Methods of Multivariate Statistics

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

General Information on the Course

Format of the Course

We will alternate between introducing the new methods and practicing them on concrete examples (first by hand, to see how the method works, and then with the help of SPSS).

Assessment: Take-Home Assignment, Handed Out After the Course

- Submission deadline: Monday, June 04, 2012, 4:00 p.m.
- Submission by email to me or in hard-copy handed in/sent to me.
- Rules of submission: You may collaborate with your colleagues (group work allowed), but you must prepare your own individual report.
- Format of submission: a typeset report or a neatly handwritten one.
- For email submission, please email one pdf-file.

Other Information

Software: SPSS.

Apart from the computers in PC-Lab 3, you can get a free 2-weeks trial licence from SPSS. If you are an external doctoral student and do not have a SPSS license, please install the 2-weeks trial license only when you need it for the take-home assignment.

Help/Support: How to Get Help on the Take-Home Assignment

Contact me by email, phone or in person.

- Email: `kerstin.hesse@hhl.de`
- Phone: +49 (0)341 9851-820
- Office: HHL Main Building, Room 115A (I am usually there from 9:00 a.m. to 5:00 p.m., but please make an appointment by email.)

References

- K. Backhaus, B. Erichson, W. Plinke, R. Weiber: Multivariate Analysemethoden (13th edn.). Springer-Verlag, Berlin, Heidelberg, 2011.
- K. Backhaus, B. Erichson, R. Weiber: Fortgeschrittene Multivariate Analysemethoden (1st edn.). Springer-Verlag, Heidelberg et al, 2011.
- J. Bleymüller, G. Gehlert: Statistische Formeln, Tabellen und Statistik-Software (12th edn.). Verlag Franz Vahlen, München, 2011.
- J. Bleymüller, G. Gehlert, H. Gülicher: Statistik für Wirtschaftswissenschaftler (15th edn.). Verlag Franz Vahlen, München, 2008.
- J. L. Devore: Probability and Statistics for Engineering and the Sciences. Brooks/Cole (a division of Thomson Learning), Belmont, CA, USA, 2004.
- L. Fahrmeir, A. Hamerle, G. Tutz (eds.): Multivariate statistische Verfahren (2nd edn.). Walter de Gruyter, Berlin, 1996.

References

- L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz: Statistik: Der Weg zur Datenanalyse (7th edn.). Springer-Verlag, Berlin, Heidelberg, 2010.
- A. Handl: Multivariate Analysemethoden: Theorie un Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS (2nd edn.). Springer-Verlag, Heidelberg et al, 2010.
- W.K. Härdle, L. Simar: Applied Multivariate Statistical Analysis (3rd edn.). Springer-Verlag, Heidelberg et al, 2012.
- p.H. Müller: Lexikon der Stochastik (2nd edn.). Wissenschaftliche Buchgesllschaft, Darmstadt, 1975.
- S. Sharma: Applied Multivariate Techniques. John Wiley & Sons, 1996.
- H. Rinne: Taschenbuch der Statistik (4th edn.). Verlag Harri Deutsch, Frankfurt am Main, 2008.
- B.G. Tabachnick, L.S. Fidell: Using Multivariate Statistics (5th edn.). Pearson Education, 2007

Outline & Table of Contents

Methods of Multivariate Statistics

Topic 1: Revision of Background Material

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Topic 1: Revision of Background Material

- general notation
- types of data and measurement scales:
  - nominal data without order and nominal data with order,
  - data on an interval scale/metric data without a unique zero point,
  - data on a ratio scale/metric data with a unique zero point
- arithmetic mean, variance and standard deviation (of metric data describing a feature for a sample of objects)
- random variables and probability distributions
- expectation value, variance, standard deviation, covariance and correlation coefficient of random variables
- estimating expectation value, variance, standard deviation, covariance and correlation coefficient of random variables from a sample
- hypothesis testing

General Notation: Scalars and Vectors

- Scalar values (real numbers) are denotes by lowercase letters: $x, y, a, b, \ldots$.
- Random variables are denoted by uppercase letters $X, Y, Z, \ldots$.
- Vectors (of real numbers or random variables) are denoted by lowercase boldface letters and are by default column vectors $\mathbf{x}, \mathbf{y}, \mathbf{w}, \ldots$. In $\mathbf{x}'$ the $'$ denotes taking the transpose.

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = (x_1, x_2, \ldots, x_N)', \quad \mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = (Y_1, Y_2, \ldots, Y_p)'.$$

- The length of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$ is denoted $\|\mathbf{x}\|_2$ and is

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_N^2} = \sqrt{\sum_{i=1}^{N} x_i^2} = \left(\sum_{i=1}^{N} x_i^2\right)^{1/2}.$$

General Notation: Matrices

- Matrices are denoted by boldface uppercase letters $\mathbf{A}, \mathbf{B}, \mathbf{X}, \ldots$.
- An $m \times n$ matrix $\mathbf{A}$ has m rows and n columns:

$$\mathbf{A} = (a_{i,j})_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} \end{pmatrix} \tag{1}$$

- The entries of a matrix $\mathbf{A}$ are usually denoted by the corresponding lowercase letter, i.e. $a_{i,j}$, with the first index for the row and the second index for the column (e.g. see (1)).
- Occasionally we may also use $\mathbf{A}_{i,j}$ to refer to the entry in the ith row and jth column of a matrix $\mathbf{A}$.
- We may drop the comma between the indices in $a_{i,j}$ (i.e. write a_{ij} instead) if there is no ambiguity.

General Notation: Transpose of a Matrix

- For an $m \times n$ matrix $\mathbf{A}$ (given by (1)), $\mathbf{A}'$ denotes the transpose of the matrix $\mathbf{A}$ which is the $n \times m$ matrix given by

$$\mathbf{A}' = \begin{pmatrix} a_{1,1} & a_{2,1} & \cdots & a_{m,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{m,n} \end{pmatrix}. \tag{2}$$

  The entries of the ith row of $\mathbf{A}$ become the entries of the ith column of $\mathbf{A}'$ (indicated in (1) and (2) in violet for the 1st row/column).

- Example of a 2×3 matrix and its transpose:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \qquad \mathbf{A}' = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

  Here the entries of $\mathbf{A}$ are $a_{1,1} = 1$, $a_{1,2} = 2$, $a_{1,3} = 3$, $a_{2,1} = 4$, $a_{2,2} = 5$ and $a_{2,3} = 6$.

Types of Data and Measurement Scales: Nominal Data

In statistics, we discuss the statistical properties of features of objects.

The objects come from a population, and usually we will inspect a sample (randomly selected subset) from this population.

The types of data (or measurement scales) below are listed in the order of increasing properties of the data.

Nominal Data Without Order / Qualitative Data: This is data of the most general kind, describing a feature (or property) of objects.

Example: color of cars, with the values: red, blue, green, ...

Nominal Data with Order: This data describes a feature of objects and is given by numbers that can be meaningfully ordered.

Example: score from a questionnaire, with possible values 1,2,3,4,5.

Types of Data and Measurement Scales: Metric Data

Metric Data Without a Unique Zero Point / Data on an Interval Scale: This data describes a feature of objects in terms of numbers that can be meaningfully ordered. The distances between the different data values have meaning.

Example: Time measurement according to a calendar; the year zero could have been defined differently (no unique zero point).

Metric Data With a Unique Zero Point / Data on a Ratio Scale: This is data describes a feature of objects in terms of numbers that can be meaningfully ordered. The distances between the different data values have meaning, and there is a unique zero point.

Example: income, debt, height, weight.

Note: Due to the unique zero point, it makes sense to consider ratios; e.g. person A has twice the income of person B.

Arithmetic Mean, Variance and Standard Deviation

We have metric data of a feature x measured for a sample of N objects from a population: x_i = value of the feature at object e_i, $i = 1, 2, \ldots, N$.

Example: gross income per year in 1000 Euros: $x_1 = 45$, $x_2 = 55$, $x_3 = 50$

Arithmetic Mean: $$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example: $\overline{x} = \frac{1}{3}(45 + 55 + 50) = \frac{150}{3} = 50$

Variance: $$\mathrm{Var}(x) = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Standard Deviation: $$\sigma = \sqrt{\mathrm{Var}(x)}$$

Example: $\sigma^2 = \frac{1}{2}\left[(45-50)^2 + (50-50)^2 + (55-50)^2\right] = 25, \ \sigma = 5$

The standard deviation measures the average deviation from the mean.

Random Variable

Random Variable: A random variable X is a function that maps each event e (from the space of all events E) of a probability experiment onto an outcome of the event, given by a value $x = X(e)$. It is required that the values $X(e)$ of the events e are determined by chance.

Discrete Random Variable: A random variable is called discrete, if it can assume only a finite (or infinite but countable) number of values.

Continuous Random Variable: A random variable is called continuous, if it is metric and if it can assume all real values from an non-empty interval.

Example of a Discrete Random Variable (Throwing the Dice):

- probability experiment: throwing the dice
- An event e is throwing the dice.
- $X(e) =$ number of eyes on the face of the dice (values $1, 2, \ldots, 6$).

Example of a Discrete Random Variable

Example (Flipping a Coin Twice):

- probability experiment: flipping a coin twice
- event: $e =$ result from flipping the coin twice
- space of all events: $E = \{HH, HT, TH, TT\}$, where $H =$ heads, $T =$ tails
- random variable: $X(e) =$ number of heads, with values in $\{0, 1, 2\}$
- If we set $e_1 = HH$, $e_2 = HT$, $e_3 = TH$, $e_4 = TT$, then

$$X(e_1) = 2, \quad X(e_2) = X(e_3) = 1, \quad \text{and} \quad X(e_4) = 0.$$

- For a perfect coin, the probability $P(X = x)$ to obtain x heads is

$$P(X = 2) = \frac{1}{4}, \quad P(X = 1) = \frac{2}{4} = \frac{1}{2}, \quad P(X = 0) = \frac{1}{4}.$$

Example of Continuous Random Variables

Example (Age and Gross Income of a Random Person):

- probability experiment: drawing a random person from a sample
- event: $e =$ drawing of a person
- space of all events: all possible choices of a person
- random variables: $X(e) =$ the person's gross income per year in 1000 Euros, $Y(e) =$ the person's age

Note: In this example we could also identify the event

$$e = \text{drawing of a random person}$$

with the person (object) itself and thus consider

$$e = \text{random person from the population}$$

and define $X(e)$ and $Y(e)$ as properties of the person (object) e:

$$X(e) = \text{gross income per year of person } e, \qquad Y(e) = \text{age of person } e.$$

Probability Distribution of a Discrete Random Variable

Let X be a discrete random variable with values $x_1, x_2, \ldots, x_i, \ldots$.

Probability Density: The function

$$f(x_i) = P(X = x_i) = (\text{probability that } X = x_i)$$

is called the probability density of X.

Probability Distribution: The function

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i) = (\text{probability that } X \leq x)$$

is called the probability distribution of X.

Ex. 1.1 (Flipping a Coin Twice): For the example of flipping a perfect coin twice with the random variable $X(e) =$ number of heads, determine the probability density and probability distribution.

Example (Gross Income of a Random Person) continued:

What is the probability that a random person e has a yearly gross income between 50,000 and 60,000 Euros, i.e.

$$P(50 \leq X \leq 60) = P(X \leq 60) - P(X < 50) = ???$$

The answer depends on the probability distribution f of the random variable $X =$ income.

If the gross income is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 10$, then the probability density is

$$f_n(x; 40, 10) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-40}{10}\right]^2\right)$$

and the probability distribution is

$$\underbrace{F_n(x; 40, 10) = P(X \leq x)}_{\substack{= \text{probability that} \\ X \text{ has a value} \leq x}} = \int_{-\infty}^{x} \underbrace{\frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{y-40}{10}\right]^2\right)}_{= f_n(y; 40, 10)} \mathrm{d}y.$$

Probability Distribution of a Continuous Random Variable

Let X be a continuous random variable.

Probability Distribution: If X has the probability distribution $F(x)$, then

$$F(x) = P(X \leq x) = \text{ (probability that } X \text{ has a value } \leq x)$$

Probability Density: If X has the probability density $f(x)$ and probability distribution $F(x)$, then

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \text{ (probability that } X \text{ has a value } \leq x)$$

and

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(y)\,\mathrm{d}y \\ &= \text{(probability that } x_1 \leq X \leq x_2) \end{aligned}$$

(Gaussian) Normal Distribution

The (Gaussian) normal distribution has the density function

$$f_n(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$

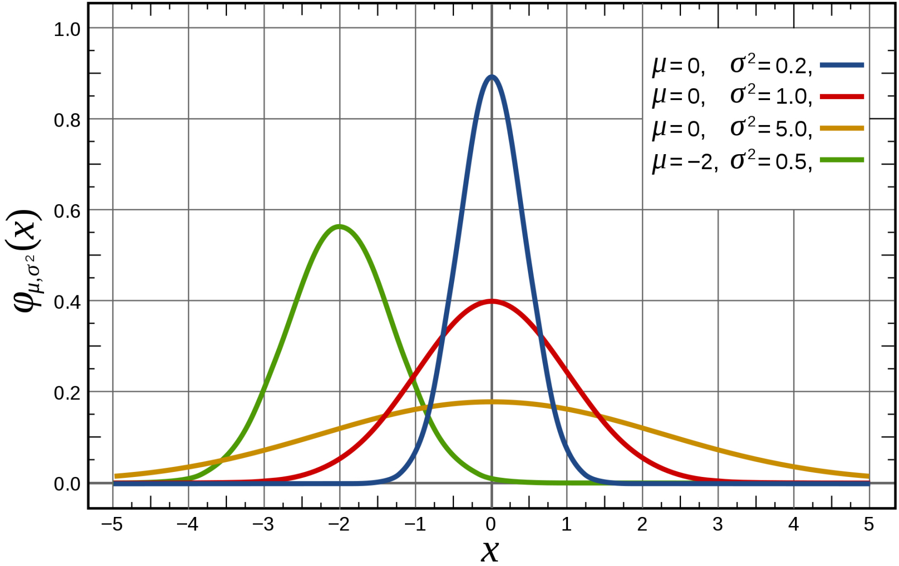
and the probability distribution

$$F_n(x;\mu,\sigma) = \int_{-\infty}^{x} f_n(y;\mu,\sigma)\,\mathrm{d}y = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2\right)\mathrm{d}y.$$

Two parameters: $\mu =$ expectation value, $\sigma =$ standard deviation.

F_n can be looked up in a table of the normal distribution (see later).

Density Function of the Normal Distribution



Expectation Value, Variance of Discrete Random Variable

Let X be a discrete random variable with values $x_1, x_2, \ldots, x_i, \ldots$, and with probability density f.

Expectation Value of X: $\mathrm{E}(X) = \sum_i x_i \cdot f(x_i)$

Variance of X: $\mathrm{Var}(X) = \mathrm{E}\big([X - \mathrm{E}(x)]^2\big) = \sum_i \big[x_i - f(x_i)\big]^2 \cdot f(x_i)$

Standard Deviation of X: $\sigma_X = \sqrt{\mathrm{Var}(X)}$

Note: The sums are over all values of X, and we have
$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.

Ex. 1.2 (Flipping a Coin Twice): Compute the expectation value and the variance of the random variable $X =$ number of heads in the probability experiment of flipping a perfect coin twice.

Expectation Value, Variance of Continuous Random Var.

Let X be a continuous random variable with probability distribution F and probability density f.

Expectation Value of X: $\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$

Variance of X: $\mathrm{Var}(X) = \mathrm{E}\big([X - \mathrm{E}(X)]^2\big) = \int_{-\infty}^{\infty} \big[x - \mathrm{E}(X)\big]^2 f(x)\, \mathrm{d}x$

Standard Deviation of X: $\sigma_X = \sqrt{\mathrm{Var}(X)}$

We have: $\mathrm{Var}(X) = \mathrm{E}\big([X - \mathrm{E}(X)]^2\big) = \mathrm{E}(X^2) - \big[\mathrm{E}(X)\big]^2.$

Note: The sums in the case of the discrete random variable have become integrals in the case of the continuous random variable.

Example: Expectation Value and Variance of Income

If the yearly gross income is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 10$, then the probability density is

$$f_n(x; 40, 10) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-40}{10}\right]^2\right).$$

Computing the expectation value $\mathrm{E}(X)$ and the Variance $\mathrm{Var}(X)$ we find

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-40}{10}\right]^2\right) \mathrm{d}x = 40 = \mu, \\ \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-40)^2 \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-40}{10}\right]^2\right) \mathrm{d}x = 100 = \sigma^2. \end{aligned}$$

Note: If a random variable X follows a normal distribution $F_n(x; \mu, \sigma)$ with parameters μ and σ, then always

$$\mathrm{E}(X) = \mu \qquad \text{and} \qquad \mathrm{Var}(X) = \mathrm{E}\big([X - \mathrm{E}(x)]^2\big) = \sigma^2.$$

Centered and Standardized Random Variables

By defining for a random variable X with $\mathrm{E}(X) = \mu_X$ and $\mathrm{Var}(X) = \sigma_X^2$

$$W = X - \mathrm{E}(X) = X - \mu_X \qquad \text{and} \qquad Z = \frac{X - \mathrm{E}(X)}{\sigma_X} = \frac{X - \mu_X}{\sigma_X} \tag{3}$$

we obtain:

- a centered random variable W with $\mathrm{E}(W) = 0$ and $\mathrm{Var}(W) = \sigma_X^2$,
- a standardized random variable Z with $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$

We can also convert back to the original variables:

$$X = W + \mathrm{E}(X) = W + \mu_X \quad \text{and} \quad X = \sigma_X \cdot Z + \mathrm{E}(X) = \sigma_X \cdot Z + \mu_X. \tag{4}$$

Statistical tables of probability distributions are often given for the standardized case $\mu = \mathrm{E}(Z) = 0$ and $\sigma^2 = \mathrm{Var}(Z) = 1$.

Standardization of Random Variables I

If our random variable X is not standardized, then we may use (3) to convert values of X to the standardized values, consult the appropriate table, and then convert with (4) back to our original variable.

We need to look up how the probability distribution for the standardized variable and the non-standardized variable are related!

For the normal distribution we have

$$F_n(x; \mu, \sigma) = F_N\left(\frac{x-\mu}{\sigma}\right) = F_N(z) \tag{5}$$

where $F_N(z) = F_n(z; 0, 1)$ is the standard normal distribution with expectation value $\mu = 0$ and standard deviation $\sigma = 1$.

Ex. 1.3: Random Variable Gross Income

If the yearly gross income X is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 10$, then the probability density is

$$f_n(x; 40, 10) = \frac{1}{10\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left[\frac{x - 40}{10} \right]^2 \right)$$

and $\mu = \mathrm{E}(X) = 40$ and $\mathrm{Var}(X) = \sigma^2 = 100$.

Use (5) and the table for the standard normal distribution F_N to determine the probability that a person has a yearly gross income between 50,000 and 60,000 Euros.

Standardization of Random Variables II

Standardization is a linear transformation: with $\mu = \mathrm{E}(X)$ and $\sigma = \mathrm{Var}(X)$,

$$Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma} = a \cdot X + b \qquad \text{with} \qquad a = \frac{1}{\sigma},\ b = -\frac{\mu}{\sigma}.$$

Linear transformations do not change the type of a probability distribution, but they change the expectation value and the variance.

For $Z = a \cdot X + b$ the expectation values and variances of X and Z are related as follows:

$$\mathrm{E}(Z) = a \cdot \mathrm{E}(X) + b \qquad \text{and} \qquad \mathrm{Var}(Z) = a^2 \cdot Var(X). \tag{6}$$

Ex. 1.4 (Standardization): Use (6) to verify that $Z = (X-\mu)/\sigma$ with $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$ does satisfy $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

More Probability Distributions

Other Probability Distributions Used in this Course:

- t-distribution or Student distribution
- χ^2-distribution
- F-distribution

The ideas and use of these distributions are analogous to the normal distribution; only the shape is somewhat different.

Probability distributions are characterized by some parameters: expectation value, standard deviation (or variance) and sometimes degrees of freedom.

Covariance and Correlation of Random Variables I

Consider the case of two discrete random variables X and Y with a joint probability density $f(x,y)$ and expectation values $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ and variances $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$.

The covariance of X and Y is computed via

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\big([X-\mathrm{E}(X)]\cdot[Y-\mathrm{E}(Y)]\big)\\ &= \sum_i \sum_j \big[x_i - \mathrm{E}(x)\big]\cdot\big[y_j - \mathrm{E}(y)\big]\cdot f(x_i,y_j)\end{aligned}$$

where the sums are taken over all values x_i of X and all values y_j of Y.

Interpretation: $f(x_i,y_j) =$ probability of the values (x_i,y_j) for (X,Y).

The covariance $\mathrm{Cov}(X,Y)$ is a measure of the correlation of X and Y. It measures whether the random variables X and Y depend on each other in a linear way, e.g. $Y = a\cdot X + b$.

Covariance and Correlation of Random Variables II

If X and Y are independent, then the $\mathrm{Cov}(X, Y) = 0$.

However, if $\mathrm{Cov}(X, Y) = 0$ then we cannot conclude that X and Y are independent.

Correlation can only measure linear relationships between random variables.

Correlations of different random variables are hard to compare as they depend on the scale of the variables. A scale-free (and thus comparable) measure is the correlation coefficient

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = \mathrm{E}\left(\underbrace{\frac{[X - \mathrm{E}(X)]}{\sigma_X}}_{=Z_X} \cdot \underbrace{\frac{[Y - \mathrm{E}(Y)]}{\sigma_Y}}_{=Z_Y} \right) = \mathrm{Cov}(Z_X, Z_Y)$$

We note that $\varrho(X, Y)$ is just the covariance $\mathrm{Cov}(Z_X, Z_Y)$ of the corresponding standardized variables Z_X and Z_Y.

Ex. 1.5 (Flipping a Coin Twice):

Consider a perfect coin, and let
$X =$ first flip of the coin,
$Y =$ second flip of the coin,
with the possible events (for both X and Y): $1 =$ heads, $0 =$ tails.

Let the joint probability density be given by $f(x, y) = 1/4$.

Do you expect that the result of the first flip of the coin has any influence on the result of the second flip of the coin and vice versa?

What do you conclude about the covariance $\text{Cov}(X, Y)$ of X and Y?

Compute the covariance $\text{Cov}(X, Y)$ of X and Y.

Covariance of Continuous Random Variables:
This can be defined analogously using integrals instead of the sums.

Estimating Parameters of a Random Var. from a Sample

In practice, a random variable X (e.g. income, height, ratings of products) is not measured on the whole population but on a large sample.

Often we have no a-priori information about the probability distribution of X or about the expectation value $\mathrm{E}(X)$ and the variance $\mathrm{Var}(X)$ of X.

Aim: Estimate the expectation value, variance and covariance of random variables from a sample.

Sampling: We draw a sample of N objects e_i (e.g. $N = 1000$ persons) and measure for each object the random variable X (e.g. the gross income): This gives values $x_1, x_2, \ldots, x_N$ for X. (x_i = value of X for object e_i)

Let Y be a second random variable (e.g. spending on foods) that is measured on the same N objects e_i of the sample drawn for measuring X: This gives values $y_1, y_2, \ldots, y_N$ for Y. (y_i = value of Y for object e_i)

Estimating Expectation Value and Variance from a Sample

Expectation Value: As the expectation value $\mu = \mathrm{E}(X)$ is the average value expected for X, it is estimated by the arithmetic mean of X in the sample:

$$\widehat{\mu} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(e.g. average gross income in the sample)}$$

Variance: As the variance $\mathrm{Var}(X)$ is the squared average deviation from $\mathrm{E}(X)$, it is estimated by:

$$\widehat{\sigma_X}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \qquad \left(\begin{array}{c}\text{e.g. squared mean deviation from the} \\ \text{average gross income in the sample}\end{array}\right)$$

Estimating Covariance & Correlation Coeff. from a Sample

Covariance: The covariance $\text{Cov}(X, Y)$ of X and Y is estimated by:

$$\widehat{\text{Cov}}(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})\,(y_i - \overline{y})$$

$\widehat{\text{Cov}}(X, Y)$ is an indicator for the strength of the correlation of X and Y.

Correlation Coefficient: The correlation coefficient $\varrho(X, Y)$ of X and Y is estimated by:

$$\widehat{\rho}(X, Y) = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\sigma_X}\,\widehat{\sigma_Y}}$$

It is a scale-free measure for the strength of the correlation of X and Y.

Ex. 1.6: Estimate Parameters of Random Var. from Sample

The gross income per month ($= X$) and the spending on foods per month ($= Y$) are sampled for $N = 4$ persons e_1, e_2, e_3, e_4:

Person	X (in Euros)	Y (in Euros)
e_1	6000	300
e_2	5000	250
e_3	6500	400
e_4	4500	250
means		

Estimate the expectation values $\mathrm{E}(X)$, $\mathrm{E}(Y)$, the variances $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, the covariance $\mathrm{Cov}(X, Y)$ and the correlation coefficient $\varrho(X, Y)$.

Comments: Estimating Parameters of a Random Variable

Notation:

- The estimates of the expectation value, the variance, ... are also called the empirical expectation value, the empirical variance, ...
- The $\widehat{\ }$ over a parameter, e.g. in $\widehat{\sigma_X}$, indicates an estimator of the parameter without the $\widehat{\ }$. So $\widehat{\sigma_X}$ denotes an estimator of σ_X.

Query: Comparing the formulas for expectation value, variance and covariance with the formulas for their estimators, why does the probability density not occur in the formulas for the estimators?

For a large sample, values of the random variable with a higher probability will be drawn more often. Thus they automatically occur with approximately the correct frequency in the sample.

Geometric Interpretation of the Covariance and Correlation

Let $\mathbf{x} = (x_1, x_2, \dots, x_N)'$ (data for X), $\mathbf{y} = (y_1, y_2, \dots, y_N)'$ (data for Y), $\overline{\mathbf{x}} = (\overline{x}, \overline{x}, \dots, \overline{x})'$ and $\overline{\mathbf{y}} = (\overline{y}, \overline{y}, \dots, \overline{y})'$ (N-vectors with mean as entries).

$$\widehat{\mathrm{Cov}}(X,Y) = \frac{1}{N-1} \cdot \underbrace{(\mathbf{x}-\overline{\mathbf{x}})'\,(\mathbf{y}-\overline{\mathbf{y}})}_{=\text{scalar product}} = \frac{1}{N-1} \cdot \underbrace{\|\mathbf{x}-\overline{\mathbf{x}}\|_2 \cdot \|\mathbf{y}-\overline{\mathbf{y}}\|_2 \cdot \cos(\alpha)}_{=\text{scalar product}},$$

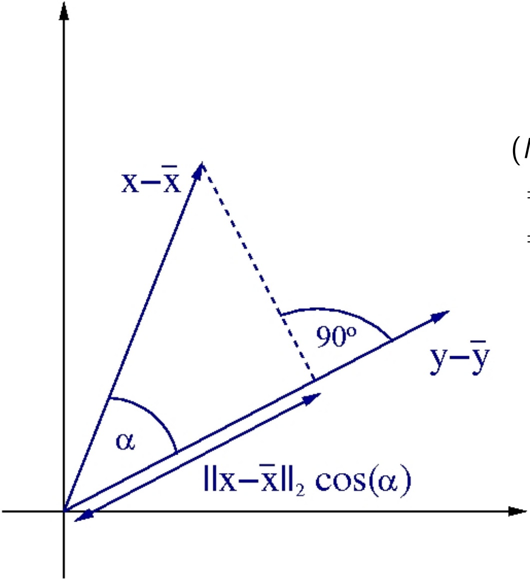
where α is the angle between the vectors $(\mathbf{x}-\overline{\mathbf{x}})$ and $(\mathbf{y}-\overline{\mathbf{y}})$.

$$\widehat{\sigma_X} = \frac{1}{\sqrt{N-1}} \cdot \|\mathbf{x}-\overline{\mathbf{x}}\|_2 \qquad \text{and} \qquad \widehat{\sigma_Y} = \frac{1}{\sqrt{N-1}} \cdot \|\mathbf{y}-\overline{\mathbf{y}}\|_2,$$

$$\widehat{\varrho}(X,Y) = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\sigma_X}\,\widehat{\sigma_Y}} = \frac{\frac{1}{N-1} \cdot (\mathbf{x}-\overline{\mathbf{x}})'\,(\mathbf{y}-\overline{\mathbf{y}})}{\frac{1}{\sqrt{N-1}} \cdot \|\mathbf{y}-\overline{\mathbf{y}}\|_2 \cdot \frac{1}{\sqrt{N-1}} \cdot \|\mathbf{x}-\overline{\mathbf{x}}\|_2} = \cos(\alpha).$$

$\widehat{\varrho}(X,Y) = \cos(\alpha)$ can only assume values in the interval $[-1, 1]$.
$\widehat{\varrho}(X,Y)$ is zero if $(\mathbf{x}-\overline{\mathbf{x}})'$ and $(\mathbf{y}-\overline{\mathbf{y}})'$ are perpendicular,
and $|\widehat{\varrho}(X,Y)|$ is 1 if the vectors are parallel or antiparallel.

Illustration of the Geometric Interpretation of $\widehat{\mathrm{Cov}}(X,Y)$



$$
\begin{aligned}
&(N-1)\cdot\widehat{\mathrm{Cov}}(X,Y)\\
&\quad= (\mathbf{x}-\overline{\mathbf{x}})'\,(\mathbf{y}-\overline{\mathbf{y}})\\
&\quad= \|\mathbf{x}-\overline{\mathbf{x}}\|_2\cdot\|\mathbf{y}-\overline{\mathbf{y}}\|_2\cdot\cos(\alpha)
\end{aligned}
$$

$$\widehat{\varrho}(X,Y)=\cos(\alpha)$$

Idea of Hypothesis Testing

- Hypothesis testing is about verifying whether statistical results are significant or not, i.e. whether they are likely to result from a true trend or whether they are due to random variations.
- Hypothesis testing does not give a definite answer but rather gives an answer with a specified margin of error.
- Hypothesis testing uses information about the probability distribution of the investigated quantity.

Application Areas of Hypothesis Testing:

- Are any of the coefficients in a (multilinear) regression significantly different from zero?
- Comparison of means (see example and later ANOVA).

Example: Hypothesis Testing I

In a geese farm, the average weight of a goose in 2010 was $\mu_1 = 5123$ g with a standard deviation of $\sigma_1 = 196$ g.

At the beginning of 2011, the fodder for fattening the geese was changed. A sample of $n = 101$ geese in 2011 (feed with the new fodder) yielded an average weight of $\overline{x} = 5151$ g.

Query: Has the new geese fodder changed the average weight μ_2 in 2011? Give an answer with a significance level of $\alpha = 0.05$.

---

1. Formulate the Null Hypothesis and Alternative Hypothesis:

$$H_0 : \mu_2 = \mu_1 = 5123 \text{ g} \qquad \left(\begin{array}{c} \text{The average weight of the geese} \\ \text{in both years is the same.} \end{array} \right)$$

$$H_1 : \mu_2 \neq \mu_1 = 5123 \text{ g} \qquad \left(\begin{array}{c} \text{The average weight of the geese} \\ \text{in both years is not the same.} \end{array} \right)$$

Example: Hypothesis Testing II

2. Find the Test Variable and its Distribution: Our random variable is the mean value $\overline{X}$ (average weight) of the geese in the sample from 2011.

   If the null hypothesis is true, then the expectation value for $\overline{X}$ is

   $$\mathrm{E}(\overline{X}) = \mu_1 = 5123 \text{ g}$$

   and its standard deviation is

   $$\sigma_{\overline{X}} \approx \frac{\sigma_1}{\sqrt{n}} = \frac{196 \text{ g}}{\sqrt{101}} = 19.50 \text{ g}.$$

   (The formula $\sigma_{\overline{X}} \approx \sigma_1/\sqrt{n}$ for the standard deviation of $\overline{X}$ will not be explained here.)

   As test variable we consider the standardized variable

   $$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sigma_{\overline{X}}} = \frac{\overline{X} - 5123 \text{ g}}{19.50 \text{ g}},$$

   which (as can be shown) follows a standard normal distribution.

Example: Hypothesis Testing III

3. Determination of the Critical Area (for Acceptance of the Null Hypothesis): Here we have a double sided test, and for $\alpha = 0.05$ we find (from the table of $f_N(z) = f_n(z;0,1)$) the two critical values

$$z_\ell = -1.96 \qquad \text{and} \qquad z_u = +1,96.$$

Hence if

$$z < z_\ell = -1.96 \qquad \text{or} \qquad z > z_u = +1,96$$

we reject the null hypothesis and if

$$-1.96 = z_\ell \leq z \leq z_u = +1,96$$

we accept the null hypothesis.

4. Computation of the Value of the Test Variable:

$$z = \frac{\mu_2 - \mathrm{E}(\overline{X})}{\sigma_{\overline{X}}} = \frac{5151\ \mathrm{g} - 5123\ \mathrm{g}}{19.50\ \mathrm{g}} = \frac{28}{19.50} \approx 1.44.$$

Example: Hypothesis Testing IV

5. Decision about the Hypotheses and Interpretation: As

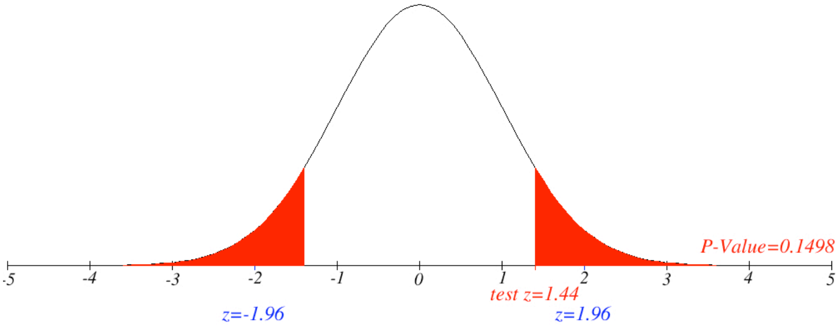
$$-1.96 = z_l \leq 1.44 \leq z_u = +1,96$$

we cannot reject the null hypothesis.

The difference of 28 g between the average weight of the geese in the sample in 2011 and the average weight μ_1 of the geese in 2012 is not statistically significant (i.e. it is likely to be caused by random variations).

Statistical Interpretation: The chance to reject the null hypothesis, when it is in fact true, is $\alpha = 0.05$ (or 5%).

Example: Area Under the Probability Distribution



As the critical values $z = \pm 1.96$ lie in the area in red, we have to accept the null hypothesis. This is also confirmed by the p-value which is

$$p = 0.1498 > 0.05 = \alpha,$$

also telling us that we must accept the null hypothesis.

Ex. 1.7: Hypothesis Testing

In our geese farm not only the average weight but the variance of the geese was sampled in 2010 and 2011, in order to determine whether the geese fodder (which was changed at the start of 2011) influenced the variance of the weight.

For a sample of $n_1 = n_2 = 101$ geese in each year we found the variance $s_1^2 = 196^2\ \mathrm{g}^2$ (2010) and $s_2^2 = 153^2\ \mathrm{g}^2$ (2011). The quotient

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2},$$

where S_1^2 and S_2^2 are the random variables for the sample variances and σ_1^2 and σ_2^2 are the variances in the population in 2010 and 2011, follows an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

Use this information to test the null hypothesis that the variances of the weight are the same with a significance level of $\alpha = 0.05$ against the alternative hypothesis that $\sigma_1^2 > \sigma_2^2$.

Methods of Multivariate Statistics

Topic 2: Analysis of Variance (ANOVA)

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Topic 2: Analysis of Variance

2.1 One-Way Analysis of Variance (1-Way ANOVA)

- definition and explanation of the idea of 1-way ANOVA, examples
- mathematical model of 1-way ANOVA
- hypothesis testing to answer the question posed by 1-way ANOVA
- examples and exercises

2.2 Two-Way Analysis of Variance (2-Way ANOVA)

- definition and explanation of the idea of 2-way ANOVA, examples
- mathematical model of 2-way ANOVA with interaction
- hypothesis testing to answer the questions posed by 2-way ANOVA
- examples and exercises

Methods of Multivariate Statistics

Part 2.1: One-Way Analysis of Variance

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Example of 1-Way ANOVA: Student Learning

Does the academic success of economics students depend on the teaching method?

Independent qualitative variable/factor A: method of teaching with three factor levels:

- $A_1 =$ traditional teaching,
- $A_2 =$ distance learning,
- $A_3 =$ blended learning.

3 subpopulations/groups of economics students:

- $P_1 =$ students taught with traditional teaching $(= A_1)$,
- $P_2 =$ students taught with distance learning $(= A_2)$,
- $P_3 =$ students taught with blended learning $(= A_3)$.

Dependent metric variable Y: academic success (measured by the mark), i.e. we propose a function/relationship $f : A \rightarrow Y$, $f(A_i) = Y_i$.

One-Way Analysis of Variance Explained

Consider a population P and a qualitative independent variable (called a factor) A with values (called factor levels) $A_1, A_2, \ldots, A_r$ defined on P.

The factor A allows us to subdivide the population P into subpopulations/groups $P_1, P_2, \ldots, P_r$, where

$$P_i = \text{set of all objects from } P \text{ for which } A \text{ has the value } A_i.$$

Let Y be a metric variable that is defined on the population P.

Research Question: For an object e_i from P, does its value A_i for A affect its value Y_i for Y? In other words, does the metric variable Y depend on the factor A?

Example (Student Learning): $A =$ teaching method, $Y =$ mark

Exercises: One-Way ANOVA Problems

Ex. 2.1 (Effect of Different Fertilizers on the Crop Yield):

The effect of four different types of fertilizer (A_1, A_2, A_3 and A_4) on the crop yield shall be investigated.

- Describe this problem in terms of one-way ANOVA.
- Given 40 fields of equal size and soil quality, suggest a way of investigating this problem empirically.

Ex. 2.2 (Effect of Shelf Placement on Margarine Sales):

How does the shelf placement (options: $A_1 =$ normal shelf or $A_2 =$ cooling shelf) effect the sales of margarine?

- Describe this problem in terms of one-way ANOVA.
- Suggest a way to investigate this problem empirically.

One-Way ANOVA: Mathematical Model – Setup

Setup and Assumptions:

- Let A be a factor with levels $A_1, A_2, \ldots, A_r$ defined on a population P.
- Let P_i denote the subpopulation of all objects from P for which A has the value A_i.
- Let Y be a metric random variable that can be sampled in P.
- Assumptions: Y is normally distributed in P and in each subpopulation P_i, and Y has the same variance in $P_1, P_2, \ldots, P_r$.

Example Student Learning:

- factor: $A =$ teaching method
- 3 populations: $P_1 =$ students taught with traditional teaching $(= A_1)$,
  $P_2 =$ students taught with distance learning $(= A_2)$,
  $P_3 =$ students taught with blended learning $(= A_3)$.
- metric variable: $Y =$ mark (of the student)

One-Way ANOVA: Mathematical Model – Means

One-way ANOVA is used to investigate whether Y depends on A, i.e. whether the factor levels A_i have an effect on the values of Y.

We investigate this by determining whether the arithmetic means of Y in the subpopulations P_i differ significantly.

Grand Mean of Y, Means of Y for the Different Subpopulations:

- $\mu =$ grand (arithmetic) mean of Y in the total population P,
- $\mu_i =$ (arithmetic) mean of Y in the population P_i, $i = 1, 2, \ldots, r$,
- $\alpha_i = \mu_i - \mu =$ effect of A_i on Y, $i = 1, 2, \ldots, r$.

Example Student Learning:

- grand mean: $\mu =$ average mark of the economics students
- means in the subpopulations: $\mu_i =$ average mark of students taught with teaching method A_i
- $\alpha_i = \mu_i - \mu =$ effect attributed solely to teaching method A_i

Comparison of Means via Hypothesis Testing

Comparison of Means in the Different Populations:
If the means $\mu_i = \mu + \alpha_i$, $i = 1, 2, \ldots, r$, are all equal, then

$$\mu = \mu_i = \mu + \alpha_i \quad \Leftrightarrow \quad \alpha_i = 0, \quad i = 1, 2, \ldots, r.$$

To investigate whether the means differ significantly (i.e. differences in the values are not solely due to random errors) we use hypothesis testing.

Hypothesis Testing: We are testing the null hypothesis

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$ (or equivalently $\mu_1 = \mu_2 = \ldots = \mu_r = \mu$).

against the alternative hypothesis

H_1: For at least one subpopulation P_i, $\alpha_i \neq 0$ (or equivalently $\mu_i \neq \mu$).

Example: Student Learning

Factor: $A =$ method of teaching; metric variable: $Y =$ mark (of student).

- $\mu =$ average mark (among all students),
- $\mu_i =$ average mark of students taught with method A_i,
- $\alpha_i = \mu_i - \mu =$ effect attributed solely to teaching method A_i.

If the teaching method has no effect on the academic success, then $\mu_1 = \mu_2 = \mu_3 = \mu$ or equivalently $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu$ (or equivalently $\alpha_1 = \alpha_2 = \alpha_3 = 0$), i.e. the average mark does not depend on the teaching method.

H_1: $\mu_i \neq \mu$ or equivalently $\alpha_i \neq 0$ for (at least) one teaching method A_i, i.e. the average mark is not the same for each teaching method (and hence depends on the teaching method).

Sampling in the Different Subpopulations

Take a sample of size n_i from population P_i and measure Y: measurements $y_{i1}, y_{i2}, \dots, y_{in_i}$ of Y (1st index for population P_i, 2nd index for number in sample). From our model,

$$y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \qquad k = 1, 2, \dots, n_i, \qquad \text{where:}$$

- ϵ_{ik} is a random error due to the variation of Y within P_i.
- The ϵ_{ik} are all normally distributed with mean value zero and the same variance σ^2.

Note: The expectation value for sampling Y in P_i is $\mu + \alpha_i$.

Example (Student Learning): $y_{i1}, y_{i2}, \dots, y_{i100}$ are the marks of 100 students taught with teaching method A_i, $i = 1, 2, 3$.

$$\underbrace{y_{ik}}_{\substack{\text{mark of student } k\\ \text{taught with } A_i}} = \underbrace{\mu}_{\substack{\text{average}\\ \text{mark}}} + \underbrace{\alpha_i}_{\substack{\text{effect on mark from}\\ \text{teaching method } A_i}} + \underbrace{\epsilon_{ik}}_{\substack{\text{random}\\ \text{error}}}$$

Sample is Used to Estimate the Means

The grand mean μ and the mean μ_i within the subpopulation P_i are estimated via the sample means: with $N = \sum_{i=1}^{r} n_i$,

$$\underbrace{\overline{y} = \frac{1}{N} \sum_{i=1}^{r} \sum_{k=1}^{n_i} y_{ik}}_{\text{estimator of } \mu} \quad \text{and} \quad \underbrace{\overline{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}}_{\text{estimator of } \mu_i}, \quad i = 1, 2, \ldots, r.$$

$\widehat{\alpha}_i = \overline{y}_i - \overline{y}$ gives then an estimator for α_i.

The random error terms ϵ_{ik} can then be estimated via

$$\epsilon_{ik} = y_{ik} - \overline{y}_i = y_{ik} - \overline{y} - \underbrace{(\overline{y}_i - \overline{y})}_{=\widehat{\alpha}_i} = y_{ik} - (\overline{y} + \widehat{\alpha}_i)$$

Example (Students Learning): $\overline{y}$ = estimator for the average mark μ,
$\overline{y}_i$ = estimator for the average mark μ_i with teaching method A_i,
$\widehat{\alpha}_i = \overline{y}_i - \overline{y}$ = estimator for the effect α_i of teaching method A_i

Decomposition of the Sum of Squares (SST)

$$\underbrace{\sum_{i=1}^{r}\sum_{k=1}^{n_i}(\underbrace{y_{ik}-\overline{y}}_{=\widehat{\alpha}_i+\epsilon_{ik}})^2}_{\substack{=\text{ SST variation}\\ \text{from grand mean}}} = \underbrace{\sum_{i=1}^{r} n_i\cdot(\underbrace{\overline{y}_i-\overline{y}}_{=\widehat{\alpha}_i})^2}_{\substack{=\text{ SSA variation}\\ \text{between groups}}} + \underbrace{\sum_{i=1}^{r}\sum_{k=1}^{n_i}(\underbrace{y_{ik}-\overline{y}_i}_{=\epsilon_{ik}})^2}_{\substack{=\text{ SSE variation}\\ \text{within groups}}}$$

Note: SSE collects the random errors due to the variation in each group.

Example (Students Learning):

- SST = (squared) variation from the average mark among all students
- SSA = (squared) variation of the average marks for the different teaching methods from the overall average mark
- SSE = sum of the (squared) variations within the groups taught with one teaching method from the average mark in that group

Mean Square Variations

We divide each sum of squares by its degrees of freedom (df):
$\mathrm{df}_{\mathrm{SST}} = N-1$, $\mathrm{df}_{\mathrm{SSA}} = r-1$, $\mathrm{df}_{\mathrm{SSE}} = N-r$ where $N = \sum_{i=1}^{r} n_i$.

$$\mathrm{MST} = \frac{\mathrm{SST}}{N-1} = \frac{1}{N-1} \sum_{i=1}^{r} \sum_{k=1}^{n_i} (y_{ik} - \overline{y})^2,$$

$$\mathrm{MSA} = \frac{\mathrm{SSA}}{r-1} = \frac{1}{r-1} \sum_{i=1}^{r} n_i \cdot (\overline{y}_i - \overline{y})^2,$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-r} = \frac{1}{N-r} \sum_{i=1}^{r} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_i)^2.$$

Motivation: If the means μ_i in the subpopulations are not all equal,

then the ratio $F_{r-1,N-r} = \dfrac{\mathrm{MSA}}{\mathrm{MSE}} = \dfrac{\mathrm{SSA}/(r-1)}{\mathrm{SSE}/(N-r)}$ should be large.

Example: Student Learning – Mean Square Variations

Student subpopulation P_i corresponds to teaching method A_i, $i = 1, 2, 3$. The sample size in P_i is $n_i = 100$.

- Overall sample (from all students) has size $N = n_1 + n_2 + n_3 = 300$
- $\mathrm{df}_{\mathrm{SST}} = N - 1 = 299$, $\mathrm{df}_{\mathrm{SSA}} = r - 1 = 2$, $\mathrm{df}_{\mathrm{SSE}} = N - r = 297$
- $\mathrm{MST} = \mathrm{SST}/(N-1) =$ squared average mark variation (from the overall average mark) among the students in the overall sample
- $\mathrm{MSA} = \mathrm{SSA}/(p-1) =$ squared average variation of the average marks for the teaching methods from the overall average mark
- $\mathrm{MSE} = \mathrm{SSE}/(N-p) =$ squared average variation of the marks from the average marks within the groups/squared average random error

If the teaching method affects the mark then $\dfrac{\mathrm{MSA}}{\mathrm{MSE}}$ should be large.

Hypothesis Test for One-Way Analysis of Variance

The null hypothesis

H_0: $\mu_1 = \mu_2 = \ldots = \mu_r = \mu$ (or equivalently $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$)

is tested under the following assumptions:

(i) The variances of Y in the populations $P_1, P_2, \ldots, P_r$ are equal.

(ii) Y is normally distributed within each subpopulation P_i and in P.

Then the random variable

$$F = F_{r-1,N-r} = \frac{\text{MSA}}{\text{MSE}} = \frac{\text{SSA}/(r-1)}{\text{SSE}/(N-r)}$$

follows an F distribution with numerator degrees of freedom $\text{df} = r - 1$ and denominator degrees of freedom $\text{df} = N - r$.

We reject H_0 with significance level α if the value $f = \frac{\text{MSA}}{\text{MSE}}$ computed for F satisfies $f > f_{r-1,N-r,\alpha}$, where $f_{r-1,N-r,\alpha}$ is the number for which

$$(\text{Probability for } F > f_{r-1,N-r,\alpha}) = P(F > f_{r-1,N-r,\alpha}) = \alpha.$$

1-Way ANOVA Table

For computing a 1-way analysis of variance (ANOVA) it is useful to employ a 1-way ANOVA table to systematically work out the required values:

Source of Variation	Degrees of Freedom (df)	Sum of Squares	Mean Sum of Squares	F
Between Groups	$r-1$	SSA	$\text{MSA} = \frac{\text{SSA}}{r-1}$	$\frac{\text{MSA}}{\text{MSE}}$
Within Groups	$N-r$	SSE	$\text{MSE} = \frac{\text{SSE}}{N-r}$	
Total	$N-1$	SST		

We will now perform a 1-way ANOVA for an example.

Ex. 2.3: Effect of Teaching Method on Student Marks

A sample of 4 students is taken from each subpopulation P_i, where $P_i =$ subpopulation taught with teaching method A_i, and where $A_1 =$ traditional teaching, $A_2 =$ distance learning, $A_3 =$ blended learning.

The random variable $Y =$ mark (of the student) is measured for each sample, giving the data in the table below.

	A_1	A_2	A_3
1	70	57	88
2	80	54	82
3	75	46	90
4	75	43	80
sum			
$\overline{y}_i = \frac{\text{sum}}{n_i}$			

Perform a 1-way ANOVA for this data:

Compute the means.

Then compute the sums of squares and the mean square deviations.

Finally use hypothesis testing with a significance level of $\alpha = 0.05$ (and $\alpha = 0.01$) to find whether the teaching method has any effect on the marks.

Methods of Multivariate Statistics

Part 2.2: Two-Way Analysis of Variance

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Example: Crop Yield Depending on Soil Quality, Fertilizer

Does the crop yield depend on the soil quality and/or the method of fertilization?

- The population P consists of all fields.
- factor A: soil quality with factor levels given by soil types A_1, A_2, A_3
- factor B: method of fertilization with factor levels given by fertilizers $B_1, B_2, \ldots, B_4$
- We observe that there are 12 different combinations $A_i \times B_j$, $i = 1, 2, 3$, $j = 1, 2, 3, 4$, of soil type A_i and fertilizer B_j.
- metric variable Y: crop yield Y measured in tons of crop per km^2
- We can measure Y in the following subpopulations:

  $P_{i\cdot} =$ all fields having soil type A_i (no assumption on fertilizer),

  $P_{\cdot j} =$ all fields fertilized with fertilizer B_j (no assumption on soil type),

  $P_{ij} =$ all fields having soil type A_i and being fertilized with fertilizer B_j

Idea of Two-Way/Two-Factor Analysis I

Consider two factors (independent qualitative variables) A and B, defined on a population P, with factor levels $A_1, A_2, \ldots, A_r$ and $B_1, B_2, \ldots, B_q$.

The levels of the factors A and B divide the population P into groups:

- $P_{i\cdot}$ = elements for which A has the factor level A_i,
- $P_{\cdot j}$ = elements for which B has the factor level B_j,
- P_{ij} = elements for which A and B have the factor levels A_i and B_j, denoted as $A_i \times B_j$.

Example (Crop Yield Depending on Soil Quality and Fertilizer):

- population: P = all fields,
- factors: A = soil quality, B = method of fertilization,
- subpopulations: $P_{i\cdot}$ = fields with soil type A_i,
  $P_{\cdot j}$ = fields fertilized with fertilizer B_j,
  P_{ij} = fields with soil type A_i and fertilized with fertilizer B_j.

Idea of Two-Way/Two-Factor Analysis II

Consider a metric variable Y defined on a population P.

We want to investigate whether Y depends on A and B and possibly on the 'interaction' $A \times B$ ('interaction' = particular combination).

The two-way/two-factor analysis (2-way ANOVA) considers the following arithmetic means of Y:

- $\mu =$ grand mean of Y in the whole population P
- $\mu_{i.} =$ mean of Y in the subset $P_{i.}$ (elements with factor level A_i),
- $\mu_{.j} =$ mean of Y in the subset $P_{.j}$ (elements with factor level B_j),
- $\mu_{ij} =$ mean of Y in the subset P_{ij} (elements with factor levels $A_i \times B_j$),

---

The 2-way ANOVA investigates whether:

- $\mu_{i.}$ depends on A_i,
- $\mu_{.j}$ depends on B_j,
- μ_{ij} depends on A_i, B_j and the 'interaction' $A_i \times B_j$

Example: Crop Yield Depending on Soil Quality, Fertilizer

metric variable: $Y =$ crop yield measured in tons of crop per km^2

The various (arithmetic) means are:

- $\mu =$ average crop yield in the population P of all fields
- $\mu_{i.} =$ average crop yield for all fields with soil type A_i
- $\mu_{.j} =$ average crop yield for all fields fertilized with B_j
- $\mu_{ij} =$ average crop yield for all fields with soil type A_i and fertilizer B_j

---

Research Questions:

- Does $\mu_{i.}$ depend on the soil type A_i?
- Does $\mu_{.j}$ depend on the fertilizer B_j?
- Does μ_{ij} depend on the soil type A_i, the fertilizer B_j and the 'interaction' (i.e. the particular combination of soil type and fertilizer) $A_i \times B_j$?

Two-Way ANOVA: Mathematical Model I – Setup

- population P of objects
- two factors/independent qualitative variables on the population P:
  factor A with factor levels $A_1, A_2, \ldots, A_r$
  factor B with factor levels $B_1, B_2, \ldots, B_q$
- subpopulations: $P_{i\cdot} =$ elements for which A has the factor level A_i,
  $P_{\cdot j} =$ elements for which B has the factor level B_j,
  $P_{ij} =$ elements for which A and B have the factor levels $A_i \times B_j$.
- $Y =$ metric random variable that we expect to depend on the factors A and B and possibly on their 'interaction' $A \times B$

Assumptions:

- Y is normally distributed with the same variance σ^2 in P and in each of the subsets $P_{i\cdot}$, $P_{\cdot j}$ and P_{ij}.
- The factors A and B and the 'interaction' $A \times B$ are independent qualitative variables. Hence they are not correlated.

Two-Way ANOVA: Mathematical Model II – Means

Research Question: Does Y depend on the factors A and/or B and possibly their 'interaction' $A \times B$?

Grand Mean of Y, Means of Y in the Different Subpopulations:

- $\mu =$ grand mean of Y for the whole population P
- $\mu_{i.} =$ mean of Y on the subset $P_{i.}$ of objects with factor level A_i
- $\mu_{.j} =$ mean of Y on the subset $P_{.j}$ of objects with factor level B_j
- $\mu_{ij} =$ mean of Y on the subset P_{ij} of objects with factor levels $A_i \times B_j$

Approach: If Y does depend on A and/or B and possibly their 'interaction' $A \times B$, then the means above should not all be the same.

We will postulate a model for the different means, and estimate the variables in the model from sampled data.

Two-Way ANOVA: Mathematical Model III – Model

The two-way analysis of variance (2-way ANOVA) postulates that

$$\underbrace{\mu_{ij}}_{\substack{\text{mean on}\\ \text{population } P_{ij}\\ \text{with } A_i \text{ and } B_j}} = \underbrace{\mu}_{\substack{\text{grand}\\ \text{mean}}} + \underbrace{\alpha_i}_{\text{effect of } A_i} + \underbrace{\beta_j}_{\text{effect of } B_j} + \underbrace{\gamma_{ij}}_{\substack{\text{effect of the}\\ \text{interaction}\\ \text{of } A_i \text{ and } B_j}}$$

where:

effect of A_i: $\alpha_i = \mu_{i.} - \mu$ from $\mu_{i.} = \mu + \alpha_i$

effect of B_j: $\beta_j = \mu_{.j} - \mu$ from $\mu_{.j} = \mu + \beta_j$

effect from the interaction of A_i and B_j:

$$\begin{aligned} \gamma_{ij} &= \mu_{ij} - \mu - \alpha_i - \beta_j \\ &= \mu_{ij} - \mu - (\mu_{i.} - \mu) - (\mu_{.j} - \mu) \\ &= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu \end{aligned}$$

Note: If the factors A and B do not interact, then we set $\gamma_{ij} = 0$ and then perform a 2-way ANOVA without interaction ($\rightarrow$ textbooks).

Example: Crop Yield Depending on Soil Quality, Fertilizer

Model: average crop yield of a field with soil quality A_i and fertilizer B_j

$$\underbrace{\mu_{ij}}_{\substack{\text{average crop yield}\\ \text{for soil type } A_j\\ \text{and fertilizer } B_j}} = \underbrace{\mu}_{\substack{\text{grand mean:}\\ \text{average crop yield}}} + \underbrace{\alpha_i}_{\substack{\text{effect of}\\ \text{soil type } A_i\\ \text{on crop yield}}} + \underbrace{\beta_j}_{\substack{\text{effect of}\\ \text{fertilizer } B_j\\ \text{on crop yield}}} + \underbrace{\gamma_{ij}}_{\substack{\text{effect of the}\\ \text{interaction}\\ \text{of soil } A_i\\ \text{and fertilizer } B_j}}$$

- effect of soil type A_i: $\alpha_i = \mu_{i\cdot} - \mu$
- effect of fertilizer B_j: $\beta_j = \mu_{\cdot j} - \mu$
- effect of 'interaction' $A_i \times B_j$ of soil type A_i and fertilizer B_j: $\gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$

If the crop yield does not depend on the soil type and the fertilizer, then:

$$\alpha_i = 0, \ \beta_j = 0 \text{ and } \gamma_{ij} = 0 \text{ for } i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, q.$$

In this case $\mu_{ij} = \mu_{i\cdot} = \mu_{\cdot j} = \mu$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, q$, i.e. the average crop yield is the same in all subpopulations.

Two-Way ANOVA: Hypothesis Testing

The null hypotheses H_0 are tested against their alternative hypotheses H_1.

(i) **Hypotheses About Factor A:**

H_0^A: $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$ (or equiv. $\mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{r\cdot}$).

H_1^A: For at least one factor level A_i we have $\alpha_i \neq 0$ (or equiv. $\mu_{i\cdot} \neq \mu_{k\cdot}$ for at least one pair i and k).

(ii) **Hypotheses About Factor B:**

H_0^B: $\beta_1 = \beta_2 = \ldots = \beta_q = 0$ (or equiv. $\mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot q}$).

H_1^B: For at least one factor level B_j we have $\beta_j \neq 0$ (or equiv. $\mu_{\cdot j} \neq \mu_{\cdot k}$ for at least one pair j and k).

(iii) **Hypotheses About the Interaction $A \times B$:**

$H_0^{A\times B}$: $\gamma_{1,1} = \gamma_{1,2} = \ldots = \gamma_{r,q-1} = \gamma_{r,q} = 0$.

$H_1^{A\times B}$: For at least one combination $A_i \times B_j$ of factor levels $\gamma_{ij} \neq 0$.

Example: Crop Yield Depending on Soil Quality, Fertilizer

What do the hypotheses say for our example?

(i) **Hypotheses About Factor A:**

H_0^A: The crop yield does not depend on the soil type.

H_1^A: The crop yield does depend on the soil type.

(ii) **Hypotheses About Factor B:**

H_0^B: The crop yield does not depend on the fertilizer.

H_1^B: The crop yield does depend on the type of fertilizer used.

(iii) **Hypotheses About the Interaction $A \times B$:**

$H_0^{A\times B}$: There is no 'interaction' between the soil type and fertilizer.

$H_1^{A\times B}$: There is an 'interaction' for at least one soil type A_i and one fertilizer B_j.

Sampling Y to Estimate the Means from Empirical Data

In each subpopulation P_{ij}, i.e. among the objects with the combination of factor levels $A_i \times B_j$, we we take a sample of size n_{ij} and measure Y:

$$\underbrace{y_{ijk}, \qquad k = 1, 2, \ldots, n_{ij}}_{n_{ij} \text{ measurements of } Y \text{ in } P_{ij}} \qquad \left(\begin{array}{ll} i = & \text{index for factor level } A_i \\ j = & \text{index for factor level } B_j \\ k = & \text{index for number in sample} \end{array} \right)$$

Orthogonal Two-Way Analysis of Variance:

We only consider the case of the orthogonal two-way analysis of variance (orthogonal 2-way ANOVA), where all samples are of the same size:

$$n_{1,1} = n_{1,2} = \ldots = n_{r,q-1} = n_{r,q} = n.$$

Example (Crop Yield Depending on Soil Quality and Fertilizer):

y_{ijk} = crop yield of kth field in sample with soil type A_i and fertilizer B_j

Model for the Sampled Data

$$y_{ijk} = \underbrace{\mu_{ij}}_{\substack{\text{mean in}\\ \text{population } P_{ij}\\ \text{with } A_i \text{ and } B_j}} + \underbrace{\epsilon_{ijk}}_{\substack{\text{random}\\ \text{error}}} = \underbrace{\mu}_{\substack{\text{grand}\\ \text{mean}}} + \underbrace{\alpha_i}_{\text{effect of } A_i} + \underbrace{\beta_j}_{\text{effect of } B_j} + \underbrace{\gamma_{ij}}_{\substack{\text{effect of the}\\ \text{interaction}\\ \text{of } A_i \text{ and } B_j}} + \underbrace{\epsilon_{ijk}}_{\substack{\text{random}\\ \text{error}}}$$

Assumption: The random error terms ϵ_{ijk} are all normally distributed with mean value zero and the same variance.

Example (Crop Yield Depending on Soil Quality and Fertilizer):

$$\begin{aligned} y_{ijk} &= \underbrace{\mu_{ij}}_{\substack{\text{average crop yield}\\ \text{for soil type } A_i\\ \text{and fertilizer } B_j}} + \underbrace{\epsilon_{ijk}}_{\substack{\text{random}\\ \text{error}}} \\ &= \underbrace{\mu}_{\substack{\text{grand mean:}\\ \text{average crop yield}}} + \underbrace{\alpha_i}_{\substack{\text{effect of}\\ \text{soil type } A_i\\ \text{on crop yield}}} + \underbrace{\beta_j}_{\substack{\text{effect of}\\ \text{fertilizer } B_j\\ \text{on crop yield}}} + \underbrace{\gamma_{ij}}_{\substack{\text{effect of the}\\ \text{interaction}\\ \text{of soil type } A_i\\ \text{and fertilizer } B_j}} + \underbrace{\epsilon_{ijk}}_{\substack{\text{random}\\ \text{error}}} \end{aligned}$$

Estimating the Means from the Empirical Data I

Size of the overall sample in the population P: $N = \sum_{i=1}^{r} \sum_{j=1}^{q} \underbrace{n_{ij}}_{=n} = r \cdot q \cdot n$

The grand mean μ of Y in P is estimated by:

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{q} \sum_{k=1}^{n} y_{ijk} = \frac{1}{N} \underbrace{\sum_{i=1}^{r}}_{\substack{\text{sum over} \\ \text{factor levels} \\ A_i \text{ of } A}} \underbrace{\sum_{j=1}^{q}}_{\substack{\text{sum over} \\ \text{factor levels} \\ B_j \text{ of } B}} \underbrace{\sum_{k=1}^{n}}_{\substack{\text{sum over} \\ \text{objects in} \\ \text{sample} \\ \text{from } P_{ij}}} y_{ijk}.$$

The mean $\mu_{i\cdot}$ of Y in $P_{i\cdot}$ (= objects with factor level A_i) is estimated by:

$$\overline{y}_{i\cdot} = \frac{1}{n \cdot q} \sum_{j=1}^{q} \sum_{k=1}^{n} y_{ijk}$$

Estimating the Means from the Empirical Data II

The mean $\mu_{.j}$ of Y in $P_{.j}$ (= objects with factor level B_j) is estimated by:

$$\overline{y}_{.j} = \frac{1}{n \cdot r} \sum_{i=1}^{r} \sum_{k=1}^{n} y_{ijk}$$

The mean μ_{ij} of Y in P_{ij} (= objects with factor level combination $A_i \times B_j$) is estimated by:

$$\overline{y}_{ij} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk}$$

The effects α_i, β_j, γ_{ij} of A_i, B_j, $A_i \times B_j$, respectively, are estimated by:

$$\widehat{\alpha_i} = \overline{y}_{i.} - \overline{y}, \qquad \widehat{\beta_j} = \overline{y}_{.j} - \overline{y}, \qquad \widehat{\gamma_{ij}} = \overline{y}_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}.$$

Decomposition of the Sum of Squares (SST)

The variation from the grand mean can be decomposed as follows

$$\mathrm{SST} = \sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{n} \big(\underbrace{y_{ijk} - \overline{y}}_{=\widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma_{ij}} + \epsilon_{ijk}} \big)^2 = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSAB} + \mathrm{SSE}$$

into the variations between the groups for the factor levels of A or of B

$$\mathrm{SSA} = n \cdot q \sum_{i=1}^{r} (\underbrace{\overline{y}_{i\cdot} - \overline{y}}_{=\widehat{\alpha}_i})^2 \qquad \text{and} \qquad \mathrm{SSB} = n \cdot r \sum_{j=1}^{q} (\underbrace{\overline{y}_{\cdot j} - \overline{y}}_{=\widehat{\beta}_j})^2,$$

the variations between the groups for the interaction levels of $A \times B$

$$\mathrm{SSAB} = n \sum_{i=1}^{r}\sum_{j=1}^{q} (\underbrace{\overline{y}_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}}_{=\widehat{\gamma_{ij}}})^2$$

and the variation within the groups due to random error

$$\mathrm{SSE} = \sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{n} (\underbrace{y_{ijk} - \overline{y}_{ij}}_{=\epsilon_{ijk}})^2.$$

Example: Crop Yield Depending on Soil Quality, Fertilizer

In our example: $r = 3$ soil types A_i and $q = 4$ types of fertilizer B_j

$$\mathrm{SST} = \underbrace{\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{n}}_{\substack{\text{sum over the soil types } A_i,\\ \text{sum over the types of fertilizer } B_j,\\ \text{and sum over the fields in each sample}}} \Big(\underbrace{y_{ijk} - \overline{y}}_{\substack{\text{difference in crop yield for}\\ k\text{th field in } P_{ij} \text{ from the}\\ \text{average crop yield } \overline{y}}} \Big)^2$$

$$\mathrm{SSA} = \underbrace{n \cdot 4}_{\substack{(\text{size } n \text{ of sample})\\ \times\ (\text{number of the}\\ \text{fertilizers } B_j)}} \underbrace{\sum_{i=1}^{3}}_{\substack{\text{sum over the}\\ \text{soil types } A_i}} \Big(\underbrace{\overline{y}_{i\cdot} - \overline{y}}_{\substack{\text{difference in the average of}\\ \text{the crop yield for fields with}\\ \text{soil type } A_i \text{ from the}\\ \text{average crop yield } \overline{y}}} \Big)^2$$

Ex. 2.4: Interpret the other sums for our example.

Mean Square (MS) Variations and 2-Way ANOVA Table

Mean square variations are computed with an ANOVA table: $N = r \cdot q \cdot n$

Source	Sum of Squares	Degrees of Freedom (df)	Mean Square Variations
Factor A	SSA	$r-1$	$\text{MSA} = \frac{\text{SSA}}{r-1}$
Factor B	SSB	$q-1$	$\text{MSB} = \frac{\text{SSB}}{q-1}$
$A \times B$	SSAB	$(r-1)\cdot(q-1)$	$\text{MSAB} = \frac{\text{SSAB}}{(r-1)\cdot(q-1)}$
Random error	SSE	$N - r\cdot q$	$\text{MSE} = \frac{\text{SSE}}{N-r\cdot q}$
Total	SST	$N-1$	$\text{MST} = \frac{\text{SST}}{N-1}$

Example: Crop Yield Depending on Soil Quality, Fertilizer

- MST is the (squared) average variation of the crop yield.
- MSA is the (squared) average variation of the (average) crop yield for the different soil types A_i.
- MSB is the (squared) average variation of the (average) crop yield for the different fertilizers B_j:

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{q-1} = \frac{n \cdot r}{q-1} \sum_{j=1}^{q} \Big(\underbrace{\overline{y}_{\cdot j} - \overline{y}}_{\substack{= \widehat{\beta}_j = \text{effect} \\ \text{from fertilizer } B_j}} \Big)^2$$

- MSAB is the (squared) average 'interaction' $A_i \times B_j$ of soil type A_i and fertilizer B_j.
- MSE is the (squared) average random variation of the crop yield within the groups corresponding to soil type A_i and fertilizer B_j. MSE is the (squared) average random error.

Hypotheses Testing with the F-Distribution

$$F_A = \frac{\text{MSA}}{\text{MSE}}, \qquad F_B = \frac{\text{MSB}}{\text{MSE}}, \qquad F_{A\times B} = \frac{\text{MSAB}}{\text{MSE}} \tag{7}$$

are random variables following an F-distribution with (numerator,denominator)-degrees of freedom $(r-1, N-r\cdot q)$, $(q-1, N-r\cdot q)$ and $((r-1)\cdot(q-1), N-r\cdot q)$, respectively.

We denote the numerical values for (7) for our data by f_A, f_B and $f_{A\times B}$.

Given a significance level α, we reject the null hypothesis H_0^A (H_0^B, $H_0^{A\times B}$) if $f_A > f_{r-1,N-rq,\alpha}$ ($f_B > f_{q-1,N-rq,\alpha}$, $f_{A\times B} > f_{(r-1)(q-1),N-rq,\alpha}$), where $f_{r-1,N-rq,\alpha}$ ($f_{q-1,N-rq,\alpha}$, $f_{(r-1)(q-1),N-rq,\alpha}$) is the number for which

$$(\text{Probability for } F_A > f_{r-1,N-rq,\alpha}) = P(F_A > f_{r-1,N-rq,\alpha}) = \alpha$$

$$\Big((\text{Probability for } F_B > f_{q-1,N-rq,\alpha}) = P(F_B > f_{q-1,N-rq,\alpha}) = \alpha,$$

$$(\text{Prob. for } F_{A\times B} > f_{(r-1)(q-1),N-rq,\alpha}) = P(F_{A\times B} > f_{(r-1)(q-1),N-rq,\alpha}) = \alpha\Big).$$

Ex. 2.5: Crop Yield Depends on Soil Quality, Fertilizer

Does the crop yield (measured in tons per km^2) depend on the soil type, the type of fertilizer and their interaction?

Here we consider 3 soil types A_1, A_2, A_3 and 2 types of fertilizer B_1 and B_2. We are given the following data for the crop yield Y:

	B_1	B_2	Means
A_1	$y_{1,1,1} = 2,\ y_{1,1,2} = 2$	$y_{1,2,1} = 3,\ y_{1,2,2} = 4$	
A_2	$y_{2,1,1} = 1,\ y_{2,1,2} = 2$	$y_{2,2,1} = 4,\ y_{2,2,2} = 5$	
A_3	$y_{3,1,1} = 3,\ y_{3,1,2} = 2$	$y_{3,2,1} = 4,\ y_{3,2,2} = 4$	
Means			

First complete the table to compute the means $\overline{y}_{i\cdot}$, $\overline{y}_{\cdot j}$ and $\overline{y}$.

Ex. 2.5: Crop Yield Depends on Soil Quality, Fertilizer

Now compute the means $\overline{y}_{ij}$ for the interaction $A_i \times B_j$ of the factors A and B.

	B_1	B_2
A_1		
A_2		
A_3		

Next compute the sums of squares.

Now complete the 2-way ANOVA table shown on the next slide.

Ex. 2.5: Crop Yield Depends on Soil Quality, Fertilizer

Source	Sum of Squares	Degrees of Freedom (df)	Mean Square Variation	F-Value
Factor A				
Factor B				
$A \times B$				
Error				
Total				

Finally formulate the three null hypotheses and alternative hypotheses.

Determine with a significance level of $\alpha = 0.05$ which of the three null hypotheses can be rejected. Interpret your result!

Methods of Multivariate Statistics

Topic 3: Measuring Distances & Investigating Data

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Topic 3: Measuring Distances and Investigating Data

- data matrix for m random variables measured on n objects
- two points of view of investigating the data to:
  1. study the relationships between the random variables
  2. study the relationships between the objects
- geometric representation of the data
- distance functions/metrics:
  - city block distance
  - Euclidean distance
  - Tschbyscheff distance/L_∞-norm
  - Mahalanobis distance

Note: We will need distances and the concepts introduced in this chapter to understand discriminant analysis and cluster analysis.

Representation of Data: The Data Matrix

Situation: m metric random variables $X_1, X_2, \ldots, X_m$ are measured on n objects $e_1, e_2, \ldots, e_n$.

$$x_{ij} = \text{observed value for } j\text{th variable } X_j \text{ on } i\text{th object}$$

The data is represented in the data matrix $\mathbf{X}$ in the following way:

$$\mathbf{X} = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \ddots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{array}{l} \leftarrow \text{object } e_1 \\ \leftarrow \text{object } e_2 \\ \\ \leftarrow \text{object } e_n \end{array}$$

$$\begin{array}{lcccc} & \uparrow & \uparrow & & \uparrow \\ \text{variable} & X_1 & X_2 & & X_m \end{array}$$

Example: objects: n persons; variables: $X_1 = \text{height}$, $X_2 = \text{weight}$

Interpretation of the Data Matrix: Two Points of View

1. The *j*th column contains the values of X_j for the objects $e_1, e_2, \ldots, e_n$:

$$\mathbf{x}_{\cdot j} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} = j\text{th column of } \mathbf{X}.$$

   If we compare the different columns of $\mathbf{X}$, then we study the relationships between the different variables $X_1, X_2, \ldots, X_m$.

   Methods: regression, factor analysis, structural equation modeling.

2. The *i*th row contains the values of $X_1, X_2, \ldots, X_m$ for the object e_i:

$$\mathbf{x}'_{i\cdot} = (x_{i1}, x_{i2}, \cdots, x_{im}) = i\text{th row of } \mathbf{X}.$$

   If we compare the different rows of $\mathbf{X}$, then we study the relationships between the different objects $e_1, e_2, \ldots, e_n$ in our sample.

   Methods: discriminant analysis, cluster analysis.

It is often useful to standardize the data:

Standardized Data:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, \quad \text{where} \quad \overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$$

The z_{ij}, $i = 1, 2, \ldots, n$, have now (arithmetic) mean $= 0$ and variance $= 1$.

Standardized Data Matrix:

$$\mathbf{Z} = (z_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \ddots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{pmatrix}$$

We have corresponding standardized random variables: $Z_j = (X_j - \mu_j)/\sigma_j$ where $\mu_j = \mathrm{E}(X_j)$ and $\sigma_j = \sqrt{\mathrm{Var}(X_j)}$.

Visualization of the Standardized Data – Method 1

We plot the columns $\mathbf{z}_{\cdot j}$ of the standardized data matrix $\mathbf{Z}$ in a coordinate system with n perpendicular axes.

$$\mathbf{z}_{\cdot j} = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{pmatrix}$$

- The ith axis in the coordinate system corresponds to object e_i.
- The column vector $\mathbf{z}_{\cdot j}$ represents the sampled data for the standardized variable Z_j (from the n objects $e_1, e_2, \ldots, e_n$).
- From the standardization, the vector $\mathbf{z}_{\cdot j}$ has length $\sqrt{n-1}$.
- If the random variables X_j and X_k are strongly positively (negatively) correlated then the corresponding data vectors we will be almost parallel (anti-parallel), i.e. their angle is close to 0° (180°).
- If X_j and X_k are uncorrelated then the corresponding data vectors will be almost perpendicular, i.e. their angle is close to 90°.

Ex. 3.1: Visualization of Height, Weight, Inseam Length

Visualize the following data with Method 1 and interpret your results.

Person	height in cm	weight in kg	inseam length in cm
e_1	180	74	78
e_2	160	50	68
e_3	170	65	73

Why is the standardization of the variables here particularly useful?

Visualization of the Standardized Data – Method 2

We plot the rows $\mathbf{x}'_{i.}$ of the non-standardized data matrix $\mathbf{X}$ in a coordinate system with m perpendicular axes

$$\mathbf{x}'_{i.} = (x_{i1}, x_{i2}, \cdots, x_{im})$$

- The jth axis in the coordinate system corresponds to the variable X_j.
- The row vector $\mathbf{x}'_{i.}$ corresponds to the data for object e_i (for the m random variables $X_1, X_2, \ldots, X_m$).
- If two objects e_i and e_k are similar, then their points in the coordinate system will be close together.

We can form groups/clusters of similar objects based on the location in the coordinate system. $\rightarrow$ We need to know how we measure distance.

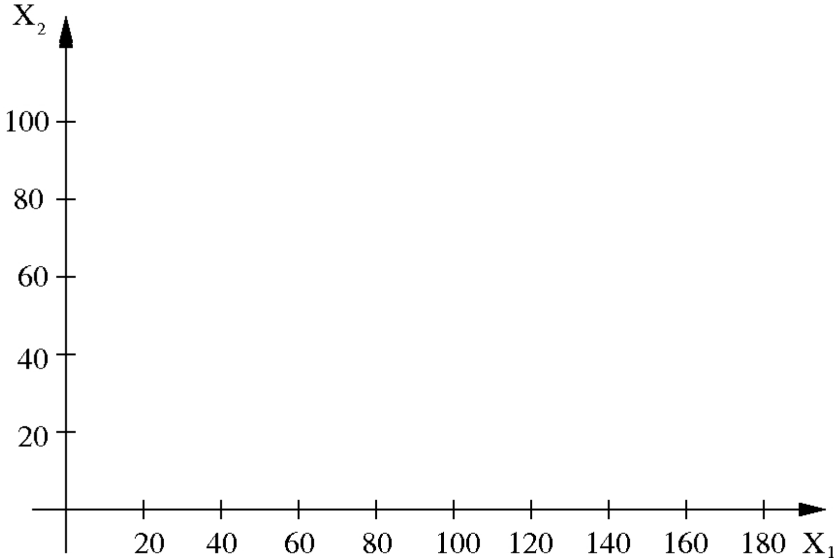
Note: Method 2 leads to discriminant analysis and cluster analysis.

Ex. 3.2: Height and Weight, Visualization with Method 2

Write down the data matrix and **X** and visualize the following data with Method 2. A suitable coordinate system has been provided on the next slide. Interpret your results.

Person	height in cm	weight in kg
e_1	180	72
e_2	181	90
e_3	182	71
e_4	181	91

Ex. 3.2: Height and Weight, Visualization with Method 2



Why Distances/Metrics are Needed

From now on we want to investigate the relationships between the different objects in our sample.

Each object e_i is represented through a row vector $\mathbf{x}'_{i\cdot}$ in our non-standardized data matrix $\mathbf{X}$, giving the values of the random variables $X_1, X_2, \ldots, X_m$ for e_i.

To investigate the relationships between objects e_i and e_k we need to measure 'how far' two objects e_i and e_k are apart. We measure this with distances.

With the help of distances we can:

- classify objects into groups/clusters $\longrightarrow$ cluster analysis
- find functions that discriminate between given groups and allows us to sort new objects into an appropriate group $\longrightarrow$ discriminant analysis

Example: Euclidean Distance/Metric

$\mathbf{x}'_{i\cdot} = (x_{i1}, x_{i2}, \ldots, x_{im})$ ith row of $\mathbf{X}$ (values of random variables for e_i)

$\mathbf{x}'_{k\cdot} = (x_{k1}, x_{k2}, \ldots, x_{km})$ kth row of $\mathbf{X}$ (values of random variables for e_k)

The Euclidean distance/metric of object e_i and object e_k is given by

$$d_{ik} = \|\mathbf{x}_{i\cdot} - \mathbf{x}_{k\cdot}\|_2 = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2}$$

Ex. 3.3: Compute the Euclidean distance between the following persons, based on the given data of their height and weight. Comment on your results.

Person	height (cm)	weight (kg)
e_1	180	72
e_2	181	90
e_3	182	71
e_4	181	91

Definition of a Distance (Function)/Metric

The distance d_{ik} between object e_i and object e_k must satisfy the following conditions:

(i) $d_{ik} \geq 0$ for all $i, k = 1, 2, \ldots, n$.

(The distance is non-negative.)

(ii) $d_{ik} = d_{ki}$ for all $i, k = 1, 2, \ldots, n$ (symmetry).

(The distance from object e_i to object e_k is the same as the distance from object e_k to object e_i.)

(iii) $d_{ii} = 0$ for all $i = 1, 2, \ldots, n$.

(The distance of an object from itself is zero.)

Example: The Euclidean distance has all these properties.

City Block Distance and Tschebyscheff Distance

$\mathbf{x}'_{i\cdot} = (x_{i1}, x_{i2}, \ldots, x_{im})$ ith row of $\mathbf{X}$ (values of random variables for e_i)

$\mathbf{x}'_{k\cdot} = (x_{k1}, x_{k2}, \ldots, x_{km})$ kth row of $\mathbf{X}$ (values of random variables for e_k)

The city block distance (L_1-norm) of the objects e_i and e_k is given by

$$d_{ik} = \|\mathbf{x}_{i\cdot} - \mathbf{x}_{k\cdot}\|_1 = \sum_{j=1}^{m} |x_{ij} - x_{kj}|.$$

The Tschebyscheff distance (L_∞-norm) of the objects e_i and e_k is given by

$$d_{ik} = \|\mathbf{x}_{i\cdot} - \mathbf{x}_{k\cdot}\|_\infty = \max_{j=1,2,\ldots,m} |x_{ij} - x_{kj}|.$$

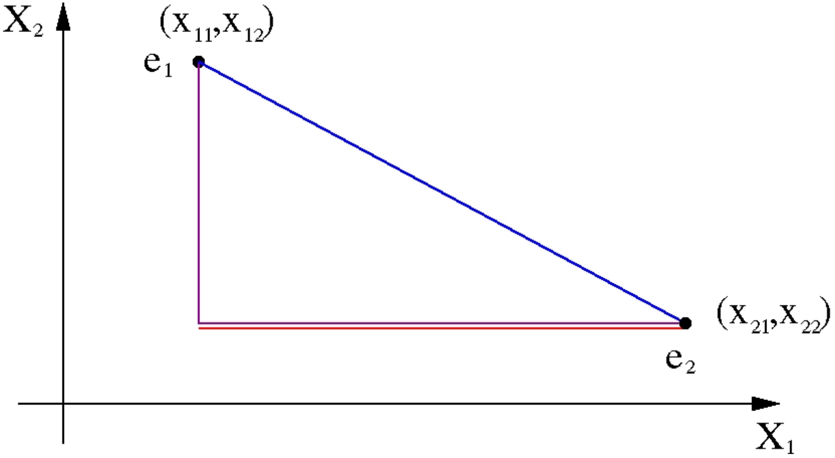
Ex. 3.4: City Block Distance and Tschebyscheff Distance

Compute the city block distance and Tschebyscheff distance between the following persons, based on the given data of their height and weight. Comment on your results.

Person	height (cm)	weight (kg)
e_1	180	72
e_2	181	90
e_3	182	71
e_4	181	91

Visualization of Different Distances

The plot below shows the Euclidean distance, the city block distance and the Tschebyscheff distance of two objects e_1 and e_2 for $m = 2$ random variables (i.e. 2 coordinate axes).



Mahalanobis Distance

Let $\mathbf{S}$ be the empirical covariance matrix of our data:

$$\mathbf{S} = (s_{jk})_{\substack{j=1,2,\ldots,m\\ k=1,2,\ldots,m}} \quad \text{with} \quad \underbrace{s_{jk} = \widehat{\mathrm{Cov}}(X_j, X_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}_{\substack{= \text{ empirical covariance for the}\\ \text{given data of } X_j \text{ and } X_k}}$$

The Mahalanobis distance between object e_i and object e_k is given by

$$d_{ik} = \sqrt{(\mathbf{x}_{i\cdot} - \mathbf{x}_{k\cdot})'\,\mathbf{S}^{-1}\,(\mathbf{x}_{i\cdot} - \mathbf{x}_{k\cdot})},$$

where $\mathbf{S}^{-1}$ is the inverse matrix of the empirical covariance matrix $\mathbf{S}$.

Note: This distance it not so easy to visualize. The intuitive idea is that it is like a 'deformed' Euclidean distance: Points with equal distance from a fixed point do no longer lie on circles but on ellipses. For $\mathbf{S} = \mathbf{I}$ (identity matrix, i.e. our data is uncorrelated), we just get the Euclidean distance.

Methods of Multivariate Statistics

Topic 4: Linear Discriminant Analysis

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Topic 4: Discriminant Analysis

Idea of Discriminant Analysis and Approaches:

- Setup: We are given g groups of objects and data for a vector $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$ of metric random variables for all objects.
- Aim: Find discriminant functions that distinguish between the groups.
- Maximum Likelihood (ML) approach, if $\mathbf{x}$ in the individual groups follows a multivariate normal distribution (not discussed).
- Fisher's linear discriminant analysis: $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$ is transformed into new random variables $Y_k = \mathbf{a}_k' \mathbf{x}$ with suitable vectors $\mathbf{a}_k$, $k = 1, 2, \ldots, r$, such that the values of Y_k distinguish well between the g groups.

4.1 Fisher's Linear Discriminant Analysis for 2 Groups

4.2 Fisher's Linear Discriminant Analysis for Multiple Groups

Idea of Fisher's Linear Discriminant Analysis

Population and its Subgroups

- A population has been subdivided into g groups $K_1, K_2, \dots, K_g$.
- The vector $\mathbf{x} = (X_1, X_2, \dots, X_m)'$ of m metric random variables is sampled in the subgroups, and its values are assumed to reflect the classification into groups.

Assumptions on the Random Variables and Their Distributions

- The probability distribution of $\mathbf{x}$ is of the same type in all groups K_ℓ (e.g. a multivariate normal distribution).
- The parameters of the distribution of $\mathbf{x}$ may differ in the groups.

Fisher's linear discriminant analysis requires no knowledge of the type of the probability distribution of $\mathbf{x} = (X_1, X_2, \dots, X_m)'$.

Methods of Multivariate Statistics

Topic 4.1: Fisher's Linear Discriminant Analysis for 2 Groups

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

- Introduce a new scalar random variable

$$Y = \mathbf{a}'\,\mathbf{x} = (a_1, a_2, \ldots, a_m)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} = a_1\,X_1 + a_2\,X_2 + \ldots + a_m\,X_m$$

- The vector $\mathbf{a}' = (a_1, a_2, \ldots, a_m)$ is determined such that the values

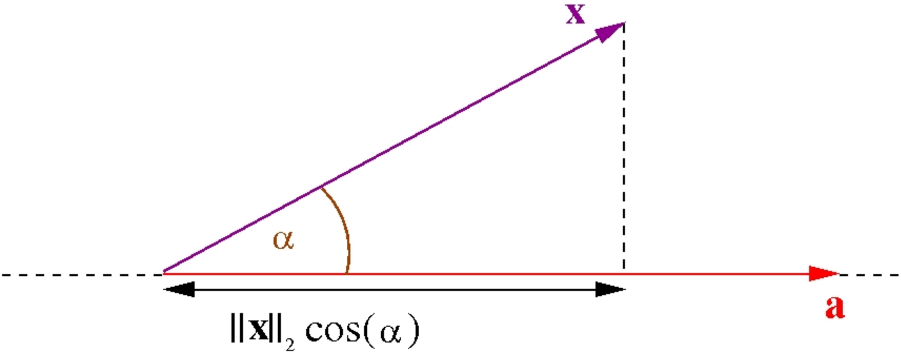
$$y = \mathbf{a}'\,\mathbf{x} = a_1\,x_1 + a_2\,x_2 + \ldots + a_m\,x_m$$

for the objects e with values $\mathbf{x} = (x_1, x_2, \ldots, x_m)'$ separate the two groups optimally.

- Normalization of $\mathbf{a}$: $\|\mathbf{a}\|_2^2 = a_1^2 + a_2^2 + \ldots + a_m^2 = 1$.

$$y = \mathbf{a}'\,\mathbf{x} = a_1\,x_1 + a_2\,x_2 + \ldots + a_m\,x_m = \underbrace{\|\mathbf{a}\|_2}_{=1}\,\|\mathbf{x}\|_2\cos(\alpha) = \|\mathbf{x}\|_2\cos(\alpha)$$

where α is the angle between $\mathbf{a}$ and $\mathbf{x}$.



$y = \mathbf{a}'\,\mathbf{x}$ is the projection of $\mathbf{x}$ onto the straight line with direction $\mathbf{a}$.

Ex. 4.1: Normal and Overweight Males

Consider the vector of random variables $\mathbf{x} = (X_1, X_2)'$, with X_1 = height in cm, X_2 = weight in kg. Given the linear function

$$Y = \mathbf{a}'\,\mathbf{x} \quad \text{with} \quad \mathbf{a}' = (2/\sqrt{5}, -1/\sqrt{5}) \approx (0.894, -0.447),$$

compute the values of Y for the data given below. Visualize the sampled data and the values for Y and also the corresponding means.

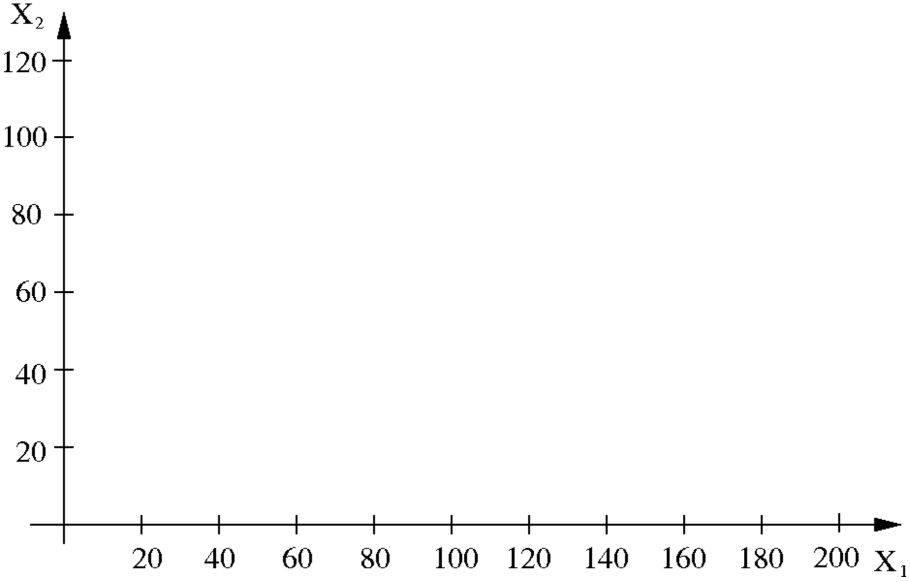
Group 1: normal weight males

Person	Height	Weight	Y
$e_{1,1}$	165	55	
$e_{1,2}$	180	70	
$e_{1,3}$	195	85	
Means			

Group 2: overweight males

Person	height	weight	Y
$e_{2,1}$	160	65	
$e_{2,2}$	170	90	
$e_{2,3}$	180	100	
Means			

Ex. 4.1: Normal and Overweight Males



Fisher's Linear Discriminant Analysis for 2 Groups: Setup

Notation (for 2 or more groups):

- group K_ℓ contains objects $e_{\ell 1}, e_{\ell 2}, \ldots, e_{\ell n_\ell}$ with vectors $\mathbf{x}_{\ell 1}, \mathbf{x}_{\ell 2}, \ldots, \mathbf{x}_{\ell n_\ell}$ for the values of the random variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$.
- Indices of $e_{\ell j}$ and $\mathbf{x}_{\ell j}$: first index ℓ for the group K_ℓ, and second index j for the number in the sample from group K_ℓ

Choosing the vector a:

- Consider a function $Y = \mathbf{a}'\,\mathbf{x}$ where $\mathbf{a}' = (a_1, a_2, \ldots, a_m)$.
- $y_{\ell j} = \mathbf{a}'\,\mathbf{x}_{\ell j}$ = value for Y for object $e_{\ell j}$ from group K_ℓ
- Aim: Choose $\mathbf{a}$ such that the values y_{1j}, $j = 1, 2, \ldots, n_1$, for the group K_1 are substantially larger (smaller) than the values y_{2j}, $j = 1, 2, \ldots, n_2$, for the group K_2.

Fisher's Linear Discriminant Analysis for 2 Groups: Model

Arithmetic Means in the 2 Groups:

$$\overline{\mathbf{x}}_\ell = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbf{x}_{\ell j} = \text{mean value vector for } \mathbf{x} \text{ in group } K_\ell,$$

$$\overline{y}_\ell = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} y_{\ell j} = \underbrace{\frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbf{a}'\,\mathbf{x}_{\ell j}}_{=\mathbf{a}'\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathbf{x}_{\ell j}} = \mathbf{a}'\,\overline{\mathbf{x}}_\ell = \text{mean value for } Y \text{ in group } K_\ell.$$

Choose $\mathbf{a}$, with $\|\mathbf{a}\|_2^2 = 1$, to maximize $\; Q(\mathbf{a}) = \dfrac{(\overline{y}_1 - \overline{y}_2)^2}{\mathrm{SS}(Y)_1 + \mathrm{SS}(Y)_2}$, where

$$\mathrm{SS}(Y)_\ell = \sum_{j=1}^{n_\ell} (y_{\ell j} - \overline{y}_\ell)^2 = \text{sum of squared deviations in group } K_\ell.$$

Motivation: At the maximum the difference of the means $\overline{y}_1 - \overline{y}_2$ is large, but the squared deviations $\mathrm{SS}(Y)_1$ and $\mathrm{SS}(Y)_2$ from the means in K_1 and K_2, respectively, are small.

Rewriting the Numerator and Denominator of $Q(\mathbf{a})$

The numerator and denominator of $Q(\mathbf{a})$ are functions of $\mathbf{a}$:

$$\overline{y}_1 - \overline{y}_2 = \mathbf{a}'\,\overline{\mathbf{x}}_1 - \mathbf{a}'\,\overline{\mathbf{x}}_2 = \mathbf{a}'\,(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2), \qquad \mathrm{SS}(Y)_1 + \mathrm{SS}(Y)_2 = \mathbf{a}'\,\mathbf{W}\,\mathbf{a}$$

with the in-group matrix

$$\mathbf{W} = \sum_{\ell=1}^{2} \mathbf{W}_\ell = \mathbf{W}_1 + \mathbf{W}_2 \qquad \text{with} \qquad \mathbf{W}_\ell = \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)\,(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)'$$

$\mathbf{W}_\ell$ is $(n_\ell - 1)$ times the covariance matrix from the data in group K_ℓ:

$$(\mathbf{W}_\ell)_{ik} = \sum_{j=1}^{n_\ell} (x_{\ell j,i} - \overline{x}_{\ell,i})\,(x_{\ell j,k} - \overline{x}_{\ell,k})' = (n_\ell - 1)\cdot \widehat{\mathrm{Cov}}(X_i, X_k) \quad \text{in } K_\ell,$$

where:

$x_{\ell j,i}$ = ith entry of $\mathbf{x}_{\ell j}$ = value for variable X_i for object $e_{\ell j}$ in group K_ℓ,
$\overline{x}_{\ell,i}$ = ith entry of $\overline{\mathbf{x}}_\ell$ = (arithmetic) mean for variable X_i in group K_ℓ.

Maximization of $Q(\mathbf{a})$ subject to $\|\mathbf{a}\|_2^2 = 1$

Optimization Problem: Maximize

$$Q(\mathbf{a}) = \frac{(\overline{y}_1 - \overline{y}_2)^2}{\mathrm{SS}(Y)_1 + \mathrm{SS}(Y)_2} = \frac{\left[\mathbf{a}'\,(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\right]^2}{\mathbf{a}'\,\mathbf{W}\,\mathbf{a}}$$

subject to the constraint $\|\mathbf{a}\|_2^2 = a_1^2 + a_2^2 + \ldots + a_m^2 = 1$.

The maximization is performed with the method of Lagrange multipliers:

- We find a minimum $Q(\mathbf{a}) = 0$ for vectors $\mathbf{a}$, with $\|\mathbf{a}\|_2^2 = 1$, that are perpendicular to $(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.
- We find a maximum for

$$\mathbf{a} = \pm \frac{1}{\|\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\|_2}\,\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2).$$

We may choose the positive sign for the vector.

The factor $1/\|\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\|_2$ guarantees that $\|\mathbf{a}\|_2^2 = 1$.

Ex. 4.2: Normal and Overweight Males

Given the data in the tables below, find the vector **a** for the function $Y = \mathbf{a}'\mathbf{x}$ and compute the values of $Y = \mathbf{a}'\mathbf{x}$ for the given data and visualize them on the Y-axis.

Group 1: K_1 = normal weight males

Person	Height (cm)	Weight (kg)
$e_{1,1}$	165	55
$e_{1,2}$	180	70
$e_{1,3}$	195	85

Group 2: K_2 = overweight males

Person	Height (cm)	Weight (kg)
$e_{2,1}$	160	65
$e_{2,2}$	170	90
$e_{2,3}$	180	100

Classification Rule for the 2 Group Case

Allocate an new unclassified object e with vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)'$ for the values of the random variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$ to the group K_1 if $y = \mathbf{a}'\,\mathbf{x}$ is closer to the mean $\overline{y}_1$ than to the mean $\overline{y}_2$.

In formulas, allocate e to K_1 if

$$|y - \overline{y}_1| < |y - \overline{y}_2| \qquad \Leftrightarrow \qquad [y - \overline{y}_1]^2 < [y - \overline{y}_2]^2$$

Otherwise allocate e to the group K_2.

Ex. 4.3: Given the function

$$Y = \mathbf{a}'\,\mathbf{x} = (0.792, -0.611) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0.792 \cdot X_1 - 0.611 \cdot X_2$$

and the groups means $\overline{y}_1 = 99.79$ and $\overline{y}_2 = 82.71$ computed in Ex. 4.2, classify a male person with height $= 190$ cm and weight $= 120$ kg.

Methods of Multivariate Statistics

Topic 4.2: Fisher's Linear Discriminant Analysis for Multiple Groups

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Linear Discriminant Analysis for Multiple Groups: Idea

Setup and Notation:

- Given are g groups $K_1, K_2, \ldots, K_g$.
- K_ℓ contains objects $e_{\ell 1}, e_{\ell 2}, \ldots, e_{\ell n_\ell}$ with vectors $\mathbf{x}_{\ell 1}, \mathbf{x}_{\ell 2}, \ldots, \mathbf{x}_{\ell n_\ell}$ for the values of the random variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$.
- Notation for $e_{\ell j}$ and $\mathbf{x}_{\ell j}$: first index ℓ for the group K_ℓ, and second index j for the number in the sample from group K_ℓ

Idea and Aim: We are looking for r vectors $(\mathbf{a}_k)' = (a_{k,1}, a_{k,2}, \ldots, a_{k,m})$ and linear functions

$$Y_k = \mathbf{a}_k' \, \mathbf{x} = a_{k,1} \, X_1 + a_{k,2} \, X_2 + \ldots + a_{k,m} \, X_m, \qquad k = 1, 2, \ldots, r,$$

with $\|\mathbf{a}_k\|_2^2 = 1,\ k = 1, 2, \ldots, r$, such that the random variables $\mathbf{y} = (Y_1, Y_2, \ldots, Y_r)'$ optimally distinguish between the groups.

Linear Discriminant Analysis for Multiple Groups: Model

For $k = 1, 2 \ldots, r$, $\mathbf{a}^k = (a_{k,1}, a_{k,2}, \ldots, a_{k,m})'$ is determined such that

$$Q(\mathbf{a}_k) = \frac{\sum_{\ell=1}^{g} n_\ell \left(\overline{y_{k,\ell}} - \overline{y_k}\right)^2}{\sum_{\ell=1}^{g} \mathrm{SS}(Y_k)_\ell} \tag{8}$$

is maximized subject to the constraint $\|\mathbf{a}_k\|_2^2 = 1$, where

$$y_{k,\ell j} = \mathbf{a}_k' \, \mathbf{x}_{\ell j} = \text{value of } Y_k \text{ for the } j\text{th object } e_{\ell j} \text{ in group } K_\ell, \tag{9}$$

$$\overline{y_k} = \frac{1}{\sum_{\ell=1}^{g} n_\ell} \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} y_{k,\ell j} = \left(\begin{array}{c} \text{mean value of } Y_k \text{ in} \\ \text{the union of all groups} \end{array} \right), \tag{10}$$

$$\overline{y_{k,\ell}} = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} y_{k,\ell j} = \text{mean value of } Y_k \text{ in the group } K_\ell, \tag{11}$$

and where $\mathrm{SS}(Y_k)_\ell$ is the sum of squared deviations for Y_k in K_ℓ

$$\mathrm{SS}(Y_k)_\ell = \sum_{j=1}^{n_\ell} (y_{k,\ell j} - \overline{y_{k,\ell}})^2.$$

Relating the Means for $\mathbf{x}$ and the Y^k

With the means for $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$,

$$\overline{\mathbf{x}} = \frac{1}{\sum_{\ell=1}^{g} n_\ell} \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \mathbf{x}_{\ell j} = \text{mean for } \mathbf{x} \text{ in the union of all groups},$$

$$\overline{\mathbf{x}}_\ell = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbf{x}_{\ell j} = \text{mean for } \mathbf{x} \text{ in the groups } K_\ell,$$

we have, from substituting (9) into (10) and (11)

$$\overline{y_k} = \mathbf{a}_k' \, \overline{\mathbf{x}} \qquad \text{and} \qquad \overline{y_{k,\ell}} = \mathbf{a}_k' \, \overline{\mathbf{x}}_\ell. \tag{12}$$

Hence, from (9) and (12),

$$\overline{y_{k,\ell}} - \overline{y_k} = \mathbf{a}_k' \, (\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}) = \begin{cases} (\text{mean for } Y_k \text{ in group } K_\ell) \\ -(\text{grand mean for } Y_k), \end{cases}$$

$$y_{k,\ell j} - \overline{y_{k,\ell}} = \mathbf{a}_k' \, (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell) = \begin{cases} (\text{value for } Y_k \text{ for the } j\text{th object in} \\ \text{group } K_\ell) - (\text{mean for } Y_k \text{ in group } K_\ell). \end{cases}$$

Rewriting the Numerator of $Q(\mathbf{a}^k)$

Substituting $\overline{y_k} = \mathbf{a}_k' \, \overline{\mathbf{x}}$ and $\overline{y_{k,\ell}} = \mathbf{a}_k' \, \overline{\mathbf{x}}_\ell$ (from (12)) into the numerator of $Q(\mathbf{a}_k)$ we find

$$
\begin{aligned}
\sum_{\ell=1}^{g} n_\ell \left(\overline{y_{k,\ell}} - \overline{y_k}\right)^2 &= \sum_{\ell=1}^{g} n_\ell \left[\mathbf{a}_k' \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)\right]^2 \\
&= \sum_{\ell=1}^{g} n_\ell \left[\mathbf{a}_k' \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)\right] \left[\mathbf{a}_k' \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)\right]' = \sum_{\ell=1}^{g} n_\ell \, \mathbf{a}_k' \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)' \mathbf{a}^k \\
&= \mathbf{a}_k' \left(\sum_{\ell=1}^{g} n_\ell \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)' \right) \mathbf{a}_k = \mathbf{a}_k' \, \mathbf{B} \, \mathbf{a}^k
\end{aligned}
$$

with the between-group matrix

$$
\mathbf{B} = \sum_{\ell=1}^{g} n_\ell \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}\right)'.
$$

Rewriting the Denominator of $Q(\mathbf{a}^k)$

Substituting $y_{k,\ell j} = \mathbf{a}_k' \, \mathbf{x}_{\ell j}$ and $\overline{y_{k,\ell}} = \mathbf{a}_k' \, \overline{\mathbf{x}}_\ell$ (from (9) and (12)) into the denominator of $Q(\mathbf{a}_k)$: we find (analogous computation)

$$\sum_{\ell=1}^{g} \mathrm{SS}(Y_k)_\ell = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (y_{k,\ell j} - \overline{y_{k,\ell}})^2 = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \left[\mathbf{a}_k' \, (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell) \right]^2$$

$$= \mathbf{a}_k' \left(\sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)' \right) \mathbf{a}_k = \mathbf{a}_k' \, \mathbf{W} \, \mathbf{a}_k,$$

with the in-group matrix

$$\mathbf{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \overline{\mathbf{x}}_\ell)'.$$

Note: For $g = 2$ this is just the in-group matrix for the case of 2 groups.

Maximization of $Q(\mathbf{a}^k)$ subject to $\|\mathbf{a}^k\|_2^2 = 1$

Optimization Problem: Maximize

$$Q(\mathbf{a}_k) = \frac{\sum_{\ell=1}^{g} n_\ell \, (\overline{y_{k,\ell}} - \overline{y_k})^2}{\sum_{\ell=1}^{g} \mathrm{SS}(Y_k)_\ell} = \frac{\mathbf{a}_k' \, \mathbf{B} \, \mathbf{a}_k}{\mathbf{a}_k' \, \mathbf{W} \, \mathbf{a}_k}$$

subject to the constraint $\|\mathbf{a}_k\|_2^2 = (a_{k,1})^2 + (a_{k,2})^2 + \ldots + (a_{k,m})^2 = 1$.

The maximization is performed with the method of Lagrange multipliers and leads to the eigenvalue-eigenvector equation:

$$\mathbf{W}^{-1} \, \mathbf{B} \, \mathbf{a}_k = \underbrace{\frac{\mathbf{a}_k' \, \mathbf{B} \, \mathbf{a}_k}{\mathbf{a}_k' \, \mathbf{W} \, \mathbf{a}_k}}_{=\lambda_k} \, \mathbf{a}_k \qquad \text{where} \qquad \|\mathbf{a}_k\|_2^2 = 1$$

We see that $\mathbf{a}_k$ is an eigenvector of $\mathbf{W}^{-1} \, \mathbf{B}$ with eigenvalue λ_k.

Computation of the Direction Vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^q$

Because $\mathrm{rank}(\mathbf{W}) = m$ and $t = \mathrm{rank}(\mathbf{B}) \leq \min\{m-1, g\}$ we find t non-zero eigenvalues of $\mathbf{W}^{-1}\,\mathbf{B}$.

Computation of the Eigenvalues: To find the $q \leq t$ positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$ of $\mathbf{W}^{-1}\,\mathbf{B}$, we solve

$$\det(\mathbf{W}^{-1}\,\mathbf{B} - \lambda\,\mathbf{I}) = 0.$$

Explanation: $\mathbf{W}^{-1}\,\mathbf{B}\,\mathbf{a}_k = \lambda_k\,\mathbf{a}_k \;\Leftrightarrow\; (\mathbf{W}^{-1}\,\mathbf{B} - \lambda_k\,\mathbf{I})\,\mathbf{a}_k = \mathbf{0}$ has only a non-zero solution $\mathbf{a}_k$ if $\det(\mathbf{W}^{-1}\,\mathbf{B} - \lambda_k\,\mathbf{I}) = 0$.

Computation of the $\mathbf{a}^k$: Solving the linear system

$$(\mathbf{W}^{-1}\,\mathbf{B} - \lambda_k\,\mathbf{I})\,\mathbf{a}_k = \mathbf{0} \qquad \Leftrightarrow \qquad \mathbf{W}^{-1}\,\mathbf{B}\,\mathbf{a}_k = \lambda_k\,\mathbf{a}_k$$

yields the eigenvector $\mathbf{a}_k$ to the eigenvalue λ_k, where we impose the normalization $\|\mathbf{a}_k\|_2^2 = 1$.

Dimension Reduction

Only use the eigenvectors $\mathbf{a}_k$ with eigenvalues λ_k that satisfy $\lambda_k > 1$.
We find $r \leq q$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 1$.

Motivation: Eigenvalues λ_k with $\lambda_k < 1$ will only provide a minor and not very distinct separation of the groups K_ℓ. Therefore they are omitted.

We distinguish the groups $K_1, K_2, \ldots, K_g$ with the r linear functions

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_r \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \, \mathbf{x} \\ \mathbf{a}_2' \, \mathbf{x} \\ \vdots \\ \mathbf{a}_r' \, \mathbf{x} \end{pmatrix}, \qquad \text{or} \qquad Y_k = \mathbf{a}_k' \, \mathbf{x}, \quad k = 1, 2, \ldots, r.$$

Classification Rule

Given a new object e with values $\mathbf{x}$ for the variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$, sort e into the group K_{ℓ^*}, where ℓ^* is such that

$$\sum_{k=1}^{r} \Big[\underbrace{\mathbf{a}_k' \left(\mathbf{x} - \overline{\mathbf{x}}_{\ell^*}\right)}_{= y_k - \overline{y_{k,\ell^*}}} \Big]^2 \le \sum_{k=1}^{r} \Big[\underbrace{\mathbf{a}_k' \left(\mathbf{x} - \overline{\mathbf{x}}_{\ell}\right)}_{= y_k - \overline{y_{k,\ell}}} \Big]^2 \qquad \text{for all } \ell \neq \ell^*.$$

Here $y_k = \mathbf{a}_k' \, \mathbf{x}$ is the value of Y_k for the new object e.

We will only test the multiple group case with SPSS as the computations by hand are (even for very simple examples) very lengthy and elaborate.

Methods of Multivariate Statistics

Topic 5: Cluster Analysis

Dr. Kerstin Hesse

Email: kerstin.hesse@hhl.de; *Phone:* +49 (0)341 9851-820; *Office:* HHL Main Building, Room 115A

HHL – Leipzig Graduate School of Management, Jahnallee 59, 04109 Leipzig, Germany

Doctoral Program at HHL, May 4-5, 2012

Topic 5: Cluster Analysis

Idea of Cluster Analysis and Classification Types

- aim: subdivision of a population into groups/clusters based on several metric variables
- types of classification

Hierarchical Classification

- distance matrix
- agglomerative and divisive methods of hierarchical classification
- agglomerative hierarchical classification (discussed in detail)
- determining the number of groups/clusters

Evaluating the Quality of a Classification

- measures of homogeneity within the groups/clusters
- measures of heterogeneity between the groups/clusters

Outlook: Non-Hierarchical Classification

Idea and Aim of Cluster Analysis

Aim: Given a (usually large) population P with elements $e_1, e_2, \ldots, e_n$, the aim of cluster analysis (automatic classification) is to optimally structure the population by forming homogeneous subgroups/clusters.

- Each group/cluster shall contain only elements that are very similar (homogeneous groups).
- The different groups/clusters shall be very dissimilar (heterogeneity between the different groups).
- The number of the groups/clusters is not known but will be determined during the process of forming the clusters.

Idea: Distances based on suitable metric variables can be used to separate P into groups/clusters. These distances can also measure homogeneity within groups and heterogeneity between groups.

Examples where Cluster Analysis is Applied

Example (Marketing): Data on a product collected via a questionnaire.

- metric variables: gross income, money spent on the product, ...
- Cluster Analysis is used to identify customer/buyer groups.
- This information can then be used to target the different costumer groups with different advertising strategies.

Example (Classifying Products): For introducing a new microscope on the market and determining its price and marketing strategy it is necessary to position it in relation to existing microscopes already on the market.

- Metric data on prices, technical information (size, resolution, ...) of microscopes on the market is collected.
- Cluster analysis is used to form groups of similar microscopes.
- Based on its technical specifications, the new microscope is allocated to one of these groups, and its price can be determined.

Different Types of Classification/Clustering

For illustration, consider a population $P = \{e_1, e_2, \ldots, e_9\}$

1. **Partition**: The groups have to be disjoint, i.e. each element belongs to exactly one group.

   Example: $K_1 = \{e_1, e_2, e_9\}$, $K_2 = \{e_4, e_5, e_8\}$, $K_3 = \{e_3, e_6, e_7\}$

2. **Hierarchy**: A hierarchy is a sequence of partitions (e.g. see page 135).

   By going from a coarser to a finer partition, each group of the finer partition has to be contained in a group from the coarser partition.

3. **Covering**: The groups may overlap, i.e. have elements in common.

   Example: $K_1 = \{e_1, e_2, e_4, e_5\}$, $K_2 = \{e_3, e_4, e_6, e_7\}$, $K_3 = \{e_7, e_8, e_9\}$

4. **Quasi-Hierarchy**: A quasi-hierarchy is a sequence of coverings.

Note: The union of the groups $K_1, K_2, \ldots, K_g$ is always the population P.

Visualization of a Hierarchical Classification via a Tree

The diagram below shows an agglomerative hierarchical classification:

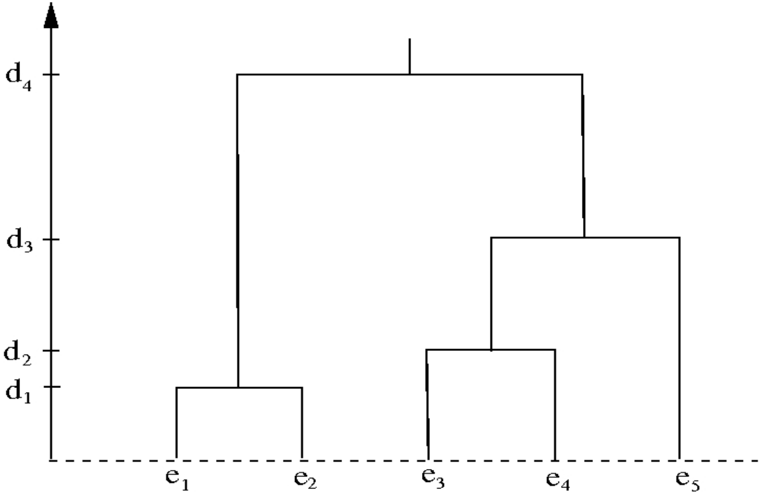
We start with the finest partition where every element forms its own group.

Then we unity in each step exactly two groups. We still have to discuss the criterions for uniting groups.

Step 4	$\{e_1, e_2, e_3, e_4, e_5\}$				
Step 3	$\{e_1, e_2\}$		$\{e_3, e_4, e_5\}$		
Step 2	$\{e_1, e_2\}$		$\{e_3, e_4\}$		$\{e_5\}$
Step 1	$\{e_1, e_2\}$		$\{e_3\}$	$\{e_4\}$	$\{e_5\}$
Step 0	$\{e_1\}$	$\{e_2\}$	$\{e_3\}$	$\{e_4\}$	$\{e_5\}$

Dendrogram for a Hierarchical Classification

d_i = distance of the groups that are united in step i,
e.g. d_3 = distance of $\{e_3, e_4\}$ and $\{e_5\}$



Distance Matrix

Distance Matrix: The starting point for any hierarchical classification of a population of n objects $e_1, e_2, \ldots, e_n$ is the distance matrix

$$\mathbf{D} = (d_{ik})_{i,k=1,2,\ldots,n} \qquad \text{where} \qquad d_{ik} = \text{distance of } e_i \text{ and } e_k.$$

Measuring Distances: The distance is measured with the help of a vector of random variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$ that characterizes the objects in the population. The distance of e_i and e_k, with $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{im})$ and $\mathbf{x}_k' = (x_{k1}, x_{k2}, \ldots, x_{km})$ for the values of the random variables, is

$$d_{ik} = \text{distance of } \mathbf{x}_i \text{ and } \mathbf{x}_k.$$

Examples of Distances:

- Euclidean distance
  $d_{ik} = \|\mathbf{x}_i - \mathbf{x}_k\|_2 = \sqrt{(x_{i1} - x_{k1})^2 + \ldots + (x_{im} - x_{km})^2}$
- City-block distance: $d_{ik} = \|\mathbf{x}_i - \mathbf{x}_k\|_1 = |x_{i1} - x_{k1}| + \ldots + |x_{im} - x_{km}|$

Hierarchical Classification: Agglomerative Approach

Agglomerative Approach: (see Example on page 135)

- We start with the finest partition:
  Each object forms an individual subgroup.
- By successively uniting subgroups we obtain larger and more heterogeneous groups.
- In the last step we end up with one group that contains all objects.

Rules for Agglomerative Hierarchical Classification:

- In each step exactly two subgroups are united.
- Once a group has been formed by uniting two subgroups, this group cannot be split up again.

Note: Different variants of the method can lead to different classifications.

Hierarchical Classification: Divisive Approach

Divisive Approach:

- We start with the coarsest partition: All objects are in one group.
- Then we successively subdivide into subgroups which are each more homogeneous.
- The last step gives only subgroups that contain one object each.

Rules for Divisive Hierarchical Classification:

- In each step exactly one group is split up into two.
- Once a group has been split up into two subgroups the new subgroups cannot again be reunited.

Note: The agglomerative approach and the divisive approach do not necessarily yield the same classification.

Method of Agglomerative Hierarchical Classification

Data: population P of n objects $e_1, e_2, \ldots, e_n$ and a distance matrix $\mathbf{D} = (d_{ik})_{i,k=1,2,\ldots,n}$ for these objects (d_{ik} = distance of e_i and e_k).

1. Start with the finest partition $\mathcal{P}^{(0)} = \{K_1, K_2, \ldots, K_n\}$ where each object forms one group $K_j = \{e_j\}$, $j = 1, 2, \ldots, n$.
2. Find the indices p and q such that $d_{pq} = \min_{i \neq j} d_{ij}$.
3. Unite the groups K_p and K_q so that we now have one group less.
4. Compute the new distance matrix: Distances from all groups to the new group $K_p^{\text{new}} = K_p \cup K_q$ need to be computed.
   (i) Compute the distances from all other groups to the new group K_p^{new} and replace the entries of the pth row and pth column accordingly:
   $$d_{pj}^{\text{new}} = d_{jp}^{\text{new}} = \text{distance of group } K_j \text{ from group } K_p^{\text{new}}.$$
   (ii) Delete the qth row and qth column of the distance matrix.
5. Return to step 2 and repeat the process with the new distance matrix until there is only one group.

Computation of the New Distance Matrix in Step 4

The distance d_{pj}^{new} between group K_j and the new group $K_p^{\text{new}} = K_p \cup K_q$ can be computed with the following schemes:

- **Single Linkage (Nearest Neighbor):** $d_{pj}^{\text{new}} = \min\{d_{pj}, d_{qj}\}$

  Interpretation: This is the distance of the two objects from K_j and K_p^{new}, respectively, that are closest together (nearest neighbors).

- **Complete Linkage (Furthest Neighbor):** $d_{pj}^{\text{new}} = \max\{d_{pj}, d_{qj}\}$

  Interpretation: This is the distance of the two objects from K_j and K_p^{new}, respectively, that are furthest apart (furthest neighbors).

- **Average Linkage:** $d_{pj}^{\text{new}} = \frac{1}{2}\left(d_{pj} + d_{qj}\right)$

- **And more:** There are other schemes, but these are the simplest ones.

Ex. 5.1: Classifying Digital Cameras

We are given the data on 5 digital cameras below.

Use agglomerative hierarchical classification with the city block distance and the nearest neighbor rule to form groups of similar digital cameras.

Draw a dendrogram of your hierarchical classification.

Camera	Price in 100 Euros	Resolution in Pixels
e_1	1	6
e_2	1.5	8
e_3	0.5	3
e_4	5	12
e_5	6	12

How Do We Determine the Number of Groups (Clusters)?

Rule of Thumb: The number of groups g is approximately $g \approx \sqrt{n/2}$.

Clearly the rule of thumb gives only a rough idea.

Inspecting our Dendrogram:

As the distances between the groups are shown, we can see by inspection in which step we should stop with uniting groups (i.e. when even larger groups would be too heterogeneous).

Ex. 5.2 (Classifying Digital Cameras): Determine the number of groups for the digital cameras from your results for Ex. 5.1.

Measures of Homogeneity Within the Groups

1. Average of the distances of the objects in K_ℓ:

$$g_1(K_\ell) = \frac{2}{n_\ell(n_\ell - 1)} \sum_{\substack{i<j,\\ e_i, e_j \in K_\ell}} d_{ij}$$

2. Distance of the least similar objects in K_ℓ:

$$g_2(K_\ell) = \max_{e_i, e_j \in K_\ell} d_{ij}$$

3. Distance of the most similar objects in K_ℓ:

$$g_3(K_\ell) = \min_{\substack{e_i, e_j \in K_\ell,\\ i \neq j}} d_{ij}$$

Note: The smaller the $g_i(K_\ell)$, the more homogeneous is the group K_ℓ.

Measures of Heterogeneity Between the Groups I

1. Complete linkage (furthest neighbor): $v_1(K_\ell, K_{\ell*})$ is the distance of the objects from the two groups that are furthest apart, i.e.

$$v_1(K_\ell, K_{\ell*}) = \max_{e_i \in K_\ell, e_j \in K_{\ell*}} d_{ij}.$$

2. Single linkage (nearest neighbor): $v_2(K_\ell, K_{\ell*})$ is the distance of the objects from the two groups that are closest together, i.e.

$$v_2(K_\ell, K_{\ell*}) = \min_{e_i \in K_\ell, e_j \in K_{\ell*}} d_{ij}.$$

3. Average linkage: $v_3(K_\ell, K_{\ell*})$ is the average distance of objects from the two groups, i.e.

$$v_3(K_\ell, K_{\ell*}) = \frac{1}{n_\ell \cdot n_{\ell*}} \sum_{i=1}^{n_\ell} \sum_{j=1}^{n_{\ell*}} d_{ij}.$$

4. Squared Euclidean distance of the means: $v_4(K_\ell, K_{\ell*}) = \|\overline{\mathbf{x}}_\ell - \overline{\mathbf{x}}_{\ell*}\|_2^2$,

   where $\overline{\mathbf{x}}_\ell$ and $\overline{\mathbf{x}}_{\ell*}$ are the means of $\mathbf{x}$ in the groups K_ℓ and $K_{\ell*}$:

$$\overline{\mathbf{x}}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \mathbf{x}_{\ell i} \quad \text{and} \quad \overline{\mathbf{x}}_{\ell*} = \frac{1}{n_{\ell*}} \sum_{j=1}^{n_{\ell*}} \mathbf{x}_{\ell* j},$$

   and $K_\ell = \{e_{\ell 1}, e_{\ell 2}, \ldots, e_{\ell n_\ell}\}$ and $\mathbf{x}_{\ell i}$ is the vector of the values of the metric variables $\mathbf{x}$ for $e_{\ell i}$ from group K_ℓ (likewise for $K_{\ell*}$).

Note: The larger the $v_i(K_\ell, K_{\ell*})$, the more dissimilar are K_ℓ and $K_{\ell*}$.

Ex. 5.3 (Quality of the Classification of Digital Cameras): Apply the criteria for the quality of a hierarchical classification in our digital camera example for the classification

$$K_1 = \{e_1, e_2, e_3\} \qquad \text{and} \qquad K_2 = \{e_4, e_5\}.$$

Improving a Classification: Non-Hierarchical Classification

Situation: We have already determined a fixed number g of groups, e.g. with an agglomerative hierarchical classification.

Aim: We want to improve this classification by moving suitable objects from one group into another.

Variance Criterion for Improving the Classification:

Move objects between groups such that for the final classification $\mathcal{K} = \{K_1, K_2, \ldots, K_g\}$ the following function is minimized

$$z(\mathcal{K}) = \sum_{\ell=1}^{g} \underbrace{\sum_{i=1}^{n_\ell} \|\mathbf{x}_{\ell i} - \overline{\mathbf{x}}_\ell\|_2^2}_{=\text{variation in group } K_\ell} , \qquad \text{where} \qquad \underbrace{\overline{\mathbf{x}}_\ell = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbf{x}_{\ell i}}_{=\text{mean for } \mathbf{x} \text{ in } K_\ell}$$

and $K_\ell = \{e_{\ell 1}, e_{\ell 2}, \ldots, e_{\ell n_\ell}\}$ and $\mathbf{x}_{\ell i}$ is the vector of the values of the metric variables $\mathbf{x} = (X_1, X_2, \ldots, X_m)'$ for $e_{\ell i}$ from group K_ℓ.

General Advice on the Application of Cluster Analysis

- You should perform cluster analyses with different distances and different schemes for computing the new distances in the hierarchical classification, as they will yield different classifications.
- Some classifications will be better suited to your application than others.
- You should use non-hierarchical classification (with different starting classifications) to improve your classifications.