Categorical Data

Categorical data consists of counts of observations falling into specified classes.

We can distinguish between various types of categorical data:

- Binary, characterizing the presence or absence of a property;
- Unordered multicategorical (also called "nominal");
- Ordered multicategorical (also called "ordinal");
- Whole numbers.

We represent the categorical data in the form of a contingency table.

DOMAINS AND LIMITATIONS

Variables that are essentially continuous can also be presented as categorical variables. One example is "age", which is a continuous variable, but ages can still be grouped into classes so it can still be presented as categorical data.

EXAMPLES

In a public opinion survey for approving or disapproving a new law, the votes cast can be either "yes" or "no". We can represent the results in the form of a contingency table:

	Yes	No
Votes	8546	5455

If we divide up the employees of a business into professions (and at least three professions are presented), the data we obtain is unordered multicategorical data (there is no natural ordering of the professions).

In contrast, if we are interested in the number of people that have achieved various levels of education, there will probably be a natural ordering of the categories: "primary, secondary" and then university. Such data would therefore be an example of ordered multicategorical data.

Finally, if we group employees into categories based on the size of each employee's family (that is, the number of family members), we obtain categorical data where the categories are whole numbers.

FURTHER READING

- ▸ **Analysis of categorical data**
- ▸ **Binary data**
- ▸ **Category**
- ▸ **Data**
- ▸ **Dichotomous variable**
- ▸ **Qualitative categorical variable**
- ▸ **Random variable**

REFERENCES

See **analysis of categorical data**.

Category

A category represents a set of people or objects that have a common characteristic.

If we want to study the people in a **population**, we can sort them into "natural" categories, by gender (men and women) for example, or into categories defined by other criteria, such as vocation (managers, secretaries, farmers ...).

FURTHER READING

- ▸ **Binary data**
- ▸ **Categorical data**
- ▸ **Dichotomous variable**
- ▸ **Population**
- ▸ **Random variable**
- ▸ **Variable**

Cauchy Distribution

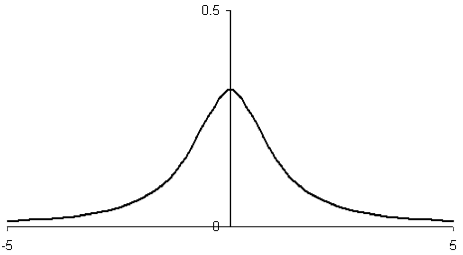
A **random variable** X follows a Cauchy distribution if its **density function** is of the form:

$$f(x) = \frac{1}{\pi\theta} \cdot \left[1 + \left(\frac{x-\alpha}{\theta}\right)^2\right]^{-1}, \quad \theta > 0.$$

The **parameters** α and θ are the location and dispersion parameters, respectively.

The Cauchy distribution is symmetric about $x = \alpha$, which represents the **median**. The first **quartile** and the third quartile are given by $\alpha \pm \theta$.

The Cauchy distribution is a **continuous probability distribution**.



Cauchy distribution, $\theta = 1$, $\alpha = 0$

MATHEMATICAL ASPECTS

The **expected value** $E[X]$ and the **variance** $\mathrm{Var}(X)$ do not exist.

If $\alpha = 0$ and $\theta = 1$, the Cauchy distribution is identical to the **Student distribution** with one **degree of freedom**.

DOMAINS AND LIMITATIONS

Its importance in physics is mainly due to the fact that it is the solution to the differential equation describing force resonance.

FURTHER READING

- ▸ **Continuous probability distribution**
- ▸ **Student distribution**

REFERENCES

Cauchy, A.L.: Sur les résultats moyens d'observations de même nature, et sur les résultats les plus probables. C.R. Acad. Sci. **37**, 198–206 (1853)

Causal Epidemiology

The aim of causal epidemiology is to identify how cause is related to effect with regard to human health.

In other words, it is the study of causes of illness, and involves attempting to find statis-

tical evidence for a causal relationship or an association between the illness and the factor proposed to cause the illness.

HISTORY

See **epidemiology**.

MATHEMATICAL ASPECTS

See **cause and effect in epidemiology, odds and odds ratio, relative risk, attributable risk, avoidable risk, incidence rate, prevalence rate**.

DOMAINS AND LIMITATIONS

Studies of the relationship between tobacco smoking and the development of lungial cancers and the relationship between HIV and the AIDS are examples of causal epidemiology. Research into a causal relation is often very complex, requiring many studies and the incorporation and combination of various data sets from biological and animal experiments, to clinical trials.

While causes cannot always be identified precisely, a knowledge of the risk factors associated with an illness and therefore the groups of people at risk allows us to intervene with preventative measures that could preserve health.

EXAMPLES

As an example, we can investigate the relationship between smoking and the development of lung cancer. Consider a study of 2000 subjects: 1000 smokers and 1000 nonsmokers. The age distributions and male/female proportions are identical in both groups. Let us analyze a summary of data obtained over many years. The results are presented in the following table:

	Smokers	Non-smokers
Number of cases of lung cancer (/1000)	50	10
Proportion of people in this group that contracted lung cancer	5%	1%

If we compare the proportion of smokers that contract lung cancer to the proportion of nonsmokers that do, we get $5\%/1\% = 5$, so we can conclude that the risk of developing lung cancer is five times higher in smokers than in nonsmokers. We generally also evaluate the significance of this result computing ***p*-value**. Suppose that the **chi-square test** gave $p < 0.001$. It normally accepted that if the ***p*-value** is smaller then 0.05, then the results obtained are statistically significant.

Suppose that we perform the same study, but the dimension of each group (smokers and nonsmokers) is 100 instead of 1000, and we observe the same proportions of people that contract lung cancer:

	Smokers	Non-smokers
Number of cases of lung cancer (/100)	5	1
Proportion of individuals in this group that contracted lungs cancer	5%	1%

If we perform the same statistical test, the ***p*-value** is found to be 0.212. Since this is greater than 0.05, we can cannot draw any solid conclud that there is about the existence of a significant statistical relation between smoking and lung cancer. This illustrates that, in order to have a statistically signifi-

cant level of difference between the results for different populations obtained from epidemiological studies, it is usually necessary to study large samples.

FURTHER READING

See **epidemiology**.

REFERENCES

See **epidemiology**.

Cause and Effect in Epidemiology

In epidemiology, the "cause" is an agent (microbial germs, polluted water, smoking, etc.) that modifies health, and the "effect" describes the the way that the health is changed by the agent. The agent is often potentially pathogenic (in which case it is known as a "risk factor").

The effect is therefore effectively a risk comparison. We can define two different types of risk in this context:

- The absolute effect of a cause expresses the increase in the risk or the additional number of cases of illness that result or could result from exposure to this cause. It is measured by the **attributable risk** and its derivatives.
- The relative effect of a cause expresses the strength of the association between the causal agent and the illness.

A cause that produces an effect by itself is called *sufficient*.

HISTORY

The terms "cause" and "effect" were defined at the birth of **epidemiology**, which occurred in the seventeenth century.

MATHEMATICAL ASPECTS

Formally, we have:

$$\text{Absolute effect} = \text{Risk for exposed} - \text{risk for unexposed}.$$

The absolute effect expresses the excess risk or cases of illness that result (or could result) from exposure to the cause.

$$\text{Relative effect} = \frac{\text{risk for exposed}}{\text{risk for unexposed}}.$$

The relative effect expresses the strength of the association between the illness and the cause. It is measured using the **relative risk** and the **odds ratio**.

DOMAINS AND LIMITATIONS

Strictly speaking, the strength of an association between a particular factor and an illness is not enough to establish a causal relationship between them. We also need to consider:

- The "temporality criterion" (we must be sure that the supposed cause precedes the effect), and;
- Fundamental and experimental research elements that allow us to be sure that the supposed causal factor is not actually a "confusion factor" (which is a factor that is not causal, but is statistically related to the unidentified real causal factor).

Two types of causality correspond to these relative and absolute effects. Relative causality is independent of the clinical or public health impact of the effect; it generally does not allow us to prejudge the clinical or public health impact of the associated effect. It can be strong when the risk of being exposed is very high or when the risk of being unexposed is very low.

Absolute causality expresses the clinical or public health impact of the associated effect,

and therefore enables us to answer the question: if we had suppressed the cause, what level of impact on the population (in terms of cases of illness) would have been avoided? If the patient had stopped smoking, what would the reduction in his risk of developing lung cancer or having a myocardial infarction be? A risk factor associated with a high relative effect, but which concerns only a small number of individuals, will cause fewer illnesses and deaths then a risk factor that is associated with a smaller relative effect but where many more individuals are exposed. It is therefore clear that the importance of a causal relation varies depending on whether we are considering relative or absolute causality.

We should also make an important point here about causal interactions. There can be many causes for the same illness. While all of these causes contribute to the same result, they can also interact. The main consequence of this causal interaction is that we cannot prejudge the effect of simultaneous exposure to causes A and B (denoted $A+B$) based on what we know about the effect of exposure to only A or only B. In contrast to the case for independent causes, we must estimate the joint effect, not restrict ourselves with the isolated analyses of the interacting causes.

EXAMPLES

The relative effect and the absolute effect are subject to different interpretations, as the following example shows.

Suppose we have two populations P_1 and P_2, each comprising 100000 individuals. In population P_1, the risk of contracting a given illness is 0.2% for the exposed and 0.1% for the unexposed. In population P_2, the risk for the exposed is 20% and that for the unexposed is 10%, as shown in the following table:

Population	Risk for the exposed (%)	Risk for the unexposed (%)
	A	B
P_1	0.2	0.1
P_2	20	10

Population	Relative effect $\frac{A}{B}$	Absolute effect (%) $C = A - B$	Avoidable cases $C \times 100000$
P_1	2.0	0.1	100
P_2	2.0	10	10000

The relative effect is the same for populations P_1 and P_2 (the ratio of the risk for the exposed to the risk for the unexposed is 2), but the impact of the same prevention measures would be very different in the two populations, because the absolute effect is ten times more important in P_2: the number of potentially avoidable cases is therefore 100 in population P_1 and 10000 in population P_2.

Now consider the **incidence rate** of lung cancer in a population of individuals who smoke 35 or more cigarettes per day: 3.15/1000/year. While this rate may seem small, it masks the fact that there is a strong relative effect (the risk is 45 times bigger for smokers then for nonsmokers) due to the fact that lung cancer is very rare in nonsmokers (the incidence rate for nonsmokers is 0.07/1000/year).

FURTHER READING

- ▸ Attributable risk
- ▸ Avoidable risk
- ▸ Incidence rate
- ▸ Odds and odds ratio
- ▸ Prevalence rate
- ▸ Relative risk

REFERENCES

Lilienfeld, A.M., Lilienfeld, D.E.: Foundations of Epidemiology, 2nd edn. Clarendon, Oxford (1980)

MacMahon, B., Pugh, T.F.: Epidemiology: Principles and Methods. Little Brown, Boston, MA (1970)

Morabia, A.: Epidemiologie Causale. Editions Médecine et Hygiène, Geneva (1996)

Rothmann, J.K.: Epidemiology. An Introduction. Oxford University Press (2002)

Census

A census is an operation that consists of observing all of the individuals in a **population**. The word census can refer to a population census, in other words a population count, but it can also refer to inquiries (called "exhaustive" inquiries) where we retrieve information about a population by observing all of the individuals in the population. Clearly, such inquiries will be very expensive for very large populations. That is why exhaustive inquiries are rarely performed; **sampling**, which consists of observing of only a portion of the population (called the **sample**), is usually preferred instead.

HISTORY

Censuses originated with the great civilizations of antiquity, when the large areas of empires and complexity associated with governing them required knowledge of the populations involved.

Among the most ancient civilizations, it is known that censuses were performed in Sumeria (between 5000 and 2000 BC), where the people involved reported lists of men and goods on clay tables in cuneiform characters.

Censuses were also completed in Mesopotamia (about 3000 BC), as well as in ancient Egypt from the first dynasty onwards; these censuses were performed due to military and fiscal objectives. Under Amasis II, everybody had to (at the risk of death) declare their profession and source(s) of revenue.

The situation in Israel was more complex: censuses were sometimes compulsary and sometimes forbidden due to the Old Testament. This influence on Christian civilization lasted quite some time; in the Middle Ages, St. Augustin and St. Ambroise were still condemning censuses.

In China, censuses have been performed since at least 200 BC, in different forms and for different purposes. Hecht, J. (1987) reported the main censuses:

1. *Han Dynasty (200 years BC to 200 years AD):* **population** censuses were related to the system of conscription.
2. *Three Kingdoms Period to Five Dynasties (221–959 AD):* related to the system of territorial distribution.
3. *Song and Yuan Dynasties (960–1368 AD):* censuses were performed for fiscal purposes.
4. *Ming Dynasty (1368–1644 AD):* "yellow registers" were established for ten-year censuses. They listed the name, profession, gender and age of every person.
5. *Qing Dynasty (since 1644 AD):* censuses were performed in order to survey population migration.

In Japan during the Middle Age, different types of census have been used. The first census was probably performed under the rule of Emperor Sujin (86 BC).

Finally, in India, a political and economic science treatise entitled "Arthasastra" (profit treaty) gives some information about the use

of an extremely detailed record. This treatise was written by Kautilya, Prime Minister in the reign of Chandragupta Maurya (313–289 BC).

Another very important civilization, the Incans, also used censuses. They used a statistics system called "quipos". Each quipo was both an instrument and a registry of information. Formed from a series of cords, the colors, combinations and knots on the cords had precise meanings. The quipos were passed to specially initiated guards that gathered together all of these statistics.

In Europe, the ancient Greeks and Romans also practiced censuses. Aristotle reported that the Greeks donated a measure of wheat per birth and a measure of barley per death to the goddess Athéna. In Rome, the first census was performed at the behest of King Servius Tullius (578–534 BC) in order to monitor revenues, and consequently raise taxes.

Later, depending on the country, censuses were practiced with different frequencies and on different scales.

In 786, Charlemagne ordered a count of all of his subjects over twelve years old; population counts were also initiated in Italy in the twelfth century; many cities performed censuses of their inhabitants in the fifteenth century, including Nuremberg (in 1449) and Strasbourg (in 1470). In the sixteenth century, France initiated marital status registers.

The seventeenth century saw the development of three different schools of thought: a German school associated with **descriptive statistics**, a French school associated with census ideology and methodology, and an English school that led to modern statistics.

In the history of censuses in Europe, there is a country that occupies a special place. In 1665, Sweden initiated registers of parishioners that were maintained by pastors; in 1668, a decree made it obligatory to be counted in these registers, and instead of being religious, the registers became administrative. 1736 saw the appearance of another decree, stating that the governor of each province had to report any changes in the population of the province to parliament. Swedish population statistics were officially recognized on the 3rd of February 1748 due to creation of the "Tabellverket" (administrative tables). The first summary for all Swedish provinces, realized in 1749, can be considered to be the first proper census in Europe, and the 11th of November 1756 marked the creation of the "Tabellkommissionen," the first official Division of Statistics. Since 1762, these tables of figures have been maintained by the Academy of Sciences.

Initially, the Swedish censuses were organized annually (1749–1751), and then every three years (1754–1772), but since 1775 they have been conducted every five years.

At the end of the eighteenth century an official institute for censuses was created in France (in 1791). In 1787 the principle of census has been registered in the Constitutional Charter of the USA (C. C. USA).

See also **official statistics**.

FURTHER READING

- ▸ **Data collection**
- ▸ **Demography**
- ▸ **Official statistics**
- ▸ **Population**
- ▸ **Sample**
- ▸ **Survey**

REFERENCES

Hecht, J.: L'idée du dénombrement jusqu'à la Révolution. Pour une histoire de la

statistique, tome 1, pp. 21–82 . Economica/INSEE (1978)

Central Limit Theorem

The central limit theorem is a fundamental theorem of statistics. In its simplest form, it prescribes that the sum of a sufficiently large number of independent identically distributed random variables approximately follows a **normal distribution**.

HISTORY

The central limit theorem was first established within the framework of **binomial distribution** by **Moivre, Abraham de** (1733). **Laplace, Pierre Simon de** (1810) formulated the proof of the theorem.

Poisson, Siméon Denis (1824) also worked on this theorem, and **Chebyshev, Pafnutii Lvovich** (1890–1891) gave a rigorous demonstration of it in the middle of the nineteenth century.

At the beginning of the twentieth century, the Russian mathematician Liapounov, Aleksandr Mikhailovich (1901) created the generally recognized form of the central limit theorem by introducing its characteristic functions. **Markov, Andrei Andreevich** (1908) also worked on it and was the first to generalize the theorem to the case of independent variables.

According to Le Cam, L. (1986), the qualifier "central" was given to it by George Polyà (1920) due to the essential role that it plays in probability theory.

MATHEMATICAL ASPECTS

Let $X_1, X_2, \ldots, X_n$ be n independent random variables that are identically distributed (with any distribution) with a **mean** μ and a finite **variance** σ^2.

We define the sum $S_n = X_1 + X_2 + \ldots + X_n$ and we establish the ratio:

$$\frac{S_n - n \cdot \mu}{\sigma \cdot \sqrt{n}},$$

where $n \cdot \mu$ and $\sigma \cdot \sqrt{n}$ represent the mean and the **standard deviation** of S_n, respectively. The central limit theorem establishes that the distribution of this ratio tends to the standard normal distribution when n tends to infinity. This means that:

$$P\left(\frac{S_n - n \cdot \mu}{\sigma\sqrt{n}} \leq x\right) \underset{n \to +\infty}{\longrightarrow} \Phi(x)$$

where $\Phi(x)$ is the **distribution function** of the standard **normal distribution**, expressed by:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x\,,$$
$$-\infty < x < \infty\,.$$

DOMAINS AND LIMITATIONS

The central limit theorem provides a simple method of approximately calculating probabilities related to the sums of random variables.

Besides its interest in relation to the **sampling** theorem, where the sums and the means play an important role, the central limit theorem is used to approximate **normal distributions** derived from summing identical distributions. We can for example, with the help of the central limit theorem, use the **normal distribution** to approximate the **binomial distribution**, the **Poisson distribution**, the **gamma distribution**, the **chi-square distribution**, the **Student distribution**, the **hypergeometric distribution**, the **Fisher distribution** and the **lognormal distribution**.

EXAMPLES

In a large batch of electrical items, the **probability** of choosing a defective item equals $p = \frac{1}{8}$. What is the probability that 4 defective items are chosen when 25 items are selected?

Let X the **dichotomous variable** be the result of a trial:

$$X = \begin{cases} 1 & \text{if the selected item is defective;} \\ 0 & \text{if it is not.} \end{cases}$$

The **random variable** X follows a **Bernoulli distribution** with **parameter** p. Consequently, the sum $S_n = X_1 + X_2 + \ldots + X_n$ follows a **binomial distribution** with a **mean**, np and a **variance** $np(1-p)$, which, following the central limit theorem, can be approximated by a **normal distribution** with a mean $\mu = np$ and a variance $\sigma^2 = np(1-p)$.

We evaluate these values:

$$\mu = n \cdot p = 25 \cdot \tfrac{1}{8} = 3.125$$

$$\sigma^2 = n \cdot p(1-p) = 25 \cdot \tfrac{1}{8} \cdot \left(1 - \tfrac{1}{8}\right) = 2.734\,.$$

We then calculate $P(S_n > 4)$ in two different ways:

1. With the binomial distribution:
   From the **binomial table**, the probability of $P(S_n \leq 4) = 0.8047$. The probability of $P\,(S_n > 4)$ is then:

   $$P(S_n > 4) = 1 - P(S_n \leq 4) = 0.1953\,,$$

2. With the normal approximation (obtained from the central limit theorem):
   In order to account for the discrete character of the random variable S_n, we must make a continuity correction; that is, we calculate the probability that S_n is greater then $4 + \frac{1}{2} = 4.5$

   We have:

   $$z = \frac{S_n - n \cdot p}{\sqrt{n \cdot p\,(1-p)}} = \frac{4.5 - 3.125}{1.654} = 0.832\,.$$

   From the **normal table**, we obtain the probability:

   $$P(Z > z) = P(Z > 0.832) = 1 - P(Z \leq 0.832) = 1 - 0.7967 = 0.2033\,.$$

FURTHER READING

- **Binomial distribution**
- **Chi-square distribution**
- **Convergence**
- **Convergence theorem**
- **Fisher distribution**
- **Gamma distribution**
- **Hypergeometric distribution**
- **Law of large numbers**
- **Lognormal distribution**
- **Normal distribution**
- **Poisson distribution**
- **Probability**
- **Probability distribution**
- **Student distribution**

REFERENCE

Laplace, P.S. de: Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. Mémoires de l'Académie Royale des Sciences de Paris, 10. Reproduced in: Œuvres de Laplace **12**, 301–347 (1810)

Le Cam, L.: The Central Limit Theorem around 1935. Stat. Sci. **1**, 78–96 (1986)

Liapounov, A.M.: Sur une proposition de la théorie des probabilités. Bulletin de l'Academie Imperiale des Sciences de St.-Petersbourg **8**, 1–24 (1900)

Markov, A.A.: Extension des théorèmes limites du calcul des probabilités aux sommes des quantités liées en chaîne. Mem. Acad. Sci. St. Petersburg **8**, 365–397 (1908)

Moivre, A. de: Approximatio ad summam terminorum binomii $(a + b)^n$, in seriem expansi. Supplementum II to Miscellanae Analytica, pp. 1–7 (1733). Photographically reprinted in a rare pamphlet on Moivre and some of his discoveries. Published by Archibald, R.C. Isis **8**, 671–683 (1926)

Poisson, S.D.: Sur la probabilité des résultats moyens des observations. Connaissance des temps pour l'an 1827, pp. 273–302 (1824)

Polyà, G.: Ueber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem. Mathematische Zeitschrift **8**, 171–181 (1920)

Tchebychev, P.L. (1890–1891). Sur deux théorèmes relatifs aux probabilités. Acta Math. **14**, 305–315

Chebyshev, Pafnutii Lvovich

Chhebyshev, Pafnutii Lvovich (1821–1894) began studying at Moscow University in 1837, where he was influenced by Zernov, Nikolai Efimovich (the first Russian to get a doctorate in mathematical sciences) and Brashman, Nikolai Dmitrievich. After gaining his degree he could not find any teaching work in Moscow, so he went to St. Petersbourg where he organized conferences on algebra and probability theory. In 1859, he took the probability course given by Buniakovsky, Viktor Yakovlevich at St. Petersbourg University.

His name lives on through the **Chebyshev inequality** (also known as the Bienaymé–Chebyshev inequality), which he proved. This was published in French just after **Bienaymé, Irénée-Jules** had an article published on the same topic in the *Journal de Mathématiques Pures et Appliquées* (also called the Journal of Liouville).

He initiated rigorous work into establishing a general version of the **central limit theorem** and is considered to be the founder of the mathematical school of St. Petersbourg.

Some principal works and articles of Chebyshev, Pafnutii Lvovich:

1845 An Essay on Elementary Analysis of the Theory of Probabilities (thesis) Crelle's Journal.

1867 Preuve de l'inégalité de Tchebychev. J. Math. Pure. Appl., 12, 177–184.

FURTHER READING

▸ **Central limit theorem**

REFERENCES

Heyde, C.E., Seneta, E.: I.J. Bienaymé. Statistical Theory Anticipated. Springer, Berlin Heidelberg New York (1977)

Chebyshev's Inequality

See **law of large numbers**.

Chi-Square Distance

Consider a **frequency table** with n rows and p columns, it is possible to calculate row profiles and column profiles. Let us then plot

the n or p points from each profile. We can define the **distances** between these points. The Euclidean distance between the components of the profiles, on which a weighting is defined (each term has a weight that is the inverse of its **frequency**), is called the chi-square distance. The name of the distance is derived from the fact that the mathematical expression defining the distance is identical to that encountered in the elaboration of the **chi square goodness of fit test**.

MATHEMATICAL ASPECTS

Let (f_{ij}), be the **frequency** of the ith row and jth column in a frequency table with n rows an p columns. The chi-square distance between two rows i and i' is given by the formula:

$$d(i,i') = \sqrt{\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2 \cdot \frac{1}{f_{.j}}},$$

where

- $f_{i.}$ is the sum of the components of the ith row;
- $f_{.j}$ is the sum of the components of the jth column;
- $[\frac{f_{ij}}{f_{i.}}]$ is the ith row profile for $j = 1, 2, \ldots, p$.

Likewise, the distance between two columns j and j' is given by:

$$d(j,j') = \sqrt{\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2 \cdot \frac{1}{f_{i.}}},$$

where $[\frac{f_{ij}}{f_{.j}}]$ is the jth column profile for $j = 1, \ldots, n$.

DOMAINS AND LIMITATIONS

The chi-square distance incorporates a weight that is inversely proportional to the total of each row (or column), which increases the importance of small deviations in the rows (or columns) which have a small sum with respect to those with more important sum package.

The chi-square distance has the property of distributional equivalence, meaning that it ensures that the **distances** between rows and columns are invariant when two columns (or two rows) with identical profiles are aggregated.

EXAMPLES

Consider a **contingency table** charting how satisfied employees working for three different businesses are. Let us establish a **distance table** using the chi-square distance.

Values for the studied **variable** X can fall into one of three **categories**:

- X_1: high satisfaction;
- X_2: medium satisfaction;
- X_3: low satisfaction.

The **observations** collected from **samples** of individuals from the three businesses are given below:

	Business 1	Business 2	Business 3	Total
X_1	20	55	30	105
X_2	18	40	15	73
X_3	12	5	5	22
Total	50	100	50	200

The relative **frequency table** is obtained by dividing all of the elements of the table by 200, the total number of observations:

	Business 1	Business 2	Business 3	Total
X_1	0.1	0.275	0.15	0.525
X_2	0.09	0.2	0.075	0.365
X_3	0.06	0.025	0.025	0.11
Total	0.25	0.5	0.25	1

We can calculate the difference in employee satisfaction between the the 3 enterprises. The column profile matrix is given below:

	Business 1	Business 2	Business 3	Total
X_1	0.4	0.55	0.6	1.55
X_2	0.36	0.4	0.3	1.06
X_3	0.24	0.05	0.1	0.39
Total	1	1	1	3

This allows us to calculate the **distances** between the different columns:

$$\begin{aligned} d^2(1,2) &= \frac{1}{0.525}\cdot(0.4-0.55)^2 \\ &\quad + \frac{1}{0.365}\cdot(0.36-0.4)^2 \\ &\quad + \frac{1}{0.11}\cdot(0.24-0.05)^2 \\ &= 0.375423 \\ d(1,2) &= 0.613 \end{aligned}$$

We can calculate $d(1, 3)$ and $d(2, 3)$ in a similar way. The distances obtained are summarized in the following **distance table**:

	Business 1	Business 2	Business 3
Business 1	0	0.613	0.514
Business 2	0.613	0	0.234
Business 3	0.514	0.234	0

We can also calculate the **distances** between the rows, in other words the difference in employee satisfaction; to do this we need the line profile table:

	Business 1	Business 2	Business 3	Total
X_1	0.19	0.524	0.286	1
X_2	0.246	0.548	0.206	1
X_3	0.546	0.227	0.227	1
Total	0.982	1.299	0.719	3

This allows us to calculate the **distances** between the different rows:

$$\begin{aligned} d^2(1,2) &= \frac{1}{0.25}\cdot(0.19-0.246)^2 \\ &\quad + \frac{1}{0.5}\cdot(0.524-0.548)^2 \\ &\quad + \frac{1}{0.25}\cdot(0.286-0.206)^2 \\ &= 0.039296 \\ d(1,2) &= 0.198 \end{aligned}$$

We can calculate $d(1, 3)$ and $d(2, 3)$ in a similar way. The differences between the degrees of employee satisfaction are finally summarized in the following **distance table**:

	X_1	X_2	X_3
X_1	0	0.198	0.835
X_2	0.198	0	0.754
X_3	0.835	0.754	0

FURTHER READING

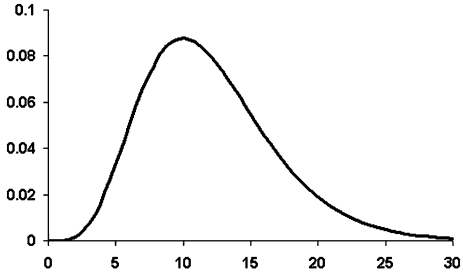
- Contingency table
- Distance
- Distance table
- Frequency

Chi-square Distribution

A **random variable** X follows a chi-square distribution with n **degrees of freedom** if its **density function** is:

$$f(x) = \frac{x^{\frac{n}{2}-1}\exp\left(-\frac{x}{2}\right)}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}, \quad x \geq 0\,,$$

where Γ is the gamma function (see **Gamma distribution**).



χ^2 distribution, $\nu = 12$

The chi-square distribution is a **continuous probability distribution**.

HISTORY

According to Sheynin (1977), the chi-square distribution was discovered by Ernst Karl Abbe in 1863. Maxwell obtained it for three degrees of freedom a few years before (1860), and Boltzman discovered the general case in 1881.

However, according to Lancaster (1966), Bienaymé obtained the chi-square distribution in 1838 as the limit of the discrete **random variable**

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

if $(N_1, N_2, \ldots, N_k)$ follow a joint multinomial distribution of **parameters** $n, p_1, p_2, \ldots, p_k$.

Ellis demonstrated in 1844 that the sum of k **random variables** distributed according to a chi-square distribution with two **degrees of freedom** follows a chi-square distribution with $2k$ degrees of freedom. The general result was demonstrated in 1852 by Bienaymé.

The works of **Pearson, Karl** are very important in this field. In 1900 he used the chi-square distribution to approximate the chi-square **statistic** used in different tests based on **contingency tables**.

MATHEMATICAL ASPECTS

The chi-square distribution appears in the theory of **random variables** distributed according to a **normal distribution**. In this, it is the distribution of the sum of squares of normal, centered and reduced random variables (with a **mean** equal to 0 and a **variance** equal to 1).

Consider $Z_1, Z_2, \ldots, Z_n$, n independent, standard normal **random variables**. Their sum of squares:

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

is a random variable distributed according to a chi-square distribution with n **degrees of freedom**.

The **expected value** of the chi-square distribution is given by:

$$E[X] = n\,.$$

The **variance** is equal to:

$$\mathrm{Var}(X) = 2n\,.$$

The chi-square distribution is related to other **continuous probability distributions**:

- The chi-square distribution is a particular case of the **gamma distribution**.
- If two **random variables** X_1 and X_2 follow a chi-square distribution with, respectively, n_1 and n_2 **degrees of freedom**, then the random variable

  $$Y = \frac{X_1/n_1}{X_2/n_2}$$

  follows a **Fisher distribution** with n_1 and n_2 degrees of freedom.
- When the number of **degrees of freedom** n tends towards infinity, the chi-square distribution tends (relatively slowly) towards a **normal distribution**.

DOMAINS AND LIMITATIONS

The chi-square distribution is used in many approaches to **hypothesis testing**, the most important being the **goodness of fit test** which involves comparing the observed **frequencies** and the hypothetical frequencies of specific classes.

It is also used for comparisons between the observed **variance** and the hypothetical variance of normally distributed **samples**, and to test the **independence** of two **variables**.

FURTHER READING

- ▸ Chi-square goodness of fit test
- ▸ Chi-square table
- ▸ Chi-square test
- ▸ Continuous probability distribution
- ▸ Fisher distribution
- ▸ Gamma distribution
- ▸ Normal distribution

REFERENCES

Lancaster, H.O.: Forerunners of the Pearson chi-square. Aust. J. Stat. **8**, 117–126 (1966)

Pearson, K.: On the criterion, that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, pp. 339–357. First published in 1900 in Philos. Mag. (5th Ser) **50**, 157–175 (1948)

Sheynin, O.B.: On the history of some statistical laws of distribution. In: Kendall, M., Plackett, R.L. (eds.) Studies in the History of Statistics and Probability, vol. II. Griffin, London (1977)

Chi-square Goodness of Fit Test

The chi-square goodness of fit test is, along with the **Kolmogorov–Smirnov test**, one of the most commonly used goodness of fit tests.

This test aims to determine whether it is possible to approximate an observed distribution by a particular **probability distribution** (**normal distribution**, **Poisson distribution**, etc).

HISTORY

The chi-square goodness of fit test is the oldest and most well-known of the goodness of fit tests. It was first presented in 1900 by **Pearson, Karl**.

MATHEMATICAL ASPECTS

Let $X_1, \ldots, X_n$ be a **sample** of n observations. The steps used to perform the chi-square goodness of fit test are then as follows:

1. State the hypothesis. The **null hypothesis** will take the following form:

$$H_0: \quad F = F_0\,,$$

where F_0 is the presumed **distribution function** of the distribution.

2. Distribute the observations in k disjoint classes:

$$[a_{i-1}, a_i]\,.$$

We denote the number of observations contained in the ith class, $i = 1, \ldots, k$, by n_i.

3. Calculate the theoretical probabilities for every class on the base of the presumed distribution function F_0:

$$p_i = F_0(a_i) - F_0(a_{i-1})\,, \quad i = 1, \ldots, k\,.$$

4. Obtain the expected frequencies for every class

$$e_i = n \cdot p_i\,, \quad i = 1, \ldots, k\,,$$

where n is the size of the **sample**.

5. Calculate the χ^2 (chi-square) statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}\,.$$

If H_0 is true, the χ^2 statistic follows a **chi-square distribution** with v degrees of freedom, where:

$$v = \left(k - 1 - \begin{array}{c}\text{number of estimated}\\ \text{parameters}\end{array}\right).$$

For example, when testing the goodness of fit to a **normal distribution**, the number of degrees of freedom equals:

- $k - 1$ if the **mean** μ and the **standard deviation** σ of the **population** are known;
- $k-2$ if one out of μ or σ is unknown and will be estimated in order to proceed with the test;
- $k - 3$ if both parameters μ and σ are unknown and both are estimated from the corresponding values of the sample.

6. Reject H_0 if the deviation between the observed and estimated frequencies is big; that is:

$$\text{if } \chi^2 > \chi^2_{v,\alpha}\,,$$

where $\chi^2_{v,\alpha}$ is the **value** given in the **chi-square table** for a particular **significance level** α.

DOMAINS AND LIMITATIONS

To apply the chi-square goodness of fit test, it is important that n is big enough and that the estimated frequencies, e_i, are not too small. We normally state that the estimated frequencies must be greater then 5, except for extreme classes, where they can be smaller then 5 but greater then 1. If this constraint is not satisfied, we must regroup the classes in order to satisfy this rule.

EXAMPLES

Goodness of Fit to the Binomial Distribution

We throw a coin four times and count the number of times that "heads" appears. This **experiment** is performed 160 times. The observed frequencies are as follows:

Number of "heads"	Number of experiments
x_i	(n_i)
0	17
1	52
2	54
3	31
4	6
Total	160

1. If the **experiment** was performed correctly and the coin is not forged, the distribution of the number of "heads" obtained should follow the **binomial distribution**. We then state a **null hypothesis** that the observed distribution can be approximated by the binomial distribution, and we will proceed with a **goodness of fit test** in order to determine whether this **hypothesis** can be accepted or not.
2. In this example, the different number of "heads" that can be obtained per experiment (0, 1, 2, 3 and 4) are each considered to be a class.

3. The **random variable** X (number of "heads" obtained after four throws of a coin) follows a binomial distribution if

$$P(X = x) = C_n^x \cdot p^x \cdot q^{n-x},$$

where:

- n is the number of independent trials = 4;
- p is the **probability** of a success ("heads") = 0.5;
- q is the probability of a failure ("tails") = 0.5;
- C_n^x is the number of combinations of x objects from n.

We then have the following theoretical probabilities for four throws:

$$P(X = 0) = \frac{1}{16}$$
$$P(X = 1) = \frac{4}{16}$$
$$P(X = 2) = \frac{6}{16}$$
$$P(X = 3) = \frac{4}{16}$$
$$P(X = 4) = \frac{1}{16}$$

4. After the experiment has been performed 160 times, the expected number of heads for each possible **value** of X is given by:

$$e_i = 160 \cdot P(X = x_i).$$

We obtain the following table:

Number of "heads" x_i	Observed frequency (n_i)	Expected frequency (e_i)
0	17	10
1	52	40
2	54	60
3	31	40
4	6	10
Total	160	160

5. The χ^2 (chi-square) statistic is then:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i},$$

where k is the number of possible values of X.

$$\chi^2 = \frac{(17-10)^2}{10} + \ldots + \frac{(6-10)^2}{10} = 12.725.$$

6. Choosing a **significance level** α of 5%, we find that the value of $\chi^2_{\nu,\alpha}$ for $k - 1 = 4$ degrees of freedom is:

$$\chi^2_{4,0.05} = 9.488.$$

Since the calculated value of χ^2 is greater than the value obtained from the table, we reject the **null hypothesis** and conclude that the binomial distribution does not give a good approximation to our observed distribution. We can then conclude that the coins were probably forged, or that they were not correctly thrown.

Goodness of Fit to the Normal Distribution

The diameters of cables produced by a factory were studied.

A **frequency table** of the observed distribution of diameters is given below:

Cable diameter (in mm)	Observed frequency n_i
19.70–19.80	5
19.80–19.90	12
19.90–20.00	35
20.00–20.10	42
20.10–20.20	28
20.20–20.30	14
20.30–20.40	4
Total	140

1. We perform a **goodness of fit test** for a **normal distribution**. The null hypothesis is therefore that the observed distribution can be approximated by a normal distribution.
2. The previous table shows the observed diameters divided up into classes.
3. If the random variable X (cable diameter) follows the normal distribution, the random variable

$$Z = \frac{X - \mu}{\sigma}$$

follows a standard normal distribution. The **mean** μ and the **standard deviation** σ of the **population** are unknown and are estimated using the mean $\bar{x}$ and the standard deviation S of the **sample**:

$$\bar{x} = \frac{\sum_{i=1}^{7} \delta_i \cdot n_i}{n} = \frac{2806.40}{140} = 20.05\,,$$

$$S = \sqrt{\frac{\sum_{i=1}^{7} n_i \cdot (\delta_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2.477}{139}} = 0.134$$

where the δ_i are the centers of the classes (in this example, the mean diameter of a class; for $i = 1$: $\delta_1 = 19.75$) and n is the total number of observations.
We can then calculate the theoretical probabilities associated with each class. The detailed calculations for the first two classes are:

$$\begin{aligned} p_1 &= P(X \le 19.8) \\ &= P(Z \le \frac{19.8 - \bar{x}}{S}) \\ &= P(Z \le -1.835) \\ &= 1 - P(Z \le 1.835) \end{aligned}$$

$$\begin{aligned} p_2 &= P(19.8 \le X \le 19.9) \\ &= P\left(\frac{19.8 - \bar{x}}{S} \le Z \le \frac{19.9 - \bar{x}}{S}\right) \\ &= P(-1.835 \le Z \le -1.089) \\ &= P(Z \le 1.835) - P(Z \le 1.089) \end{aligned}$$

These probabilities can be found by consulting the **normal table**. We get:

$$\begin{aligned} p_1 &= P(X \le 19.8) = 0.03325 \\ p_2 &= P(19.8 \le X \le 19.9) = 0.10476 \\ p_3 &= P(19.9 \le X \le 20.0) = 0.22767 \\ p_4 &= P(20.0 \le X \le 20.1) = 0.29083 \\ p_5 &= P(20.1 \le X \le 20.2) = 0.21825 \\ p_6 &= P(20.2 \le X \le 20.3) = 0.09622 \\ p_7 &= P(X > 20.3) = 0.02902 \end{aligned}$$

4. The expected frequencies for the classes are then given by:

$$e_i = n \cdot p_i\,,$$

which yields the following table:

Cable diameter (in mm)	Observed frequency n_i	Expected frequency e_i
19.70–19.80	5	4.655
19.80–19.90	12	14.666
19.90–20.00	35	31.874
20.00–20.10	42	40.716
20.10–20.20	28	30.555
20.20–20.30	14	13.471
20.30–20.40	4	4.063
Total	140	140

5. The χ^2 (chi-square) statistic is then:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}\,,$$

where $k = 7$ is the number of classes.

$$\chi^2 = \frac{(5-4.655)^2}{4.655} + \frac{(12-14.666)^2}{14.666} + \ldots + \frac{(4-4.063)^2}{4.063} = 1.0927\,.$$

6. Choosing a significance level $\alpha = 5\%$, we find that the value of $\chi^2_{\nu,\alpha}$ with $k-3 = 7-3 = 4$ degrees of freedom in the chi-square table is:

$$\chi^2_{4,0.05} = 9.49\,.$$

Since the value calculated from χ^2 is smaller then the value obtained from the chi-square table, we do not reject the null hypothesis and we conclude that the difference between the observed distribution and the normal distribution is not significant at a **significance level** of 5%.

FURTHER READING

- ▸ **Chi-square distribution**
- ▸ **Chi-square table**
- ▸ **Goodness of fit test**
- ▸ **Hypothesis testing**
- ▸ **Kolmogorov–Smirnov test**

REFERENCE

Pearson, K.: On the criterion, that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, pp. 339–357. First published in 1900 in Philos. Mag. (5th Ser) **50**, 157–175 (1948)

Chi-Square Table

The chi-square table gives the values obtained from the **distribution function** of a **random variable** that follows a **chi-square distribution**.

HISTORY

One of the first chi-square tables was published in 1902, by Elderton. It contains **distribution function** values that are given to six decimal places.

In 1922, **Pearson, Karl** reported a table of values for the incomplete gamma function, down to seven decimals.

MATHEMATICAL ASPECTS

Let X be a **random variable** that follows a **chi-square distribution** with ν degrees of freedom. The **density function** of the random variable X is given by:

$$f(t) = \frac{t^{\frac{\nu}{2}-1}\exp\left(-\frac{t}{2}\right)}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\,, \quad t \geq 0\,,$$

where Γ represents the gamma function (see **gamma distribution**).

The **distribution function** of the random variable X is defined by:

$$F(x) = P(X \leq x) = \int_0^x f(t)\,\mathrm{d}t\,.$$

The chi-square table gives the values of the distribution function $F(x)$ for different values of ν.

We often use the chi-square table in the opposite way, to find the value of x that corresponds to a given **probability**.

We generally denote as $\chi^2_{\nu,\alpha}$ the value of the random variable X for which

$$P(X \leq \chi^2_{\nu,\alpha}) = 1-\alpha\,.$$

Note: the notation χ^2 is read "chi-square".

EXAMPLES

See Appendix F.

The chi-square table allows us, for a given number of degrees of freedom v, to determine:

1. The **value** of the **distribution function** $F(x)$, given x.
2. The value of $\chi^2_{v,\alpha}$, given the **probability** $P(X \leq \chi^2_{v,\alpha})$.

FURTHER READING

- ▸ **Chi-square distribution**
- ▸ **Chi-square goodness of fit test**
- ▸ **Chi-square test**
- ▸ **Chi-square test of independence**
- ▸ **Statistical table**

REFERENCES

Elderton, W.P.: Tables for testing the goodness of fit of theory to observation. Biometrika **1**, 155–163 (1902)

Pearson, K.: Tables of the Incomplete Γ-function. H.M. Stationery Office (Cambridge University Press, Cambridge since 1934), London (1922)

Chi-Square Test

There are a number of chi-square tests, all of which involve comparing the test results to the values from the **chi-square distribution**. The most well-known of these tests are introduced below:

- The **chi-square test of independence** is used to determine whether two qualitative categorical variables associated with a **sample** are independent.
- The **chi-square goodness of fit test** is used to determine whether the distribution observed for a sample can be approximated by a theoretical distribution. We might want to know, for example, whether the distribution observed for the sample corresponds to a particular **probability distribution** (**normal distribution**, **Poisson distribution**, etc).
- The chi-square test for an unknown **variance** is used when we want to test whether this variance takes a particular constant **value**.
- The chi-square test is used to test for homogeneity of the variances calculated for many samples drawn from a normally distributed **population**.

HISTORY

In 1937, Bartlett, M.S. proposed a method of testing the homogeneity of the **variance** for many samples drawn from a normally distributed **population**.

See also **chi-square test of independence** and **chi-square goodness of fit test**.

MATHEMATICAL ASPECTS

The mathematical aspects of the **chi-square test of independence** and those of the **chi-square goodness of fit test** are dealt with in their corresponding entries.

The chi-square test used to check whether an unknown **variance** takes a particular constant **value** is the following:

Let $(x_1, \ldots, x_n)$ be a random **sample** coming from a normally distributed **population** of unknown **mean** μ and of unknown **variance** σ^2.

We have good reason to believe that the variance of the population equals a presumed value σ_0^2. The hypotheses for each case are described below.

A: Two-sided case:

$$H_0: \quad \sigma^2 = \sigma_0^2$$
$$H_1: \quad \sigma^2 \neq \sigma_0^2$$

B: One-sided test:

$$H_0: \ \sigma^2 \leq \sigma_0^2$$
$$H_1: \ \sigma^2 > \sigma_0^2$$

C: One-sided test:

$$H_0: \ \sigma^2 \geq \sigma_0^2$$
$$H_1: \ \sigma^2 < \sigma_0^2$$

We then determine the **statistic** of the given chi-square test using:

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}.$$

This statistic is, under H_0, chi-square distributed with $n-1$ degrees of freedom. In other words, we look for the value of $\chi^2_{n-1,\alpha}$ in the **chi-square table**, and we then compare that value to the calculated value χ^2. The decision rules depend on the case, and are as follows.

Case A

If $\chi^2 \geq \chi^2_{n-1,\alpha_1}$ or if $\chi^2 \leq \chi^2_{n-1,1-\alpha_2}$ we reject the **null hypothesis** H_0 for the **alternative hypothesis** H_1, where we have split the **significance level** α into α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha$. Otherwise we do not reject the null hypothesis H_0.

Case B

If $\chi^2 < \chi^2_{n-1,\alpha}$ we do not reject the null hypothesis H_0. If $\chi^2 \geq \chi^2_{n-1,\alpha}$ we reject the null hypothesis H_0 for the alternative hypothesis H_1.

Case C

If $\chi^2 > \chi^2_{n-1,1-\alpha}$ we do not reject the null hypothesis H_0. If $\chi^2 \leq \chi^2_{n-1,1-\alpha}$ we reject the null hypothesis H_0 for the alternative hypothesis H_1.

Other chi-square tests are proposed in the work of Ostle, B. (1963).

DOMAINS AND LIMITATIONS

χ^2 (chi-square) **statistic** must be calculated using absolute frequencies and not relative ones.

Note that the chi-square test can be unreliable for small samples, especially when some of the estimated frequencies are small (< 5). This issue can often be resolved by grouping categories together, if such grouping is sensible.

EXAMPLES

Consider a batch of items produced by a machine. They can be divided up into classes depending on their diameters (in mm), as in the following table:

Diameter (mm) x_i	Number of items n_i
59.5	2
59.6	6
59.7	7
59.8	15
59.9	17
60.0	25
60.1	15
60.2	7
60.3	3
60.4	1
60.5	2
Total	$N = 100$

We have a random **sample** drawn from a normally distributed **population** where the **mean** and **variance** are not known. The vendor of these items would like the variance σ^2 to be smaller than or equal to 0.05. We test

the following hypotheses:

null hypothesis H_0: $\sigma^2 \leq 0.05$

alternative hypothesis H_1: $\sigma^2 > 0.05$.

In this case we use the one-tailed hypothesis test.

We start by calculating the mean for the sample:

$$\bar{x} = \frac{\sum_{i=1}^{11} n_i \cdot x_i}{N} = \frac{5995}{100} = 59.95\,.$$

We can then calculate the χ^2 **statistics**:

$$\begin{aligned}\chi^2 &= \frac{\sum_{i=1}^{11} n_i \cdot (x_i - \bar{x})^2}{\sigma_0^2} \\ &= \frac{2 \cdot (-0.45)^2 + \ldots + 2 \cdot (0.55)^2}{0.05} \\ &= \frac{3.97}{0.05} = 79.4\,.\end{aligned}$$

Using a **significance level** of $\alpha = 5\%$, we then find the **value** of $\chi^2_{99,0.05}$ $(= 123.2)$ in the **chi-square table**.

As $\chi^2 = 79.4 < \chi^2_{99,0.05}$, we do not reject the null hypothesis, which means that the vendor should be happy to sell these items since they are not significantly different in diameter.

FURTHER READING

- ▸ **Chi-square distribution**
- ▸ **Chi-square goodness of fit test**
- ▸ **Chi-square table**
- ▸ **Chi-square test of independence**

REFERENCES

Bartlett, M.S.: Some examples of statistical methods of research in agriculture and applied biology. J. Roy. Stat. Soc. (Suppl.) **4**, 137–183 (1937)

Ostle, B.: Statistics in Research: Basic Concepts and Techniques for Research Workers. Iowa State College Press, Ames, IA (1954)

C

Chi-square Test of Independence

The chi-square test of independence aims to determine whether two variables associated with a **sample** are independent or not. The variables studied are categorical qualitative variables.

The chi-square independence test is performed using a **contingency table**.

HISTORY

The first contingency tables were used only for enumeration. However, encouraged by the work of **Quetelet, Adolphe** (1849), statisticians began to take an interest in the associations between the variables used in the tables. For example, **Pearson, Karl** (1900) performed fundamental work on contingency tables.

Yule, George Udny (1900) proposed a somewhat different approach to the study of contingency tables to Pearson's, which lead to a disagreement between them. Pearson also argued with **Fisher, Ronald Aylmer** about the number of degrees of freedom to use in the chi-square test of independence. Everyone used different numbers until Fisher, R.A. (1922) was eventually proved to be correct.

MATHEMATICAL ASPECTS

Consider two qualitative categorical variables X and Y. We have a **sample** containing n observations of these variables.

These observations can be presented in a **contingency table**.

We denote the observed **frequency** of the **category** i of the variable X and the category j of the variable Y as n_{ij}.

		Categories of variable Y			
		Y_1	...	Y_c	Total
Categories of variable X	X_1	n_{11}	...	n_{1c}	$n_{1.}$
	...	...	...	...	...
	X_r	n_{r1}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	...	$n_{.c}$	$n_{..}$

The hypotheses to be tested are:

Null hyp. H_0: the two variables are independent,

Alternative hyp. H_1: the two variables are not independent.

Steps Involved in the Test

1. Compute the expected frequencies, denoted by e_{ij}, for each case in the **contingency table** under the **independence hypothesis**:

$$e_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}},$$

$$n_{i.} = \sum_{k=1}^{c} n_{ik} \text{ and } n_{.j} = \sum_{k=1}^{r} n_{kj},$$

where c represents the number of columns (or number of categories of variable X in the contingency table) and r the number of rows (or the number of categories of variable Y).

2. Calculate the **value** of the χ^2 (chi-square) statistic, which is really a measure of the deviation of the observed frequencies n_{ij} from the expected frequencies e_{ij}:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

3. Choose the **significance level** α to be used in the test and compare the calculated **value** of χ^2 with the value obtained from the **chi-square table**, $\chi^2_{v,\alpha}$. The number of degrees of freedom correspond to the number of cases in the table that can take arbitrary values; the values taken by the other cases are imposed on them by the row and column totals. So, the number of degrees of freedom is given by:

$$v = (r-1)(c-1).$$

4. If the calculated χ^2 is smaller then the $\chi^2_{v,\alpha}$ from the table, we do not reject the null hypothesis. The two variables can be considered to be independent.
   However, if the calculated χ^2 is greater then the $\chi^2_{v,\alpha}$ from the table, we reject the null hypothesis for the alternative hypothesis. We can then conclude that the two variables are not independent.

DOMAINS AND LIMITATIONS

Certain conditions must be fulfilled in order to be able to apply the chi-square test of independence:

1. The **sample**, which contains n observations, must be a random sample;
2. Each individual observation can only appear in one **category** for each **variable**. In other words, each individual observation can only appear in one line and one column of the **contingency table**.

Note that the chi-square test of independence is not very reliable for small samples, especially when the estimated frequencies are small (< 5). To avoid this issue we can group categories together, but only when this groups obtained are sensible.

EXAMPLES

We want to determine whether the proportion of smokers is independent of gender.
The two variables to be studied are categorical and qualitative and contain two categories each:

- Variable "gender:" M or F;
- Variable "smoking status:" "smokes" or "does not smoke."

The hypotheses are then:

H_0: chance of being a smoker is independent of gender

H_1: chance of being a smoker is not independent of gender.

The **contingency table** obtained from a **sample** of 100 individuals ($n = 100$) is shown below:

		Smoking status		
		"smokes"	"does not smoke"	Total
Gender	M	21	44	65
	F	10	25	35
	Total	31	69	100

We now denote the observed frequencies as n_{ij} $(i = 1, 2, j = 1, 2)$.
We then estimate all of the frequencies in the table based on the **hypothesis** that the two variables are independent of each other. We denote these estimated frequencies by e_{ij}:

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

We therefore obtain:

$$e_{11} = \frac{65 \cdot 31}{100} = 20.15$$
$$e_{12} = \frac{65 \cdot 69}{100} = 44.85$$
$$e_{21} = \frac{35 \cdot 31}{100} = 10.85$$
$$e_{12} = \frac{35 \cdot 69}{100} = 24.15 .$$

The estimated **frequency table** is given below:

		Smoking status		
		"smokes"	"does not smoke"	Total
Gender	M	20.15	44.85	65
	F	10.85	24.15	35
	Total	31	69	100

If the **null hypothesis** H_0 is true, the statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

is chi-square-distributed with $(r-1)(c-1) = (2-1)(2-1) = 1$ degree of freedom and

$$\chi^2 = 0.036 + 0.016 + 0.066 + 0.030$$
$$= 0.148 .$$

If a **significance level** of 5% is selected, the **value** of $\chi^2_{1,0.05}$ is 3.84, from the **chi-square table**.
Since the calculated value of χ^2 is smaller then the value found in the chi-square table, we do not reject the null hypothesis and we conclude that the two variables studied are independent.

FURTHER READING

- ▸ Chi-square distribution
- ▸ Chi-square table
- ▸ Contingency table
- ▸ Test of independence

REFERENCE

Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of P. J. Roy. Stat. Soc. Ser. A **85**, 87–94 (1922)

Pearson, K.: On the criterion, that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, pp. 339–357. First published in 1900 in Philos. Mag. (5th Ser) **50**, 157–175 (1948)

Quetelet, A.: Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences. (French translation by Downs, Olinthus Gregory). Charles & Edwin Layton, London (1849)

Yule, G.U.: On the association of attributes in statistics: with illustration from the material of the childhood society. Philos. Trans. Roy. Soc. Lond. Ser. A **194**, 257–319 (1900)

Classification

Classification is the grouping together of similar objects. If each object is characterized by p **variables**, classification can be performed according to rational criteria. Depending on the criteria used, an object could potentially belong to several classes.

HISTORY

Classifying the residents of a locality or a country according to their sex and other physical characteristics is an activity that dates back to ancient times. The Hindus, the ancient Greeks and the Romans all developed multiple typologies for human beings. The oldest comes from Galen (129–199 A.D.).

Later on, the concept of classification spread to the fields of biology and zoology; the works of Linné (1707–1778) should be mentioned in this regard.

The first ideas regarding actual methods of **cluster analysis** are attributed to Adanson (eighteenth century). Zubin (1938), Tryon (1939) and Thorndike (1953) also attempted to develop some methods, but the true development of classification methods coincides with the advent of the computer.

MATHEMATICAL ASPECTS

Classification methods can be divided into two large **categories**, one based on **probabilities** and the other not.

The first category contains, for example, discriminating analysis. The second category can be further subdivided into two groups. The first group contains what are known as optimal classification methods. In the second group, we can distinguish between several subtypes of classification method:

- Partition methods that consist of distributing n objects among g groups in such a way that each object exclusively belongs to just one group. The number of groups g is fixed beforehand, and the partition applied most closely satisfies the classification criteria.
- Partition methods incorporating infringing classes, where an object is allowed to belong to several groups simultaneously.
- Hierarchical classification methods, where the structure of the **data** at different levels of classification is taken into account. We can then take account for the relationships that exist between the dif-

ferent groups created during the partition. Two different kinds of hierarchical techniques exist: agglomerating techniques and dividing techniques.

Agglomerating methods start with separated objects, meaning that the n objects are initially distributed into n groups. Two groups are agglomerated in each subsequent step until there is only one group left.

In contrast, dividing methods start with all of the objects grouped together, meaning that all of the n objects are in one single group to start with. New groups are created at each subsequent step until there are n groups.

- Geometric classification, in which the objects are depicted on a **scatter plot** and then grouped according to position on the plot. In a **graphical representation**, the proximities of the objects to each other in the graphic correspond to the similarities between the objects.

The first three types of classification are generally grouped together under the term **cluster analysis**. What they have in common is the fact that the objects to be classified must present a certain amount of structure that allows us to measure the degree of similarity between the objects.

Each type of classification contains a multitude of methods that allow us to create classes of similar objects.

DOMAINS AND LIMITATIONS

Classification can be used in two cases:

- Description cases;
- Prediction cases.

In the first case, the classification is done on the basis of some generally accepted standard characteristics. For example, professions can be classified into freelance, managers, workers, and so on, and one can calculate average salaries, average frequency of health problems, and so on, for each class.

In the second case, classification will lead to a prediction and then to an action. For example, if the foxes in a particular region exhibit apathetic behavior and excessive salivation, we can conclude that there is a new rabies epidemic. This should then prompt a vaccine campaign.

FURTHER READING

- ▸ **Cluster analysis**
- ▸ **Complete linkage method**
- ▸ **Data analysis**

REFERENCES

Everitt, B.S.: Cluster Analysis. Halstead, London (1974)

Gordon, A.D.: Classification. Methods for the Exploratory Analysis of Multivariate Data. Chapman & Hall, London (1981)

Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)

Thorndike, R.L.: Who belongs in a family? Psychometrika **18**, 267–276 (1953)

Tryon, R.C.: Cluster Analysis. McGraw-Hill, New York (1939)

Zubin, J.: A technique for measuring like-mindedness. J. Abnorm. Social Psychol. **33**, 508–516 (1938)

Cluster Analysis

Clustering is the partitioning of a data set into subsets or clusters, so that the degree of association is strong between members of the same cluster and weak between members of

different clusters according to some defined distance measure.

Several methods of performing **cluster analysis** exist:

- Partitional clustering
- Hierarchical clustering.

HISTORY

See **classification** and **data analysis**.

MATHEMATICAL ASPECTS

To carry out **cluster analysis** on a set of n objects, we need to define a **distance** between the objects (or more generally a measure of the similarity between the objects) that need to be classified. The existence of some kind of structure within the set of objects is assumed.

To carry out a hierarchical classification of a set E of objects $\{x_1, x_2, \ldots, x_n\}$, it is necessary to define a **distance** associated with E that can be used to obtain a **distance table** between the objects of E. Similarly, a distance must also be defined for any subsets of E.

One approach to hierarchical clustering is to use the agglomerating method. It can be summarized in the following algorithm:

1. Locate the pair of objects (x_i, x_j) which have the smallest **distance** between each other.
2. Aggregate the pair of objects (x_i, x_j) into a single element α and re-establish a new **distance table**. This is achieved by suppressing the lines and columns associated with x_i and x_j and replacing them with a line and a column associated with α. The new distance table will have a line and a column less than the previous table.
3. Repeat these two operations until the desired number of classes are obtained or until all of the objects are gathered into the same class.

Note that the distance between the group formed from aggregated elements and the other elements can be defined in different ways, leading to different methods. Examples include the single link method and the **complete linkage method**.

The single link method is a hierarchical classification method that uses the Euclidean **distance** to establish a **distance table,** and the distance between two classes is given by the Euclidean distance between the two closest elements (the minimum distance).

In the **complete linkage method**, the **distance** between two classes is given by the Euclidean distance between the two elements furthest away (the maximum distance).

Given that the only difference between these two methods is that the distance between two classes is either the minimum and the maximum distance, only the single link method will be considered here.

For a set $E = \{X_1, X_2, \ldots, X_n\}$, the **distance table** for the elements of E is then established.

Since this table is symmetric and null along its diagonal, only one half of the table is considered:

$$\begin{array}{cccc} d(X_1, X_2) & d(X_1, X_3) & \ldots & d(X_1, X_n) \\ & d(X_2, X_3) & \ldots & d(X_2, X_n) \\ & & \ddots & \ldots \\ & & & d(X_{n-1}, X_n) \end{array}$$

where $d(X_i, X_j)$ is the Euclidean distance between X_i and X_j for $i < j$, where the **values** of i and j are between 1 and n.

The **algorithm** for the single link method is as follows:

- Search for the minimum $d(X_i, X_j)$ for $i < j$;

- The elements X_i and X_j are aggregated into a new group $C_k = X_i \cup X_j$;
- The set E is then partitioned into

$$\{X_1\}, \ldots, \{X_{i-1}\}, \{X_i, X_j\}, \{X_{i+1}\}, \ldots, \{X_{j-1}\}, \{X_{j+1}\}, \ldots, \{X_n\};$$

- The **distance table** is then recreated without the lines and columns associated with X_i and X_j, and with a line and a column representing the **distances** between X_m and $C_k, m = 1, 2, \ldots, n\ m \neq i$ and $m \neq j$, given by:

$$d(C_k, X_m) = \min\{d(X_i, X_m); d(X_j, X_m)\}.$$

The algorithm is repeated until the desired number of groups is attained or until there is only one group containing all of the elements.

In the general case, the **distance** between two groups is given by:

$$d(C_k, C_m) = \min\{d(X_i, X_j) \text{ with } X_i \text{ belonging to } C_k \text{ and } X_j \text{ to } C_m\},$$

The formula quoted previously applies to the particular case when the groups are composed, respectively, of two elements and one single element.

This series of agglomerations can be represented by a **dendrogram**, where the abscissa shows the distance separating the objects. Note that we could find more than one pair when we search for the pair of closest elements. In this case, the pair that is selected for aggregation in the first step does not influence later steps (provided the algorithm does not finish at this step), because the other pair of closest elements will be aggregated in the following step. The aggregation order is not shown on the dendrogram because it reports the distance that separates two grouped objects.

DOMAINS AND LIMITATIONS

The choice of the **distance** between the group formed of aggregated elements and the other elements can be operated in several ways, according to the method that is used, as for example in the single link method and in the **complete linkage method**.

EXAMPLES

Let us illustrate how the **single link method** of cluster analysis can be applied to the examination grades obtained by five students each studying four courses: English, French, maths and physics.

We want to divide these five students into two groups using the single link method.

The grades obtained in the examinations, which range from 1 to 6, are summarized in the following table:

	English	French	Maths	Physics
Alain	5.0	3.5	4.0	4.5
Jean	5.5	4.0	5.0	4.5
Marc	4.5	4.5	4.0	3.5
Paul	4.0	5.5	3.5	4.0
Pierre	4.0	4.5	3.0	3.5

We then work out the Euclidian **distances** between the students and use them to create a **distance table**:

	Alain	Jean	Marc	Paul	Pierre
Alain	0	1.22	1.5	2.35	2
Jean	1.22	0	1.8	2.65	2.74
Marc	1.5	1.8	0	1.32	1.12
Paul	2.35	2.65	1.32	0	1.22
Pierre	2	2.74	1.12	1.22	0

By only considering the upper part of this symmetric table, we obtain:

	Jean	Marc	Paul	Pierre
Alain	1.22	1.5	2.35	2
Jean		1.8	2.65	2.74
Marc			1.32	1.12
Paul				1.22

The minimum **distance** is 1.12, between Marc and Pierre; we therefore form the first group from these two students. We then calculate the new distances.
For example, we calculate the new distance between Marc and Pierre on one side and Alain on the other by taking the minimum distance between Marc and Alain and the minimum distance between Pierre and Alain:

$$d(\{\text{Marc,Pierre}\},\text{Alain}) = \min\{d(\text{Marc,Alain});d(\text{Pierre,Alain})\} = \min\{1.5; 2\} = 1.5\,,$$

also

$$d(\{\text{Marc,Pierre}\},\text{Jean}) = \min\{d(\text{Marc,Jean});d(\text{Pierre,Jean})\} = \min\{1.8; 2.74\} = 1.8\,,$$

and

$$d(\{\text{Marc,Pierre}\},\text{Paul}) = \min\{d(\text{Marc,Paul});d(\text{Pierre,Paul})\} = \min\{1.32; 1.22\} = 1.22\,.$$

The new **distance table** takes the following form:

	Jean	Marc and Pierre	Paul
Alain	1.22	1.5	2.35
Jean		1.8	2.65
Marc and Pierre			1.22

The minimum distance is now 1.22, between Alain and Jean and also between the group of Marc and Pierre on the one side and Paul on the other side (in other words, two pairs exhibit the minimum distance); let us choose to regroup Alain and Jean first. The other pair will be aggregated in the next step. We rebuild the **distance table** and obtain:

$$d(\{\text{Alain,Jean}\},\{\text{Marc,Pierre}\}) = \min\{d(\text{Alain},\{\text{Marc,Pierre}\}), d(\text{Jean},\{\text{Marc,Pierre}\})\} = \min\{1.5; 1.8\} = 1.5$$

as well as:

$$d(\{\text{Alain,Jean}\},\text{Paul}) = \min\{d(\text{Alain,Paul});d(\text{Jean,Paul})\} = \min\{2.35; 2.65\} = 2.35\,.$$

This gives the following **distance table**:

	Marc and Pierre	Paul
Alain and Jean	1.5	2.35
Marc and Pierre		1.22

Notice that Paul must now be integrated in the group formed from Marc and Pierre, and the new **distance** will be:

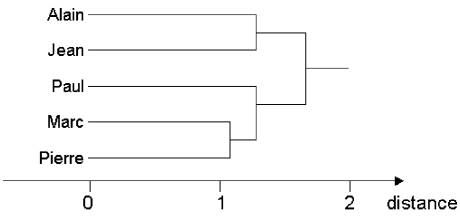
$$d(\{(\text{Marc,Pierre}),\text{Paul}\},\{\text{Alain,Jean}\}) = \min\{d(\{\text{Marc,Pierre}\},\{\text{Alain,Jean}\}), d(\text{Paul},\{\text{Alain,Jean}\})\} = \min\{1.5; 2.35\} = 1.5$$

which gives the following **distance table**:

	Alain and Jean
Marc, Pierre and Paul	1.5

We finally obtain two groups:
{Alain and Jean} and {Marc, Pierre and Paul} which are separated by a **distance** of 1.5.

The following **dendrogram** illustrates the successive aggregations:



FURTHER READING

- ▸ Classification
- ▸ Complete linkage method
- ▸ Dendrogram
- ▸ Distance
- ▸ Distance table

REFERENCES

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., Ralambondrainy, H.: Classification automatique des données—aspects statistiques et informatiques. Dunod, Paris (1989)

Everitt, B.S.: Cluster Analysis. Halstead, London (1974)

Gordon, A.D.: Classification. Methods for the Exploratory Analysis of Multivariate Data. Chapman & Hall, London (1981)

Jambu, M., Lebeaux, M.O.: Classification automatique pour l'analyse de données. Dunod, Paris (1978)

Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)

Lerman, L.C.: Classification et analyse ordinale des données. Dunod, Paris (1981)

Tomassone, R., Daudin, J.J., Danzart, M., Masson, J.P.: Discrimination et classement. Masson, Paris (1988)

Cluster Sampling

In cluster **sampling**, the first step is to divide the **population** into subsets called clusters. Each cluster consists of individuals that are supposed to be representative of the population.

Cluster **sampling** then involves choosing a random **sample** of clusters and then observing all of the individuals that belong to each of them.

HISTORY

See **sampling**.

MATHEMATICAL ASPECTS

Cluster sampling is the process of randomly extracting representative sets (known as clusters) from a larger population of units and then applying a questionnaire to all of the units in the clusters. The clusters often consist of geographical units, like city districts. In this case, the method involves dividing a city into districts, and then selecting the districts to be included in the sample. Finally, all of the people or households in the chosen district are questioned.

There are two principal reasons to perform cluster sampling. In many inquiries, there is no complete and reliable list of the population units on which to base the sampling, or it may be that it is too expensive to create such a list. For example, in many countries, including industrialized ones, it is rare to have complete and up-to-date lists of all of the members of the population, households or rural estates. In this situation, sampling can be achieved in a geographical manner: each urban region is divided up into districts and each rural region into rural estates. The districts and the agricultural areas are considered to be clusters and we use the com-

plete list of clusters because we do not have a complete and up-to-date list of all population units. Therefore, we sample a requisite number of clusters from the list and then question all of the units in the selected cluster.

DOMAINS AND LIMITATIONS

The advantage of cluster **sampling** is that it is not necessary to have a complete, up-to-date list of all of the units of the **population** to perform analysis.

For example, in many countries, there are no updated lists of people or housing. The costs of creating such lists are often prohibitive. It is therefore easier to analyze subsets of the **population** (known as clusters).

In general, cluster **sampling** provides **estimations** that are not as precise as **simple random sampling**, but this drop in accuracy is easily offset by the far lower cost of cluster sampling.

In order to perform cluster **sampling** as efficiently as possible:

- The clusters should not be too big, and there should be a large enough number of clusters,
- cluster sizes should be as uniform as possible;
- The individuals belonging to each cluster must be as heterogenous as possible with respect to the parameter being observed.

Another reason to use cluster sampling is cost. Even when a complete and up-to-date list of all population units exists, it may be preferable to use cluster sampling from an economic point of view, since it is completed faster, involves fewer workers and minimizes transport costs. It is therefore more appropriate to use cluster sampling if the money saved by doing so is far more significant than the increase in sampling variance that will result.

EXAMPLES

Consider N, the size of the **population** of town X. We want to study the distribution of "Age" for town X without performing a census. The population is divided into G parts. **Simple random sampling** is performed amongst these G parts and we obtain g parts. The final **sample** will be composed of all the individuals in the g selected parts.

FURTHER READING

- ▸ **Sampling**
- ▸ **Simple random sampling**

REFERENCES

Hansen, M.H., Hurwitz, W.N., Madow, M.G.: Sample Survey Methods and Theory. Vol. I. Methods and Applications. Vol. II. Theory. Chapman & Hall, London (1953)

Coefficient of Determination

The coefficient of determination, denoted R^2, is the quotient of the explained variation (sum of squares due to regression) to the total variation (total sum of squares total SS (TSS)) in a **model** of simple or **multiple linear regression**:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}.$$

It equals the square of the **correlation coefficient**, and it can take values between 0 and 1. It is often expressed as a **percentage**.

HISTORY

See **correlation coefficient**.

MATHEMATICAL ASPECTS

Consider the following **model** for multiple regression:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

for $i = 1, \ldots, n$, where

- Y_i are the dependent variables,
- X_{ij} $(i = 1, \ldots, n, j = 1, \ldots, p)$ are the independent variables,
- ε_i are the random nonobservable **error** terms,
- β_j $(j = 1, \ldots, p)$ are the parameters to be estimated.

Estimating the parameters $\beta_0, \beta_1, \ldots, \beta_p$ yields the estimation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} \, .$$

The coefficient of determination allows us to measure the quality of fit of the regression equation to the measured values.

To determine the quality of the fit of the **regression** equation, consider the gap between the observed **value** and the estimated value for each **observation** of the **sample**. This gap (or **residual**) can also be expressed in the following way:

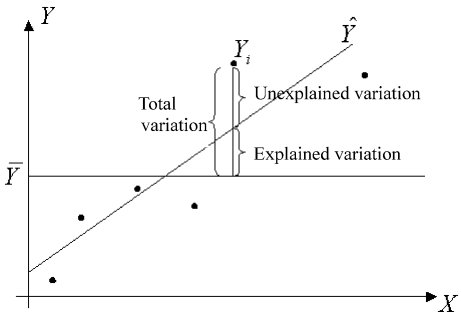
$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$\text{TSS} = \text{RSS} + \text{REGSS}$$

where

- **TSS** is the total sum of squares,
- **RSS** the residual sum of squares and
- **REGSS** the sum of the squares of the regression.

These concepts and the relationships between them are presented in the following graph:



Using these concepts, we can define R^2, which is the determination coefficient. It measures the proportion of variation in **variable** Y, which is described by the **regression** equation as:

$$R^2 = \frac{\text{REGSS}}{\text{TSS}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

If the **regression** function is to be used to make predictions about subsequent observations, it is preferable to have a high **value** of R^2, because the higher the value of R^2, the smaller the unexplained variation.

EXAMPLES

The following table gives values for the Gross National Product (GNP) and the demand for domestic products covering the 1969–1980 period for a particular country.

Year	GNP *X*	Demand for domestic products *Y*
1969	50	6
1970	52	8
1971	55	9
1972	59	10

Year	GNP X	Demand for domestic products Y
1973	57	8
1974	58	10
1975	62	12
1976	65	9
1977	68	11
1978	69	10
1979	70	11
1980	72	14

We will try to estimate the demand for small goods as a function of GNP according to the **model**

$$Y_i = a + b \cdot X_i + \varepsilon_i\,, \quad i = 1, \ldots, 12\,.$$

Estimating the **parameters** a and b by the **least squares** method yields the following **estimators**:

$$\hat{b} = \frac{\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{12}(X_i - \bar{X})^2} = 0.226\,,$$

$$\hat{a} = \bar{Y} - \hat{b} \cdot \bar{X} = -4.047\,.$$

The estimated line is written

$$\hat{Y} = -4.047 + 0.226 \cdot X\,.$$

The quality of the fit of the measured points to the regression line is given by the determination coefficient:

$$R^2 = \frac{\text{REGSS}}{\text{TSS}} = \frac{\sum_{i=1}^{12}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{12}(Y_i - \bar{Y})^2}\,.$$

We can calculate the **mean** using:

$$\bar{Y} = \frac{\sum_{i=1}^{12} Y_i}{n} = 9.833\,.$$

Y_i	$\hat{Y}_i$	$(Y_i - \bar{Y})^2$	$(\hat{Y}_i - \bar{Y})^2$
6	7.260	14.694	6.622
8	7.712	3.361	4.500
9	8.390	0.694	2.083
10	9.294	0.028	0.291
8	8.842	3.361	0.983
10	9.068	0.028	0.586
12	9.972	4.694	0.019
9	10.650	0.694	0.667
11	11.328	1.361	2.234
10	11.554	0.028	2.961
11	11.780	1.361	3.789
14	12.232	17.361	4.754
Total		47.667	30.489

We therefore obtain:

$$R^2 = \frac{30.489}{47.667} = 0.6396$$

or, in percent:

$$R^2 = 63.96\%\,.$$

We can therefore conclude that, according to the **model** chosen, 63.96% of the variation in the demand for small goods is explained by the variation in the GNP.

Obviously the **value** of R^2 cannot exceed 100%. While 63.96% is relatively high, it is not close enough to 100% to rule out trying to modify the **model** further.

This analysis also shows that other variables apart from the GNP should be taken into account when determining the function corresponding to the demand for small goods, since the GNP only partially explains the variation.

FURTHER READING

- ▸ Correlation coefficient
- ▸ Multiple linear regression
- ▸ Regression analysis
- ▸ Simple linear regression

Coefficient of Kurtosis

The coefficient of kurtosis is used to measure the peakness or flatness of a curve. It is based on the **moments** of the distribution. This coefficient is one of the **measures of kurtosis**.

HISTORY

See **coefficient of skewness**.

MATHEMATICAL ASPECTS

The coefficient of kurtosis (β_2) is based on the centered fourth-order **moment** of a **distribution** which is equal to:

$$\mu_4 = E\left[(X-\mu)^4\right].$$

In order to obtain a coefficient of kurtosis that is independent of the units of measurement, the fourth-order **moment** is divided by the **standard deviation** of the **population** σ raised to the fourth power. The coefficient of kurtosis then becomes equal to:

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

For a **sample** $(x_1, x_2, \ldots, x_n)$, the **estimator** of this coefficient is denoted by b_2. It is equal to:

$$b_2 = \frac{m_4}{S^4},$$

where m_4 is the centered fourth-order **moment** of the sample, given by:

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^4,$$

where $\bar{x}$ is the **arithmetic mean**, n is the total number of **observations** and S^4 is the **standard deviation** of the sample raised to the fourth power.

For the case where a **random variable** X takes **values** x_i with **frequencies** f_i, $i = 1, 2, \ldots, h$, the centered fourth-order **moment** of the **sample** is given by the formula:

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^4.$$

DOMAINS AND LIMITATIONS

For a **normal distribution**, the coefficient of kurtosis is equal to 3. Therefore a curve will be called platikurtic (meaning flatter than the normal distribution) if it has a kurtosis coefficient smaller than 3. It will be leptokurtic (meaning sharper than the normal distribution) if β_2 is greater than 3.

Let us now prove that the coefficient of kurtosis is equal to 3 for the **normal distribution**. We know that $\beta_2 = \frac{\mu_4}{\sigma^4}$, meaning that the centered fourth-order **moment** is divided by the **standard deviation** raised to the fourth power. It can be proved that the centered sth order moment, denoted μ_s, satisfies the following relation for a **normal distribution**:

$$\mu_s = (s-1)\sigma^2 \cdot \mu_{s-2}.$$

This formula is a recursive formula which expresses higher order **moments** as a function of lower order moments.

Given that $\mu_0 = 1$ (the zero-order **moment** of any **random variable** is equal to 1, since it is the **expected value** of this **variable** raised to the power zero) and that $\mu_1 = 0$ (the centered first-order **moment** is zero for any random variable), we have:

$$\mu_2 = \sigma^2$$

$$\mu_3 = 0$$

$$\mu_4 = 3\sigma^4$$

etc. .

The coefficient of kurtosis is then equal to:

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{3\sigma^4}{\sigma^4} = 3\,.$$

EXAMPLES

We want to calculate the kurtosis of the distribution of daily turnover for 75 bakeries. Let us calculate the coefficient of kurtosis β_2 using the following **data**:

Table categorizing the daily turnovers of 75 bakeries

Turnover	Frequencies
215–235	4
235–255	6
255–275	13
275–295	22
295–315	15
315–335	6
335–355	5
355–375	4

The fourth-order **moment** of the **sample** is given by:

$$m_4 = \frac{1}{n}\sum_{i=1}^{h} f_i(x_i - \bar{x})^4\,,$$

where $n = 75$, $\bar{x} = 290.60$ and x_i is the center of class **interval** i. We can summarize the calculations in the following table:

x_i	$x_i - \bar{x}$	f_i	$f_i(x_i - \bar{x})^4$
225	−65.60	4	74075629.16
245	−45.60	6	25942428.06
265	−25.60	13	5583457.48
285	−5.60	22	21635.89
305	14.40	15	644972.54
325	34.40	6	8402045.34
345	54.40	5	43789058.05
365	74.40	4	122560841.32
			281020067.84

Since $S = 33.88$, the coefficient of kurtosis is equal to:

$$\beta_2 = \frac{\frac{1}{75}(281020067.84)}{(33.88)^4} = 2.84\,.$$

Since β_2 is smaller than 3, we can conclude that the distribution of the daily turnover in 75 bakeries is platikurtic, meaning that it is flatter than the **normal distribution**.

FURTHER READING

- ▸ **Measure of kurtosis**
- ▸ **Measure of shape**

REFERENCES

See **coefficient of skewness β_1 de Pearson**.

Coefficient of Skewness

The coefficient of skewness measures the skewness of a distribution. It is based on the notion of the **moment** of the distribution. This coefficient is one of the **measures of skewness**.

HISTORY

Between the end of the nineteenth century and the beginning of the twentieth century, **Pearson, Karl** studied large sets of **data** which sometimes deviated significantly from normality and exhibited considerable skewness.

He first used the following coefficient as a **measure of skewness**:

$$\text{skewness} = \frac{\bar{x} - \text{mode}}{S}\,,$$

where $\bar{x}$ represents the **arithmetic mean** and S the **standard deviation**.

This measure is equal to zero if the data are distributed symmetrically.

He discovered empirically that for a moderately asymmetric distribution (the gamma distribution):

$$M_o - \bar{x} \approx 3 \cdot (M_d - \bar{x}) ,$$

where M_o and M_d denote the mode and the median of data set. By substituting this expression into the previous coefficient, the following alternative formula is obtained:

$$\text{skewness} = \frac{3 \cdot (\bar{x} - M_d)}{S} .$$

Following this, **Pearson, K.** (1894,1895) introduced a coefficient of skewness, known as the β_1 coefficient, based on calculations of the centered **moments**. This coefficient is more difficult to calculate but it is more descriptive and better adapted to large numbers of **observations**.

Pearson, K. also created the **coefficient of kurtosis** (β_2), which is used to measure the oblateness of a curve. This coefficient is also based on the **moments** of the distribution being studied.

Tables giving the limit values of the coefficients β_1 and β_2 can be found in the works of Pearson and Hartley (1966, 1972). If the sample estimates a fall outside the limit for β_1, β_2, we conclude that the population is significantly curved or skewed.

MATHEMATICAL ASPECTS

The skewness coefficient is based on the centered third-order **moment** of the distribution in question, which is equal to:

$$\mu_3 = E\left[(X - \mu)^3\right] .$$

To obtain a coefficient of skewness that is independent of the measuring unit, the third-order **moment** is divided by the **standard deviation** of the **population** σ raised to the third power. The coefficient obtained, designated by $\sqrt{\beta_1}$, is equal to:

$$\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} .$$

The **estimator** of this coefficient, calculated for a **sample** $(x_1, x_2, \ldots x_n)$, is denoted by $\sqrt{b_1}$. It is equal to:

$$\sqrt{b_1} = \frac{m_3}{S^3} ,$$

where m_3 is the centered third-order **moment** of the sample, given by:

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3 .$$

Here $\bar{x}$ is the **arithmetic mean**, n is the total number of **observations** and S^3 is the **standard deviation** raised to the third power.

For the case where a **random variable** X takes **values** x_i with **frequencies** f_i, $i = 1, 2, \ldots, h$, the centered third-order **moment** of the **sample** is given by the formula:

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^3 .$$

If the coefficient is positive, the distribution spreads to the the right. If it is negative, the distribution expands to the left. If it is close to zero, the distribution is approximately symmetric.

If the **sample** is taken from a normal **population**, the **statistic** $\sqrt{b_1}$ roughly follows a **normal distribution** with a **mean** of 0 and a **standard deviation** of $\sqrt{\frac{6}{n}}$. If the **size of the sample** n is bigger than 150, the **normal table** can be used to test the skewness **hypothesis**.

DOMAINS AND LIMITATIONS

This coefficient (similar to the other **measures of skewness**) is only of interest if it

can be used to compare the shapes of two or more distributions.

EXAMPLES

Suppose that we want to compare the shapes of the daily turnover distributions obtained for 75 bakeries for two different years. We then calculate the skewness coefficient in both cases.

The **data** are categorized in the table below:

Turnover	Frequencies for year 1	Frequencies for year 2
215–235	4	25
235–255	6	15
255–275	13	9
275–295	22	8
295–315	15	6
315–335	6	5
335–355	5	4
355–375	4	3

The third-order **moment** of the **sample** is given by:

$$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{h} f_i \cdot (x_i - \bar{x})^3 .$$

For year 1, $n = 75$, $\bar{x} = 290.60$ and x_i is the center of each class **interval** i. The calculations are summarized in the following table:

x_i	$x_i - \bar{x}$	f_i	$f_i(x_i - \bar{x})^3$
225	−65.60	4	−1129201.664
245	−45.60	6	−568912.896
265	−25.60	13	−218103.808
285	−5.60	22	−3863.652
305	14.40	15	44789.760
325	34.40	6	244245.504
345	54.40	5	804945.920
365	74.40	4	1647323.136
			821222.410

Since $S = 33.88$, the coefficient of skewness is equal to:

$$\sqrt{b_1} = \frac{\frac{1}{75}(821222.41)}{(33.88)^3} = \frac{10949.632}{38889.307} = 0.282 .$$

For year 2, $n = 75$ and $\bar{x} = 265.27$. The calculations are summarized in the following table:

x_i	$x_i - \bar{x}$	f_i	$f_i(x_i - \bar{x})^3$
225	−40.27	25	−1632213.81
245	−20.27	15	−124864.28
265	−0.27	9	−0.17
285	19.73	8	61473.98
305	39.73	6	376371.09
325	59.73	5	1065663.90
345	79.73	4	2027588.19
365	99.73	3	2976063.94
			4750082.83

Since $S = 42.01$, the coefficient of skewness is equal to:

$$\sqrt{b_1} = \frac{\frac{1}{75}(4750082.83)}{(42.01)^3} = \frac{63334.438}{74140.933} = 0.854 .$$

The coefficient of skewness for year 1 is close to zero ($\sqrt{b_1} = 0.282$), so the daily turnover distribution for the 75 bakeries for year 1 is very close to being a symmetrical distribution. For year 2, the skewness coefficient is higher; this means that the distribution spreads towards the right.

FURTHER READING

- ▸ Measure of shape
- ▸ Measure of skewness

REFERENCES

Pearson, E.S., Hartley, H.O.: Biometrika Tables for Statisticians, vols. I and II. Cambridge University Press, Cambridge (1966,1972)

Pearson, K.: Contributions to the mathematical theory of evolution. I. In: Karl Pearson's Early Statistical Papers. Cambridge University Press, Cambridge, pp. 1–40 (1948). First published in 1894 as: On the dissection of asymmetrical frequency curves. Philos. Trans. Roy. Soc. Lond. Ser. A **185**, 71–110

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

Coefficient of Variation

The coefficient of variation is a **measure of relative dispersion**. It describes the **standard deviation** as a percentage of the **arithmetic mean**.

This coefficient can be used to compare the dispersions of **quantitative variables** that are not expressed in the same units (for example, when comparing the salaries in different countries, given in different currencies), or the dispersions of variables that have very different **means**.

MATHEMATICAL ASPECTS

The coefficient of variation CV is defined as the ratio of the **standard deviation** to the **arithmetic mean** for a set of **observations**; in other words:

$$CV = \frac{S}{\bar{x}} \cdot 100$$

for a **sample**, where:

- S is the standard deviation of the sample, and
- $\bar{x}$ is the arithmetic mean of the sample,

or:

$$CV = \frac{\sigma}{\mu} \cdot 100$$

for a **population**, where

- σ is the standard deviation of the population, and
- μ is the mean of the population.

This coefficient is independent of the unit of measurement used for the variable.

EXAMPLES

Let us study the salary distributions for two companies from two different countries. According to a **survey**, the **arithmetic means** and the **standard deviations** of the salaries are as follows:

Company A:

$$\bar{x} = 2500\,\text{CHF},$$
$$S = 200\,\text{CHF},$$
$$CV = \frac{200}{2500} \cdot 100 = 8\%.$$

Company B:

$$\bar{x} = 1000\,\text{CHF},$$
$$S = 50\,\text{CHF},$$
$$CV = \frac{50}{1000} \cdot 100 = 5\%.$$

The **standard deviation** represents 8% of the **arithmetic mean** for company A, and 5% for company B. The salary distribution is a bit more homogeneous for company B than for company A.

FURTHER READING

- ▸ Arithmetic mean
- ▸ Measure of dispersion
- ▸ Standard deviation

REFERENCES

Johnson, N.L., Leone, F.C.: Statistics and experimental design in engineering and the physical sciences. Wiley, New York (1964)

Collinearity

Variables are known to be mathematically collinear if one of them is a linear combination of the other variables. They are known as statistically collinear if one of them is approximately a linear combination of other variables. In the case of a regression model where the explanatory variables are strongly correlated to each other, we say that there is collinearity (or multicollinearity) between the explanatory variables. In the first case, it is simply impossible to define least squares estimators, and in the second case, these estimators can exhibit considerable variance.

HISTORY

The term "collinearity" was first used in mathematics at the beginning of the twentieth century, due to the rediscovery of the theorem of Pappus of Alexandria (a third-century mathematician). Let A, B, C be three points on a line and A', B', C' be three points on a different line. If we relate the pairs using AB'.A'B, CA'.AC' and BC'.B'C, their intersections will occur in a line; in other words, the three intersection points will be collinear.

MATHEMATICAL ASPECTS

In the case of a matrix of explanatory variables $\mathbf{X}$, collinearity means that one of the columns of $\mathbf{X}$ is (approximately) a linear combination of the other columns. This implies that $\mathbf{X}'\mathbf{X}$ is almost singular. Consequently, the **estimator** obtained by the least squares method $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ is obtained by inverting an almost singular **matrix**, which causes its components to become unstable. The **ridge regression** technique was created in order to deal with these collinearity problems.

A collinear relation between more than two variables will not always be the result of observing the pairwise correlations between the variables. A better indication of the presence of a collinearity problem is provided by variance inflation factors, *VIF*. The variance inflation factor of an explanatory variable X_j is defined by:

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the **coefficient of determination** of the model

$$X_j = \beta_0 + \sum_{k \neq j} \beta_k X_k + \varepsilon .$$

The coefficient *VIF* takes values of between 1 and ∞. If the X_j are mathematically collinear with other variables, we get $R_j^2 = 1$ and $VIF_j = \infty$. On the other hand, if the X_j are reciprocally independent, we have $R_j^2 = 0$ and $VIF_j = 1$. In practice, we consider that there is a real problem with collinearity when VIF_j is greater then 100, which corresponds to a R_j^2 that is greater then 0.99.

DOMAINS AND LIMITATIONS

Inverting a singular matrix, similar to inverting 0, is not a valid operation. Using the

same principle, inverting an almost singular matrix is similar to inverting a very small number. Some of the elements of the matrix must therefore be very big. Consequently, when the explanatory variables are collinear, some elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ of $\widehat{\boldsymbol{\beta}}$ will probably be very large. This is why collinearity leads to unstable regression estimators. Aside from this problem, collinearity also results in a calculation problem; it is difficult to precisely calculate the inverse of an almost singular matrix.

EXAMPLES

Thirteen portions of cement are examined in the following example. Each portion contains four ingredients, as described in the table. The goal of the experiment is to determine how the quantities X_1, X_2, X_3 and X_4, corresponding to the quantities of these four ingredients, affect the quantity Y of heat given out as the cement hardens.

- Y_i quantity of heat given out during the hardening of the ith portion (in joules);
- X_{i1} quantity of ingredient 1 (tricalcium aluminate) in the ith portion;
- X_{i2} quantity of ingredient 2 (tricalcium silicate) in the ith portion;
- X_{i3} quantity of the ingredient 3 (tetracalcium aluminoferrite) in the ith portion;
- X_{i4} quantity of ingredient 4 (dicalcium silicate) in the ith portion.

Table: Heat given out by the cement portions during hardening

Portion	Ingredient	Ingredient	Ingredient	Ingredient	Heat
i	1 X_1	2 X_2	3 X_3	4 X_4	Y
1	7	26	6	60	78.5
2	1	29	12	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	54	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	48	4	26	115.9
11	1	40	23	34	83.9
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Source: Birkes & Dodge (1993)

We start with a **simple linear regression**. The model used for the linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon\,.$$

We obtain the following results:

Variable	$\hat{\beta}_i$	S.d.	t_c
Constant	62.41	70.07	0.89
X_1	1.5511	0.7448	2.08
X_2	0.5102	0.7238	0.70
X_3	0.1019	0.7547	0.14
X_4	−0.1441	0.7091	−0.20

We can see that only coefficient X_1 is significantly greater then zero; in the other cases $t_c < t_{(\alpha/2,n-2)}$ (value taken from the Student table) for a **significance level** of $\alpha = 0.1$. Moreover, the standard deviations of the estimated coefficients $\widehat{\beta}_2$, $\widehat{\beta}_3$ and $\widehat{\beta}_4$ are greater then the coefficients themselves. When the

variables are strongly correlated, it is known that the effect of one can mask the effect of another. Because of this, the coefficients can appear to be insignificantly different from zero.

To verify the presence of *multicollinearity* for a couple of variables, we calculate the correlation matrix.

	X_1	X_2	X_3	X_4
X_1	1	0.229	−0.824	0.245
X_2	0.229	1	−0.139	−0.973
X_3	−0.824	−0.139	1	0.030
X_4	0.245	−0.973	0.030	1

We note that a strong negative correlation (−0.973) exists between X_2 and X_4. Looking at the data, we can see the reason for that. Aside from portions 1 to 5, the total quantity of silicates ($X_2 + X_4$) is almost constant across the portions, and is approximately 77; therefore, X_4 is approximately $77 - X_2$. this situation does not allow us to distinguish between the individual effects of X_2 and those of X_4. For example, we see that the four largest values of X_4 (60, 52, 47 and 44) correspond to values of Y smaller then the mean heat 95.4. Therefore, at first sight it seems that large quantities of ingredient 4 will lead to small amounts of heat. However, we also note that the four largest values of X_4 correspond to the four smallest values of X_2, giving a negative correlation between X_2 and X_4. This suggests that ingredient 4 taken alone has a small effect on variable Y, and that the small quantities of 2 taken alone can explain the small amount of heat emitted. Hence, the linear dependence between two explanatory variables (X_2 and X_4) makes it more complicated to see the effect of each variable alone on the response variable Y.

FURTHER READING

- ▸ **Multiple linear regression**
- ▸ **Ridge regression**

REFERENCES

Birkes, D., Dodge, Y.: Alternative Methods of Regression. Wiley, New York (1993)

Bock, R.D.: Multivariate Statistical Methods in Behavioral Research. McGraw-Hill, New York (1975)

Combination

A combination is an un-ordened collection of unique elements or objects.

A k-combination is a subset with k elements. The number of k-combinations from a set of n elements is the number of **arrangements**.

HISTORY

See **combinatory analysis**.

MATHEMATICAL ASPECTS

1. *Combination without repetition*
   Combination without repetition describe the situation where each object drawn is not placed back for the next drawing. Each object can therefore only occur once in each group.
   The number of combination without repetition of k objects among n is given by:

$$C_n^k = \binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} .$$

2. *Combination with repetitions*
   Combination with repetitions (or with remittance) are used when each drawn object is placed back for the next drawing. Each object can then occur r times in each group, $r = 0, 1, \ldots, k$.

The number of combinations with repetitions of k objects among n is given by:

$$K_n^k = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k! \cdot (n-1)!}.$$

EXAMPLES

1. *Combinations without repetition*
   Consider the situation where we must choose a committee of three people from an assembly of eight people. How many different committees could potentially be picked from this assembly, if each person can only be selected once in each group? Here we need to calculate the number of possible combinations of three people from eight:

$$C_n^k = \frac{n!}{k! \cdot (n-k)!} = \frac{8!}{3! \cdot (8-3)!} = \frac{40320}{6 \cdot 120} = 56$$

   Therefore, it is possible to form 56 different committees containing three people from an assembly of eight people.
2. *Combinations with repetitions* Consider an urn containing six numbered balls. We carry out four successive drawings, and place the drawn ball back into the urn after each drawing. How many different combinations could occur from this drawing? In this case we want to find the number of combinations with repetition (because each drawn ball is placed back in the urn before the next drawing). We obtain

$$K_n^k = \frac{(n+k-1)!}{k! \cdot (n-1)!} = \frac{9!}{4! \cdot (6-1)!} = \frac{362880}{24 \cdot 120} = 126$$

   different combinations.

FURTHER READING

- ▸ Arrangement
- ▸ Binomial distribution
- ▸ Combinatory analysis
- ▸ Permutation

C

Combinatory Analysis

Combinatory analysis refers to a group of techniques that can be used to determine the number of elements in a particular set without having to count them one-by-one.
The elements in question could be the results from a scientific **experiment** or the different potential outcomes of a random **event**.
Three particular concepts are important in combinatory analysis:

- Permutations;
- Combinations;
- Arrangements.

HISTORY

Combinatory analysis has interested mathematicians for centuries. According to Takacs (1982), such analysis dates back to ancient Greece. However, the Hindus, the Persians (including the poet and mathematician Khayyâm, Omar) and (especially) the Chinese also studied such problems. A 3000 year-old Chinese book "I Ching" describes the possible **arrangements** of a set of n elements, where $n \leq 6$. In 1303, Chu, Shih-chieh published a work entitled "Ssu Yuan Yü Chien" (Precious mirror of the four elements). The cover of the book depictsa triangle that shows the combinations of k elements taken from a set of size n where $0 \leq k \leq n$. This arithmetic triangle was also explored by several European mathematicians such as Stifel, Tartaglia and Hérigone, and especially Pascal, who wrote the "Trai-

té du triangle arithmétique" (Treatise of the arithmetic triangle) in 1654 (although it was not published until after his death in 1665). Another document on **combinations** was published in 1617 by Puteanus, Erycius called "Erycii Puteani Pretatis Thaumata in Bernardi Bauhusii è Societate Jesu Proteum Parthenium". However, combinatory analysis only revealed its true power with the works of Fermat (1601–1665) and Pascal (1623–1662). The term "combinatory analysis" was introduced by Leibniz (1646–1716) in 1666. In his work "Dissertatio de Arte Combinatoria," he systematically studied problems related to **arrangements**, **permutations** and **combinations**.

Other works in this field should be mentioned here, such as those of Wallis, J. (1616–1703), reported in "The Doctrine of Permutations and Combinations" (an essential and fundamental part of the "Doctrines of Chances"), or those of **Bernoulli**, J., **Moivre, A. de**, Cardano, G. (1501–1576), and Galileo (1564–1642).

In the second half of the nineteenth century, Cayley (1829–1895) solved some problems related to this type of analysis via graphics that he called "trees". Finally, we should also mention the important work of MacMahon (1854–1929), "Combinatory Analysis" (1915–1916).

EXAMPLES

See **arrangement**, **combination** and **permutation**.

FURTHER READING

- ▸ **Arithmetic triangle**
- ▸ **Arrangement**
- ▸ **Binomial**
- ▸ **Binomial distribution**
- ▸ **Combination**
- ▸ **Permutation**

REFERENCES

MacMahon, P.A.: Combinatory Analysis, vols. I and II. Cambridge University Press, Cambridge (1915–1916)

Stigler, S.: The History of Statistics, the Measurement of Uncertainty Before 1900. Belknap, London (1986)

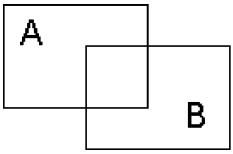
Takács, L.: Combinatorics. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York (1982)

Compatibility

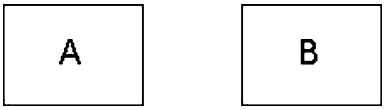
Two **events** are said to be compatible if the occurrence of the first event does not prevent the occurrence of the second (or in other words, if the **intersection** between the two events is not null):

$$P(A \cap B) \neq 0\,.$$

We can represent two compatible **events** A and B schematically in the following way:



Two **events** A and B are incompatible (or mutually exclusive) if the occurrence of A prevents the occurrence of B, or vice versa. We can represent this in the following way:



This means that the **probability** that these two events happen at the same time is zero:

$$A \cap B = \phi \longrightarrow P(A \cap B) = 0\,.$$

MATHEMATICAL ASPECTS

If two **events** A and B are compatible, the **probability** that at least one of the events occurs can be obtained using the following formula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

On the other hand, if the two **events** A and B are incompatible, the **probability** that the two events A and B happen at the same time is zero:

$$P(A \cap B) = 0\,.$$

The **probability** that at least one of the events occurs can be obtained by simply adding the individual probabilities of A and B:

$$P(A \cup B) = P(A) + P(B)\,.$$

EXAMPLES

Consider a **random experiment** that involves drawing a card from a pack of 52 cards. We are interested in the three following **events**:

$$A = \text{“draw a heart”}$$
$$B = \text{“draw a queen”}$$
$$C = \text{“draw a club”}\,.$$

The **probabilities** associated with each of these **events** are:

$$P(A) = \frac{13}{52}$$
$$P(B) = \frac{4}{52}$$
$$P(C) = \frac{13}{52}\,.$$

The **events** A and B are compatible, because it is possible to draw both a heart and a queen at the same time (the queen of hearts). Therefore, the intersection between A and B is the queen of hearts. The probability of this event is given by:

$$P(A \cap B) = \frac{1}{52}\,.$$

The **probability** of the union of the two **events** A and B (drawing either a heart or a queen) is then equal to:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{4}{13}\,. \end{aligned}$$

On the other hand, the **events** A and C are incompatible, because a card cannot be both a heart and a club! The **intersection** between A and C is an empty set.

$$A \cap C = \phi\,.$$

The **probability** of the union of the two events A and C (drawing a heart or a club) is simply given by the sum of the probabilities of each event:

$$\begin{aligned} P(A \cup C) &= P(A) + P(C) \\ &= \frac{13}{52} + \frac{13}{52} \\ &= \frac{1}{2}\,. \end{aligned}$$

FURTHER READING

Event
Independence
Probability

Complementary

Consider the **sample space** Ω for a **random experiment**.
For any **event** A, an element of Ω, we can determine a new event B that contains all of the elements of the **sample space** Ω that are not included in A.
This event B is called the complement of A with respect to Ω and is obtained by the negation of A.

MATHEMATICAL ASPECTS

Consider an **event** A, which is an element of the **sample space** Ω.
The compliment of A with respect to Ω is denoted $\bar{A}$. It is given by the negation of A:

$$\begin{aligned}\bar{A} &= \Omega - A \\ &= \{w \in \Omega;\ w \notin A\}\ .\end{aligned}$$

EXAMPLES

Consider a **random experiment** that consists of flipping a coin three times.
The **sample space** of this experiment is

$$\begin{aligned}\Omega = \{&TTT, TTH, THT, THH,\\ &HTT, HTH, HHT, HHH\}\ .\end{aligned}$$

Consider the **event**

$$\begin{aligned}A &= \text{“Heads } (H) \text{ occurs twice”} \\ &= \{THH, HTH, HHT\}\ .\end{aligned}$$

The compliment of A with respect to Ω is equal to

$$\begin{aligned}\bar{A} &= \{TTT, TTH, THT, HTT, HHH\} \\ &= \text{“Heads } (H) \text{ does not occur twice”}\ .\end{aligned}$$

FURTHER READING

- ▸ Event
- ▸ Random experiment
- ▸ Sample space

Complete Linkage Method

The complete linkage method is a hierarchical **classification** method where the distance between two classes is defined as the greatest distance that could be obtained if we select one element from each class and measure the distance between these elements. In other words, it is the distance between the most distant elements from each class.
For example, the **distance** used to construct the **distance table** is the Euclidian distance. Using the complete linkage method, the distance between two classes is given by the Euclidian distance between the most distant elements (the maximum distance).:

MATHEMATICAL ASPECTS

See **cluster analysis**.

FURTHER READING

- ▸ Classification
- ▸ Cluster analysis
- ▸ Dendrogram
- ▸ Distance
- ▸ Distance table

Completely Randomized Design

A completely randomized design is a type of experimental design where the experimental units are randomly assigned to the different treatments.
It is used when the experimental units are believed to be “uniform;” that is, when there is no uncontrolled **factor** in the **experiment**.

HISTORY

See **design of experiments**.

EXAMPLES

We want to test five different drugs based on aspirin. To do this, we randomly distribute the five types of drug to 40 patients. Denoting the five drugs by A, B, C, D and E, we obtain the following random distribution:

A is attributed to 10 people;

B is attributed to 12 people;

C is attributed to 4 people;

D is attributed to 7 people;

E is attributed to 7 people.

We have then a completely randomized design where the treatments (drugs) are randomly attributed to the experimental units (patients), and each patient receives only one treatment. We also assume that the patients are "uniform:" that there are no differences between them. Moreover, we assume that there is no uncontrolled **factor** that intervenes during the treatment.

In this example, the completely randomized design is a **factorial experiment** that uses only one factor: the aspirin. The five types of aspirin are different levels of the factor.

FURTHER READING

- ▸ **Design of experiments**
- ▸ **Experiment**

Composite Index Number

Composite **index numbers** allow us to measure, with a single number, the relative variations within a group of **variables** upon moving from one situation to another.

The consumer price index, the wholesale price index, the employment index and the Dow-Jones index are all examples of composite index numbers.

The aim of using composite index numbers is to summarize all of the simple index numbers contained in a complex number (a value formed from a set of simple values) in just one index.

The most commonly used composite index numbers are:

- The **Laspeyres index**;
- The **Paasche index**;
- The **Fisher index**.

HISTORY

See **index number**.

MATHEMATICAL ASPECTS

There are several methods of creating composite index numbers.

To illustrate these methods, let us use a scenario where a price index is determined for the current period n with respect to a reference period 0.

1. *Index number of the arithmetic means* (the sum method):

$$I_{n/0} = \frac{\sum P_n}{\sum P_0} \cdot 100\,,$$

where $\sum P_n$ is the sum of the prices of the items at the current period, and $\sum P_0$ is the sum of the prices of the items at the reference period.

2. *Arithmetic mean of simple index numbers:*

$$I_{n/0} = \frac{1}{N} \cdot \sum \left(\frac{P_n}{P_0} \right) \cdot 100\,,$$

where N is the number of goods considered and $\frac{P_n}{P_0}$ is the **simple index number** of each item.

In these two methods, each item has the same importance. This is a situation

which often does not correspond to reality.

3. *Index number of weighted arithmetic means* (the weighted sum method):
   The general formula for an index number calculated by the weighted sum method is as follows:
$$I_{n/0} = \frac{\sum P_n \cdot Q}{\sum P_0 \cdot Q} \cdot 100\,.$$
   Choosing the quantity Q for each item considered could prove problematic here: Q must be the same for both the numerator and the denominator when calculating a price index.
   In the **Laspeyres index**, the value of Q corresponding to the reference year is used. In the **Paasche index**, the value of Q for the current year is used. Other statisticians have proposed using the value of Q for a given year.

EXAMPLES

Consider the following table indicating the(fictitious) prices of three consumer goods in the reference year (1970) and their current prices.

	Price (francs)	
Goods	1970 (P_0)	Now (P_n)
Milk	0.20	1.20
Bread	0.15	1.10
Butter	0.50	2.00

Using these numbers, we now examine the three main methods of constructing composite **index numbers**.

1. *Index number of arithmetic means* (the sum method):
$$I_{n/0} = \frac{\sum P_n}{\sum P_0} \cdot 100 = \frac{4.30}{0.85} \cdot 100 = 505.9\,.$$
   According to this method, the price index has increased by 405.9% (505.9 − 100) between the reference year and the current year.
2. *Arithmetic mean of simple index numbers:*
$$I_{n/0} = \frac{1}{N} \cdot \sum \left(\frac{P_n}{P_0}\right) \cdot 100 = \frac{1}{3} \cdot \left(\frac{1.20}{0.20} + \frac{1.10}{0.15} + \frac{2.00}{0.50}\right) \cdot 100 = 577.8\,.$$
   This method gives a slightly different result from the previous one, since we obtain an increase of 477.8% (577.8 − 100) in the price index between the reference year and the current year.
3. *Index number of weighted arithmetic means* (the weighted sum method):
$$I_{n/0} = \frac{\sum P_n \cdot Q}{\sum P_0 \cdot Q} \cdot 100\,.$$
   This method is used in conjunction with the **Laspeyres index** or the **Paasche index**.

FURTHER READING

- ▸ Fisher index
- ▸ Index number
- ▸ Laspeyres index
- ▸ Paasche index
- ▸ Simple index number

Conditional Probability

The probability of an event given that another event is known to have occured.

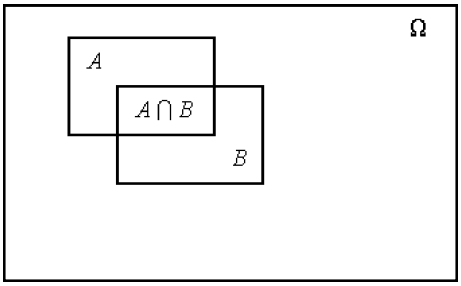
The conditional probability is denoted $P(A|B)$, which is read as the "probability of A conditioned by B."

HISTORY

The concept of **independence** dominated probability theory until the beginning of the twentieth century. In 1933, Kolmogorov, Andrei Nikolaievich introduced the concept of conditional probability; this concept now plays an essential role in theoretical and applied probability and statistics.

MATHEMATICAL ASPECTS

Consider a **random experiment** for which we know the **sample space** Ω. Consider two events A and B from this space. The **probability** of A, $P(A)$ depends on the set of possible events in the **experiment** (Ω).



Now consider that we have supplementary information concerning the experiment: that the event B has occurred. The probability of the event A occurring will then be a function of the space B rather than a function of Ω. The probability of A conditioned by B is calculated as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} .$$

If A and B are two incompatible events, the intersection between A and B is an empty space. We will then have $P(A|B) = P(B|A) = 0$.

DOMAINS AND LIMITATIONS

The concept of conditional probability is one of the most important ones in probability theory. This importance is mainly due to the following points:

1. We are often interested in calculating the probability of an **event** when some information about the result is already known. In this case, the probability required is a conditional probability.
2. Even when partial information on the result is not known, conditional probabilities can be useful when calculating the probabilities required.

EXAMPLES

Consider a group of 100 cars distributed according to two criteria, comfort and speed. We will make the following distinctions:

$$\text{a car can be } \begin{cases} \text{fast} \\ \text{slow} \end{cases} ,$$

$$\text{a car can be } \begin{cases} \text{comfortable} \\ \text{uncomfortable} \end{cases} .$$

A partition of the 100 cars based on these criteria is provided in the following table:

	fast	slow	total
comfortable	40	10	50
uncomfortable	20	30	50
total	60	40	100

Consider the following events:

$A =$ "a fast car is chosen"

and $B =$ "a comfortable car is chosen."

The probability of these two events are:

$$P(A) = 0.6 ,$$
$$P(B) = 0.5 .$$

The probability of choosing a fast car is then of 0.6, and that of choosing a comfortable car is 0.5.

Now imagine that we are given supplementary information: a fast car was chosen. What, then, is the probability that this car is also comfortable?
We calculate the probability of B knowing that A has occurred, or the conditional probability of B depending on A:
We find in the table that

$$P(A \cap B) = 0.4\,.$$

$$\Rightarrow P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.4}{0.6} = 0.667\,.$$

The probability that the car is comfortable, given that we know that it is fast, is therefore $\frac{2}{3}$.

FURTHER READING

- ▸ **Event**
- ▸ **Probability**
- ▸ **Random experiment**
- ▸ **Sample space**

REFERENCES

Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Heidelberg New York (1933)

Kolmogorov, A.N.: Foundations of the Theory of Probability. Chelsea Publishing Company, New York (1956)

Confidence Interval

A confidence interval is any **interval** constructed around an **estimator** that has a particular **probability** of containing the true **value** of the corresponding **parameter** of a **population**.

HISTORY

According to Desrosières, A. (1988), Bowley, A.L. was one of the first to be interested in the concept of the confidence interval. Bowley presented his first confidence interval calculations to the Royal Statistical Society in 1906.

MATHEMATICAL ASPECTS

In order to construct a confidence interval that contains the true **value** of the **parameter** θ with a given **probability**, an equation of the following form must be solved:

$$P(L_i \leq \theta \leq L_s) = 1 - \alpha\,,$$

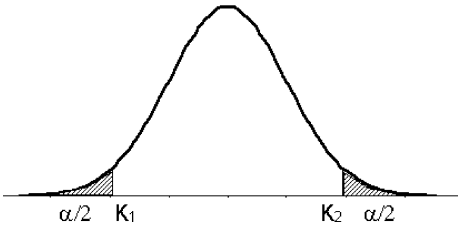
where

$\boldsymbol{\theta}$	is the parameter to be estimated,
$\boldsymbol{L_i}$	is the lower limit of the **interval**,
$\boldsymbol{L_s}$	is the upper limit of the interval and
$\boldsymbol{1-\alpha}$	is the given probability, called the **confidence level** of the interval.

The probability α measures the **error** risk of the interval, meaning the probability that the interval does not contain the true **value** of the parameter θ.
In order to solve this equation, a function $f(t, \theta)$ must be defined where t is an **estimator** of θ, for which the **probability distribution** is known.
Defining this **interval** for $f(t, \theta)$ involves writing the equation:

$$P(k_1 \leq f(t, \theta) \leq k_2) = 1 - \alpha\,,$$

where the constants k_1 and k_2 are given by the **probability distribution** of the function $f(t, \theta)$. Generally, the error risk α is divided into two equal parts at $\frac{\alpha}{2}$ distributed on each side of the distribution of $f(t, \theta)$. If, for example, the function $f(t, \theta)$ follows a centered and reduced **normal distribution**, the constants k_1 and k_2 will be symmetric and can be represented by $-z_{\frac{\alpha}{2}}$ and $+z_{\frac{\alpha}{2}}$, as shown in the following figure.



Once the constants k_1 and k_2 have been found, the **parameter** θ must be isolated in the equation given above. The confidence interval θ is found in this way for the **confidence level** $1-\alpha$.

DOMAINS AND LIMITATIONS

One should be very careful when interpreting a confidence interval. If, for a **confidence level** of 95%, we find a confidence interval for a **mean** of μ where the lower and upper limits are k_1 and k_2 respectively, we can conclude the following (for example):

"On the basis of the studied **sample**, we can affirm that it is probable that the **mean** of the **population** can be found in the **interval** established."

It would not be correct to conclude that there is a 95% chance of finding the **mean** of the **population** in the **interval**. Indeed, since μ and the limits k_1 and k_2 of the interval are constants, the interval may or may not contain μ. However, if the statistician has the ability to repeat the **experiment** (which consists of drawing a **sample** from the population) several times, 95% of the intervals obtained will contain the true **value** of μ.

EXAMPLES

A business that fabricates lightbulbs wants to test the average lifespan of its lightbulbs. The distribution of the **random variable** X, which represents the life span in hours, is a **normal distribution** with **mean** μ and **standard deviation** $\sigma = 30$.

In order to estimate μ, the business burns out $n = 25$ lightbulbs.

It obtains an average lifespan of $\bar{x} = 860$ hours. It wants to establish a confidence interval around the **estimator** $\bar{x}$ at a **confidence level** of 0.95. Therefore, the first step is to obtain a function $f(t,\theta) = f(\bar{x},\mu)$ for the known distribution. Here we use:

$$f(t,\theta) = f(\bar{x},\mu) = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}},$$

which follows a centered and reduced **normal distribution**. The equation $P(k_1 \leq f(t,\theta) \leq k_2) = 1-\alpha$ becomes:

$$P\left(-z_{0.025} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0.025}\right) = 0.95,$$

because the **error** risk α has been divided into two equal parts at $\frac{\alpha}{2} = 0.025$.

The table for the centered and reduced **normal distribution**, the **normal table**, gives $z_{0.025} = 1.96$. Therefore:

$$P\left(-1.96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95.$$

To obtain the confidence interval for μ at a **confidence level** of 0.95, μ must be isolated in the equation above:

$$P\left(-1.96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

By replacing $\bar{x}$, σ and n with their respective **values**, we obtain:

$$P\left(860 - 1.96\frac{30}{\sqrt{25}} \leq \mu \leq 860 + 1.96\frac{30}{\sqrt{25}}\right) = 0.95$$

$$P(848.24 \leq \mu \leq 871.76) = 0.95\,.$$

The confidence interval for μ at the confidence level 0.95 is therefore:

$$[848.24,\, 871.76]\,.$$

This means that we can affirm with a **probability** of 95% that this **interval** contains the true **value** of the **parameter** μ that corresponds to the average lifespan of the lightbulbs.

FURTHER READING

- ▸ **Confidence level**
- ▸ **Estimation**
- ▸ **Estimator**

REFERENCES

Bowley, A.L.: Presidential address to the economic section of the British Association. J. Roy. Stat. Soc. **69**, 540–558 (1906)

Desrosières, A. La partie pour le tout: comment généraliser? La préhistoire de la contrainte de représentativité. Estimation et sondages. Cinq contributions à l'histoire de la statistique. Economica, Paris (1988)

Confidence Level

The confidence level is the **probability** that the **confidence interval** constructed around an **estimator** contains the true **value** of the corresponding **parameter** of the **population**.

We designate the confidence level by $(1 - \alpha)$, where α corresponds to the risk of **error**; that is, to the probability that the confidence interval does not contain the true value of the parameter.

HISTORY

The first example of a confidence interval appears in the work of Laplace (1812). According to Desrosières, A. (1988), Bowley, A.L. was one of the first to become interested in the concept of the confidence interval.

See **hypothesis testing**.

MATHEMATICAL ASPECTS

Let θ be a parameter associated with a population. θ is to be estimated and T is its estimator from a random sample. We evaluate the precision of T as the estimator of θ by constructing a **confidence interval** around the estimator, which is often interpreted as an error margin.

In order to construct this confidence interval, we generally proceed in the following manner. From the distribution of the estimator T, we determine an interval that is likely to contain the true value of the parameter. Let us denote this interval by $(T - \varepsilon, T + \varepsilon)$ and the probability of true value of the paramter being in this interval as $(1 - \alpha)$. We can then say that the error margin ε is related to α by the probability:

$$P(T - \varepsilon \leq \theta \leq T + \varepsilon) = 1 - \alpha\,.$$

The level of probability associated with an interval of estimation is called the confidence level or the confidence degree.

The interval $T - \varepsilon \leq \theta \leq T + \varepsilon$ is called the confidence interval for θ at the confidence level $1 - \alpha$. Let us use, for example, $\alpha = 5\%$,

which will give the confidence interval of the parameter θ to a probability level of 95%. This means that, if we use T as an estimator of θ, then the interval indicated will on average contain the true value of the parameter 95 times out of 100 samplings, and it will not contain it 5 times.

The quantity ε of the confidence interval corresponds to half od the length of the interval. This parameter therefore gives us an idea of the error margin for the estimator. For a given confidence level $1-\alpha$, the smaller the confidence interval, more efficient the estimator.

DOMAINS AND LIMITATIONS

The most commonly used confidence levels are 90%, 95% and 99%. However, if necessary, other levels can be used instead.

Although we would like to use the highest confidence level in order to maximize the **probability** that the confidence interval contains the true value of the **parameter**, but if we increase the confidence level,the **interval** increases as well. Therefore, what we gain in terms of confidence is lost in terms of precision, so have to find a compromise.

EXAMPLES

A company that produces ligthtbulbs wants to study the mean lifetime of its bulbs. The distribution of the **random variable** X that represents the lifetime in hours is a **normal distribution** of **mean** μ and **standard deviation** $\sigma = 30$.

In order to estimate μ, the company burns out a random **sample** of $n = 25$ bulbs.

The company obtains an average bulb lifetime of $\bar{x} = 860$ hours. It then wants to construct a 95% **confidence interval** for μ around the **estimator** $\bar{x}$.

The standard deviation σ of the population is known; the value of ε is $z_{\alpha/2} \cdot \sigma_{\bar{X}}$. The value of $z_{\alpha/2}$ is obtained from the normal table, and it depends on the probability attributed to the parameter α. We then deduce the confidence interval of the estimator of μ at the probability level $1-\alpha$:

$$\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}\,.$$

From the **hypothesis** that the bulb lifetime X follows a normal distribution with mean μ and standard deviation $\sigma = 30$, we deduce that the expression

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

follows a standard normal distribution.

From this we obtain:

$$P\left[-z_{0.025} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0.025}\right] = 0.95\,,$$

where the risk of **error** α is divided into two parts that both equal $\frac{\alpha}{2} = 0.025$.

The table of the standard normal distribution (the **normal table**) gives $z_{0.025} = 1.96$. We then have:

$$P\left(-1.96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95\,.$$

To get the confidence interval for μ, we must isolate μ in the following equation via the following transformations:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95\,,$$

$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95\,.$$

Substituting $\bar{x}$, σ and n for their respective values, we evaluate the confidence interval:

$$860 - 1.96 \cdot \frac{30}{\sqrt{25}} \leq \mu \leq 860 + 1.96 \cdot \frac{30}{\sqrt{25}}$$

$$848.24 \leq \mu \leq 871.76 .$$

We can affirm with a confidence level of 95% that this **interval** contains the true value of the **parameter** μ, which corresponds to the mean bulb lifetime.

FURTHER READING

- ► **Confidence interval**
- ► **Estimation**
- ► **Estimator**
- ► **Hypothesis testing**

Contingency Table

A contingency table is a crossed table containing various attributes of a population or an observed sample. Contingency table analysis consists of discovering and studying the relations (if they exist) between these attributes.

A contingency table can be a two-dimensional table with r lines and c columns relating to two qualitative categorical variables possessing, respectively, r and c categories. It can also be multidimensional when the number of qualitative variables is greater then two: if, for example, the elements of a **population** or a **sample** are characterized by three attributes, the associated contingency table has the dimensions $I \times J \times K$, where I represents the number of categories defining the first attribute, J the number of categories of the second attribute and K the number of the categories of the third attribute.

HISTORY

The term "contingency," used in relation to a crossed table of categorical data, seems to have originated with **Pearson, Karl** (1904), who used he term "contingency" to mean a measure of the total **deviation** relative to the **independence**.

See also **chi-square test of independence**.

MATHEMATICAL ASPECTS

If we consider a two-dimensional table, containing entries for two qualitative categorical variables X and Y that have, respectively, r and c categories, the contingency table is:

		Categories of the variable Y			
		Y_1	...	Y_c	Total
Categories of the variable X	X_1	n_{11}	...	n_{1c}	$n_{1.}$
	...	...	...	...	...
	X_r	n_{r1}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	...	$n_{.c}$	$n_{..}$

where

- n_{ij} represents the observed **frequency** for category i of **variable** X and category j of variable Y;
- $n_{i.}$ represents the sum of the frequencies observed for category i of variable X,
- $n_{.j}$ represents the sum of the observed frequencies for category j of variable Y, and
- $n_{..}$ indicates the total number of observations.

In the case of a multidimensional table, the elements of the table are denoted by n_{ijk}, representing the observed frequency for category i of variable X, category j of variable Y and category k of variable Z.

DOMAINS AND LIMITATIONS

The **independence** of two categorical qualitative variables represented in the contingen-

cy table can be assessed by performing a **chi-square test of independence**.

FURTHER READING

- ▸ **Chi-square test of independence**
- ▸ **Frequency distribution**

REFERENCES

Fienberg, S.E.: The Analysis of Cross-Classified Categorical Data, 2nd edn. MIT Press, Cambridge, MA (1980)

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)

Continuous Distribution Function

The **distribution function** of a continuous **random variable** is defined to be the **probability** that the random variable takes a **value** less than or equal to a real number.

HISTORY

See **probability**.

MATHEMATICAL ASPECTS

The function defined by

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x\,.$$

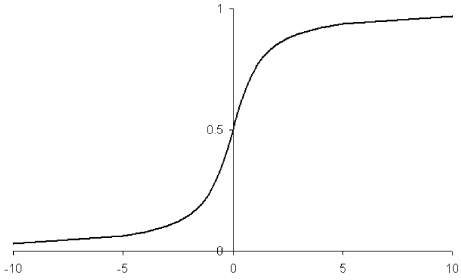
is called the **distribution function** of a continuous **random variable**.

In other words, the **density function** f is the derivative of the continuous distribution function.

Properties of the Continuous Distribution Function

1. $F(x)$ is a continually increasing function for all x;
2. F takes its **values** in the **interval** $[0, 1]$;
3. $\lim_{b\to-\infty} F(b) = 0$;
4. $\lim_{b\to\infty} F(b) = 1$;
5. $F(x)$ is a continuous and differentiable function.

This distribution function can be graphically represented on a system of axes. The different **values** of the **random variable** X are plotted on the abscissa and the corresponding values of $F(x)$ on the ordinate.



DOMAINS AND LIMITATIONS

The **probability** that the continuous **random variable** X takes a **value** in the **interval** $]a, b]$ for all $a < b$, meaning that $P(a < X \leq b)$, is equal to $F(b) - F(a)$, where F is the distribution function of the random variable X.

Demonstration: The **event** $\{X \leq b\}$ can be written as the union of two mutually exclusive events: $\{X \leq a\}$ and $\{a < X \leq b\}$:

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}\,.$$

By finding the **probability** on each side of the equation, we obtain:

$$\begin{aligned} P(X \leq b) &= P(\{X \leq a\} \cup \{a < X \leq b\}) \\ &= P(X \leq a) + P(a < X \leq b)\,. \end{aligned}$$

The sum of probabilities result from the fact that the two events are exclusive.
By subtracting $P(X \leq a)$ on each side, we have:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)\,.$$

Finally, from the definition of the **distribution function**, we obtain:

$$P(a < X \leq b) = F(b) - F(a)\,.$$

EXAMPLES

Consider a continuous **random variable** X for which the **density function** is given by:

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{if not} \end{cases}\,.$$

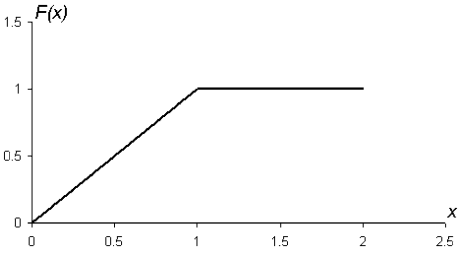
The **probability** that X takes a **value** in the **interval** $[a, b]$, with $0 < a$ and $b < 1$, is as follows:

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)\,\mathrm{d}x \\ &= \int_a^b 1\,\mathrm{d}x \\ &= x\,\Big|_a^b \\ &= b - a\,. \end{aligned}$$

Therefore, for $0 < x < 1$ the distribution function is:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(0 \leq X \leq x) \\ &= x\,. \end{aligned}$$

This function is presented in the following figure:



FURTHER READING

- ▸ Density function
- ▸ Probability
- ▸ Random experiment
- ▸ Random variable
- ▸ Value

Continuous Probability Distribution

Every **random variable** has a corresponding **frequency distribution**. For a continuous random variable, this distribution is continuous too.
A continuous probability distribution is a **model** that represents the **frequency distribution** of a continuous **variable** in the best way.

MATHEMATICAL ASPECTS

The **probability distribution** of a continuous **random variable** X is given by its **density function** $f(x)$ or its **distribution function** $F(x)$.
It can generally be characterized by its **expected value**:

$$E[X] = \int_D x \cdot f\,(x)\,\mathrm{d}x = \mu$$

and its **variance**:

$$\begin{aligned} \mathrm{Var}(X) &= \int_D (x - \mu)^2 \cdot f(x)\,\mathrm{d}x \\ &= E\left[(X - \mu)^2\right] \\ &= E\left[X^2\right] - E[X]^2\,, \end{aligned}$$

where D represents the **interval** covering the range of **values** that X can take.
One essential property of a continuous **random variable** is that the **probability** that it will take a specific numerical **value** is zero, whereas the probability that it will take a value over an **interval** (finite or infinite) is usually nonzero.

DOMAINS AND LIMITATIONS

The most famous continuous probability distribution is the **normal distribution**.
Continuous probability distributions are often used to approximate **discrete probability distributions**. They are used in model construction just as much as they are used when applying statistical techniques.

FURTHER READING

- ▸ Beta distribution
- ▸ Cauchy distribution
- ▸ Chi-square distribution
- ▸ Continuous distribution function
- ▸ Density function
- ▸ Discrete probability distribution
- ▸ Expected value
- ▸ Exponential distribution
- ▸ Fisher distribution
- ▸ Gamma distribution
- ▸ Laplace distribution
- ▸ Lognormal distribution
- ▸ Normal distribution
- ▸ Probability
- ▸ Probability distribution
- ▸ Random variable
- ▸ Student distribution
- ▸ Uniform distribution
- ▸ Variance of a random variable

REFERENCES

Johnson, N.L., Kotz, S.: Distributions in Statistics: Continuous Univariate Distributions, vols. 1 and 2. Wiley, New York (1970)

C

Contrast

In analysis of variance a contrast is a linear combination of the **observations** or factor levels or treatments in a **factorial experiment**, where the sum of the coefficients is zero.

HISTORY

According to Scheffé, H. (1953), Tukey, J.W. (1949 and 1951) was the first to propose a method of simultaneously **estimating** all of the contrasts.

MATHEMATICAL ASPECTS

Consider $T_1, T_2, \ldots, T_k$, which are the sums of $n_1, n_2, \ldots, n_k$ **observations**. The linear function

$$c_j = c_{1j} \cdot T_1 + c_{2j} \cdot T_2 + \cdots + c_{kj} \cdot T_k$$

is a contrast if and only if

$$\sum_{i=1}^{k} n_i \cdot c_{ij} = 0\,.$$

If each $n_i = n$, meaning that if T_i is the sum of the same number of **observations**, the condition is reduced to:

$$\sum_{i=1}^{k} c_{ij} = 0\,.$$

DOMAINS AND LIMITATIONS

In most **experiments** involving several **treatments**, it is interesting for the experimenter to make comparisons between the different treatments. The statistician uses contrasts to carry out this type of comparison.

EXAMPLES

When an **analysis of variance** is carried out for a three-**level factor**, some contrasts of interest are:

$$c_1 = T_1 - T_2$$
$$c_2 = T_1 - T_3$$
$$c_3 = T_2 - T_3$$
$$c_4 = T_1 - 2 \cdot T_2 + T_3 .$$

FURTHER READING

- ▸ **Analysis of variance**
- ▸ **Experiment**
- ▸ **Factorial experiment**

REFERENCES

Ostle, B.: Statistics in Research: Basic Concepts and Techniques for Research Workers. Iowa State College Press, Ames, IA (1954)

Scheffé, H.: A method for judging all contrasts in the analysis of variance. Biometrika **40**, 87–104 (1953)

Tukey, J.W.: Comparing individual means in the analysis of variance. Biometrics **5**, 99–114 (1949)

Tukey, J.W.: Quick and Dirty Methods in Statistics. Part II: Simple Analyses for Standard Designs. Quality Control Conference Papers 1951. American Society for Quality Control, New York, pp. 189–197 (1951)

Convergence

In statistics, the term "convergence" is related to **probability** theory. This statistical convergence is often termed stochastic convergence in order to distinguish it from classical convergence.

MATHEMATICAL ASPECTS

Different types of stochastic convergence can be defined. Let $\{x_n\}_{n \in N}$ be a set of random variables. The most important types of stochastic convergence are:

1. $\{X_n\}_{n \in \mathbb{N}}$ converges *in distribution* to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(z) = F_X(z) \quad \forall z ,$$

where F_{X_n} and F_X are the distribution functions of X_n and X, *respectively.* This convergence is simply the point convergence (well-known in mathematics) of the set of the distribution functions of the X_n.

2. $\{X_n\}_{n \in \mathbb{N}}$ converges *in probability* to a random variable X if:

$$\lim_{n \to \infty} P\left(|X_n - X| > \varepsilon\right) = 0 ,$$
$$\text{for every } \varepsilon > 0 .$$

3. $\{X_n\}_{n \in \mathbb{N}}$ exhibits *almost sure* convergence to a random variable X if:

$$P\left(\left\{w | \lim_{n \to \infty} X_n(w) = X(w)\right\}\right) = 1 .$$

4. Suppose that all elements of X_n have a finite **expectancy**. The set $\{X_n\}_{n \in \mathbb{N}}$ converges *in mean square* to X if:

$$\lim_{n \to \infty} E\left[(X_n - X)^2\right] = 0 .$$

Note that:

- Almost sure convergence and mean square convergence both imply a convergence in probability;
- Convergence in probability (weak convergence) implies convergence in distribution.

EXAMPLES

Let X_i be independent random variables uniformly distributed over [0, 1]. We define the following set of random variables from X_i:

$$Z_n = n \cdot \min_{i=1,\ldots,n} X_i \,.$$

We can show that the set $\{Z_n\}_{n\in\mathbb{N}}$ converges in distribution to an **exponential distribution** Z with a parameter of 1 as follows:

$$\begin{aligned} 1 - F_{Z_n}(t) &= P(Z_n > t) \\ &= P\left(\min_{i=1,\ldots,n} X_i > \frac{t}{n}\right) \\ &= P\left(X_1 > \frac{t}{n} \text{ and } X_2 > \frac{t}{n} \text{ and } \ldots X_n > \frac{t}{n}\right) \\ &\overset{\text{ind.}}{=} \prod_{i=1}^{n} P\left(X_i > \frac{t}{n}\right) = \left(1 - \frac{t}{n}\right)^n . \end{aligned}$$

Now, for $\lim_{n\to\infty}$, $\left(1 - \frac{t}{n}\right)^n = \exp(-t)$ since:

$$\begin{aligned} \lim_{n\to\infty} F_{Z_n} &= \lim_{n\to\infty} P(Z_n \le t) \\ &= 1 - \exp(-t) = F_Z \,. \end{aligned}$$

Finally, let S_n be the number of successes obtained during n Bernoulli trials with a probability of success p. **Bernoulli's theorem** tells us that $\frac{S_n}{n}$ converges in probability to a "random" variable that takes the value p with probability 1.

FURTHER READING

- ▸ Bernoulli's theorem
- ▸ Central limit theorem
- ▸ Convergence theorem
- ▸ De Moivre–Laplace Theorem
- ▸ Law of large numbers
- ▸ Probability
- ▸ Random variable
- ▸ Stochastic process

REFERENCES

Le Cam, L.M., Yang, C.L.: Asymptotics in Statistics: Some Basic Concepts. Springer, Berlin Heidelberg New York (1990)

Staudte, R.G., Sheater, S.J.: Robust Estimation and Testing. Wiley, New York (1990)

Convergence Theorem

The convergence theorem leads to the most important theoretical results in probability theory. Among them, we find the **law of large numbers** and the **central limit theorem**.

EXAMPLES

The **central limit theorem** and the **law of large numbers** are both convergence theorems.

The law of large numbers states that the **mean** of a sum of identically distributed random variables converges to their common **mathematical expectation**.

On the other hand, the central limit theorem states that the distribution of the sum of a sufficiently large number of random variables tends to approximate the **normal distribution**.

FURTHER READING

- ▸ Central limit theorem
- ▸ Law of large numbers

Correlation Coefficient

The simple correlation coefficient is a measure of the strength of the linear relation between two **random variables**.

The correlation coefficient can take **values** that occur in the interval [−1; 1]. The two extreme values of this interval represent a perfectly linear relation between the **variables**, "positive" in the first case and "negative" in the other. The value 0 (zero) implies the absence of a linear relation.

The correlation coefficient presented here is also called the Bravais–Pearson correlation coefficient.

HISTORY

The concept of correlation originated in the 1880s with the works of **Galton, F.**. In his autobiography *Memories of My Life* (1890), he writes that he thought of this concept during a walk in the grounds of Naworth Castle, when a rain shower forced him to find shelter. According to Stigler, S.M. (1989), Porter, T.M. (1986) was carrying out historical research when he found a forgotten article written by Galton in 1890 in The North American Review, under the title "Kinship and correlation". In this article, which he published right after its discovery, Galton (1908) explained the nature and the importance of the concept of correlation.

This discovery was related to previous works of the mathematician, notably those on heredity and linear regression. Galton had been interested in this field of study since 1860. He published a work entitled "Natural inheritance" (1889), which was the starting point for his thoughts on correlation.

In 1888, in an article sent to the Royal Statistical Society entitled "Co-relations and their measurement chiefly from anthropometric data," Galton used the term "correlation" for the first time, although he was still alternating between the terms "co-relation" and "correlation" and he spoke of a "co-relation index." On the other hand, he invoke the concept of a negative correlation. According to Stigler (1989), Galton only appeared to suggest that correlation was a positive relationship.

Pearson, Karl wrote in 1920 that correlation had been discovered by Galton, whose work "Natural inheritance" (1889) pushed him to study this concept too, along with two other researchers, Weldon and Edgeworth. Pearson and Edgeworth then developed the theory of correlation.

Weldon thought the correlation coefficient should be called the "Galton function." However, Edgeworth replaced Galton's term "co-relation index" and Weldon's term "Galton function" by the term "correlation coefficient."

According to Mudholkar (1982), **Pearson, K.** systemized the analysis of correlation and established a theory of correlation for three **variables**. Researchers from University College, most notably his assistant Yule, G.U., were also interested in developing multiple correlation.

Spearman published the first study on rank correlation in 1904.

Among the works that were carried out in this field, it is worth highlighting those of Yule, who in an article entitled "Why do we sometimes get non-sense-correlation between time-series" (1926) discussed the problem of correlation analysis interpretation. Finally, correlation robustness was investigated by Mosteller and Tukey (1977).

MATHEMATICAL ASPECTS

Simple Linear Correlation Coefficient

Simple linear correlation is the term used to describe a linear dependence between two **quantitative variables** X and Y (see **simple linear regression**).

If X and Y are **random variables** that follow an unknown **joint distribution**, then the simple linear correlation coefficient is equal to the **covariance** between X and Y divided by the product of their **standard deviations**:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} .$$

Here $\mathrm{Cov}(X, Y)$ is the measured **covariance** between X and Y; σ_X and σ_Y are the respective **standard deviations** of X and Y.

Given a sample of size n, (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ from the joint distribution, the quantity

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

is an **estimation** of ρ; it is the **sampling** correlation.

If we denote $(X_i - \bar{X})$ by x_i and $(Y_i - \bar{Y})$ by y_i, we can write this equation as:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}} .$$

Test of Hypothesis

To test the **null hypothesis**

$$H_0 \colon \rho = 0$$

against the **alternative hypothesis**

$$H_1 \colon \rho \neq 0 ,$$

we calculate the **statistic** t:

$$t = \frac{r}{S_r} ,$$

where S_r is the **standard deviation** of the **estimator** r:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} .$$

Under H_0, the **statistic** t follows a **Student distribution** with $n-2$ **degrees of freedom**. For a given significance level α, H_0 is rejected if $|t| \geq t_{\frac{\alpha}{2}, n-2}$; the **value** of $t_{\frac{\alpha}{2}, n-2}$ is the **critical value** of the test given in the **Student table**.

Multiple Correlation Coefficient

Known as the coefficient of determination denoted by R^2, determines whether the hyperplane estimated from a **multiple linear regression** is correctly adjusted to the data points.

The **value** of the multiple **determination coefficient** R^2 is equal to:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} .$$

It corresponds to the square of the multiple correlation coefficient. Notice that

$$0 \leq R^2 \leq 1 .$$

In the case of **simple linear regression**, the following relation can be derived:

$$r = \mathrm{sign}(\hat{\beta}_1)\sqrt{R^2} ,$$

where $\hat{\beta}_1$ is the **estimator** of the **regression** coefficient β_1, and it is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} .$$

DOMAINS AND LIMITATIONS

If there is a linear relation between two variables, the correlation coefficient is equal to 1 or −1.

A positive relation (+) means that the two variables vary in the same direction. If the individuals obtain high scores in the first variable (for example the **independent variable**), they will have a tendency to obtain high scores in the second variable (the **dependant variable**). The opposite is also true.

A negative relation (−) means that the individuals that obtain high scores in the first **variable** will have a tendency to obtain low scores in the second one, and vice versa.

Note that if the **variables** are independent the correlation coefficient is equal to zero. The reciprocal conclusion is not necessarily true. The fact that two or more variables are related in a statistical way is not sufficient to conclude that a cause and effect relation exists. The existence of a statistical correlation is not a proof of causality.

Statistics provides numerous correlation coefficients. The choice of which to use for a particular set of **data** depends on different factors, such as:

- The type of scale used to express the variable;
- The nature of the underlying distribution (continuous or discrete);
- The characteristics of the distribution of the scores (linear or nonlinear).

EXAMPLES

The **data** for two **variables** X and Y are shown in the table below:

No of			$x =$	$y =$			
order	X	Y	$X - \bar{X}$	$Y - \bar{Y}$	xy	x^2	y^2
1	174	64	−1.5	−1.3	1.95	2.25	1.69
2	175	59	−0.5	−6.3	3.14	0.25	36.69
3	180	64	4.5	−1.3	−5.85	20.25	1.69
4	168	62	−7.5	−3.3	24.75	56.25	10.89
5	175	51	−0.5	−14.3	7.15	0.25	204.49
6	170	60	−5.5	−5.3	29.15	30.25	28.09
7	170	68	−5.5	2.7	−14.85	30.25	7.29
8	178	63	2.5	−2.3	−5.75	6.25	5.29
9	187	92	11.5	26.7	307.05	132.25	712.89
10	178	70	2.5	4.7	11.75	6.25	22.09
Total	1755	653			358.5	284.5	1034.1

$$\bar{X} = 175.5 \quad \text{and} \quad \bar{Y} = 65.3 .$$

We now perform the necessary calculations to obtain the correlation coefficient between the two variables. Applying the formula gives:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}} = \frac{358.5}{\sqrt{284.5 \cdot 1034.1}} = 0.66 .$$

Test of Hypothesis

We can calculate the estimated **standard deviation** of r:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{0.56}{10 - 2}} = 0.266 .$$

Calculating the **statistic** t gives:

$$t = \frac{r - 0}{S_r} = \frac{0.66}{0.266} = 2.485\,.$$

If we choose a significance level α of 5%, the **value** from the **Student table**, $t_{0.025,8}$, is equal to 2.306.

Since $|t| = 2.485 > t_{0.025,8} = 2.306$, the **null hypothesis**

$$H_0 \colon \rho = 0$$

is rejected.

FURTHER READING

- ▸ **Coefficient of determination**
- ▸ **Covariance**
- ▸ **Dependence**
- ▸ **Kendall rank correlation coefficient**
- ▸ **Multiple linear regression**
- ▸ **Regression analysis**
- ▸ **Simple linear regression**
- ▸ **Spearman rank correlation coefficient**

REFERENCES

Galton, F.: Co-relations and their measurement, chiefly from anthropological data. Proc. Roy. Soc. Lond. **45**, 135–145 (1888)

Galton, F.: Natural Inheritance. Macmillan, London (1889)

Galton, F.: Kinship and correlation. North Am. Rev. **150**, 419–431 (1890)

Galton, F.: Memories of My Life. Methuen, London (1908)

Mosteller, F., Tukey, J.W.: Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA (1977)

Mudholkar, G.S.: Multiple correlation coefficient. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 5. Wiley, New York (1982)

Pearson, K.: Studies in the history of statistics and probability. Biometrika **13**, 25–45 (1920). Reprinted in: Pearson, E.S., Kendall, M.G. (eds.) Studies in the History of Statistics and Probability, vol. I. Griffin, London

Porter, T.M.: The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton, NJ (1986)

Stigler, S.: Francis Galton's account of the invention of correlation. Stat. Sci. **4**, 73–79 (1989)

Yule, G.U.: On the theory of correlation. J. Roy. Stat. Soc. **60**, 812–854 (1897)

Yule, G.U. (1926) Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. J. Roy. Stat. Soc. (**2**) 89, 1–64

Correspondence Analysis

Correspondence analysis is a **data analysis** technique that is used to describe **contingency tables** (or crossed tables). This analysis takes the form of a **graphical representation** of the associations and the "correspondence" between rows and columns.

HISTORY

The theoretical principles of correspondence analysis date back to the works of Hartley, H.O. (1935) (published under his original name Hirschfeld) and of **Fisher, R.A.** (1940) on **contingency tables**. They were first presented in the framework of inferential statistics.

The term "correspondence analysis" first appeared in the autumn of 1962, and the first presentation of this method that referred to

this term was given by Benzécri, J.P. in the winter of 1963. In 1976 the works of Benzécri, J.P., which retraced twelve years of his laboratory work, were published, and since then the algebraic and geometrical properties of this descriptive analytical tool have become more widely known and used.

MATHEMATICAL ASPECTS

Consider a contingency table relating to two **categorial qualitative variables** X and Y that have, respectively, r and c **categories**:

	Y_1	Y_2	...	Y_c	Total
X_1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
X_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	$n_{..}$

where

- n_{ij} represents the **frequency** that category i of **variable** X and category j of variable Y is observed,
- $n_{i.}$ represents the sum of the observed frequencies for category i of variable X,
- $n_{.j}$ represents the sum of the observed frequencies for category j of variable Y,
- $n_{..}$ represents the total number of **observations**.

We will assume that $r \geq c$; if not we take the **transpose** of the initial table and use this transpose as the new **contingency table**. The correspondence analysis of a contingency table with more lines than columns, is performed as follows:

1. Tables of row profiles X_I and column profiles X_J are constructed..
   For a fixed line (column), the line (column) profile is the line (column) obtained by dividing each element in this row (column) by the sum of the elements in the line (column).
   The line profile of row i is obtained by dividing each term of row i by $n_{i.}$, which is the sum of the observed **frequencies** in the row.
   The table of row profiles is constructed by replacing each row of the **contingency table** with its profile:

	Y_1	Y_2	...	Y_c	Total
X_1	$\frac{n_{11}}{n_{1.}}$	$\frac{n_{12}}{n_{1.}}$	...	$\frac{n_{1c}}{n_{1.}}$	1
X_2	$\frac{n_{21}}{n_{2.}}$	$\frac{n_{22}}{n_{2.}}$	...	$\frac{n_{2c}}{n_{2.}}$	1
...	...	...	...	...	...
X_r	$\frac{n_{r1}}{n_{r.}}$	$\frac{n_{r2}}{n_{r.}}$	...	$\frac{n_{rc}}{n_{r.}}$	1
Total	$n'_{.1}$	$n'_{.2}$	...	$n'_{.c}$	r

   It is also common to multiply each element of the table by 100 in order to convert the terms into percentages and to make the sum of terms in each row 100%.
   The column profile matrix is constructed in a similar way, but this time each column of the **contingency table** is replaced with its profile: the column profile of column j is obtained by dividing each term of column j by $n_{.j}$, which is the sum of **frequencies** observed for the category corresponding to this column.

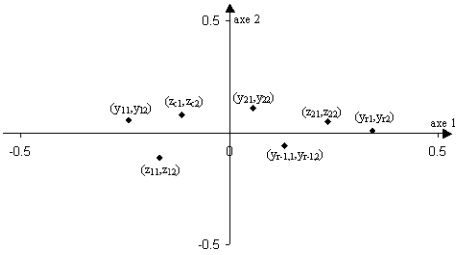
	Y_1	Y_2	...	Y_c	Total
X_1	$\frac{n_{11}}{n_{.1}}$	$\frac{n_{12}}{n_{.2}}$	...	$\frac{n_{1c}}{n_{.c}}$	$n'_{1.}$
X_2	$\frac{n_{21}}{n_{.1}}$	$\frac{n_{22}}{n_{.2}}$	...	$\frac{n_{2c}}{n_{.c}}$	$n'_{2.}$
...	...	...	...	...	...
X_r	$\frac{n_{r1}}{n_{.1}}$	$\frac{n_{r2}}{n_{.2}}$	...	$\frac{n_{rc}}{n_{.c}}$	$n'_{r.}$
Total	1	1	...	1	c

   The tables of row profiles and column profiles correspond to a transformation of the **contingency table** that is used to make the rows and columns comparable.

2. Determine the **inertia matrix** V.
   This is done in the following way:
   - The weighted mean of the r column coordinates is calculated:
   $$g_j = \sum_{i=1}^{r} \frac{n_{i.}}{n_{..}} \cdot \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}}, j = 1, \ldots, c\,.$$
   - The c obtained values g_j are written r times in the rows of a matrix G;
   - The diagonal matrix D_I is constructed with diagonal elements of $\frac{n_{i.}}{n_{..}}$;
   - Finally, the **inertia matrix** V is calculated using the following formula:
   $$V = (X_I - G)' \cdot D_I \cdot (X_I - G)\,.$$
3. Using the **matrix** M, which consists of $\frac{n_{..}}{n_{.j}}$ terms on its diagonal and zero terms elsewhere, we determine the matrix C:
   $$C = \sqrt{M} \cdot V \cdot \sqrt{M}\,.$$
4. Find the **eigenvalues** (denoted k_l) and the **eigenvectors** (denoted v_l) of this matrix C.
   The c **eigenvalues** $k_c, k_{c-1}, \ldots, k_1$ (written in decreasing order) are the **inertia**. The corresponding **eigenvectors** are called the **factorial axes** (or axes of inertia).
   For each **eigenvalue**, we calculate the corresponding inertia explained by the **factorial axis**. For example, the first factorial axis explains:
   $$\frac{100 \cdot k_1}{\sum_{l=1}^{c} k_l} \quad \text{(in \%) of } \textbf{inertia.}$$
   In the same way, the two first **factorial axes** explain:
   $$\frac{100 \cdot (k_1 + k_2)}{\sum_{l=1}^{c} k_l} \quad \text{(in \%) of } \textbf{inertia.}$$
   If we want to know, for example, the number of **eigenvalues** and therefore the factorial axes that explain at least 3/4 of the total inertia, we sum the explained inertia from each of the eigenvalues until we obtain 75%.
   We then calculate the main axes of inertia from these factorial axes.
5. The main axes of **inertia**, denoted u_l, are then given by:
   $$u_l = \sqrt{M^{-1}} \cdot v_l\,,$$
   meaning that its jth component is:
   $$u_{jl} = \sqrt{\frac{n_{.j}}{n_{..}}} \cdot v_{jl}\,.$$
6. We then calculate the main components, denoted by y_k, which are the orthogonal projections of the row coordinates on the main axes of inertia: the ith **coordinate** of the lth main component takes the following **value**:
   $$y_{il} = x_i \cdot M \cdot u_l\,,$$
   meaning that
   $$y_{il} = \sum_{j=1}^{c} \frac{n_{ij}}{n_{i.}} \sqrt{\frac{n_{..}}{n_{.j}}} \cdot v_{jl}$$
   is the **coordinate** of row i on the lth axis.
7. After the main components y_l (of the row coordinates) have been calculated, we determine the main components of the column coordinates (denoted by z_l) using the y_l, thanks to the transaction formulae:
   $$z_{jl} = \frac{1}{\sqrt{k_l}} \sum_{i=1}^{r} \frac{n_{ij}}{n_{.j}} \cdot y_{il}\,, \quad \text{for } j = 1, 2, \ldots, c;$$
   $$y_{il} = \frac{1}{\sqrt{k_l}} \sum_{i=1}^{c} \frac{n_{ij}}{n_{i.}} \cdot z_{jl}\,, \quad \text{for } i = 1, 2, \ldots, r\,.$$

z_{jl} is the **coordinate** of the column j on the lth **factorial axis**.

8. The simultaneous representation of the row coordinates and the column coordinates on a **scatter plot** with two **factorial axes**, gives a signification to the axis depending on the points it is related to.
   The quality of the representation is related to the proportion of the inertia explained by the two main axes used. The closer the explained **inertia** is to 1, the better the quality.



9. It can be useful to insert additional point-rows or column coordinates (illustrative **variables**) along with the active variables used for the correspondence analysis.
   Consider an extra row-coordinate:

$$(n_{s1}, n_{s2}, n_{s3}, \ldots, n_{sc})$$

of profile-row

$$\left(\frac{n_{s1}}{n_{s.}}, \frac{n_{s2}}{n_{s.}}, \ldots, \frac{n_{sc}}{n_{s.}}\right).$$

The lth main component of the extra row-dot is given by the second formula of point 7):

$$y_{il} = \frac{1}{\sqrt{k_l}} \sum_{j=1}^{c} \frac{n_{sj}}{n_{s.}} \cdot z_{jl},$$

$$\text{for } i = 1, 2, \ldots, r.$$

We proceed the same way for a column coordinate, but apply the first formula of point 7 instead.

DOMAINS AND LIMITATIONS

When simultaneously representing the row coordinate and the column coordinate, it is not advisable to interpret eventual proximities crossed between lines and columns because the two points are not in the same initial space. On the other hand it is interesting to interpret the position of a row coordinate by comparing it to the set of column coordinate (and vice versa).

EXAMPLES

Consider the example of a company that wants to find out how healthy its staff. One of the subjects covered in the questionnaire concerns the number of medical visits (dentists included) per year. Three possible answers are proposed:

- Between 0 and 6 visits;
- Between 7 and 12 visits;
- More than 12 visits per year.

The staff questioned are distributed into five **categories**:

- Managerial staff members over 40;
- Managerial staff members under 40;
- Employees over 40;
- Employees under 40;
- Office personnel.

The results are reported in the following **contingency table**, to which margins have been added:

	Number of visits per year			
	0 to 6	7 to 12	> 12	Total
Managerial staff members > 40	5	7	3	15
Managerial staff members < 40	5	5	2	12

	Number of visits per year			
	0 to 6	7 to 12	> 12	Total
Employees > 40	2	12	6	20
Employees < 40	24	18	12	54
Office personnel	12	8	4	24
Total	48	50	27	125

Using these **data** we will describe the eight main steps that should be followed to obtain a **graphical representation** of employee health via correspondence analysis.

1. We first determine the table of line profiles matrix X_I, by dividing each element by the sum of the elements of the line in which it is located:

$$\begin{array}{cccc} 0.333 & 0.467 & 0.2 & 1 \\ 0.417 & 0.417 & 0.167 & 1 \\ 0.1 & 0.6 & 0.3 & 1 \\ 0.444 & 0.333 & 0.222 & 1 \\ 0.5 & 0.333 & 0.167 & 1 \\ 1.794 & 2.15 & 1.056 & 5 \end{array}$$

and the table of column profiles by dividing each element by the sum of the elements in the corresponding column:

0.104	0.14	0.111	0.355
0.104	0.1	0.074	0.278
0.042	0.24	0.222	0.504
0.5	0.36	0.444	1.304
0.25	0.16	0.148	0.558
1	1	1	3

2. We then calculate the matrix of inertia V for the **frequencies** (given by $\frac{n_{i.}}{n_{..}}$ for values of i from 1 to 5) by proceeding in the following way:

- We start by calculating the weighted **mean** of the five line-dots:

$$g_j = \sum_{i=1}^{5} \frac{n_{i.}}{n_{..}} \cdot \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}}$$

for $j = 1, 2$ and 3;

- We then write these three values five times in the **matrix** G, as shown below:

$$G = \begin{bmatrix} 0.384 & 0.400 & 0.216 \\ 0.384 & 0.400 & 0.216 \\ 0.384 & 0.400 & 0.216 \\ 0.384 & 0.400 & 0.216 \\ 0.384 & 0.400 & 0.216 \end{bmatrix}.$$

- We then construct a diagonal **matrix** D_I containing the $\frac{n_{i.}}{n_{..}}$;
- Finally, we calculate the **matrix of inertia** V, as given by the following formula:

$$V = (X_I - G)' \cdot D_I \cdot (X_I - G),$$

which, in this case, gives:

$$V = \begin{bmatrix} 0.0175 & -0.0127 & -0.0048 \\ -0.0127 & 0.0097 & 0.0029 \\ -0.0048 & 0.0029 & 0.0019 \end{bmatrix}.$$

3. We define a third-order square **matrix** M that contains $\frac{n_{..}}{n_{.j}}$ terms on its diagonal and zero terms everywhere else:

$$M = \begin{bmatrix} 2.604 & 0 & 0 \\ 0 & 2.5 & 0 \\ 0 & 0 & 4.630 \end{bmatrix}.$$

The square root of M, denoted $\sqrt{M}$, is obtained by taking the square root of each diagonal element of M. Using this new **matrix**, we determine

$$C = \sqrt{M} \cdot V \cdot \sqrt{M}$$

$$C = \begin{bmatrix} 0.0455 & -0.0323 & -0.0167 \\ -0.0323 & 0.0243 & 0.0100 \\ -0.0167 & 0.0100 & 0.0087 \end{bmatrix}.$$

4. The **eigenvalues** of C are obtained by diagonalizing the **matrix**. Arranging these values in decreasing order gives:

$$k_1 = 0.0746$$
$$k_2 = 0.0039$$
$$k_3 = 0\,.$$

The explained inertia is determined for each of these values. For example:

$$k_1 : \frac{0.0746}{0.0746 + 0.0039} = 95.03\%\,,$$

$$k_2 : \frac{0.0039}{0.0746 + 0.0039} = 4.97\%\,.$$

For the last one, k_3, the explained inertia is zero.
The first two **factorial axes**, associated with the **eigenvalues** k_1 and k_2, explain all of the inertia. Since the third eigenvalue is zero it is not necessary to calculate the **eigenvector** that is associated with it. We focus on calculating the first two normalized eigenvectors then:

$$v_1 = \begin{bmatrix} 0.7807 \\ -0.5576 \\ -0.2821 \end{bmatrix} \quad \text{and}$$

$$v_2 = \begin{bmatrix} 0.0807 \\ 0.5377 \\ -0.8393 \end{bmatrix}\,.$$

5. We then calculate the main **axes** of inertia by:

$$u_i = \sqrt{M^{-1}} \cdot v_i\,, \quad \text{for} \quad i = 1 \text{ and } 2\,,$$

where $\sqrt{M^{-1}}$ is obtained by inverting the diagonal elements of $\sqrt{M}$.

We find:

$$u_1 = \begin{bmatrix} 0.4838 \\ -0.3527 \\ -0.1311 \end{bmatrix} \text{ and}$$

$$u_2 = \begin{bmatrix} 0.0500 \\ 0.3401 \\ -0.3901 \end{bmatrix}\,.$$

6. We then calculate the main components by projecting the rows onto the main axes of inertia. Constructing an auxiliary **matrix** U formed from the two **vectors** u_1 and u_2, we define:

$$Y = X_I \cdot M \cdot U$$
$$= \begin{bmatrix} -0.1129 & 0.0790 \\ 0.0564 & 0.1075 \\ -0.5851 & 0.0186 \\ 0.1311 & -0.0600 \\ 0.2349 & 0.0475 \end{bmatrix}\,.$$

We can see the coordinates of the five rows written horizontally in this **matrix** Y; for example, the first column indicates the components related to the first **factorial** axis and the second indicates the components related to the second axis.
7. We then use the coordinates of the rows to find those of the columns via the following transition formulae written as matrices:

$$Z = K \cdot Y' \cdot X_J\,,$$

where:

- $\boldsymbol{K}$ is the second-order diagonal **matrix** that has $1/\sqrt{k_i}$ terms on its diagonal and zero terms elsewhere;
- $\boldsymbol{Y'}$ is the transpose of the matrix containing the coordinates of the rows, and;
- $\boldsymbol{X_J}$ is the column profile matrix.

We obtain the matrix:

$$Z = \begin{bmatrix} 0.3442 & -0.2409 & -0.1658 \\ 0.0081 & 0.0531 & -0.1128 \end{bmatrix},$$

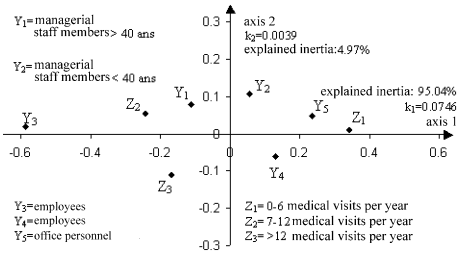
where each column contains the components of one of the three column coordinates; for example, the first line corresponds to each coordinate on the first **factorial axis** and the second line to each coordinate on the second axis.

We can verify the transition formula that gives Y from Z:

$$Y = X_I \cdot Z' \cdot K$$

using the same notation as seen previously.

8. We can now represent the five **categories** of people questioned and the three categories of answers proposed on the same factorial plot:



We can study this factorial plot at three different levels of analysis, depending:

- The set of **categories** for the people questioned;
- The set of modalities for the medical visits;
- Both at the same time.

In the first case, close proximity between two rows (between two **categories** of personnel) signifies similar medical visit profiles. On the factorial plot, this is the case for the employees under 40 (Y_4) and the office personnel (Y_5). We can verify from the table of line profiles that the **percentages** for these two categories are indeed very similar.

Similarly, the proximity between two columns (representing two **categories** related to the number of medical visits) indicates similar distributions of people within the business for these categories. This can be seen for the modalities Z_2 (from 7 to 12 medical visits per year) and Z_3 (more than 12 visits per year).

If we consider the rows and the columns simultaneously (and not separately as we did previously), it becomes possible to identify similarities between categories for certain modalities. For example, the employees under 40 (Y_4) and the office personnel (Y_5) seem to have the same behavior towards health: high proportions of them (0.44 and 0.5 respectively) go to the doctor less than 6 times per year (Z_1).

In conclusion, axis 1 is confronted on one side with the categories indicating an average or high number of visits (Z_2 and Z_3)—the employees or the managerial staff members over 40 (Y_1 and Y_3)—and on the other side with the modalities associated with low numbers of visits (Z_1): Y_2, Y_4 and Y_5. The first factor can then be interpreted as the importance of medical control according to age.

FURTHER READING

- ▸ **Contingency table**
- ▸ **Data analysis**
- ▸ **Eigenvalue**
- ▸ **Eigenvector**
- ▸ **Factorial axis**
- ▸ **Graphical representation**
- ▸ **Inertia matrix**
- ▸ **Matrix**
- ▸ **Scatterplot**

REFERENCES

Benzécri, J.P.: L'Analyse des données. Vol. 1: La Taxinomie. Vol. 2: L'Analyse factorielle des correspondances. Dunod, Paris (1976)

Benzécri, J.P.: Histoire et préhistoire de l'analyse des données, les cahiers de l'analyse des données 1, no. 1–4. Dunod, Paris (1976)

Fisher, R.A.: The precision of discriminant functions. Ann. Eugen. (London) **10**, 422–429 (1940)

Greenacre, M.: Theory and Applications of Correspondence Analysis. Academic, London (1984)

Hirschfeld, H.O.: A connection between correlation and contingency. Proc. Camb. Philos. Soc. **31**, 520–524 (1935)

Lebart, L., Salem, A.: Analyse Statistique des Données Textuelles. Dunod, Paris (1988)

Saporta, G.: Probabilité, analyse de données et statistiques. Technip, Paris (1990)

Covariance

The covariance between two **random variables** X and Y is the measure of how much two random variables vary together.

If X and Y are independent **random variables**, the covariance of X and Y is zero. The converse, however, is not true.

MATHEMATICAL ASPECTS

Consider X and Y, two **random variables** defined in the same **sample space** Ω. The covariance of X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined by

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]\,,$$

where $E[.]$ is the **expected value**.

Developing the right side of the equation gives:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY - E[X]Y - XE[Y] \\ &\quad + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] \\ &\quad - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\,.\end{aligned}$$

Properties of Covariance

Consider X, Y and Z, which are **random variables** defined in the same **sample space** Ω, and a, b, c and d, which are constants. We find that:

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
2. $\mathrm{Cov}(X, c) = 0$
3. $\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$
4. $\mathrm{Cov}(X, cY + dZ) = c\,\mathrm{Cov}(X, Y) + d\mathrm{Cov}(X, Z)$
5. $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$.

Consequences of the Definition

1. If X and Y are independent **random variables**,
$$\mathrm{Cov}(X, Y) = 0\,.$$
In fact $E[XY] = E[X]E[Y]$, meaning that:
$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0\,.$$
The reverse is not generally true: $\mathrm{Cov}(X, Y) = 0$ does not necessarily imply that X and Y are independent.
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
where $\mathrm{Var}(X)$ represents the **variance** of X.
In fact:
$$\begin{aligned}\mathrm{Cov}(X, X) &= E[XX] - E[X]E[X] \\ &= E[X^2] - (E[X])^2 \\ &= \mathrm{Var}(X)\,.\end{aligned}$$

DOMAINS AND LIMITATIONS

Consider two **random variables** X and Y, and their sum $X + Y$. We then have:

$$E[X + Y] = E[X] + E[Y] \quad \text{and}$$
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) .$$

We now show these results for discrete variables. If $P_{ji} = P(X = x_i, Y = y_j)$ we have:

$$\begin{aligned} E[X + Y] &= \sum_i \sum_j (x_i + y_j) P_{ji} . \\ &= \sum_i \sum_j x_i P_{ji} + \sum_i \sum_j y_j P_{ji} \\ &= \sum_i x_i \left(\sum_j P_{ji} \right) \\ &\quad + \sum_j y_j \left(\sum_i P_{ji} \right) \\ &= \sum_i x_i P_i + \sum_j y_j P_j \\ &= E[X] + E[Y] . \end{aligned}$$

Moreover:

$$\begin{aligned} &\text{Var}(X + Y) = E[(X + Y)^2] - (E[X + Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &\quad - (E[X + Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &\quad - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &\quad - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2 \\ &= \text{Var}(X) + \text{Var}(Y) + 2(E[XY] \\ &\quad - E[X]E[Y]) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) . \end{aligned}$$

These results can be generalized for n **random variables** $X_1, X_2, \ldots, X_n$, with x_i having an **expected value** equal to $E(X_i)$ and a **variance** equal to $\text{Var}(X_i)$. We then have:

$$\begin{aligned} &E[X_1 + X_2 + \cdots + X_n] \\ &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= \sum_{i=1}^{n} E[X_i] . \\ &\text{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) \\ &\quad + 2[\text{Cov}(X_1, X_2) + \cdots + \text{Cov}(X_1, X_n) \\ &\quad + \text{Cov}(X_2, X_3) + \cdots + \text{Cov}(X_2, X_n) \\ &\quad + \cdots + \text{Cov}(X_{n-1}, X_n)] \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j>i} \text{Cov}(X_i, X_j) . \end{aligned}$$

EXAMPLES

Consider two psychological tests carried out in succession. Each subject receives a grade X of between 0 and 3 for the first test and a grade Y of between 0 and 2 for the second test. Given that the **probabilities** of X being equal to 0, 1, 2 and 3 are respectively 0.16, 0.3, 0.41 and 0.13, and that the probabilities of Y being equal to 0, 1 and 2 are respectively 0.55, 0.32 and 0.13, we have:

$$\begin{aligned} E[X] &= 0 \cdot 0.16 + 1 \cdot 0.3 + 2 \cdot 0.41 \\ &\quad + 3 \cdot 0.13 \\ &= 1.51 \\ E[Y] &= 0 \cdot 0.55 + 1 \cdot 0.32 + 2 \cdot 0.13 \\ &= 0.58 \\ E[XY] &= 0 \cdot 0 \cdot 0.16 \cdot 0.55 \\ &\quad + 0 \cdot 1 \cdot 0.16 \cdot 0.32 + \ldots \\ &\quad + 3 \cdot 2 \cdot 0.13 \cdot 0.13 \\ &= 0.88 . \end{aligned}$$

We can then calculate the covariance of X and Y:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= 0.88 - (1.51 \cdot 0.58) \\ &= 0.00428\,.\end{aligned}$$

FURTHER READING

- ▸ Correlation coefficient
- ▸ Expected value
- ▸ Random variable
- ▸ Variance of a random variable

Covariance Analysis

Covariance analysis is a method used to estimate and test the effects of **treatments**. It checks whether there is a significant difference between the **means** of several treatments by taking into account the observed **values** of the **variable** before the treatment. **Covariance analysis** is a precise way of performing **treatment** comparisons because it involves adjusting the response variable Y to a concomitant variable X which corresponds to the **values** observed before the treatment.

HISTORY

Covariance analysis dates back to 1930. It was first developed by **Fisher, R.A.** (1932). After that, other authors applied covariance analysis to agricultural and medical problems. For example, Bartlett, M.S. (1937) applied covariance analysis to his studies on cotton cultivation in Egypt and on milk yields from cows in winter.

Delurry, D.B. (1948) used covariance analysis to compare the effects of different medications (atropine, quinidine, atrophine) on rat muscles.

MATHEMATICAL ASPECTS

We consider here a covariance analysis of a **completely randomized design** implying one **factor**.

The linear **model** that we will consider is the following:

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

where

- Y_{ij} represents **observation** j, receiving **treatment** i,
- μ is the general **mean** common to all treatments,
- τ_i is the actual effect of treatment i on th observation,
- X_{ij} is the **value** of the concomitant variable, and
- ε_{ij} is the experimental error in observation Y_{ij}.

Calculations

In order to calculate the F ratio that will help us to determine whether there is a significant difference between **treatments**, we need to work out sums of squares and sums of products. Therefore, if $\bar{X}_{i.}$ and $\bar{Y}_{i.}$ are respectively the **means** of the X values and the Y values for treatment i, and if $\bar{X}_{..}$ and $\bar{Y}_{..}$ are respectively the means of all the **values** of X and Y, we obtain the formulae given below.

1. The total sum of squares for X:

$$S_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2\,.$$

2. The total sum of squares for Y (S_{YY}) is calculated in the same way as S_{XX}, but X is replaced by Y.

3. The total sum of products of X and Y:

$$S_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) .$$

4. The sum of squares of the treatments for X:

$$T_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 .$$

5. The sum of squares of the treatments for Y (T_{YY}) is calculated in the same way as T_{XX}, but X is replaced by Y.
6. The sum of the products of the treatments of X and Y:

$$T_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..}) .$$

7. The sum of squares of the **errors** for X:

$$E_{XX} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 .$$

8. The sum of the squares of the **errors** for Y: is calculated in the same way as E_{XX}, but X is replaced by Y.
9. The sum of products of the **errors** X and Y:

$$E_{XY} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.}) .$$

Substituting in appropriate values and calculating these formulae corresponds to an **analysis of variance** for each of X, Y and XY. The **degrees of freedom** associated with these different formulae are as follows:

1. For the total sum: $\sum_{i=1}^{t} n_i - 1$.
2. For the sum of the **treatments**: $t - 1$.
3. For the sum of the **errors**: $\sum_{i=1}^{t} n_i - t$.

Adjustment of the **variable** Y to the concomitant variable X yields two new sums of squares:

1. The adjusted total sum of squares:

$$SS_{\text{tot}} = S_{YY} - \frac{S_{XY}^2}{S_{XX}} .$$

2. The adjusted sum of the squares of the errors:

$$SS_{\text{err}} = E_{YY} - \frac{E_{XY}^2}{E_{XX}} .$$

The new **degrees of freedom** for these two sums are:

1. $\sum_{i=1}^{t} n_i - 2$;
2. $\sum_{i=1}^{t} n_i - t - 1$, where a **degree of freedom** is subtracted due to the adjustment.

The third adjusted sum of squares, the adjusted sum of the squares of the **treatments**, is given by:

$$SS_{\text{tr}} = SS_{\text{tot}} - SS_{\text{err}} .$$

This has the same number of **degrees of freedom** $t - 1$ as before.

Covariance Analysis Table

We now have all of the elements needed to establish the covariance analysis table. The sums of squares divided by the number of **degrees of freedom** gives the **means** of the squares.

Source of variation	Degrees of freedom	Sum of squares and of products $\sum_{i=1}^{t} x_i^2$	$\sum_{i=1}^{t} x_i y_i$	$\sum_{i=1}^{t} y_i^2$
Treatments	$t-1$	T_{XX}	T_{XY}	T_{YY}
Errors	$\sum_{i=1}^{t} n_i - t$	E_{XX}	E_{XY}	E_{YY}
Total	$\sum_{i=1}^{t} n_i - 1$	S_{XX}	S_{XY}	S_{YY}

Note: the numbers in the $\sum x_i^2$ and $\sum y_i^2$ columns cannot be negative; on the other hand, the numbers in the $\sum_{i=1}^{t} x_i y_i$ column can be negative.

Adjustment

Source of variation	Degrees of freedom	Sum of squares	Mean of squares
Treatments	$t-1$	SS_{tr}	MC_{tr}
Errors	$\sum_{i=1}^{t} n_i - t - 1$	SS_{err}	MC_{err}
Total	$\sum_{i=1}^{t} n_i - 2$	TSS	

F Ratio: Testing the Treatments

The *F* **ratio,** used to test the **null hypothesis** that there is a significant difference between the **means** of the **treatments** once adjusted to the variable X, is given by:

$$F = \frac{MC_{tr}}{MC_{err}}.$$

The ratio follows a **Fisher distribution** with $t-1$ and $\sum_{i=1}^{t} n_i - t - 1$ **degrees of freedom**. The **null hypothesis**

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

will be rejected at the significance level α if the F ratio is superior or equal to the **value** of the **Fisher table**, in other words if:

$$F \geq F_{t-1, \sum_{i=1}^{t} n_i - t - 1, \alpha}.$$

It is clear that we assume that the β coefficient is different from zero when performing covariance analysis. If this is not the case, a simple **analysis of variance** is sufficient.

Test Concerning the β Slope

So, we would like to know whether there is a significant effect of the concomitant variables before the application of the treatment. To test this **hypothesis**, we will assume the **null hypothesis** to be

$$H_0 : \beta = 0$$

and the **alternative hypothesis** to be

$$H_1 : \beta \neq 0.$$

The F ratio can be established:

$$F = \frac{E_{XY}^2 / E_{XX}}{MC_{err}}.$$

It follows a **Fisher distribution** with 1 and $\sum_{i=1}^{t} n_i - t - 1$ **degrees of freedom**. The **null hypothesis** will be rejected at the significance level α if the F ratio is superior to or equal to the value in the **Fisher table**, in other words if:

$$F \geq F_{1, \sum_{i=1}^{t} n_i - t - 1, \alpha}.$$

DOMAINS AND LIMITATIONS

The basic **hypotheses** that need to be constructed before initiating a covariance analysis are the same as those used for an **analysis of variance** or a **regression analysis**. These are the hypotheses of normality, homogeneity (homoscedasticity), **variances** and independence.

In covariance analysis, as in an analysis of variance, the **null hypothesis** stipulates that the independent samples come from different **populations** that have identical **means**. Moreover, since there are always conditions associated with any statistical technique, those that apply to covariance analysis are as follows:

1. The **population** distributions must be approximately normal, if not completely normal.

2. The populations from which the **samples** are taken must have the same **variance** σ^2, meaning:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 ,$$

where k is the number of populations to be compared.

3. The samples must be chosen randomly and all of the samples must be independent.

We must also add a basic **hypothesis** specific to covariance analysis, which is that the **treatments** that were carried out must not influence the **values** of the concomitant **variable** X.

EXAMPLES

Consider an **experiment** consisting of comparing the effects of three different diets on a **population** of cows.

The **data** are presented in the form of a table, which includes the three different diets, each of which was administered to five porch. The initial weights are denoted by the concomitant variable X (in kg), and the gains in weight (after **treatment**) are denoted by Y:

Diets

1		2		3	
X	Y	X	Y	X	Y
32	167	26	182	36	158
29	172	33	171	34	191
22	132	22	173	37	140
23	158	28	163	37	192
35	169	22	182	32	162

We first calculate the various sums of squares and products:

1. The total sum of squares for X:

$$S_{XX} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_{..})^2$$
$$= (32 - 29.8667)^2 + \ldots + (32 - 29.8667)^2$$
$$= 453.73 .$$

2. The total sum of squares for Y:

$$S_{YY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(Y_{ij} - \bar{Y}_{..})^2$$
$$= (167 - 167.4667)^2 + \ldots + (162 - 167.4667)^2$$
$$= 3885.73 .$$

3. The total sum of the products of X and Y:

$$S_{XY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..})$$
$$= (32 - 29.8667) \cdot (167 - 167.4667) + \ldots + (32 - 29.8667) \cdot (162 - 167.4667)$$
$$= 158.93 .$$

4. The sum of the squares of the treatments for X:

$$T_{XX} = \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{X}_{i.} - \bar{X}_{..})^2$$
$$= 5(28.2 - 29.8667)^2 + 5(26.2 - 29.8667)^2 + 5(35.2 - 29.8667)^2$$
$$= 223.33 .$$

5. The sum of the squares of the treatments for Y:

$$T_{YY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$= 5(159.6 - 167.4667)^2 + 5(174.2 - 167.4667)^2 + 5(168.6 - 167.4667)^2 = 542.53\,.$$

6. The sum of the products of the treatments of X and Y:

$$T_{XY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..}) = 5(28.2 - 29.8667) \cdot (159.6 - 167.4667) + \ldots + 5(35.2 - 29.8667) \cdot (168.6 - 167.4667) = -27.67\,.$$

7. The sum of the squares of the errors for X:

$$E_{XX} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_{i.})^2 = (32 - 28.2)^2 + \ldots + (32 - 35.2)^2 = 230.40\,.$$

8. The sum of the squares of the errors for Y:

$$E_{YY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(Y_{ij} - \bar{Y}_{i.})^2 = (167 - 159.6)^2 + \ldots + (162 - 168.6)^2 = 3343.20\,.$$

9. The sum of the products of the errors of X and Y:

$$E_{XY} = \sum_{i=1}^{3}\sum_{j=1}^{5}(X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.}) = (32 - 28.2)(167 - 159.6) + \ldots + (32 - 35.2)(162 - 168.6) = 186.60\,.$$

The **degrees of freedom** associated with these different calculations are as follows:

1. For the total sums:

$$\sum_{i=1}^{3} n_i - 1 = 15 - 1 = 14\,.$$

2. For the sums of **treatments**:

$$t - 1 = 3 - 1 = 2\,.$$

3. For the sums of **errors**:

$$\sum_{i=1}^{3} n_i - t = 15 - 3 = 12\,.$$

Adjusting the **variable** Y to the concomitant variable X yields two new sums of squares:

1. The total adjusted sum of squares:

$$SS_{\text{tot}} = S_{YY} - \frac{S_{XY}^2}{S_{XX}} = 3885.73 - \frac{158.93^2}{453.73} = 3830.06\,.$$

2. The adjusted sum of the squares of the errors:

$$SS_{\text{err}} = E_{YY} - \frac{E_{XY}^2}{E_{XX}} = 3343.20 - \frac{186.60^2}{230.40} = 3192.07\,.$$

The new **degrees of freedom** for these two sums are:

1. $\sum_{i=1}^{3} n_i - 2 = 15 - 2 = 13$;
2. $\sum_{i=1}^{3} n_i - t - 1 = 15 - 3 - 1 = 11$.

The adjusted sum of the squares of the **treatments** is given by:

$$SS_{\text{tr}} = SS_{\text{tot}} - SS_{\text{err}} = 3830.06 - 3192.07 = 637.99\,.$$

This has the same number of **degrees of freedom** as before:

$$t - 1 = 3 - 1 = 2\,.$$

We now have all of the elements required in order to establish a covariance analysis table. The sums of squares divided by the **degrees of freedom** gives the **means** of the squares.

Source of variation	Degrees of freedom	Sum of squares and of products		
		$\sum_{i=1}^{3} x_i^2$	$\sum_{i=1}^{3} x_i y_i$	$\sum_{i=1}^{3} y_i^2$
Treatments	2	223.33	−27.67	543.53
Errors	12	230.40	186.60	3343.20
Total	14	453.73	158.93	3885.73

Adjustment

Source of variation	Degrees of freedom	Sum of squares	Mean of squares
Treatments	2	637.99	318.995
Errors	11	3192.07	290.188
Total	13	3830.06	

The F **ratio**, which is used to test the **null hypothesis** that there is no significant difference between the **means** of the **treatments** once adjusted to the variable Y, is given by:

$$F = \frac{MC_{\text{tr}}}{MC_{\text{err}}} = \frac{318.995}{290.188} = 1.099\,.$$

If we choose a significance level of $\alpha = 0.05$, the value of F in the **Fisher table** is equal to:

$$F_{2,11,0.05} = 3.98\,.$$

Since $F < F_{2,11,0.05}$, we cannot reject the **null hypothesis** and so we conclude that there is no significant difference between the responses to the three diets once the **variable** Y is adjusted to the initial weight X.

FURTHER READING

- ▸ Analysis of variance
- ▸ Design of experiments
- ▸ Missing data analysis
- ▸ Regression analysis

C

REFERENCES

Bartlett, M.S.: Some examples of statistical methods of research in agriculture and applied biology. J. Roy. Stat. Soc. (Suppl.) **4**, 137–183 (1937)

DeLury, D.B.: The analysis of covariance. Biometrics **4**, 153–170 (1948)

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

Huitema, B.E.: The Analysis of Covariance and Alternatives. Wiley, New York (1980)

Wildt, A.R., Ahtola, O.: Analysis of Covariance (Sage University Papers Series on Quantitative Applications in the Social Sciences, Paper 12). Sage, Thousand Oaks, CA (1978)

Covariation

It is often interesting, particularly in economics, to compare two **time series**.

Since we wish to measure the level of **dependence** between two **variables**, this is somewhat reminiscent of the concept of correlation. However, in this case, since the **time series** are bound by a third variable, time, finding the **correlation coefficient** would only lead to an artificial relation.

Indeed, if two **time series** are considered, x_t and y_t, which represent completely independent phenomena and are linear functions

of time:

$$x_t = a \cdot t + b\,,$$
$$y_t = c \cdot t + d\,,$$

where a, b, c and d are constants, it is always possible to eliminate the time factor t between the two equations and to obtain a functional relation of the type $y = e \cdot x + f$. This relation states that there is a linear dependence between the two time series, which is not the case.

Therefore, measuring the correlation between the evolutions of two phenomena over time does not imply the existence of a link between them. The term covariation is therefore used instead of correlation, and this dependence is measured using a covariation coefficient. We can distinguish between:

- The linear covariation coefficient;
- The tendency covariation coefficient.

HISTORY

See **correlation coefficient** and **time series**.

MATHEMATICAL ASPECTS

In order to compare two **time series** y_t and x_t, the first step is to attempt to represent them on the same graphic.

However, visual comparison is generally difficult. The following change of **variables** is performed:

$$Y_t = \frac{y_t - \bar{y}}{S_y} \quad \text{and} \quad X_t = \frac{x_t - \bar{x}}{S_x}\,,$$

which are the centered and reduced variables where S_y and S_x are the **standard deviations** of the respective **time series**.

We can distinguish between the following covariation coefficients:

- *The linear covariation coefficient*
  The form of this expression is identical to the one for the **correlation coefficient** r, but here the calculations do not have the same grasp because the goal is to detect the eventual existence of relation between variations that are themselves related to time and to measure the order of magnitude

$$C = \frac{\sum_{t=1}^{n}(x_t - \bar{x}) \cdot (y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \cdot \sum_{t=1}^{n}(y_t - \bar{y})^2}}\,.$$

This yields values of between -1 and $+1$. If it is close to ± 1, there is a linear **relation** between the time evolutions of the two **variables**.

Notice that:

$$C = \frac{\sum_{t=1}^{n} X_t \cdot Y_t}{n}\,.$$

Here n is the number of **observations**, while Y_t and X_t are the centered and reduced series obtained by a change of **variable**, respectively.

- *The tendency covariation coefficient*
  The influence exerted by the **means** is eliminated by calculating:

$$K = \frac{\sum_{t=1}^{n}(x_t - T_{x_t}) \cdot (y_t - T_{y_t})}{\sqrt{\sum_{t=1}^{n}(x_t - T_{x_t})^2 \cdot \sum_{t=1}^{n}(y_t - T_{y_t})^2}}\,.$$

The means $\bar{x}$ and $\bar{y}$ have simply been replaced with the **values** of the **secular trends** T_{x_t} and T_{y_t} of each **time series**.

The tendency covariation coefficient K also takes values between -1 to $+1$, and the closer it gets to ± 1, the stronger the covariation between the time series.

DOMAINS AND LIMITATIONS

There are many examples of the need to compare two **time series** in economics: for example, when comparing the evolution of the price of a product to the evolution of the quantity of the product on the market, or the evolution of the national revenue to the evolution of real estate transactions. It is important to know whether there is some kind of dependence between the two phenomena that evolve over time: this is the goal of measuring the covariation.

Visually comparing two **time series** is an important operation, but this is often a difficult task because:

- The **data** undergoing comparison may come from very different domains and present very different orders of magnitude, so it is preferable to study the deviations from the **mean**.
- The peaks and troughs of two time series may have very different amplitudes; it is then preferable to homogenize the dispersions by linking the variations back to the **standard deviation** of the time series.

Visual comparison is simplified if we consider the centered and reduced **variables** obtained via the following variable changes:

$$Y_t = \frac{y_t - \bar{y}}{S_y} \quad \text{and} \quad X_t = \frac{x_t - \bar{x}}{S_x}.$$

Also, in a similar way to the **correlation coefficient**, nonlinear relations can exist between two variables that give a C **value** that is close to zero.

It is therefore important to be cautious during interpretation.

The tendency covariation coefficient is preferentially used when the relation between the time series is not linear.

EXAMPLES

Let us study the covariation between two **time series**.

The **variable** x_t represents the annual production of an agricultural product; the variable y_t is its average annual price per unit in constant euros.

t	x_t	y_t
1	320	5.3
2	660	3.2
3	300	2.2
4	190	3.4
5	320	2.7
6	240	3.5
7	360	2.0
8	170	2.5

$$\sum_{t=1}^{8} x_t = 2560\,, \qquad \sum_{t=1}^{8} y_t = 24.8\,,$$

giving $\bar{x} = 320$ and $\bar{y} = 3.1$.

$x_t - \bar{x}$	$(x_t - \bar{x})^2$	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
0	0	2.2	4.84
340	115600	0.1	0.01
−20	400	−0.9	0.81
−130	16900	0.3	0.09
0	0	−0.4	0.16
−80	6400	0.4	0.16
40	1600	−1.1	1.21
−150	22500	0.6	0.36

$$\sum_{t=1}^{8} (x_t - \bar{x})^2 = 163400\,,$$

$$\sum_{t=1}^{8} (y_t - \bar{y})^2 = 7.64\,,$$

giving $\sigma_x = 142.9$ and $\sigma_y = 0.98$.

The centered and reduced **values** X_t and Y_t are then calculated.

X_t	Y_t	$X_t \cdot Y_t$	$X_t \cdot Y_{t-1}$
0.00	2.25	0.00	5.36
2.38	0.10	0.24	−0.01
−0.14	−0.92	0.13	0.84
−0.91	0.31	−0.28	0.00
0.00	−0.41	0.00	0.23
−0.56	0.41	−0.23	0.11
0.28	−1.13	−0.32	1.18
−1.05	−0.61	0.64	

The linear covariation coefficient is then calculated:

$$C = \frac{\sum_{t=1}^{8} X_t \cdot Y_t}{8} = \frac{0.18}{8} = 0.0225\,.$$

If the **observations** X_t are compared with Y_{t-1}, meaning the production this year is compared with that of the previous year, we obtain:

$$C = \frac{\sum_{t=2}^{8} X_t \cdot Y_{t-1}}{8} = \frac{7.71}{8} = 0.964\,.$$

The linear covariation coefficient for a shift of one year is very strong. There is a strong (positive) covariation with a shift of one year between the two **variables**.

FURTHER READING

- ▸ **Correlation coefficient**
- ▸ **Moving average**
- ▸ **Secular trend**
- ▸ **Standard deviation**
- ▸ **Time series**

REFERENCES

Kendall, M.G.: Time Series. Griffin, London (1973)

Py, B.: Statistique Déscriptive. Economica, Paris (1987)

Cox, David R.

Cox, David R. was born in 1924 in Birmingham in England. He studied mathematics at the University of Cambridge and obtained his doctorate in applied mathematics at the University of Leeds in 1949. From 1966 to 1988, he was a professor of statistics at Imperial College London, and then from 1988 to 1994 he taught at Nuffield College, Oxford.

Cox, David is an eminent statistician. He was knighted by Queen Elizabeth II in 1982 in gratitude for his contributions to statistical science, and has been named Doctor Honoris Causa by many universities in England and elsewhere. From 1981 to 1983 he was President of the Royal Statistical Society; he was also President of the Bernoulli Society from 1972 to 1983 and President of the International Statistical Institute from 1995 to 1997.

Due to the variety of subjects that he has studied and developed, Professor Cox has had a profound impact in his field. He was named Doctor Honoris Causa of the University of Neuchâtel in 1992.

Cox, Sir David is the author and the coauthor of more then 250 articles and 16 books, and between 1966 and 1991 he was the editor of Biometrika.

Some principal works and articles of Cox, Sir David:

1964 (and Box, G.E.P.) An analysis of transformations (with discussion) J. Roy. Stat. Soc. Ser. B 26, 211–243.

1970 The Analysis of Binary Data. Methuen, London.

1973 (and Hinkley, D.V.) Theoretical Statistics. Chapman & Hall, London.

1974 (and Atkinson, A.C.) Planning experiments for discriminating between models. J. Roy. Stat. Soc. Ser. B 36, 321–348.

1978 (and Hinkley, D.V.) Problems and Solutions in Theoretical Statistics. Chapman & Hall, London.

1981 (and Snell, E.J.) Applied Statistics: Principles and Examples. Chapman & Hall.

1981 Theory and general principles in statistics. J. Roy. Stat. Soc. Ser. A 144, 289–297.

2000 (and Reid, N.) Theory of Design Experiments. Chapman & Hall, London.

Cox, Mary Gertrude

Cox, Gertrude Mary was born in 1900, near Dayton, Iowa, USA. Her ambition was to help people, and so she initially studied a social sciences course for two years. Then she worked in orphanage for young boys in Montana for two years. In order to become a director of the orphanage, she decided to continue her education at Iowa State College. She graduated from Iowa State College in 1929. To pay for her studies, Cox, Gertrude worked with **Snedecor, George Waddel**, her professor, in his statistical laboratory, which led to her becoming interested in statistics. After graduating, she started studying for a doctorate in psychology. In 1933, before finishing her doctorate, Snedecor, George, then the director of the Iowa State Statistical Laboratory, convinced her to become his assistant, which she agreed to, although she remained in the field of psychology because she worked on the evaluation of statistical test in psychology and the analysis of psychological data.

On 1st November 1940, she became the director of the Department of Experimental Statistics of the State of North Carolina.

In 1945, the General Education Board gave her permission to create an institute of statistics at the University of North Carolina, with a department of mathematical statistics at Chapel Hill.

She was a founder member of the Journal of the International Biometric Society in 1947, and she was a director of it from 1947 to 1955 and president of it from 1968 to 1969. She was president of the American Statistical Association (ASA) in 1956. She died in 1978.

Principal work of Cox, M. Gertrude:

1957 (and Cochran, W.) Experimental Designs, 2nd edn. Wiley, New York

C_p Criterion

The C_p citerion is a model selection citerion in linear regression. For a linear regression model with p parameters including any constant term, in the model, a rule of thumb is to select a model in which the value of C_p is close to the number of terms in the model.

HISTORY

Introduced by Mallows, Colin L. in 1964, the model selection criterion C_p has been used ever since as a criterion for evaluating the goodness of fit of a regression model.

MATHEMATICAL ASPECTS

Let $Y_i = \beta_0 + \sum_{j=1}^{p-1} X_{ji}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n$ be a **multiple linear regression** model. Denote the **mean square error** as $MSE(\widehat{y}_i)$. The criterion introduced in this section can be used to choose the model with the minimal sum of mean square errors:

$$\begin{aligned}\sum MSE\,(\widehat{y}_i) &= \sum \left(E\left(\widehat{y}_i - y_i\right)^2 \right. \\ &\quad \left. - \sigma^2\,(1 - 2h_{ii}) \right) \\ &= E\left(\sum \left(\widehat{y}_i - y_i\right)^2\right) \\ &\quad - \sigma^2 \sum (1 - 2h_{ii}) \\ &= E(\mathrm{RSS}) - \sigma^2(n - 2p)\,,\end{aligned}$$

where

- $\boldsymbol{n}$ is the number of observations,
- $\boldsymbol{p}$ is the number of estimated parameters,
- $\boldsymbol{h_{ii}}$ are the diagonal elements of the **hat matrix**, and
- $\boldsymbol{\widehat{y}_i}$ is the estimator of y_i.

Recall the following property of the h_{ii}:

$$\sum h_{ii} = p\,.$$

Define the coefficient

$$\begin{aligned}\Gamma_p &= \frac{\sum MSE\,(\widehat{y}_i)}{\sigma^2} \\ &= \frac{E\,(\mathrm{RSS}) - \sigma^2\,(n - 2p)}{\sigma^2} \\ &= \frac{E\,(\mathrm{RSS})}{\sigma^2} - n + 2p\,.\end{aligned}$$

If the model is correct, we must have:

$$E\,(\mathrm{RSS}) = (n - p)\,\sigma^2\,,$$

which implies

$$\Gamma_p = p\,.$$

In practice, we estimate Γ_p by

$$C_p = \frac{\mathrm{RSS}}{\widehat{\sigma}^2} - n + 2p\,,$$

where $\widehat{\sigma}^2$ is an estimator of σ^2. Here we estimate σ^2 using the s^2 of the full model. For this full model, we actually obtain $C_p = p$, which is not an interesting result. However, for all of the other models, where we use only a *subset* of the explanatory variables, the coefficient C_p can have values that are different from p. From the models that incorporate only a subset of the explanatory variables, we then choose those for which the value of C_p is the closest to p.

If we have k explanatory variables, we can also define the coefficient C_p for a model that incorporates a subset $X_1, \ldots, X_{p-1}$, $p \leq k$ of the k explanatory variables in the following manner:

$$\begin{aligned}C_p = &\frac{(n - k)\,\mathrm{RSS}\left(X_1, \ldots, X_{p-1}\right)}{\mathrm{RSS}\,(X_1, \ldots, X_k)} \\ &- n + 2p\,,\end{aligned}$$

where $\mathrm{RSS}\left(X_1, \ldots, X_{p-1}\right)$ is the sum of the squares of the residuals related to the model with $p - 1$ explanatory variables, and $\mathrm{RSS}\,(X_1, \ldots, X_k)$ is the sum of the squares of the residuals related to the model with all k explanatory variables. Out of the two models that incorporate $p - 1$ explanatory variables, we choose, according to the criterion, the one for which the value of the coefficient C_p is the closest to p.

DOMAINS AND LIMITATIONS

The C_p criterion is used when selecting variables. When used with the R^2 criterion, this criterion can tell us about the goodness of fit of the chosen model. The underlying idea of such a procedure is the following: instead

of trying to explain one variable using all of the available explanatory variables, it is sometimes possible to determine an underlying model with a subset of these variables, and the explanatory power of this model is almost the same as that of the model containing all of the explanatory variables. Another reason for this is that the collinearity of the explanatory variables tends to decrease estimator precision, and so it can be useful to delete some superficial variables.

EXAMPLES

Consider some data related to the Chicago fires of 1975. We denote the variable corresponding to the logarithm of the number of fires per 1000 households per district i of Chicago in 1975 by Y, and the variables corresponding to the proportion of households constructed before 1940, to the number of thefts and to the median revenue per district i by X_1, X_2 and X_3, respectively.

Since this set contains three explanatory variables, we have $2^3 = 8$ possible models. We divide the eight possible equations into four sets:

1. Set A contains the only equation without explanatory variables:

$$Y = \beta_0 + \varepsilon\,.$$

2. Set B contains three equations with one explanatory variable:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_3 X_3 + \varepsilon\,.$$

3. Set C contains three equations with two explanatory variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon\,.$$

4. Set D contains one equation with three explanatory variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon\,.$$

District	X_1	X_2	X_3	Y
i	x_{i1}	x_{i2}	x_{i3}	y_i
1	0.604	29	11.744	1.825
2	0.765	44	9.323	2.251
3	0.735	36	9.948	2.351
4	0.669	37	10.656	2.041
5	0.814	53	9.730	2.152
6	0.526	68	8.231	3.529
7	0.426	75	21.480	2.398
8	0.785	18	11.104	1.932
9	0.901	31	10.694	1.988
10	0.898	25	9.631	2.715
11	0.827	34	7.995	3.371
12	0.402	14	13.722	0.788
13	0.279	11	16.250	1.740
14	0.077	11	13.686	0.693
15	0.638	22	12.405	0.916
16	0.512	17	12.198	1.099
17	0.851	27	11.600	1.686
18	0.444	9	12.765	0.788
19	0.842	29	11.084	1.974
20	0.898	30	10.510	2.715
21	0.727	40	9.784	2.803
22	0.729	32	7.342	2.912
23	0.631	41	6.565	3.589
24	0.830	147	7.459	3.681
25	0.783	22	8.014	2.918
26	0.790	29	8.177	3.148
27	0.480	46	8.212	2.501
28	0.715	23	11.230	1.723
29	0.731	4	8.330	3.082
30	0.650	31	5.583	3.073
31	0.754	39	8.564	2.197
32	0.208	15	12.102	1.281
33	0.618	32	11.876	1.609
34	0.781	27	9.742	3.353

District	X_1	X_2	X_3	Y
i	x_{i1}	x_{i2}	x_{i3}	y_i
35	0.686	32	7.520	2.856
36	0.734	34	7.388	2.425
37	0.020	17	13.842	1.224
38	0.570	46	11.040	2.477
39	0.559	42	10.332	2.351
40	0.675	43	10.908	2.370
41	0.580	34	11.156	2.380
42	0.152	19	13.323	1.569
43	0.408	25	12.960	2.342
44	0.578	28	11.260	2.747
45	0.114	3	10.080	1.946
46	0.492	23	11.428	1.960
47	0.466	27	13.731	1.589

Source: Andrews and Herzberg (1985)

We denote the resulting models in the following way: **1** for X_1, **2** for X_2, **3** for X_3, **12** for X_1 and X_2, **13** for X_1 and X_3, **23** for X_2 and X_3, and **123** for the full model. The following table represents the results obtained for the C_p criterion for each model:

Model	C_p
1	32.356
2	29.937
3	18.354
12	18.936
13	14.817
23	3.681
123	4.000

If we consider a model from group B (containing one explanatory variable), the number of estimated parameters is $p = 2$ and none of the C_p values for the three models approaches 2. If we now consider a model from group C (with two explanatory variables), p equals 3 and the C_p of model **23** approaches this. Finally, for the complete model we find that $C_p = 4$, which is also the number of estimated parameters, but this is not an interesting result as previously explained. Therefore, the most reasonable choice for the model appears to be:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon\,.$$

FURTHER READING

- ▸ Coefficient of determination
- ▸ Collinearity
- ▸ Hat matrix
- ▸ Mean squared error
- ▸ Regression analysis

REFERENCES

Andrews D.F., Hertzberg, A.M.: Data: A Collection of Problems from Many Fields for Students and Research Workers. Springer, Berlin Heidelberg New York (1985)

Mallows, C.L.: Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS, 7–9 May 1964 (1964)

Mallows, C.L.: Some comments on *Cp*. Technometrics **15**, 661–675 (1973)

Cramér, Harald

Cramér, Harald (1893–1985) entered the University of Stockholm in 1912 in order to study chemistry and mathematics; he became a student of Leffler, Mittag and Riesz, Marcel. In 1919, Cramér was named assistant professor at the University of Stockholm. At the same time, he worked as an actuary for an insurance company, Svenska Life Assurance, which allowed him to study probability and statistics.

His main work in actuarial mathematics is *Collective Risk Theory*. In 1929, he was asked to create a new department in Stockholm, and he became the first Swedish professor of actuarial and statistical mathematics. At the end of the Second World War he wrote his principal work *Mathematical Methods of Statistics*, which was published for the first time in 1945 and was recently (in 1999) republished.

Some principal works and articles of Cramér, Harald

1946 Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.

1946 Collective Risk Theory: A Survey of the Theory from the Point of View of the Theory of Stochastic Processes. Skandia Jubilee Volume, Stockholm.

Criterion Of Total Mean Squared Error

The criterion of total mean squared error is a way of comparing estimations of the parameters of a biased or unbiased model.

MATHEMATICAL ASPECTS

Let

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1})$$

be a vector of estimators for the parameters of a regression model. We define the total mean square error, *TMSE*, of the vector $\widehat{\boldsymbol{\beta}}$ of estimators as being the sum of the mean squared errors (*MSE*) of its components.

We recall that

$$\begin{aligned} MSE\left(\widehat{\beta}_j\right) &= E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right) \\ &= V\left(\widehat{\beta}_j\right) + \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2 , \end{aligned}$$

where $E(.)$ and $V(.)$ are the usual symbols used for the **expected value** and the **variance**. We define the total mean squared error as:

$$\begin{aligned} TMSE(\widehat{\beta}_j) &= \sum_{j=1}^{p-1} MSE\left(\widehat{\beta}_j\right) \\ &= \sum_{j=1}^{p-1} E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right) \\ &= \sum_{j=1}^{p-1} \left[\mathrm{Var}\left(\widehat{\beta}_j\right) + \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2\right] \\ &= (p-1)\sigma^2 \cdot \mathrm{Trace}\,(\mathbf{V}) \\ &\quad + \sum_{j=1}^{p-1} \left(E\left(\widehat{\beta}_j\right) - \beta_j\right)^2 . \end{aligned}$$

where $\mathbf{V}$ is the **variance-covariance matrix** of $\widehat{\boldsymbol{\beta}}$.

DOMAINS AND LIMITATIONS

Unfortunately, when we want to calculate the total mean squared error

$$TMSE\left(\widehat{\boldsymbol{\beta}}\right) = \sum_{j=1}^{p-1} E\left(\left(\widehat{\beta}_j - \beta_j\right)^2\right)$$

of a vector of estimators

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1})$$

for the parameters of the model

$$Y_i = \beta_0 + \beta_1 X_{i1}^s + \ldots + \beta_{p-1} X_{ip-1}^s + \varepsilon_i ,$$

we need to know the values of the model parameters β_j, which are obviously unknown for all real data. The notation X_j^s means that the data for the jth explanatory variable were standardized (see **standardized data**).

Therefore, in order to estimate these *TMSE* we generate data from a structural similarly

model. Using a generator of pseudo-random numbers, we can simulate all of the data for the artificial model, analyze it with different models of regression (such as simple regression or **ridge regression**), and calculate what we call the *total squared error*, *TSE*, of the vectors of the estimators $\widehat{\boldsymbol{\beta}}$ related to each method:

$$TSE(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{p-1} (\widehat{\beta}_j - \beta_j)^2 .$$

We repeat this operation 100 times, ensuring that the 100 data sets are pseudo-independent. For each model, the average of 100 *TSE* gives a good estimation of the *TMSE*. Note that some statisticians prefer the model obtained by selecting variables due to its simplicity. On the other hand, other statisticians prefer the ridge method because it uses all of the available information.

EXAMPLES

We can generally compare the following regression methods: **linear regression** by mean squares, ridge regression, or the variable selection method.

In the following example, thirteen portions of cement have been examined. Each portion is composed of four ingredients, given in the table. The aim is to determine how the quantities x_{i1}, x_{i2}, x_{i3} and x_{i4} of these four ingredients influence the quantity y_i, the heat given out due to the hardening of the cement.

Heat given out by the cement

Portion	Ingredient				Heat
	1	2	3	4	
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	y_i
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.9
12	11	66	9	12	113.3
13	10	68	8	12	109.4

y_i quantity of heat given out due to the hardening of the ith portion (in joules);

x_{i1} quantity of ingredient 1 (tricalcium aluminate) in the ith portion;

x_{i2} quantity of ingredient 2 (tricalcium silicate) in the ith portion;

x_{i3} quantity of ingredient 3 (tetracalcium alumino-ferrite) in the ith portion;

x_{i4} quantity of ingredient 4 (dicalcium silicate) in the ith portion.

In this paragraph we will compare the estimators obtained by **least squares** (LS) regression with those obtained by ridge regression (R) and those obtained with the variable selection method (SV) via the total mean squared error *TMSE*. The three estimation vectors obtained from each method are:

$$\widehat{\mathbf{Y}}_{\mathrm{LS}} = 95.4 + 9.12\mathbf{X_1}^s + 7.94\mathbf{X_2}^s + 0.65\mathbf{X_3}^s - 2.41\mathbf{X_4}^s$$

$$\widehat{\mathbf{Y}}_{\mathrm{R}} = 95.4 + 7.64\mathbf{X_1}^s + 4.67\mathbf{X_2}^s - 0.91\mathbf{X_3}^s - 5.84\mathbf{X_4}^s$$

$$\widehat{\mathbf{Y}}_{\mathrm{SV}} = 95.4 + 8.64\mathbf{X_1}^s + 10.3\mathbf{X_2}^s$$

We note here that all of the estimations were obtained from standardized explanato-

ry variables. We compare these three estimations using the total mean squared errors of the three vectors:

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \left(\hat{\beta}_{LS_1}, \hat{\beta}_{LS_2}, \hat{\beta}_{LS_3}, \hat{\beta}_{LS_4}\right)'$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = \left(\hat{\beta}_{R_1}, \hat{\beta}_{R_2}, \hat{\beta}_{R_3}, \hat{\beta}_{R_4}\right)'$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{SV}} = \left(\hat{\beta}_{SV_1}, \hat{\beta}_{SV_2}, \hat{\beta}_{SV_3}, \hat{\beta}_{SV_4}\right)' .$$

Here the subscript LS corresponds to the method of least squares, R to the ridge method and SV to the variable selection method, respectively. For this latter method, the estimations for the coefficients of the unselected variables in the model are considered to be zero. In our case we have:

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (9.12, 7.94, 0.65, -2.41)'$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (7.64, 4.67, -0.91, -5.84)'$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{SV}} = (8.64, 10.3, 0, 0)' .$$

We have chosen to approximate the underlying process that results in the cement data by the following least squares equation:

$$Y_{iMC} = 95.4 + 9.12X^s_{i1} + 7.94X^s_{i2} + 0.65X^s_{i3} - 2.41X^s_{i4} + \varepsilon_i .$$

The procedure consists of generating 13 random error terms $\varepsilon_1, \ldots, \varepsilon_{13}$ 100 times based on a normal distribution with mean 0 and standard deviation 2.446 (recall that 2.446 is the least squares estimator of σ for the cement data). We then calculate $Y_{1LS}, \ldots, Y_{13LS}$ using the X^s_{ij} values in the data table. In this way, we generate 100 $Y_{1LS}, \ldots, Y_{13LS}$ samples from 100 $\varepsilon_1, \ldots, \varepsilon_{13}$ samples.

The three methods are applied to each of these 100 $Y_{1LS}, \ldots, Y_{13LS}$ samples(always using the same values for X^s_{ij}), which yields 100 estimators $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SV}}$. Note that these three methods are applied to these 100 samples without any influence from the results from the equations

$$\widehat{Y}_{iLS} = 95.4 + 9.12X^s_{i1} + 7.94X^s_{i2} + 0.65X^s_{i3} - 2.41X^s_{i4} ,$$
$$\widehat{Y}_{iR} = 95.4 + 7.64X^s_{i1} + 4.67X^s_{i2} - 0.91X^s_{i3} - 5.84X^s_{i4} \text{ and}$$
$$\widehat{Y}_{iSV} = 95.4 + 8.64X^s_{i1} + 10.3X^s_{i2}$$

obtained for the original sample. Despite the fact that the variable selection method has chosen the variables X_{i1} and X_{i2} in $\widehat{Y}_{iSV} = 95.4 + 8.64X^s_{i1} + 10.3X^s_{i2}$, it is possible that, for one of these 100 samples, the method has selected better X_{i2} and X_{i3} variables, or only X_{i3}, or all of the subset of the four available variables. In the same way, despite the fact that the equation $\widehat{Y}_{iR} = 95.4 + 7.64X^s_{i1} + 4.67xX^s_{i2} - 0.91X^s_{i3} - 5.84X^s_{i4}$ was obtained with $k = 0.157$, the value of k is recalculated for each of these 100 samples (for more details refer to the ridge regression example). From these 100 estimations of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{SV}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$, we can calculate 100 *ETC*s for each method, which we label as:

$$ETC_{\mathrm{LS}} = \left(\hat{\beta}_{LS_1} - 9.12\right)^2 + \left(\hat{\beta}_{LS_2} - 7.94\right)^2 + \left(\hat{\beta}_{LS_3} - 0.65\right)^2 + \left(\hat{\beta}_{LS_4} + 2.41\right)^2$$
$$ETC_{\mathrm{R}} = \left(\hat{\beta}_{R_1} - 9.12\right)^2 + \left(\hat{\beta}_{R_2} - 7.94\right)^2 + \left(\hat{\beta}_{R_3} - 0.65\right)^2 + \left(\beta_{R_4} + 2.41\right)^2$$

$$ETC_{\mathrm{SV}} = \left(\hat{\beta}_{SV_1} - 9.12\right)^2 + \left(\hat{\beta}_{SV_2} - 7.94\right)^2 + \left(\hat{\beta}_{SV_3} - 0.65\right)^2 + \left(\hat{\beta}_{SV_4} + 2.41\right)^2 .$$

The means of the 100 values of ETC_{LS}, ETC_{R} and ETC_{SV} are the estimations of $TMSE_{\mathrm{LS}}$, $TMSE_{\mathrm{SV}}$ and $TMSE_{\mathrm{R}}$: the *TMSE* estimations for the three considered methods.
After this simulation was performed, the following estimations were obtained:

$$TMSE_{\mathrm{LS}} = 270 ,$$
$$TMSE_{\mathrm{R}} = 75 .$$
$$TMSE_{\mathrm{SV}} = 166 .$$

These give the following differences:

$$TMSE_{\mathrm{LS}} - TMSE_{\mathrm{SV}} = 104 ,$$
$$TMSE_{\mathrm{LS}} - TMSE_{\mathrm{R}} = 195 ,$$
$$TMSE_{\mathrm{SV}} - TMSE_{\mathrm{R}} = 91 .$$

Since the standard deviations of 100 observed differences are respectively 350, 290 and 280, we can calculate the approximate 95% confidence intervals for the differences between the *TMSE*s of two methods

$$TMSE_{\mathrm{LS}} - TMSE_{\mathrm{SV}} = 104 \pm \frac{2 \cdot 350}{\sqrt{100}} ,$$
$$TMSE_{\mathrm{LS}} - TMSE_{\mathrm{R}} = 195 \pm \frac{2 \cdot 290}{\sqrt{100}} ,$$
$$TMSE_{\mathrm{SV}} - TMSE_{\mathrm{R}} = 91 \pm \frac{2 \cdot 280}{\sqrt{100}} .$$

We get

$$34 < TMSE_{\mathrm{LS}} - TMSE_{\mathrm{SV}} < 174 ,$$
$$137 < TMSE_{\mathrm{LS}} - TMSE_{\mathrm{R}} < 253 ,$$
$$35 < TMSE_{\mathrm{SV}} - TMSE_{\mathrm{R}} < 147 .$$

We can therefore conclude (at least for the particular model used to generate the simulated data, and taking into account our aim—to get a small *TMSE*), that the ridge method is the best of the methods considered, followed by the varible selection procedure.

FURTHER READING

- ▸ **Bias**
- ▸ **Expected value**
- ▸ **Hat matrix**
- ▸ **Mean squared error**
- ▸ **Ridge regression**
- ▸ **Standardized data**
- ▸ **Variance**
- ▸ **Variance–covariance matrix**
- ▸ **Weighted least-squares method**

REFERENCES

Box, G.E.P, Draper, N.R.: A basis for the selection of response surface design. J. Am. Stat. Assoc. **54**, 622–654 (1959)

Dodge, Y.: Analyse de régression appliquée. Dunod, Paris (1999)

Critical Value

In **hypothesis testing**, the critical value is the limit **value** at which we take the decision to reject the **null hypothesis** H_0, for a given **significance level**.

HISTORY

The concept of a critical value was introduced by **Neyman, Jerzy** and **Pearson, Egon Sharpe** in 1928.

MATHEMATICAL ASPECTS

The critical value depends on the type of the test used (**two-sided test** or **one-sided**

test on the right or the left), the **probability distribution** and the **significance level** α.

DOMAINS AND LIMITATIONS

The critical value is determined from the **probability distribution** of the **statistic** associated with the test. It is determined by consulting the **statistical table** corresponding to this probability distribution (**normal table**, **Student table**, **Fisher table**, **chi-square table**, etc).

EXAMPLES

A company produces steel cables. Using a **sample** of size $n = 100$, it wants to verify whether the diameters of the cables conform closely enough to the required diameter 0.9 cm in general.

The **standard deviation** σ of the **population** is known and equals 0.05 cm.

In this case, **hypothesis testing** involves a **two-sided test**. The hypotheses are the following:

null hypothesis H_0: $\mu = 0.9$

alternative hypothesis H_1: $\mu \neq 0.9$.

To a **significance level** of $\alpha = 5\%$, by looking at the **normal table** we find that the critical value $z_{\frac{\alpha}{2}}$ equals 1.96.

FURTHER READING

- ▸ Confidence interval
- ▸ Hypothesis testing
- ▸ Significance level
- ▸ Statistical table

REFERENCE

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)

C

Cyclical Fluctuation

Cyclical fluctuations is a term used to describe oscillations that occur over long periods about the **secular trend** line or curve of a **time series**.

HISTORY

See **time series**.

MATHEMATICAL ASPECTS

Consider Y_t, a **time series** given by its components; Y_t can be written as:

- $Y_t = T_t \cdot S_t \cdot C_t \cdot I_t$ (multiplicative **model**), or;
- $Y_t = T_t + S_t + C_t + I_t$ (additive model).

where

$\boldsymbol{Y_t}$ is the **data** at time t;
$\boldsymbol{T_t}$ is the **secular trend** at time t;
$\boldsymbol{S_t}$ is the **seasonal variation** at time t;
$\boldsymbol{C_t}$ is the cyclical fluctuation at time t, and;
$\boldsymbol{I_t}$ is the **irregular variation** at time t.

The first step when investigating a time series is always to determine the secular trend T_t, and then to determine the seasonal variation S_t. It is then possible to adjust the initial data of the **time series** Y_t according to these two components:

- $\dfrac{Y_t}{S_t \cdot T_t} = C_t \cdot I_t$ (multiplicative **model**);
- $Y_t - S_t - T_t = C_t + I_t$ (additive model).

To avoid cyclical fluctuations, a weighted **moving average** is established over a few months only. The use of moving averages allows use to smooth the **irregular variations** I_t by preserving the cyclical fluctuations C_t.

The choice of a weighted **moving average** allows us to give more weight to the central **values** compared to the extreme values, in order to reproduce cyclical fluctuations in a more accurate way. Therefore, large weights will be given to the central values and small weights to the extreme values.

For example, for a **moving average** considered for an interval of five months, the weights $-0.1, 0.3, 0.6, 0.3$ and -0.1 can be used; since their sum is 1, there will be no need for normalization.

If the **values** of $C_t \cdot I_t$ (resulting from the adjustments performed with respect to the **secular trend** and to the **seasonal variations**) are denoted by X_t, the value of the cyclical fluctuation for the month t is determined by:

$$C_t = -0.1 \cdot X_{t-2} + 0.3 \cdot X_{t-1} + 0.6 \cdot X_t + 0.3 \cdot X_{t+1} - 0.1 \cdot X_{t+2}\,.$$

DOMAINS AND LIMITATIONS

Estimating cyclical fluctuations allows us to:

- Determine the maxima or minima that a **time series** can attain.
- Perform short- or medium-term **forecasting**.
- Identify the cyclical components.

The limitations and advantages of the use of weighted **moving averages** when evaluating cyclical fluctuations are the following:

- Weighted moving averages can smooth a curve with cyclical fluctuations which still retaining most of the original fluctuation, because they preserve the amplitudes of the cycles in an accurate way.
- The use of an odd number of months to establish the moving average facilitates better centering of the **values** obtained.
- It is difficult to study the cyclical fluctuation of a **time series** because the cycles usually vary in length and amplitude. This is due to the presence of a multitude of factors, where the effects of these factors can change from one cycle to the other. None of the models used to explain and predict such fluctuations have been found to be completely satisfactory.

EXAMPLES

Let us establish a **moving average** considered over five months of **data** adjusted according to the **secular trend** and **seasonal variations**.

Let us also use the weights $-0.1, 0.3, 0.6, 0.3, -0.1$; since their sum is equal to 1, there is no need for normalization.

Let X_i be the adjusted **values** of $C_i \cdot I_i$:

$$C_i = -0.1 \cdot X_{i-2} + 0.3 \cdot X_{i-1} + 0.6 \cdot X_i + 0.3 \cdot X_{i+1} - 0.1 \cdot X_{i+2}\,.$$

The table below shows the electrical power consumed by street lights every month in millions of kilowatt hours during the years 1952 and 1953. The **data** have been adjusted according to the **secular trend** and **seasonal variations**.

Year	Month	Data X_i	Moving average for 5 months C_i
1952	J	99.9	
	F	100.4	
	M	100.2	100.1
	A	99.0	99.1
	M	98.1	98.4
	J	99.0	98.7
	J	98.5	98.3
	A	97.8	98.3
	S	100.3	99.9
	O	101.1	101.3
	N	101.2	101.1
	D	100.4	100.7

Year	Month	Data X_i	Moving average for 5 months C_i
1953	J	100.6	100.3
	F	100.1	100.4
	M	100.5	100.1
	A	99.2	99.5
	M	98.9	98.7
	J	98.2	98.2
	J	98.4	98.5
	A	99.7	99.5
	S	100.4	100.5
	O	101.0	101.0
	N	101.1	
	D	101.1	

No significant cyclical effects appear in these data and non significant effects do not mean no effects. The beginning of an economic cycle is often sought, but this only appears every 20 years.

FURTHER READING

- ▸ Forecasting
- ▸ Irregular variation
- ▸ Moving average
- ▸ Seasonal variation
- ▸ Secular trend
- ▸ Time series

REFERENCE

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

Wold, H.O. (ed.): Bibliography on Time Series and Stochastic Processes. Oliver & Boyd, Edinburgh (1965)