurements that are complementary to the visual techniques presented above, which include those given by Anscombe (1961) and Anscombe and Tukey (1963).

EXAMPLES

In the nineteenth century, a Scottish physicist named Forbe, James D. wanted to estimate the altitude above sea level by measuring the boiling point of water. He knew that the altitude could be determined from the atmospheric pressure; he then studied the relation between pressure and the boiling point of water. Forbe suggested that for an **interval** of observed **values**, a plot of the logarithm of the pressure as a function of the boiling point of water should give a straight line. Since the logarithm of these pressures is small and varies little, we have multiplied these values by 100 below.

X boiling point	*Y* 100 · log (pressure)
194.5	131.79
194.3	131.79
197.9	135.02
198.4	135.55
199.4	136.46
199.9	136.83
200.9	137.82
201.1	138.00
201.4	138.06
201.3	138.05
203.6	140.04
204.6	142.44
209.5	145.47
208.6	144.34
210.7	146.30
211.9	147.54
212.2	147.80

The **simple linear regression model** for this problem is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i , \quad i = 1, \ldots, 17 .$$

Using the **least squares** method, we can find the following estimation function:

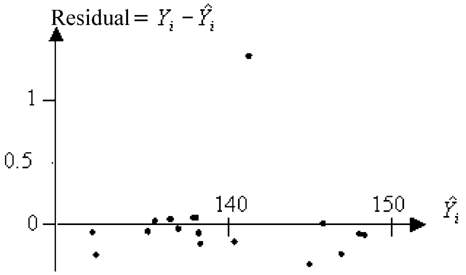
$$\hat{Y}_i = -42.131 + 0.895X_i$$

where $\hat{Y}_i$ is the estimated **value** of **variable** Y for a given X.
For each of these 17 values of X_i, we have an estimated value $\hat{Y}_i$. We can calculate the **residuals**:

$$e_i = Y_i - \hat{Y}_i\,.$$

These results are presented in the following table:

i	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
1	194.5	131.79	132.037	−0.247
2	194.3	131.79	131.857	−0.067
3	197.9	135.02	135.081	−0.061
4	198.4	135.55	135.529	0.021
5	199.4	136.46	136.424	0.036
6	199.9	136.83	136.872	−0.042
7	200.9	137.82	137.768	0.052
8	201.1	138.00	137.947	0.053
9	201.4	138.06	138.215	−0.155
10	201.3	138.05	138.126	−0.076
11	203.6	140.04	140.185	−0.145
12	204.6	142.44	141.081	1.359
13	209.5	145.47	145.469	0.001
14	208.6	144.34	144.663	−0.323
15	210.7	146.30	146.543	−0.243
16	211.9	147.54	147.618	−0.078
17	212.2	147.80	147.886	−0.086



Plotting the **residuals** as a function of the estimated **values** $\hat{Y}_i$ gives the previous graph.
It is apparent from this graph that, except for one **observation** (the 12th), where the value of the **residual** seems to indicate an **outlier**, the residuals are distributed in a very thin horizontal strip. In this case the **residuals** do not provide any reason to doubt the validity of the chosen **model**. By analyzing the standardized residuals we can determine whether the 12th observation is an **outlier** or not.

FURTHER READING

- Anderson–Darling test
- Least squares
- Multiple linear regression
- Outlier
- Regression analysis
- Residual
- Scatterplot
- Simple linear regression

REFERENCES

Anscombe, F.J.: Examination of residuals. Proc. 4th Berkeley Symp. Math. Statist. Prob. **1**, 1–36 (1961)

Anscombe, F.J.: Graphs in statistical analysis. Am. Stat. **27**, 17–21 (1973)

Anscombe, F.J., Tukey, J.W.: Analysis of residuals. Technometrics **5**, 141–160 (1963)

Cook, R.D., Weisberg, S.: Residuals and Influence in Regression. Chapman & Hall, London (1982)

Cook, R.D., Weisberg, S.: An Introduction to Regression Graphics. Wiley, New York (1994)

Cook, R.D., Weisberg, S.: Applied Regression Including Computing and Graphics. Wiley, New York (1999)

A

Draper, N.R., Smith, H.: Applied Regression Analysis, 3rd edn. Wiley, New York (1998)

Euler, L.: Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, pièce ayant remporté le prix de l'année 1748, par l'Académie royale des sciences de Paris. Republié en 1960, dans Leonhardi Euleri, Opera Omnia, 2ème série. Turici, Bâle, 25, pp. 47–157 (1749)

Mayer, T.: Abhandlung über die Umwälzung des Monds um seine Achse und die scheinbare Bewegung der Mondflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 1748 **1**, 52–183 (1750)

Analysis of Variance

The analysis of variance is a technique that consists of separating the total variation of **data** set into logical components associated with specific sources of variation in order to compare the **mean** of several **populations**. This analysis also helps us to test certain **hypotheses** concerning the parameters of the model, or to estimate the components of the **variance**. The sources of variation are globally summarized in a component called **error variance**, sometime called within-treatment mean square and another component that is termed "effect" or treatment, sometime called between-treatment mean square.

HISTORY

Analysis of variance dates back to **Fisher, R.A.** (1925). He established the first fundamental principles in this field. Analysis of variance was first applied in the fields of biology and agriculture.

MATHEMATICAL ASPECTS

The analysis of **variance** compares the **means** of three or more random **samples** and determines whether there is a significant difference between the **populations** from which the samples are taken. This technique can only be applied if the random samples are independent, if the population distributions are approximately normal and all have the same variance σ^2.

Having established that the **null hypothesis**, assumes that the **means** are equal, while the **alternative hypothesis** affirms that at least one of them is different, we fix a **significant level**. We then make two **estimates** of the unknown **variance** σ^2:

- The first, denoted s_{E}^2, corresponds to the mean of the variances of each sample;
- The second, s_{Tr}^2, is based on the variation between the means of the samples.

Ideally, if the **null hypothesis** is verified, these two estimations will be equal, and the F ratio ($F = s_{\mathrm{Tr}}^2 / s_{\mathrm{E}}^2$, as used in the **Fisher test** and defined as the quotient of the second estimation of σ^2 to the first) will be equal to 1.

The **value** of the F ratio, which is generally more than 1 because of the variation from the **sampling**, must be compared to the value in the **Fisher table** corresponding to the fixed **significant level**. The decision rule consists of either rejecting the null hypothesis if the calculated value is greater than or equal to the tabulated value, or else the **means** are equal, which shows that the **samples** come from the same **population**.

Consider the following **model**:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \,,$$
$$i = 1, 2, \ldots, t \,, \quad j = 1, 2, \ldots, n_i \,.$$

Here

$\boldsymbol{Y_{ij}}$ represents the **observation** j receiving the **treatment** i,

μ is the general **mean** common to all treatments,

τ_i is the actual effect of treatment i on the observation,

ε_{ij} is the experimental error for observation Y_{ij}.

In this case, the **null hypothesis** is expressed in the following way:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_t ,$$

which means that the t treatments are identical.

The **alternative hypothesis** is formulated in the following way:

$$H_1 : \text{the } \textbf{values} \text{ of } \tau_i (i = 1, 2, \ldots, t) \\ \text{are not all identical} .$$

The following formulae are used:

$$SS_{\mathrm{Tr}} = \sum_{i=1}^{t} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 , \qquad s^2_{\mathrm{Tr}} = \frac{SS_{\mathrm{Tr}}}{t-1} ,$$

$$SS_{\mathrm{E}} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 , \qquad s^2_{\mathrm{E}} = \frac{SS_{\mathrm{E}}}{N-t} ,$$

and

$$SS_T = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$$

or

$$SS_T = SS_{\mathrm{Tr}} + SS_{\mathrm{E}} .$$

where

$$\bar{Y}_{i.} = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$ is the **mean** of the ith set

$$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n_i} Y_{ij}$$ is the global mean taken on all the **observations**, and

$$N = \sum_{i=1}^{t} n_i$$ is the total number of observations.

and finally the value of the F ratio

$$F = \frac{s^2_{\mathrm{Tr}}}{s^2_{\mathrm{E}}} .$$

It is customary to summarize the information from the analysis of variance in an analysis of variance table:

Source of variation	Degrees of freedom	Sum of squares	Mean of squares	F
Among treatments	$t-1$	SS_{Tr}	s^2_{Tr}	$\frac{s^2_{\mathrm{Tr}}}{s^2_{\mathrm{E}}}$
Within treatments	$N-t$	SS_{E}	s^2_{E}	
Total	$N-1$	SS_T		

DOMAINS AND LIMITATIONS

An analysis of variance is always associated with a **model**. Therefore, there is a different analysis of variance in each distinct case. For example, consider the case where the analysis of variance is applied to **factorial experiments** with one or several **factors**, and these factorial experiments are linked to several **designs of experiment**.

We can distinguish not only the number of **factors** in the **experiment** but also the type of **hypotheses** linked to the effects of the **treatments**. We then have a **model** with fixed effects, a model with variable effects and a model with mixed effects. Each of these requires a specific analysis, but whichever model is used, the basic assumptions of additivity, normality, homoscedasticity and independence must be respected. This means that:

1. The experimental **errors** of the **model** are **random variables** that are independent of each other;

A

2. All of the errors follow a **normal distribution** with a **mean** of zero and an unknown **variance** σ^2.

All **designs of experiment** can be analyzed using analysis of variance. The most common designs are **completely randomized designs**, **randomized block designs** and **Latin square designs**.

An analysis of variance can also be performed with simple or multiple **linear regression**.

If during an analysis of variance the null hypothesis (the case for equality of means) is rejected, a **least significant difference test** is used to identify the **populations** that have significantly different means, which is something that an analysis of variance cannot do.

EXAMPLES

See **two-way analysis of variance**, **one-way analysis of variance**, **linear multiple regression** and **simple linear regression**.

FURTHER READING

- ▸ Design of experiments
- ▸ Factor
- ▸ Fisher distribution
- ▸ Fisher table
- ▸ Fisher test
- ▸ Least significant difference test
- ▸ Multiple linear regression
- ▸ One-way analysis of variance
- ▸ Regression analysis
- ▸ Simple linear regression
- ▸ Two-way analysis of variance

REFERENCES

Fisher, R.A.: Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh (1925)

Rao, C.R.: Advanced Statistical Methods in Biometric Research. Wiley, New York (1952)

Scheffé, H.: The Analysis of Variance. Wiley, New York (1959)

Anderson, Oskar

Anderson, Oskar (1887–1960) was an important member of the Continental School of Statistics; his contributions touched upon a wide range of subjects, including correlation, time series analysis, nonparametric methods and sample survey, as well as econometrics and statistical applications in social sciences.

Anderson, Oskar received a bachelor degree with distinction from the Kazan Gymnasium and then studied mathematics and physics for a year at the University of Kazan. He then entered the Faculty of Economics at the Polytechnic Institute of St. Petersburg, where he studied mathematics, statistics and economics.

The publications of Anderson, Oskar combine the traditions of the Continental School of Statistics with the concepts of the English Biometric School, particularly in two of his works: "Einführung in die mathematische Statistik" and "Probleme der statistischen Methodenlehre in den Sozialwissenschaften".

In 1949, he founded the journal *Mitteilungsblatt für Mathematische Statistik* with Kellerer, Hans and Münzner, Hans.

Some principal works of Anderson, Oskar:

1935 Einführung in die Mathematische Statistik. Julius Springer, Wien

1954 Probleme der statistischen Methodenlehre in den Sozialwissenschaften. Physica-Verlag, Würzberg

Anderson, Theodore W.

Anderson, Theodore Wilbur was born on the 5th of June 1918 in Minneapolis, in the state of Minnesota in the USA. He became a Doctor of Mathematics in 1945 at the University of Princeton, and in 1946 he became a member of the Department of Mathematical Statistics at the University of Columbia, where he was named Professor in 1956. In 1967, he was named Professor of Statistics and Economics at Stanford University. He was, successively: Fellow of the Guggenheim Foundation between 1947 and 1948; Editor of the *Annals of Mathematical Statistics* from 1950 to 1952; President of the Institute of Mathematical Statistics in 1963; and Vice-President of the American Statistical Association from 1971 to 1973. He is a member of the American Academy of Arts and Sciences, of the National Academy of Sciences, of the Institute of Mathematical Statistics and of the Royal Statistical Society. Anderson's most important contribution to statistics is surely in the domain of multivariate analysis. In 1958, he published the book entitled *An Introduction to Multivariate Statistical Analysis*. This book was the reference work in this domain for over forty years. It has been even translated into Russian.

Some of the principal works and articles of Theodore Wilbur Anderson:

1952 (with Darling, D.A.) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann. Math. Stat. 23, 193–212.

1958 An Introduction to Multivariate Statistical Analysis. Wiley, New York.

1971 The Statistical Analysis of Time Series. Wiley, New York.

1989 Linear latent variable models and covariance structures. J. Econometrics, 41, 91–119.

1992 (with Kunitoma, N.) Asymptotic distributions of regression and autoregression coefficients with Martingale difference disturbances. J. Multivariate Anal., 40, 221–243.

1993 Goodness of fit tests for spectral distributions. Ann. Stat. 21, 830–847.

FURTHER READING

► **Anderson–Darling test**

Anderson–Darling Test

The Anderson–Darling test is a goodness-of-fit test which allows to control the hypothesis that the distribution of a random variable observed in a sample follows a certain theoretical distribution. In particular, it allows us to test whether the empirical distribution obtained corresponds to a **normal distribution**.

HISTORY

Anderson, Theodore W. and Darling D.A. initially used Anderson–Darling statistics, denoted A^2, to test the conformity of a distribution with perfectly specified parameters (1952 and 1954). Later on, in the 1960s and especially the 1970s, some other authors (mostly Stephens) adapted the test to a wider range of distributions where some of the parameters may not be known.

MATHEMATICAL ASPECTS

Let us consider the **random variable** X, which follows the normal distribution with an expectation μ and a variance σ^2, and has a distribution function $F_X(x;\theta)$, where θ is a parameter (or a set of parameters) that determine, F_X. We furthermore assume θ to be known.

An observation of a sample of size n issued from the variable X gives a distribution function $F_n(x)$. The Anderson–Darling statistic, denoted by A^2, is then given by the weighted sum of the squared deviations $F_X(x;\theta) - F_n(x)$:

$$A^2 = \frac{1}{n}\left(\sum_{i=1}^{n}\left(F_X(x;\theta) - F_n(x)\right)^2\right).$$

Starting from the fact that A^2 is a random variable that follows a certain distribution over the interval $[0; +\infty[$, it is possible to test, for a significance level that is fixed a priori, whether $F_n(x)$ is the realization of the random variable $F_X(X;\theta)$; that is, whether X follows the probability distribution with the distribution function $F_X(x;\theta)$.

Computation of A^2 Statistic

Arrange the observations $x_1, x_2, \ldots, x_n$ in the sample issued from X in ascending order i.e., $x_1 < x_2 < \ldots < x_n$. Note that $z_i = F_X(x_i;\theta)$, $(i = 1, 2, \ldots, n)$. Then compute, A^2 by:

$$A^2 = -\frac{1}{n}\left(\sum_{i=1}^{n}(2i-1)\left(\ln(z_i) + \ln(1 - z_{n+1-i})\right)\right) - n\,.$$

For the situation preferred here (X follows the normal distribution with expectation μ and variance σ^2), we can enumerate four cases, depending on the known parameters μ and σ^2 (F is the distribution function of the standard normal distribution):

1. μ and σ^2 are known, so $F_X(x;(\mu,\sigma^2))$ is perfectly specified. Naturally we then have $z_i = F(w_i)$ where $w_i = \frac{x_i-\mu}{\sigma}$.
2. σ^2 is known but μ is unknown and is estimated using $\bar{x} = \frac{1}{n}\left(\sum_i x_i\right)$, the mean of the sample. Then, let $z_i = F(w_i)$, where $w_i = \frac{x_i-\bar{x}}{\sigma}$.
3. μ is known but σ^2 is unknown and is estimated using $s'^2 = \frac{1}{n}\left(\sum_i (x_i - u)^2\right)$. In this case, let $z_i = F(w_i)$, where $w_i = \frac{x_{(i)}-\mu}{s'}$.
4. μ and σ^2 are both unknown and are estimated respectively using $\bar{x}$ and $s^2 = \frac{1}{n-1}\left(\sum_i (x_i - \bar{x})^2\right)$. Then, let $z_i = F(w_i)$, where $w_i = \frac{x_i-\bar{x}}{s}$.

Asymptotic distributions were found for A^2 by Anderson and Darling for the first case, and by Stephens for the next two cases. For last case, Stephens determined an asymptotic distribution for the transformation: $A^* = A^2\left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$.

Therefore, as shown below, we can construct a table that gives, depending on the case and the significance level (10%, 5%, 2.5% or 1% below), the limiting values of A^2 (and A^* for the case 4) beyond which the normality hypothesis is rejected:

	Significance level			
Case:	0.1	0.050	0.025	0.01
1: $A^2 =$	1.933	2.492	3.070	3.857
2: $A^2 =$	0.894	1.087	1.285	1.551
3: $A^2 =$	1.743	2.308	2.898	3.702
4: $A^* =$	0.631	0.752	0.873	1.035

DOMAINS AND LIMITATIONS

As the distribution of A^2 is expressed asymptotically, the test needs the sample size n to be large. If this is not the case then, for the first two cases, the distribution of A^2 is not known and it is necessary to perform a transformation of the type $A^2 \longmapsto A^*$, from which A^* can be determined. When $n > 20$, we can avoid such a transformation and so the data in the above table are valid.

The Anderson–Darling test has the advantage that it can be applied to a wide range of distributions (not just a **normal distribution** but also exponential, logistic and gamma distributions, among others). That allows us to try out a wide range of alternative distributions if the initial test rejects the null hypothesis for the distribution of a random variable.

EXAMPLES

The following data illustrate the application of the Anderson–Darling test for the normality hypothesis:

Consider a sample of the heights (in cm) of 25 male students. The following table shows the observations in the sample, and also w_i and z_i. We can also calculate $\bar{x}$ and s from these data: $\bar{x} = 177.36$ and $s = 4.98$. Assuming that F is a standard normal distribution function, we have:

Obs:	x_i	$w_i = \frac{x_i - \bar{x}}{s}$	$z_i = F(w_i)$
1	169	−1.678	0.047
2	169	−1.678	0.047
3	170	−1.477	0.070
4	171	−1.277	0.100
5	173	−0.875	0.191
6	173	−0.875	0.191
7	174	−0.674	0.250
8	175	−0.474	0.318
9	175	−0.474	0.318
10	175	−0.474	0.318
11	176	−0.273	0.392
12	176	−0.273	0.392
13	176	−0.273	0.392
14	179	0.329	0.629
15	180	0.530	0.702
16	180	0.530	0.702
17	180	0.530	0.702
18	181	0.731	0.767
19	181	0.731	0.767
20	182	0.931	0.824
21	182	0.931	0.824
22	182	0.931	0.824
23	185	1.533	0.937
24	185	1.533	0.937
25	185	1.533	0.937

We then get $A^2 \cong 0.436$, which gives

$$A^* = A^2 \cdot \left(1.0 + \frac{0.75}{25} + \frac{0.25}{625}\right)$$
$$= A^2 \cdot (1.0336) \cong 0.451 .$$

Since we have case 4, and a significance level fixed at 1%, the calculated value of A^* is much less then the value shown in the table (1.035). Therefore, the normality hypothesis cannot be rejected at a significance level of 1%.

FURTHER READING

- ▸ Goodness of fit test
- ▸ Histogram
- ▸ Nonparametric statistics
- ▸ Normal distribution
- ▸ Statistics

REFERENCES

Anderson, T.W., Darling, D.A.: Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann. Math. Stat. **23**, 193–212 (1952)

Anderson, T.W., Darling, D.A.: A test of goodness of fit. J. Am. Stat. Assoc. **49**, 765–769 (1954)

Durbin, J., Knott, M., Taylor, C.C.: Components of Cramer-Von Mises statistics, II. J. Roy. Stat. Soc. Ser. B **37**, 216–237 (1975)

Stephens, M.A.: EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**, 730–737 (1974)

Arithmetic Mean

The arithmetic **mean** is a **measure of central tendency**. It allows us to characterize the center of the **frequency distribution** of a **quantitative variable** by considering all of the observations with the same weight afforded to each (in contrast to the **weighted arithmetic mean**).

It is calculated by summing the observations and then dividing by the number of observations.

HISTORY

The arithmetic mean is one of the oldest methods used to combine observations in order to give a unique approximate value. It appears to have been first used by Babylonian astronomers in the third century BC. The arithmetic mean was used by the astronomers to determine the positions of the sun, the moon and the planets. According to Plackett (1958), the concept of the arithmetic mean originated from the Greek astronomer Hipparchus.

In 1755 Thomas Simpson officially proposed the use of the arithmetic mean in a letter to the President of the Royal Society.

MATHEMATICAL ASPECTS

Let $x_1, x_2, \ldots, x_n$ be a set of n quantities or n observations relating to a **quantitative variable** X.

The arithmetic mean $\bar{x}$ of $x_1, x_2, \ldots, x_n$ is the sum of these observations divided by the number n of observations:

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

When the observations are ordered in the form of a **frequency distribution**, the arithmetic mean is calculated in the following way:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} ,$$

where x_i are the different values of the **variable**, f_i are the frequencies associated with these values, k is the number of different values, and the sum of the frequencies equals the number of observations:

$$\sum_{i=1}^{k} f_i = n .$$

To calculate the **mean** of a frequency distribution where values of the quantitative variable X are grouped in classes, we consider that all of the observations belonging to a certain class take the central value of the class, assuming that the observations are uniformly distributed inside the classes (if this **hypothesis** is not correct, the arithmetic mean obtained will only be an approximation.)

Therefore, in this case we have:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} ,$$

where the x_i are the class centers, the f_i are the frequencies associated with each class, and k is the number of classes.

Properties of the Arithmetic Mean

- The algebraic sum of deviations between every value of the set and the arithmetic mean of this set equals 0:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 .$$

- The sum of square deviations from every value to a given number "a" is smallest when "a" is the arithmetic mean:

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

Proof:
We can write:

$$x_i - a = (x_i - \bar{x}) + (\bar{x} - a) \ .$$

Finding the squares of both members of the equality, summarizing them and then simplifying gives:

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n \cdot (\bar{x} - a)^2 \ .$$

As $n \cdot (\bar{x} - a)^2$ is not negative, we have proved that:

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

- The arithmetic mean $\bar{x}$ of a **sample** $(x_1, \ldots, x_n)$ is normally considered to be an **estimator** of the **mean** μ of the **population** from which the sample was taken.
- Assuming that x_i are independent random variables with the same **distribution function** for the mean μ and the **variance** σ^2, we can show that
  1. $E[\bar{x}] = \mu$,
  2. $\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$,

  if these moments exist.
  Since the **mathematical expectation** of $\bar{x}$ equals μ, the arithmetic mean is an estimator without **bias** of the mean of the population.
- If the x_i result from the **random sampling** without replacement of a finite population with a mean μ, the identity

$$E[\bar{x}] = \mu$$

is still valid, but the variance of $\bar{x}$ must be adjusted by a factor that depends on the size N of the population and the size n of the sample:

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n} \cdot \left[\frac{N-n}{N-1}\right] ,$$

where σ^2 is the variance of the population.

Relationship Between the Arithmetic Mean and Other Measures of Central Tendency

- The arithmetic mean is related to two principal measures of central tendency: the **mode** M_o and the **median** M_d.
  If the distribution is symmetric and unimodal:

$$\bar{x} = M_d = M_o \ .$$

  If the distribution is unimodal, it is normally true that:
  $\bar{x} \geq M_d \geq M_o$ if the distribution is stretched to the right,
  $\bar{x} \leq M_d \leq M_o$ if the distribution is stretched to the left.
  For a unimodal, slightly asymmetric distribution, these three measures of the central tendency often approximately satisfy the following relation:

$$(\bar{x} - M_o) = 3 \cdot (\bar{x} - M_d) \ .$$

- In the same way, for a unimodal distribution, if we consider a set of positive numbers, the **geometric mean** G is always smaller than or equal to the **arithmetic mean** $\bar{x}$, and is always greater than or equal to the **harmonic mean** H. So we have:

$$H \leq G \leq \bar{x}\,.$$

These three means are identical only if all of the numbers are equal.

DOMAINS AND LIMITATIONS

The arithmetic mean is a simple measure of the central **value** of a set of quantitative observations. Finding the mean can sometimes lead to poor data interpretation:

If the monthly salaries (in Euros) of 5 people are 3000, 3200, 2900, 3500 and 6500, the arithmetic mean of the salary is $\frac{19100}{5} = 3820$. This **mean** gives us some idea of the sizes of the salaries sampled, since it is situated between the biggest and the smallest one. However, 80% of the salaries are smaller then the mean, so in this case it is not a particularly good representation of a typical salary.

This case shows that we need to pay attention to the form of the distribution and the reliability of the observations before we use the arithmetic mean as the **measure of central tendency** for a particular set of values. If an absurd observation occurs in the distribution, the arithmetic mean could provide an unrepresentative value for the central tendency.
If some observations are considered to be less reliable then others, it could be useful to make them less important. This can be done by calculating a **weighted arithmetic mean**, or by using the **median**, which is not strongly influenced by any absurd observations.

EXAMPLES

In company A, nine employees have the following monthly salaries (in Euros):

3000 3200 2900 3440 5050
4150 3150 3300 5200

The arithmetic mean of these monthly salaries is:

$$\bar{x} = \frac{(3000 + 3200 + \cdots + 3300 + 5200)}{9} = \frac{33390}{9} = 3710\,\text{Euros}\,.$$

We now examine a case where the data are presented in the form of a **frequency distribution**.
The following **frequency table** gives the number of days that 50 employees were absent on sick leave during a period of one year:

x_i: Days of illness	f_i: Number of employees
0	7
1	12
2	19
3	8
4	4
Total	50

Let us try to calculate the mean number of days that the employees were absent due to illness.
The total number of sick days for the 50 employees equals the sum of the product of each x_i by its respective **frequency** f_i:

$$\sum_{i=1}^{5} x_i \cdot f_i = 0 \cdot 7 + 1 \cdot 12 + 2 \cdot 19 + 3 \cdot 8 + 4 \cdot 4 = 90\,.$$

The total number of employees equals:

$$\sum_{i=1}^{5} f_i = 7 + 12 + 19 + 8 + 4 = 50\,.$$

L The arithmetic mean of the number of sick days per employee is then:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i \cdot f_i}{\sum_{i=1}^{5} f_i} = \frac{90}{50} = 1.8$$

which means that, on average, the 50 employees took 1.8 days off for sickness per year.

In the following example, the data are grouped in classes.

We want to calculate the arithmetic mean of the daily profits from the sale of 50 types of grocery. The frequency distribution for the groceries is given in the following table:

Classes (profits in Euros)	Mid-points x_i	Frequencies f_i (number of groceries)	$x_i \cdot f_i$
500–550	525	3	1575
550–600	575	12	6900
600–650	625	17	10625
650–700	675	8	5400
700–750	725	6	4350
750–800	775	4	3100
Total		50	31950

The arithmetic mean of the profits is:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i \cdot f_i}{\sum_{i=1}^{6} f_i} = \frac{31950}{50} = 639\,,$$

which means that, on average, each of the 50 groceries provide a daily profit of 639 Euros.

FURTHER READING

- Geometric mean
- Harmonic mean
- Mean
- Measure of central tendency
- Weighted arithmetic mean

REFERENCES

Plackett, R.L.: Studies in the history of probability and statistics. VII. The principle of the arithmetic mean. Biometrika **45**, 130–135 (1958)

Simpson, T.: A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations in practical astronomy. Philos. Trans. Roy. Soc. Lond. **49**, 82–93 (1755)

Simpson, T.: An attempt to show the advantage arising by taking the mean of a number of observations in practical astronomy. In: Miscellaneous Tracts on Some Curious and Very Interesting Subjects in Mechanics, Physical-Astronomy, and Speculative Mathematics. Nourse, London (1757). pp. 64–75

Arithmetic Triangle

The arithmetic triangle is used to determine **binomial** coefficients $(a + b)^n$ when calculating the number of possible combinations of k objects out of a total of n objects (C_n^k).

HISTORY

The notion of finding the number of combinations of k objects from n objects in total has been explored in India since the ninth century. Indeed, there are traces of it in the

Meru Prastara written by Pingala in around 200 BC.

Between the fourteenth and the fifteenth centuries, al-Kashi, a mathematician from the Iranian city of Kashan, wrote *The Key to Arithmetic*. In this work he calls **binomial** coefficients "exponent elements".

In his work *Traité du Triangle Arithmétique*, published in 1665, Pascal, Blaise (1654) defined the numbers in the "arithmetic triangle", and so this triangle is also known as Pascal's triangle.

We should also note that the triangle was made popular by Tartaglia, Niccolo Fontana in 1556, and so Italians often refer to it as Tartaglia's triangle, even though Tartaglia did not actually study the arithmetic triangle.

MATHEMATICAL ASPECTS

The arithmetic triangle has the following form:

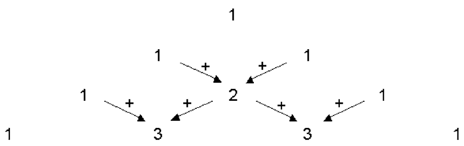
$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
 & & & & & & \dots & & & & & &
\end{array}$$

Each element is a **binomial** coefficient

$$C_n^k = \frac{n!}{k!\,(n-k)!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k}\,.$$

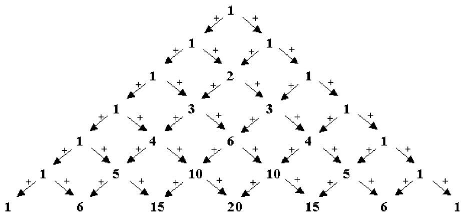
This coefficient corresponds to the element k of the line $n+1$, $k = 0, \ldots, n$.

Any particular number is obtained by adding together its neighboring numbers in the previous line.



For example:

$$C_6^4 = C_5^3 + C_5^4 = 10 + 5 = 15\,.$$



More generally, we have the **relation**:

$$C_n^k + C_n^{k+1} = C_{n+1}^{k+1}\,,$$

because:

$$\begin{aligned}
C_n^k + C_n^{k+1} &= \frac{n!}{(n-k)! \cdot k!} \\
&\quad + \frac{n!}{(n-k-1)! \cdot (k+1)!} \\
&= \frac{n! \cdot [(k+1) + (n-k)]}{(n-k)! \cdot (k+1)!} \\
&= \frac{(n+1)!}{(n-k)! \cdot (k+1)!} \\
&= C_{n+1}^{k+1}\,.
\end{aligned}$$

FURTHER READING

- ▶ **Binomial**
- ▶ **Binomial distribution**
- ▶ **Combination**
- ▶ **Combinatory analysis**

REFERENCES

Pascal, B.: Traité du triangle arithmétique (publ. posthum. in 1665), Paris (1654)

Pascal, B.: Œuvres, vols. 1–14. Brunschvicg, L., Boutroux, P., Gazier, F. (eds.)

Les Grands Ecrivains de France. Hachette, Paris (1904–1925)

Pascal, B.: Mesnard, J. (ed.) Œuvres complètes. Vol. 2. Desclée de Brouwer, Paris (1970)

Rashed, R.: La naissance de l'algèbre. In: Noël, E. (ed.) Le Matin des Mathématiciens. Belin-Radio France, Paris (1985) Chap. 12).

Youschkevitch, A.P.: Les mathématiques arabes (VIIIème-XVème siècles). Partial translation by Cazenave, M., Jaouiche, K. Vrin, Paris (1976)

ARMA Models

ARMA models (sometimes called Box-Jenkins models) are autoregressive moving average models used in time series analysis. The autoregressive part, denoted *AR*, consists of a finite linear combination of previous observations. The moving average part, *MA*, consists of a finite linear combination in t of the previous values for a white noise (a sequence of mutually independent and identically distributed random variables).

MATHEMATICAL ASPECTS

1. *AR* model (autoregressive)
   In an autoregressive process of order p, the present observation y_t is generated by a weighted mean of the past observations up to the pth period. This takes the following form:

$$AR(1): y_t = \theta_1 y_{t-1} + \varepsilon_t ,$$
$$AR(2): y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t ,$$
$$\vdots$$
$$AR(p): y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + \varepsilon_t ,$$

   where $\theta_1, \theta_2, \dots, \theta_p$ are the positive or negative parameters to be estimated and ε_t is the error factor, which follows a normal distribution.

2. *MA* model (moving average)
   In a moving average process of order q, each observation y_t is randomly generated by a **weighted arithmetic mean** until the qth period:

$$MA(1): y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1}$$
$$MA(2): y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2}$$
$$\dots$$
$$MA(p): y_t = \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} - \dots - \alpha_q \varepsilon_{t-q} ,$$

   where $\alpha_1, \alpha_2, \dots, \alpha_q$ are positive or negative parameters and ε_t is the Gaussian random error.
   The *MA* model represents a time series fluctuating about its mean in a random manner, which gives rise to the term "moving average", because it smoothes the series, subtracting the white noise generated by the randomness of the element.

3. *ARMA* model (autoregressive moving average model)
   ARMA models represent processes generated from a combination of past values and past errors. They are defined by the following equation:

$$ARMA(p, q):$$
$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + \varepsilon_t - \alpha_1 \varepsilon_{t-1} - \alpha_2 \varepsilon_{t-2} - \dots - \alpha_q \varepsilon_{t-q} ,$$

   with $\theta_p \neq 0, \alpha_q \neq 0$, and $(\varepsilon_t, t \in Z)$ is a weak white noise.

FURTHER READING
- ▶ Time series
- ▶ Weighted arithmetic mean

REFERENCES
Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control (Series in Time Series Analysis). Holden Day, San Francisco (1970)

Arrangement

Arrangements are a concept found in **combinatory analysis**.

The number of arrangements is the number of ways drawing k objects from n objects where the order in which the objects are drawn is taken into account (in contrast to **combinations**).

HISTORY
See **combinatory analysis**.

MATHEMATICAL ASPECTS
1. *Arrangements without repetitions*
   An arrangement without repetition refers to the situation where the objects drawn are not placed back in for the next drawing. Each object can then only be drawn once during the k drawings.
   The number of arrangements of k objects amongst n without repetition is equal to:
   $$A_n^k = \frac{n!}{(n-k)!}\,.$$
2. *Arrangements with repetitions*
   Arrangements with repetition occur when each object pulled out is placed back in for the next drawing. Each object can then be drawn r times from k drawings, $r = 0, 1, \ldots, k$.
   The number of arrangements of k objects amongst n with repetitions is equal to n to the power k:
   $$A_n^k = n^k\,.$$

EXAMPLES
1. *Arrangements without repetitions*
   Consider an urn containing six balls numbered from 1 to 6. We pull out four balls from the urn in succession, and we want to know how many numbers it is possible to form from the numbers of the balls drawn. We are then interested in the number of arrangements (since we take into account the order of the balls) without repetition (since each ball can be pulled out only once) of four objects amongst six. We obtain:
   $$A_n^k = \frac{n!}{(n-k)!} = \frac{6!}{(6-4)!} = 360$$
   possible arrangements. Therefore, it is possible to form 360 different numbers by drawing four numbers from the numbers 1,2,3,4,5,6 when each number can appear only once in the four-digit number formed.
   As a second example, let us investigate the arrangements without repetitions of two letters from the letters A, B and C. With $n = 3$ and $k = 2$ we have:
   $$A_n^k = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = 6\,.$$
   We then obtain:
   AB, AC, BA, BC, CA, CB.
2. *Arrangements with repetitions*
   Consider the same urn as described previously. We perform four successive drawings, but this time we put each ball drawn back in the urn.

We want to know how many four-digit numbers (or arrangements) are possible if four numbers are drawn.

In this case, we are investigating fthe number of arrangements with repetition (since each ball is placed back in the urn before the next drawing). We obtain

$$A_n^k = n^k = 6^4 = 1296$$

different arrangements. It is possible to form 1296 four-digit numbers from the numbers 1,2,3,4,5,6 if each number can appear more than once in the four-digit number.

As a second example we again take the three letters A, B and C and form an arrangement of two letters with repetitions. With $n = 3$ and $k = 2$, we have:

$$A_n^k = n^k = 3^2 = 9 .$$

We then obtain:

AA, AB, AC, BA, BB, BC, CA, CB, CC.

FURTHER READING

- ► **Combination**
- ► **Combinatory analysis**
- ► **Permutation**

REFERENCES

See **combinatory analysis**.

Attributable Risk

The attributable risk is the difference between the **risk** encountered by individuals exposed to a particular factor and the risk encountered by individuals who are not exposed to it. This is the opposite to **avoidable risk**. It measures the absolute effect of a cause (that is, the excess **risk** or cases of illness).

HISTORY

See **risk**.

MATHEMATICAL ASPECTS

By definition we have:

$$\begin{aligned}\text{attributable risk} &= \text{risk for those exposed} \\ &\quad - \text{risk for those not exposed} .\end{aligned}$$

DOMAINS AND LIMITATIONS

The confidence interval of an attributable risk is equivalent to the confidence interval of the difference between the proportions p_E and p_NE, where p_E and p_NE represent the risks encountered by individuals exposed and not exposed to the studied factor, respectively. Take n_E and n_{NE} to be, respectively, the size of the exposed and nonexposed populations. Then, for a confidence level of $(1-\alpha)$, is given by:

$$(p_\mathrm{E} - p_\mathrm{NE}) \pm z_\alpha \sqrt{\frac{p_\mathrm{E} \cdot (1-p_\mathrm{E})}{n_\mathrm{E}} + \frac{p_\mathrm{NE} \cdot (1-p_\mathrm{NE})}{n_{NE}}} ,$$

where z_α the **value** obtained from the **normal table** (for example, for a confidence interval of 95%, $\alpha = 0.05$ and $z_\alpha = 1.96$). The confidence interval for $(1-\alpha)$ for an avoidable risk has bounds given by:

$$(p_\mathrm{NE} - p_\mathrm{E}) \pm z_\alpha \sqrt{\frac{p_\mathrm{E} \cdot (1-p_\mathrm{E})}{n_\mathrm{E}} \cdot \frac{p_\mathrm{NE} \cdot (1-p_\mathrm{NE})}{n_{NE}}} .$$

Here, n_E and n_{NE} need to be large. If the confidence interval includes zero, we cannot rule out an absence of attributable risk.

EXAMPLES

As an example, we consider a study of the risk of breast cancer in women due to smoking:

A

Group	Incidence rate (/100000 /year)	Attributable to risk from smoking (*A*) (/100000 /year)
Nonexposed	57.0	57.0 − 57.0 = 0
Passive smokers	126.2	126.2 − 57.0 = 69.2
Active smokers	138.1	138.1 − 57.0 = 81.1
Total	114.7	114.7 − 57.0 = 57.7

The risks attributable to passive and active smoking are respectively 69 and 81 (/100000 year). In other words, if the exposure to tobacco was removed, the incidence rate for active smokers (138/100000 per year) could be reduced by 81/100000 per year and that for passive smokers (126/100000 per year) by 69/100000 per year. The incidence rates in both categories of smokers would become equal to the rate for nonexposed women (57/100000 per year). Note that the incidence rate for nonexposed women is not zero, due to the influence of other factors aside from smoking.

Group	No. indiv. observed over two years	Cases attrib. to smoking (for two-year period)	Cases attrib. to smoking (per year)
Nonexposed	70160	0.0	0.0
Passive smokers	110860	76.7	38.4
Active smokers	118636	96.2	48.1
Total	299656	172.9	86.5

We can calculate the number of cases of breast cancer attributable to tobacco exposure by multiplying the number of individuals observed per year by the attributable risk. By dividing the number of incidents attributable to smoking in the two-year period by two, we obtain the number of cases attributable to smoking per year, and we can then determine the risk attributable to smoking in the population, denoted PAR, as shown in the following example. The previous table shows the details of the calculus.

We describe the calculus for the passive smokers here. In the two-year study, 110860 passive smokers were observed. The risk attributable to the passive smoking was 69.2/100000 per year. This means that the number of cases attributable to smoking over the two-year period is $(110860 \cdot 69.2)/100000 = 76.7$. If we want to calculate the number of cases attributable to passive smoking per year, we must then divide the last value by 2, obtaining 38.4. Moreover, we can calculate the risk attributable to smoking per year simply by dividing the number of cases attributable to smoking for the two-year period (172.9) by the number of individuals studied during these two years (299656 persons). We then obtain the risk attributable to smoking as 57.7/100000 per year. We note that we can get the same result by taking the difference between the total incidence rate (114.7/100000 per year, see the examples under the entries for **incidence rate**, **prevalence rate**) and the incidence rate of the nonexposed group (57.0/100000 per year).

The risk of breast cancer attributable to smoking in the population (PAR) is the ratio of the number of the cases of breast cancer attributable to exposure to tobacco and the number of cases of breast cancer diag-