Solutions to the Exercises in 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

Outline & Table of Contents

This document contains the solutions to the exercises in the course material (on the slides).

I will upload this document after the end of the course, so that you have all the solutions for the assignment.

Methods of Multivariate Statistics

Solutions to 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

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*.

Solution:

- *random variable:* $X(e) =$ number of heads, with values in $\{0, 1, 2\}$
- If we set $e_1 = HH$, $e_2 = HT$, $e_3 = TH$, $e_4 = TT$, $H =$ heads, $T =$ tails, then $X(e_1) = 2$, $X(e_2) = X(e_3) = 1$ and $X(e_4) = 0$
- For a perfect coin, the *probability density* $f : \{0, 1, 2\} \to \mathbb{R}$ is

$$f(0) = P(X = 0) = \frac{1}{4},$$

$$f(1) = P(X = 1) = \frac{1}{2},$$

$$f(2) = P(X = 2) = \frac{1}{4}.$$

Ex. 1.1: Flipping a Coin Twice

- The *probability distribution* is

$$F(x) = P(X \leq x) = \sum_{\substack{k=1, \\ k \leq x}}^{3} f(k),$$

and we find

$$F(0) = f(0) = \frac{1}{4},$$

$$F(1) = f(0) + f(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4},$$

$$F(2) = f(0) + f(1) = f(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1.$$

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.

<u>Solution:</u> The expectation value and the variance are

$$\begin{aligned} \mathrm{E}(X) &= x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) \\ &= 0 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{4} = 1, \end{aligned}$$

$$\begin{aligned} Var(X) &= \mathrm{E}(X^2) - \big[\mathrm{E}(X)\big]^2 \\ &= x_1^2 \cdot f(x_1) + x_2^2 \cdot f(x_2) + x_3^2 \cdot f(x_3) - \big[\mathrm{E}(X)\big]^2 \\ &= 0^2 \cdot \tfrac{1}{4} + 1^2 \cdot \tfrac{1}{2} + 2^2 \cdot \tfrac{1}{4} - 1^2 = 0 + \tfrac{1}{2} + 1 - 1 = \tfrac{1}{2}. \end{aligned}$$

Ex. 1.3: Random Variable 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

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

where $F_N(z) = F_n(z; 0, 1)$, 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.

Solution: The probability that a person has a yearly gross income between 50,000 and 60,000 Euros is given by

$$P(50 \le X \le 60) = P(X \le 60) - P(x < 50) = F_n(60; 40, 10) - F_n(50; 40, 10).$$

Ex. 1.3: Random Variable Income

We *standardize our random variable* $X =$ *yearly gross income* and find the corresponding values for $x_1 = 50$ and $x_2 = 60$, which yields from

$$Z = \frac{X - \mathrm{E}(X)}{\sigma} = \frac{X - 40}{10}$$

the values

$$z_1 = \frac{x_1 - 40}{10} = \frac{50 - 40}{10} = 1, \qquad z_2 = \frac{x_2 - 40}{10} = \frac{60 - 40}{10} = 2.$$

The normal distribution $F_n(x; \mu, \sigma)$ is related to the standard normal distribution $F_N(z) = F_n(z; 0, 1)$ via

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

Thus we find with this formula from any table of the normal distribution:

$$F_n(50; 40, 10) = F_N(1) = 0.8413,$$
$$F_n(60; 40, 10) = F_N(2) = 0.9772.$$

Ex. 1.3: Random Variable Income

Hence

$$P(50 \leq X \leq 60) = F_n(60; 40, 10) - F_n(50; 40, 10) = 0.1359.$$

The probability that the yearly gross income is between 50,000 and 60,000 Euros bis 0.1359.

Ex. 1.4: Standardization

Use the formula

$$\mathrm{E}(Z) = a \cdot \mathrm{E}(X) + b \quad \text{and} \quad \mathrm{Var}(Z) = a^2 \cdot \mathrm{Var}(X) \quad \text{for} \quad Z = a \cdot X + b \quad (1)$$

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 $\mathrm{Var}(Z) = 1$.

Solution: For

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}$$

we find, from (1) with $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$,

$$\mathrm{E}(Z) = \frac{1}{\sigma} \cdot \mathrm{E}(X) - \frac{\mu}{\sigma} = \frac{\mathrm{E}(X) - \mu}{\sigma} = 0 \qquad \text{as} \qquad \mu = \mathrm{E}(X),$$

and

$$\mathrm{Var}(Z) = \left(\frac{1}{\sigma}\right)^2 \cdot \mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{\sigma^2} = 1 \qquad \text{as} \qquad \sigma^2 = \mathrm{Var}(X).$$

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.

Solution: We expect that the result X of the first flip of the coin has *no effect* on the result Y of the second flip of the coin and vice versa.

Hence we expect that X and Y are *uncorrelated*, i.e. $\text{Cov}(X, Y) = 0$.

Ex. 1.5: Flipping a Coin Twice

Let us consider *why the probability density* $f(x, y) = 1/4$ *makes sense*:

- For a perfect coin, we expect that heads and tails turn up with the same probability $1/2$.
- Thus for each (i.e. first or second) flip of the coin considered independently we expect the probability densities $f_X(x) = 1/2$ and $f_Y(y) = 1/2$.
- As we assume that the flips of the coin are uncorrelated, we expect

$$f(x, y) = f_X(x) \cdot f_Y(y) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

To compute $\mathrm{Cov}(X, Y)$, we need the *expectation values* $\mathrm{E}(X)$ and $\mathrm{E}(Y)$:
As $1 =$ heads and $0 =$ tails, we have:

$$\mathrm{E}(X) = \sum_{i=0}^{1} \sum_{j=0}^{1} i \cdot \underbrace{f(i, j)}_{=1/4} = \frac{1}{4} \sum_{i=0}^{1} \underbrace{\sum_{j=0}^{1} i}_{=2 \cdot i} = \frac{1}{2} \underbrace{\sum_{i=0}^{2} i}_{=1} = \frac{1}{2},$$

Ex. 1.5: Flipping a Coin Twice

$$\mathrm{E}(Y) = \sum_{i=0}^{1} \sum_{j=0}^{1} j \cdot \underbrace{f(i,j)}_{=1/4} = \frac{1}{4} \sum_{i=0}^{1} \underbrace{\sum_{j=0}^{1} j}_{=1} = \frac{1}{4} \underbrace{\sum_{i=0}^{1} 1}_{=2} = \frac{1}{4} \cdot 2 = \frac{1}{2}.$$

We note that $\mathrm{E}(X) = \mathrm{E}(Y) = 1/2$ is just the expectation value for a single flip of a perfect coin.

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \sum_{i=0}^{1} \sum_{j=0}^{1} \underbrace{\big[i - \mathrm{E}(X)\big]}_{=i-\frac{1}{2}} \cdot \underbrace{\big[j - \mathrm{E}(X)\big]}_{=j-\frac{1}{2}} \cdot \underbrace{f(i,j)}_{=1/4} \\ &= \frac{1}{4} \sum_{i=0}^{1} \sum_{j=0}^{1} \left(i - \frac{1}{2}\right) \cdot \left(j - \frac{1}{2}\right) \\ &= \frac{1}{4} \sum_{i=0}^{1} \left(i - \frac{1}{2}\right) \underbrace{\sum_{j=0}^{1} \left(j - \frac{1}{2}\right)}_{=-\frac{1}{2}+\frac{1}{2}=0} = 0 \end{aligned}$$

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 $\varrho(X, Y)$.

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

Solution: We *estimate the expectation values via the means*:

$$\widehat{\mu_X} = \overline{x} = \frac{1}{4}\left(6000 + 5000 + 6500 + 4500\right) = \frac{22000}{4} = 5500,$$
$$\widehat{\mu_Y} = \overline{y} = \frac{1}{4}\left(300 + 250 + 400 + 250\right) = \frac{1200}{4} = 300.$$

The expectation value $\mathrm{E}(X)$ of the monthly gross income X is estimated by $\widehat{\mu_X} = \overline{x} = 5500$ Euros. The expectation value $\mathrm{E}(Y)$ of the monthly spending on foods Y is estimated by $\widehat{\mu_Y} = \overline{y} = 300$ Euros.

$$\begin{aligned}\widehat{\sigma_X}^2 &= \frac{1}{3}\Big[(6000-5500)^2 + (5000-5500)^2 \\ &\quad +(6500-5500)^2 + (4500-5500)^2\Big] \\ &= \frac{1}{3}\left[500^2 + (-500)^2 + 1000^2 + (-1000)^2\right] = \frac{2500000}{3} = 833333.\overline{3}\end{aligned}$$

The *variance* $\mathrm{Var}(X) = \sigma_X^2$ is estimate by $\widehat{\sigma_X}^2 \approx 833333.33$, and the *standard deviation* σ_X of X is estimated by $\widehat{\sigma_X} = \sqrt{833333.\overline{3}} \approx 912.87$.

$$\begin{aligned}\widehat{\sigma_Y}^2 &= \frac{1}{3}\left[(300-300)^2+(250-300)^2+(400-300)^2+(250-300)^2\right]\\ &= \frac{1}{3}\left[0^2+(-50)^2+100^2+(-50)^2\right] = \frac{15000}{3} = 5000\end{aligned}$$

The *variance* $\mathrm{Var}(Y)=\sigma_Y^2$ is estimated by $\widehat{\sigma_Y}^2 = 5000$, and
the *standard deviation* σ_Y of Y is estimated by $\widehat{\sigma_Y} = \sqrt{5000} \approx 70.71$.

Next we estimate the covariance of X and Y from our sample.

$$\begin{aligned}\widehat{\mathrm{Cov}}(X,Y) &= \frac{1}{3}\big[(6000-5500)\cdot(300-300)+(5000-5500)\cdot(250-300)\\ &\quad +(6500-5500)\cdot(400-300)+(4500-5500)\,(250-300)\big]\\ &= \frac{1}{3}\big[500\cdot 0+(-500)\cdot(-50)+1000\cdot 100+(-1000)\cdot(-50)\big]\\ &= \frac{1}{3}\big[0+25000+100000+50000\big] = \frac{175000}{3} = 58333.\overline{3}\end{aligned}$$

The *covariance* $\mathrm{Cov}(X,Y)$ is estimated by $\widehat{\mathrm{Cov}}(X,Y) \approx 58333.33$.

To get a better idea of the strength of the correlation of X and Y we finally estimate the *correlation coefficient*:

$$\widehat{\varrho}(X,Y) = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\sigma_X}\,\widehat{\sigma_Y}} = \frac{58333.\overline{3}}{\sqrt{833333.\overline{\overline{3}}}\cdot\sqrt{5000}} \approx 0.904$$

The correlation coefficient $\varrho(X,Y)$ is estimated by $\widehat{\varrho}(X,Y) \approx 0.904$ which is quite close to 1 and indicates a *very strong correlation* between the monthly gross income X and the monthly spending on foods Y.

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 $s_1^2 = 196^2$ g^2 (2010) and $s_2^2 = 153^2$ 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 $\sigma_1^2 > \sigma_2^2$.

Ex. 1.7: Hypothesis Testing

Solution:

1. *Formulating the Null Hypothesis and the Alternative Hypothesis*:

$H_0 : \sigma_1^2 = \sigma_2^2$ (The variance of the weight is the same in both years.)

$H_1 : \sigma_1^2 > \sigma_2^2$ (The variance of the weight in 2010 is larger than in 2011.)

2. *Find the Test Variable and its Distribution*: The test variable is

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

and its follows an F-distribution with $\nu_1 = n_1 - 1 = 100$ numerator and $\nu_2 = n_2 - 1 = 100$ denominator degrees of freedom.

Under the null hypothesis $\sigma_1^2 = \sigma_2^2$, the variances of the geese population cancel in (2). So our test variable is

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

Ex. 1.7: Hypothesis Testing

3. *Determination of the Critical Area (for Acceptance of the Null Hypothesis)*: As the alternative hypothesis is an inequality, we have a one-sided test with $\alpha = 0.05$. Consulting the table of the F-distribution with $\nu_1 = 100$ numerator and $\nu_2 = 100$ denominator degrees of freedom, we find that the critical value is:

$$f_c = 1.39$$

If $f = s_1^2/s_2^2 > f_c$ then the null hypothesis is rejected.

If $f = s_1^2/s_2^2 \leq f_c$ then the null hypothesis cannot be rejected.

4. *Computation of the Value of the Test Variable*:

$$f = \frac{s_1^2}{s_2^2} = \frac{196^2}{153^2} = 1.641$$

Ex. 1.7: Hypothesis Testing

5. *Decision about the Hypotheses and Interpretation*: As

$$f = 1.641 > f_c = 1.39$$

the *null hypothesis is rejected*.

Interpretation: The chance to reject the null hypothesis, when it is in fact true, is 0.05 (or 5%). This means that *with* 95% *confidence* we can say that the variance of the weight σ_1^2 in 2010 is strictly larger than the variance of the weight σ_2^2 in 2011.

Methods of Multivariate Statistics

Solutions to 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

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.

<u>Solution:</u>

- *population*: $P =$ set of all fields
- *independent variable/factor*: $A =$ method of fertilization with *4 factor levels* given by the 4 types of fertilizer A_1, A_2, A_3 and A_4
- *4 subpopulations*: P_1, P_2, P_3 and P_4, where $P_i =$ fields fertilized with fertilizer A_i
- *dependent metric variable*: $Y =$ crop yield (e.g. measured in tons of crop per km^2)
- *design of empirical investigation*: Fertilize 100 fields each with fertilizer A_1, A_2, A_3 and A_4, respectively. Measure the crop yield.

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.

Solution: The *population* is the set of all supermarkets.

- *qualitative independent variable/factor*: $A =$ shelf placement with the *2 factor levels* $A_1 =$ normal shelf, $A_2 =$ cooling shelf.
- *2 subpopulations*: $P_1 =$ supermarkets with margarine in the normal shelf A_1; $P_2 =$ supermarkets with margarine in the cooling shelf A_2.
- *metric variable*: $Y =$ margarine sales, measured e.g. via kg of margarine sold per 1000 transactions at the cash register.
- *design of empirical investigation*: In 100 comparable supermarkets, place margarine in the normal self in 50 supermarkets and in the cooling shelf in the other 50 supermarkets. Measure the margarine sales over 1 month.

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.

Ex. 2.3: Effect of Teaching Method on Student Marks

Solution: The *factor* A is the teaching method with 3 *factor levels*:
A_1 = traditional teaching, A_2 = distance learning, A_3 = blended learning.
The *independent metric variable* is Y = mark (of the student).
In each subpopulation we have $n_1 = n_2 = n_3 = n = 4$ students.

ANOVA Model:

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

	A_1	A_2	A_3
1	70	57	88
2	80	54	82
3	75	46	90
4	75	43	80
sum	300	200	340
$\overline{y}_i = \frac{\text{sum}}{4}$	75	50	85

- *Means in the samples*:
  $\overline{y}_1 = 75, \ \overline{y}_2 = 50, \ \overline{y}_3 = 85$
- *Grand mean*: As the samples in each subpopulation have the same size $n = 4$:

$$\overline{y} = \frac{\overline{y}_1 + \overline{y}_2 + \overline{y}_3}{3} = \frac{75 + 50 + 85}{3} = \frac{210}{3} = 70$$

Ex. 2.3: Effect of Teaching Method on Student Marks

Computed so far: $\overline{y}_1 = 75$, $\overline{y}_2 = 50$, $\overline{y}_3 = 85$, and $\overline{y} = 70$
We complete an *ANOVA table* for $r = 3$ factor levels and for samples of the same size $n = 4$ in each subpopulation; hence $N = r \cdot n = 12$.

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

$$\begin{aligned}
\mathrm{SSA} &= 4 \cdot (75-70)^2 + 4 \cdot (50-70)^2 + 4 \cdot (85-70)^2 = 2600, \\
\mathrm{SSE} &= (70-75)^2 + (80-75)^2 + (75-75)^2 + (75-75)^2 \\
&\quad + (57-50)^2 + (54-50)^2 + (46-50)^2 + (43-50)^2 \\
&\quad + (88-85)^2 + (82-85)^2 + (90-85)^2 + (80-85)^2 = 248,
\end{aligned}$$

Ex. 2.3: Effect of Teaching Method on Student Marks

$$\text{SST} = \text{SSA} + \text{SSE} = 2600 + 248 = 2848.$$

The *ANOVA table* is shown below:

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F
Between Groups	2	2600	$\frac{2600}{2} = 1300$	$\frac{1300}{248/9} \approx 47.18$
Within Groups	9	248	$\frac{248}{9} \approx 27.56$	
Total	11	2848		

The random variable $F = \frac{\text{MSA}}{\text{MSE}}$ follows an *F-distribution with $r - 1 = 2$ numerator and $N - r = 9$ denominator degrees of freedom*. For our data we find the value:

$$f = \frac{1300}{248/9} \approx 47.18$$

Ex. 2.3: Effect of Teaching Method on Student Marks

Null Hypothesis H_0: The mark does not depend on the method of teaching, i.e. $\alpha_1 = \alpha_2 = \alpha_3 = 0$ or equivalently $\mu_1 = \mu_2 = \mu_3 = \mu$.

Alternative Hypothesis H_1: The mark does depend on the method of teaching, i.e. there is at least one $\alpha_i \neq 0$.

Hypothesis Testing with a significance level of $\alpha = 0.05$ (and $\alpha = 0.01$):
The tables for the *F-distribution for $r - 1 = 2$ numerator and $N - r = 9$ denominator degrees of freedom* for $\alpha = 0.05$ (and $\alpha = 0.01$) yield:

$$f_{2,9,0.05} = 4.26 \qquad \text{(and} \qquad f_{2,9,0.01} = 8.02\text{)}.$$

As $f \approx 47.18$ is strictly larger than these values we *reject the null hypothesis* H_0, and conclude that the teaching method affects the mark.

The chance of rejecting the null hypothesis, when it is in fact correct, is $\alpha = 0.05$ (and $\alpha = 0.01$), that is 5% (and 1%). So our conclusion has a 5% chance of error.

Ex. 2.4: 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* $\overline{y}_{i\cdot\cdot}$, $\overline{y}_{\cdot j}$ *and* $\overline{y}$.

Ex. 2.4: 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. 2.4: 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!

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

<u>Solution:</u>

	B_1	B_2	means
A_1	$y_{1,1,1} = 2,\ y_{1,1,2} = 2$	$y_{1,2,1} = 3,\ y_{1,2,2} = 4$	$\overline{y}_{1\cdot} = \frac{11}{4} = 2.75$
A_2	$y_{2,1,1} = 1,\ y_{2,1,2} = 2$	$y_{2,2,1} = 4,\ y_{2,2,2} = 5$	$\overline{y}_{2\cdot} = \frac{12}{4} = 3$
A_3	$y_{3,1,1} = 3,\ y_{3,1,2} = 2$	$y_{3,2,1} = 4,\ y_{3,2,2} = 4$	$\overline{y}_{3\cdot} = \frac{13}{4} = 3.25$
means	$\overline{y}_{\cdot 1} = \frac{12}{6} = 2$	$\overline{y}_{\cdot 2} = \frac{24}{6} = 4$	$\overline{y} = \frac{36}{12} = 3$

$$\overline{y}_{1\cdot} = \frac{1}{4} \cdot \big(y_{1,1,1} + y_{1,1,2} + y_{1,2,1} + y_{1,2,2}\big) = \frac{1}{4} \cdot \big(2 + 2 + 3 + 4\big) = \frac{11}{4} = 2.75$$

$$\overline{y}_{2\cdot} = \frac{1}{4} \cdot \big(y_{2,1,1} + y_{2,1,2} + y_{2,2,1} + y_{2,2,2}\big) = \frac{1}{4} \cdot \big(1 + 2 + 4 + 5\big) = \frac{12}{4} = 3$$

$$\overline{y}_{3\cdot} = \frac{1}{4} \cdot \big(y_{3,1,1} + y_{3,1,2} + y_{3,2,1} + y_{3,2,2}\big) = \frac{1}{4} \cdot \big(3 + 2 + 4 + 4\big) = \frac{13}{4} = 3.25$$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

$$\begin{aligned}
\overline{y}_{\cdot 1} &= \frac{1}{6}\cdot\big(y_{1,1,1}+y_{1,1,2}+y_{2,1,1}+y_{2,1,2}+y_{3,1,1}+y_{3,1,2}\big)\\
&= \frac{1}{6}\cdot\big(2+2+1+2+3+2\big)=\frac{12}{6}=2\\
\overline{y}_{\cdot 2} &= \frac{1}{6}\cdot\big(y_{1,2,1}+y_{1,2,2}+y_{2,2,1}+y_{2,2,2}+y_{3,2,1}+y_{3,2,2}\big)\\
&= \frac{1}{6}\cdot\big(3+4+4+5+4+4\big)=\frac{24}{6}=4\\
\overline{y} &= \frac{1}{12}\cdot\big(y_{1,1,1}+y_{1,1,2}+y_{1,2,1}+y_{1,2,2}+y_{2,1,1}+y_{2,1,2}\\
&\qquad +y_{2,2,1}+y_{2,2,2}+y_{3,1,1}+y_{3,1,2}+y_{3,2,1}+y_{3,2,2}\big)\\
&= \frac{1}{12}\cdot\big(2+2+3+4+1+2+4+5+3+2+4+4\big)=\frac{36}{12}=3
\end{aligned}$$

We compute the *means for the interaction of the factors*:

$$\begin{aligned}
\overline{y}_{1,1} &= \frac{1}{2}\cdot\big(y_{1,1,1}+y_{1,1,2}\big)=\frac{1}{2}\cdot(2+2)=\frac{4}{2}=2\\
\overline{y}_{1,2} &= \frac{1}{2}\cdot\big(y_{1,2,1}+y_{1,2,2}\big)=\frac{1}{2}\cdot(3+4)=\frac{7}{2}=3.5
\end{aligned}$$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

$$\overline{y}_{2,1} = \frac{1}{2} \cdot \left(y_{2,1,1} + y_{2,1,2}\right) = \frac{1}{2} \cdot (1+2) = \frac{3}{2} = 1.5$$

$$\overline{y}_{2,2} = \frac{1}{2} \cdot \left(y_{2,2,1} + y_{2,2,2}\right) = \frac{1}{2} \cdot (4+5) = \frac{9}{2} = 4.5$$

$$\overline{y}_{3,1} = \frac{1}{2} \cdot \left(y_{3,1,1} + y_{3,1,2}\right) = \frac{1}{2} \cdot (3+2) = \frac{5}{2} = 2.5$$

$$\overline{y}_{3,2} = \frac{1}{2} \cdot \left(y_{3,2,1} + y_{3,2,2}\right) = \frac{1}{2} \cdot (4+4) = \frac{8}{2} = 4$$

The means for the *interaction of two factor levels* are listed in the table below:

	B_1	B_2
A_1	$\overline{y}_{1,1} = 2$	$\overline{y}_{1,2} = \frac{7}{2} = 3.5$
A_2	$\overline{y}_{2,1} = \frac{3}{2} = 1.5$	$\overline{y}_{2,3} = \frac{9}{2} = 4.5$
A_3	$\overline{y}_{3,1} = \frac{5}{2} = 2.5$	$\overline{y}_{3,2} = 4$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

Computation of the *sums of squares*, where $r = 3$, $q = 2$ and $n = 2$:

$$\begin{aligned}
\text{SSA} &= n \cdot q \cdot \left[(\overline{y}_{1\cdot} - \overline{y})^2 + (\overline{y}_{2\cdot} - \overline{y})^2 + (\overline{y}_{3\cdot} - \overline{y})^2\right] \\
&= 4 \cdot \left[(2.75 - 3)^2 + (3 - 3)^2 + (3.25 - 3)^2\right] \\
&= 4 \cdot 2 \cdot 0.25^2 = \tfrac{8}{16} = \tfrac{1}{2} = 0.5
\end{aligned}$$

$$\begin{aligned}
\text{SSB} &= n \cdot r \cdot \left[(\overline{y}_{\cdot 1} - \overline{y})^2 + (\overline{y}_{\cdot 2} - \overline{y})^2\right] \\
&= 6 \cdot \left[(2 - 3)^2 + (4 - 3)^2\right] = 6 \cdot 2 = 12
\end{aligned}$$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

$$
\begin{aligned}
\text{SSAB} &= n \cdot \Big[(\overline{y}_{1,1} - \overline{y}_{1\cdot} - \overline{y}_{\cdot 1} + \overline{y})^2 + (\overline{y}_{1,2} - \overline{y}_{1\cdot} - \overline{y}_{\cdot 2} + \overline{y})^2 \\
&\qquad + (\overline{y}_{2,1} - \overline{y}_{2\cdot} - \overline{y}_{\cdot 1} + \overline{y})^2 + (\overline{y}_{2,2} - \overline{y}_{2\cdot} - \overline{y}_{\cdot 2} + \overline{y})^2 \\
&\qquad + (\overline{y}_{3,1} - \overline{y}_{3\cdot} - \overline{y}_{\cdot 1} + \overline{y})^2 + (\overline{y}_{3,2} - \overline{y}_{3\cdot} - \overline{y}_{\cdot 2} + \overline{y})^2\Big] \\
&= 2 \cdot \Big[(2 - 2.75 - 2 + 3)^2 + (3.5 - 2.75 - 4 + 3)^2 \\
&\qquad + (1.5 - 3 - 2 + 3)^2 + (4.5 - 3 - 4 + 3)^2 \\
&\qquad + (2.5 - 3.25 - 2 + 3)^2 + (4 - 3.25 - 4 + 3)^2\Big] \\
&= 2 \cdot \Big[(0.25)^2 + (-0.25)^2 + (-0.5)^2 + (0.5)^2 + (0.25)^2 + (-0.25)^2\Big] \\
&= 2 \cdot \Big[\tfrac{4}{16} + \tfrac{2}{4}\Big] = 2 \cdot \Big[\tfrac{1}{4} + \tfrac{1}{2}\Big] = \tfrac{3}{2} = 1.5
\end{aligned}
$$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

$$
\begin{aligned}
\text{SSE} &= (y_{1,1,1}-\overline{y}_{1,1})^2 + (y_{1,1,2}-\overline{y}_{1,1})^2 + (y_{1,2,1}-\overline{y}_{1,2})^2 \\
&\quad + (y_{1,2,2}-\overline{y}_{1,2})^2 + (y_{2,1,1}-\overline{y}_{2,1})^2 + (y_{2,1,2}-\overline{y}_{2,1})^2 \\
&\quad + (y_{2,2,1}-\overline{y}_{2,2})^2 + (y_{2,2,2}-\overline{y}_{2,2})^2 + (y_{3,1,1}-\overline{y}_{3,1})^2 \\
&\quad + (y_{3,1,2}-\overline{y}_{3,1})^2 + (y_{3,2,1}-\overline{y}_{3,2})^2 + (y_{3,2,2}-\overline{y}_{3,2})^2 \\
&= (2-2)^2 + (2-2)^2 + (3-3.5)^2 + (4-3.5)^2 \\
&\quad + (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 \\
&\quad + (3-2.5)^2 + (2-2.5)^2 + (4-4)^2 + (4-4)^2 \\
&= 8 \cdot 0.5^2 = 8 \cdot 0.25 = 2
\end{aligned}
$$

$$
\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE} = 0.5 + 12 + 1.5 + 2 = 16
$$

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

ANOVA Table:

Source	Sum of Squares	Degrees of Freedom (df)	Mean Square Variance	F-value
A	$\text{SSA} = \frac{1}{2} = 0.5$	$3 - 1 = 2$	$\text{MSA} = \frac{1}{4} = 0.25$	$\frac{\text{MSA}}{\text{MSE}} = \frac{1/4}{1/3} = \frac{3}{4} = 0.75$
B	$\text{SSB} = 12$	$2 - 1 = 1$	$\text{MSB} = 12$	$\frac{\text{MSB}}{\text{MSE}} = \frac{12}{1/3} = 36$
$A \times B$	$\text{SSAB} = \frac{3}{2} = 1.5$	$2 \cdot 1 = 2$	$\text{MSAB} = \frac{3}{4} = 0.75$	$\frac{\text{MSAB}}{\text{MSE}} = \frac{3/4}{1/3} = \frac{9}{4} = 2.25$
Error	$\text{SSE} = 2$	$12 - 2 \cdot 3 = 6$	$\text{MSE} = \frac{2}{6} = \frac{1}{3}$	
Total	$\text{SST} = 16$	$12 - 1 = 11$	$\text{MST} = \frac{16}{11}$	

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

Factor A (soil quality):

H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu$, i.e. the average crop yields $\mu_{i\cdot}$ for the different soil qualities are the same as the overall average crop yield μ. Hence the crop yield does not depend on the soil quality.

H_1: For at least one $\mu_{i\cdot}$ we have $\mu_{i\cdot} \neq \mu$, i.e. the crop yield does depend on the soil quality.

The random variable

$$F_A = \frac{\text{MSA}}{\text{MSE}}$$

follows an *F-distribution with (numerator, denominator)* $= (2, 6)$ *degrees of freedom*. From the table for the F-distribution for $\alpha = 0.05$ we find $f_{2,6,0.05} = 5.14$.

From the ANOVA table, we have the value $f_A = 0.75$ for $F_A = \frac{\text{MSA}}{\text{MSE}}$. As $f_A = 0.75 \leq f_{2,6,0.05} = 5.14$ we *cannot reject the null hypothesis*, and we conclude that the soil quality does not affect the crop yield.

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

Factor B (fertilizer):

H_0; $\mu_{\cdot 1} = \mu_{\cdot 2} = \mu$, i.e. the average crop yields $\mu_{\cdot j}$ for the different fertilizers are the same as the overall average crop yield μ. Hence the crop yield does not depend on the fertilizer.

H_1: Either $\mu_{\cdot 1} \neq \mu$ or $\mu_{\cdot 2} \neq \mu$, i.e. the crop yield depends on the fertilizer.

The random variable

$$F_B = \frac{\text{MSB}}{\text{MSE}}$$

follows an F*-distribution with (numerator, denominator)* $= (1, 6)$ *degrees of freedom*. From the table for the F-distribution for $\alpha = 0.05$ we find $f_{1,6,0.05} = 5.99$.

From the ANOVA table, we have the value $f_B = 36$ for $F_B = \frac{\text{MSB}}{\text{MSE}}$. As $f_B = 36 > f_{1,6,0.05} = 5.99$, we *reject the null hypothesis* and conclude that the crop yield does depend on the fertilizer. The chance of rejecting the null hypothesis, when it is in fact true, is $\alpha = 0.05$ or 5%.

Ex. 2.4: Crop Yield Depends on Soil Quality, Fertilizer

Interaction $A \times B$ (soil quality and fertilizer):

H_0: $\gamma_{1,1} = \gamma_{1,2} = \gamma_{2,1} = \gamma_{2,2} = \gamma_{3,1} = \gamma_{3,2}$, i.e. the average crop yield does not depend on the interaction of soil type and fertilizer.

H_1: For at least two pairs (i, j) and (k, ℓ) we have $\gamma_{i,j} \neq \gamma_{k,\ell}$, i.e. the crop yield depends on the interaction of soil type and fertilizer.

The random variable $F_{A\times B} = \frac{\text{MSAB}}{\text{MSE}}$ follows an *F-distribution with (numerator, denominator)* $= (2, 6)$ *degrees of freedom*. From the table for the F-distribution for $\alpha = 0.05$ we find $f_{2,6,0.05} = 5.14$.

From the ANOVA table, $f_{A\times B} = 2.25$ is the value for $F_{A\times B} = \frac{\text{MSAB}}{\text{MSE}}$. As $f_{A\times B} = 2.25 < f_{2,6,0.05} = 5.14$ the *null hypothesis cannot be rejected*, i.e. the crop yield does not depend on the interaction of soil type and fertilizer.

Comment: As the average crop yield does not depend on the soil type (factor A), it *does not make sense* to ask about the interaction $A \times B$.

Methods of Multivariate Statistics

Solutions to 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

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?

Solution: We start by computing the *arithmetic means* of the three random variables $X_1 =$ height, $X_2 =$ weight, and $X_3 =$ inseam length.

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

$$\overline{x_1} = \frac{1}{3} \cdot \big(180 + 160 + 170\big) = \frac{510}{3} = 170$$

$$\overline{x_2} = \frac{1}{3} \cdot \big(74 + 50 + 65\big) = \frac{189}{3} = 63$$

$$\overline{x_3} = \frac{1}{3} \cdot \big(78 + 68 + 73\big) = \frac{219}{3} = 73$$

So the *arithmetic means* are $\overline{x_1} = 170$ cm, $\overline{x_2} = 63$ kg, and $\overline{x_3} = 73$ cm.
Next we compute the *empirical variances* and *standard deviations*:

$$s_1^2 = \frac{1}{2} \cdot \left[(180-170)^2 + (160-170)^2 + (170-170)^2\right]$$

$$= \frac{1}{2} \cdot \left[10^2 + (-10)^2\right] = \frac{200}{2} = 100$$

$$s_2^2 = \frac{1}{2} \cdot \left[(74-63)^2 + (50-63)^2 + (65-63)^2\right]$$

$$= \frac{1}{2} \cdot \left[11^2 + (-13)^2 + 2^2\right] = \frac{294}{2} = 147$$

$$s_s^2 = \frac{1}{2} \cdot \left[(78-73)^2 + (68-73)^2 + (73-73)^2\right]$$

$$= \frac{1}{2} \cdot \left[5^2 + (-5)^2\right] = \frac{50}{2} = 25$$

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

The *empirical standard deviations* are given by:

$$s_1 = \sqrt{100} = 10, \qquad s_2 = \sqrt{147} \approx 12.124, \qquad s_3 = \sqrt{25} = 5.$$

Now we can compute the values for the *corresponding standardized random variables*:

$$Z_1 = \frac{X_1 - \overline{x_1}}{s_1} = \frac{X_1 - 170}{10}$$

$$Z_2 = \frac{X_2 - \overline{x_2}}{s_2} = \frac{X_1 - 63}{\sqrt{147}}$$

$$Z_3 = \frac{X_3 - \overline{x_3}}{s_3} = \frac{X_3 - 73}{5}$$

With these formulas, we compute the following *standardized data matrix*:

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

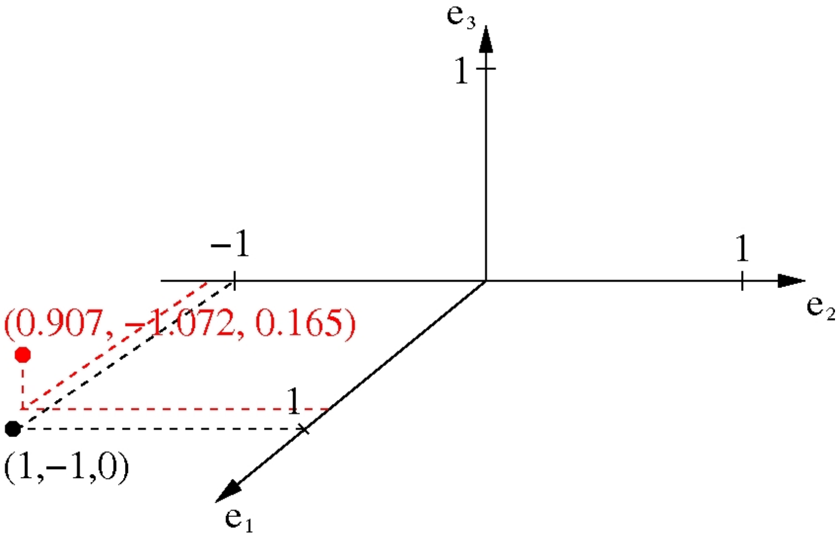
$$\mathbf{Z} = \begin{pmatrix} \frac{180-170}{10} & \frac{74-63}{\sqrt{147}} & \frac{78-73}{5} \\ \frac{160-170}{10} & \frac{50-63}{\sqrt{147}} & \frac{68-73}{5} \\ \frac{170-170}{10} & \frac{65-63}{\sqrt{147}} & \frac{73-73}{5} \end{pmatrix} \approx \begin{pmatrix} 1 & 0.907 & 1 \\ -1 & -1.072 & -1 \\ 0 & 0.165 & 0 \end{pmatrix}$$

The *columns* of the standardized data matrix are plotted on the next slide, where the axes of the coordinate system correspond to the persons e_1, e_2 and e_3. Thus a point in our coordinate system represents the values of one standardized random variable for the three persons in our sample.

The three points in our coordinate systems for the standardized random variables Z_1 (height), Z_2 (weight) and Z_3 (inseam length) are *very close together*, indicating a *strong correlation* between these variables.

Comment: The standardization of the random variables is here particularly useful, as it *removes the effect of the different scales* of the random variables and thus makes their correlation easily visible.

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



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. 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

Solution: The data matrix is given by

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

$$\mathbf{X} = \begin{pmatrix} 180 & 72 \\ 181 & 90 \\ 182 & 71 \\ 181 & 91 \end{pmatrix} \begin{matrix} \longleftarrow \text{ person } e_1 \\ \longleftarrow \text{ person } e_2 \\ \longleftarrow \text{ person } e_3 \\ \longleftarrow \text{ person } e_4 \end{matrix}$$

and we have plotted its row vectors on the next slide.

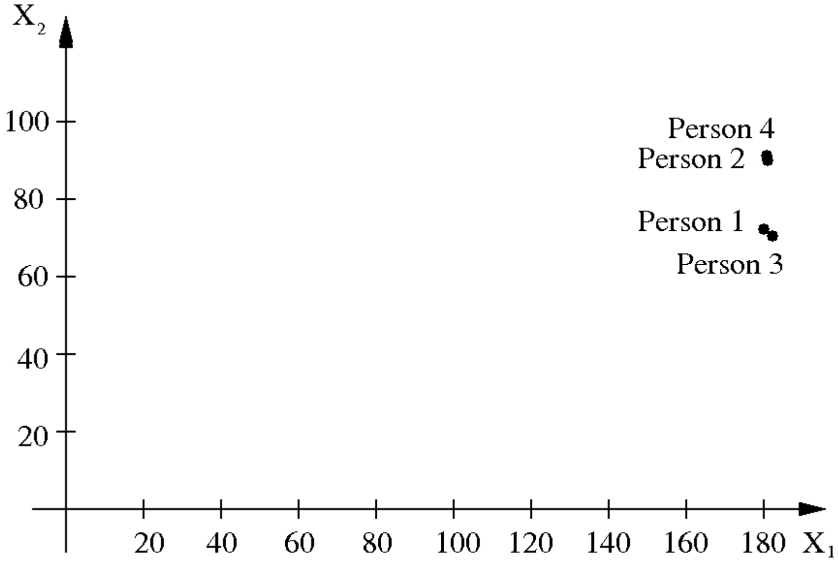
We observe two *clusters/groups* of points:

- cluster 1 contains persons e_1 and e_3
- cluster 2 contains persons e_2 and e_4

We may identify cluster 1 with normal weight persons and cluster 2 with slightly overweight persons.

Note: This way of forming clusters is still *too naive*: If we add another normal weight person with height 160 cm and weight 50 kg, then this person would lie far apart from both clusters due to her/his shorter height!

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



Ex. 3.3: Height and Weight Data, Euclidean Distance

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

Solution:

$$d_{1,1} = 0$$

$$d_{1,2} = \sqrt{(180-181)^2 + (72-90)^2} = \sqrt{(-1)^2 + (-18)^2} = \sqrt{325} \approx 18.028$$

$$d_{1,3} = \sqrt{(180-182)^2 + (72-71)^2} = \sqrt{(-2)^2 + 1^2} = \sqrt{5} \approx 2.236$$

$$d_{1,4} = \sqrt{(180-181)^2 + (72-91)^2} = \sqrt{(-1)^2 + (-19)^2} = \sqrt{362} \approx 19.026$$

Ex. 3.3: Height and Weight Data, Euclidean Distance

$$d_{2,1} = d_{1,2} = \sqrt{325} \approx 18.028$$

$$d_{2,2} = 0$$

$$d_{2,3} = \sqrt{(181 - 182)^2 + (90 - 71)^2} = \sqrt{(-1)^2 + 19^2} = \sqrt{362} \approx 19.026$$

$$d_{2,4} = \sqrt{(181 - 181)^2 + (90 - 91)^2} = \sqrt{0^2 + (-1)^2} = \sqrt{1} = 1$$

$$d_{3,1} = d_{1,3} = \sqrt{5} \approx 2.236$$

$$d_{3,2} = d_{2,3} = \sqrt{362} \approx 19.026$$

$$d_{3,3} = 0$$

$$d_{3,4} = \sqrt{(182 - 181)^2 + (71 - 91)^2} = \sqrt{1^2 + (-20)^2} = \sqrt{401} \approx 20.025$$

$$d_{4,1} = d_{1,4} = \sqrt{362} \approx 19.026$$

Ex. 3.3: Height and Weight Data, Euclidean Distance

$$d_{4,2} = d_{2,4} = 1$$

$$d_{4,3} = d_{3,4} = \sqrt{401} \approx 20.025$$

$$d_{4,4} = 0$$

From the computed distances, we find that *persons* e_1 *and* e_3 *are similar* and that *persons* e_2 *and* e_4 *are also similar*.

The persons e_1 and e_3 are *dissimilar* from the persons e_2 and e_4.

This reflects our results from the visualization in the previous question.

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

Solution: We compute the *city block distance* and the *Tschebyscheff distance*.

City block distance:

$$d_{1,1} = 0$$

$$d_{1,2} = |180 - 181| + |72 - 90| = 1 + 18 = 19 \quad \Rightarrow \quad d_{2,1} = d_{1,2} = 19$$

$$d_{1,3} = |180 - 182| + |72 - 71| = 2 + 1 = 3 \quad \Rightarrow \quad d_{3,1} = d_{1,3} = 3$$

$$d_{1,4} = |180 - 181| + |72 - 91| = 1 + 19 = 20 \quad \Rightarrow \quad d_{4,1} = d_{1,4} = 20$$

$$d_{2,2} = 0$$

$$d_{2,3} = |181 - 182| + |90 - 71| = 1 + 19 = 20 \quad \Rightarrow \quad d_{3,2} = d_{2,3} = 20$$

$$d_{2,4} = |181 - 181| + |90 - 91| = 0 + 1 = 1 \quad \Rightarrow \quad d_{4,2} = d_{2,4} = 1$$

$$d_{3,3} = 0$$

$$d_{3,4} = |182 - 181| + |71 - 91| = 1 + 20 = 21 \quad \Rightarrow \quad d_{4,3} = d_{3,4} = 21$$

$$d_{4,4} = 0$$

Ex. 3.4: City Block Distance and Tschebyscheff Distance

Tschebyscheff distance:

$$d_{1,1} = 0$$

$$d_{1,2} = \max\left\{|180 - 181|, |72 - 90|\right\} = \max\{1, 18\} = 18 \quad \Rightarrow \quad d_{2,1} = 18$$

$$d_{1,3} = \max\left\{|180 - 182|, |72 - 71|\right\} = \max\{2, 1\} = 2 \quad \Rightarrow \quad d_{3,1} = 2$$

$$d_{1,4} = \max\left\{|180 - 181|, |72 - 91|\right\} = \max\{1, 19\} = 19 \quad \Rightarrow \quad d_{4,1} = 19$$

$$d_{2,2} = 0$$

$$d_{2,3} = \max\left\{|181 - 182|, |90 - 71|\right\} = \max\{1, 19\} = 19 \quad \Rightarrow \quad d_{3,2} = 19$$

$$d_{2,4} = \max\left\{|181 - 181|, |90 - 91|\right\} = \max\{0, 1\} = 1 \quad \Rightarrow \quad d_{4,2} = 1$$

$$d_{3,3} = 0$$

$$d_{3,4} = \max\left\{|182 - 181|, |71 - 91|\right\} = \max\{1, 20\} = 20 \quad \Rightarrow \quad d_{4,3} = 20$$

$$d_{4,4} = 0$$

Ex. 3.4: City Block Distance and Tschebyscheff Distance

For both the *city block distance* and the *Tschebyscheff distance* we note from the computed distances that:

- the persons e_1 and e_3 are similar,
- the persons e_2 and e_4 are similar,
- the person e_1 and e_3 are dissimilar from the persons e_2 and e_4.

We note that we arrived at this conclusion regardless which distance was used.

Methods of Multivariate Statistics

Solutions to 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

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

Solution: We set X_1 = height and X_2 = weight. We have

$$Y = \mathbf{a}' \mathbf{x} = \frac{2}{\sqrt{5}} \cdot X_1 - \frac{1}{\sqrt{5}} \cdot X_2.$$

Group 1: K_1 = normal weight males

Person	X_1	X_2	Y
$e_{1,1}$	165	55	122.98
$e_{1,2}$	180	70	129.69
$e_{1,3}$	195	85	136.40
Means	180	70	129.69

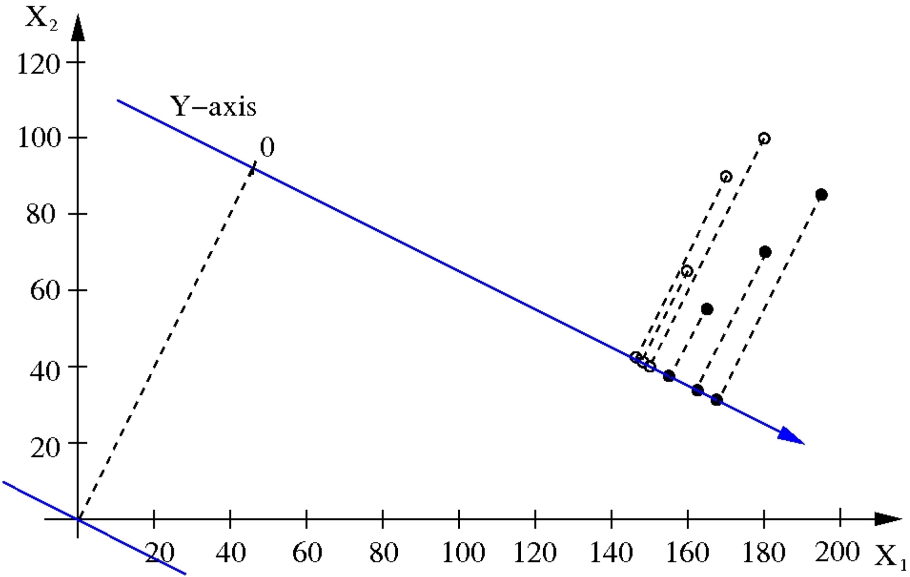
Group 2: K_2 = overweight males

Person	X_1	X_2	Y
$e_{2,1}$	160	65	114.04
$e_{2,2}$	170	90	111.80
$e_{2,3}$	180	100	116.28
Means	170	85	114.04

Means in group K_1: $\bar{\mathbf{x}}_1 = (180, 70)'$, $\bar{y}_1 = 129.69$

Means in group K_2: $\bar{\mathbf{x}}_2 = (170, 85)'$, $\bar{y}_2 = 114.04$

Ex. 4.1: Normal and Overweight Males



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

Solution: Let X_1 = height and X_2 = weight. From the calculations in Ex. 4.1, the *means for* $\mathbf{x} = (X_1, X_2)'$ are $\bar{\mathbf{x}}_1 = (180, 70)'$ in K_1 and $\bar{\mathbf{x}}_2 = (170, 85)'$ in K_2. We start with *computing the matrix* $\mathbf{W}$.

Ex. 4.2: Normal and Overweight

$$\mathbf{W} = \underbrace{\begin{pmatrix} 450 & 450 \\ 450 & 450 \end{pmatrix}}_{=\mathbf{W}_1} + \underbrace{\begin{pmatrix} 200 & 350 \\ 350 & 650 \end{pmatrix}}_{=\mathbf{W}_2} = \begin{pmatrix} 650 & 800 \\ 800 & 1100 \end{pmatrix},$$

where in group 1 ($= K_1$)

$$\begin{aligned}
(\mathbf{W}_1)_{11} &= (165-180)^2 + (180-180)^2 + (195-180)^2 = 450, \\
(\mathbf{W}_1)_{22} &= (55-70)^2 + (70-70)^2 + (85-70)^2 = 450, \\
(\mathbf{W}_1)_{12} = (\mathbf{W}_1)_{21} &= (165-180)\,(55-70) + (180-180)\,(70-70) \\
&\quad + (195-180)\,(85-70) = 450,
\end{aligned}$$

and in group 2 ($= K_2$)

$$\begin{aligned}
(\mathbf{W}_2)_{11} &= (160-170)^2 + (170-170)^2 + (180-170)^2 = 200, \\
(\mathbf{W}_2)_{22} &= (65-85)^2 + (90-85)^2 + (100-85)^2 = 650, \\
(\mathbf{W}_2)_{12} = (\mathbf{W}_1)_{21} &= (160-170)\,(65-85) + (170-170)\,(90-85) \\
&\quad + (180-170)\,(100-85) = 350.
\end{aligned}$$

Ex. 4.2: Normal and Overweight Males

$$\mathbf{W}^{-1} = \frac{1}{\det(\mathbf{W})}\begin{pmatrix} 1100 & -800 \\ -800 & 650 \end{pmatrix} = \frac{1}{75000}\begin{pmatrix} 1100 & -800 \\ -800 & 650 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{11}{750} & -\frac{4}{375} \\ -\frac{4}{375} & \frac{13}{1500} \end{pmatrix} \approx \begin{pmatrix} 0.0147 & -0.0107 \\ -0.0107 & 0.0087 \end{pmatrix},$$

with

$$\det(\mathbf{W}) = 1100 \cdot 650 - (-800) \cdot (-800) = 75000.$$

Find the vector $\mathbf{a}$: We compute $\mathbf{a} = \mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/\|\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\|_2$:

$$\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \begin{pmatrix} \frac{11}{750} & -\frac{4}{375} \\ -\frac{4}{375} & \frac{13}{1500} \end{pmatrix} \left[\begin{pmatrix} 180 \\ 70 \end{pmatrix} - \begin{pmatrix} 170 \\ 85 \end{pmatrix} \right]$$

$$= \begin{pmatrix} \frac{11}{750} & -\frac{4}{375} \\ -\frac{4}{375} & \frac{13}{1500} \end{pmatrix} \begin{pmatrix} 10 \\ -15 \end{pmatrix} = \begin{pmatrix} \frac{23}{75} \\ -\frac{71}{300} \end{pmatrix} \approx \begin{pmatrix} 0.307 \\ -0.237 \end{pmatrix},$$

Ex. 4.2: Normal and Overweight Males

$$\mathbf{a} = \frac{\mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\|\mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\|_2} \approx \frac{\begin{pmatrix} 0.307 \\ -0.237 \end{pmatrix}}{\sqrt{0.307^2 + (-0.237)^2}} \approx \begin{pmatrix} 0.792 \\ -0.611 \end{pmatrix}$$

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

Group 1: K_1 normal weight males

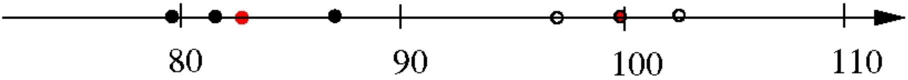
	X_1	X_2	Y
$e_{1,1}$	165	55	97.08
$e_{1,2}$	180	70	99.79
$e_{1,3}$	195	85	102.51
Means	180	70	99.79

Group 2: K_2 = overweight males

	X_1	X_2	Y
$e_{2,1}$	160	65	87.01
$e_{2,2}$	170	90	79.65
$e_{2,3}$	180	100	81.46
Means	170	85	82.71

Ex. 4.2: Normal and Overweight Males

To visualize the data for Y we only need one axis, the Y-axis representing the new variable Y.



The data for Y from group K_1 has been visualized by the unfilled dots and the data from group K_2 has been visualized by the filled dots. The dots in read represent the means of Y in the two groups.

We note that the two groups are pretty well separated.

Ex. 4.3: Classification of Normal and Overweight Males

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.

<u>Solution:</u> For the new person $x_1 = 190$ and $x_2 = 120$. Hence

$$y = 0.792 \cdot x_1 - 0.611 \cdot x_2 = 0.792 \cdot 190 - 0.611 \cdot 120 = 77.16.$$

Because

$$\begin{aligned} |77.16 - \overline{y}_1| &= |77.16 - 99.79| = 22.63 \\ > |77.16 - \overline{y}_2| &= |77.16 - 82.71| = 5.55 \end{aligned}$$

we *allocate the new person to the group* K_2 *(overweight male persons)*.

Methods of Multivariate Statistics

Solutions to 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

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

Ex. 5.1: Classifying Digital Cameras

Solution: *Initial partition*: $\mathcal{P}^{(0)} = \{K_1^{(0)}, K_2^{(0)}, K_3^{(0)}, K_4^{(0)}, K_5^{(0)}\}$ with the groups $K_i^{(0)} = \{e_i\}$ consisting of just one camera.

We compute the *initial distance matrix*

$$\mathbf{D}^{(0)} = (d_{ij}^{(0)})_{i,j=1,2,\ldots,5} = \begin{pmatrix} 0 & 2.5 & 3.5 & 10 & 11 \\ 2.5 & 0 & 6 & 7.5 & 8.5 \\ 3.5 & 6 & 0 & 13.5 & 14.5 \\ 10 & 7.5 & 13.5 & 0 & 1 \\ 11 & 8.5 & 14.5 & 1 & 0 \end{pmatrix},$$

where $(\mathbf{D}^{(0)})_{ij} = d_{ij}^{(0)}$ is the *city block distance of camera e_i and e_j*. The details of the computation of the matrix entries are shown below:

$$d_{1,1}^{(0)} = d_{2,2}^{(0)} = d_{3,3}^{(0)} = d_{4,4}^{(0)} = d_{5,5}^{(0)} = 0,$$

Ex. 5.1: Classifying Digital Cameras

$$
\begin{aligned}
d_{1,2}^{(0)} = d_{2,1}^{(0)} &= |1 - 1.5| + |6 - 8| = 2.5, \\
d_{1,3}^{(0)} = d_{3,1}^{(0)} &= |1 - 0.5| + |6 - 3| = 3.5, \\
d_{1,4}^{(0)} = d_{4,1}^{(0)} &= |1 - 5| + |6 - 12| = 10, \\
d_{1,5}^{(0)} = d_{5,1}^{(0)} &= |1 - 6| + |6 - 12| = 11, \\
d_{2,3}^{(0)} = d_{3,2}^{(0)} &= |1.5 - 0.5| + |8 - 3| = 6, \\
d_{2,4}^{(0)} = d_{4,2}^{(0)} &= |1.5 - 5| + |8 - 12| = 7.5, \\
d_{2,5}^{(0)} = d_{5,2}^{(0)} &= |1.5 - 6| + |8 - 12| = 8.5, \\
d_{3,4}^{(0)} = d_{4,3}^{(0)} &= |0.5 - 5| + |3 - 12| = 13.5, \\
d_{3,5}^{(0)} = d_{5,3}^{(0)} &= |0.5 - 6| + |3 - 12| = 14.5, \\
d_{4,5}^{(0)} = d_{5,4}^{(0)} &= |5 - 6| + |12 - 12| = 1.
\end{aligned}
$$

Ex. 5.1: Classifying Digital Cameras

Step 1: From inspecting the initial distance matrix

$$\mathbf{D}^{(0)} = (d_{ij}^{(0)})_{i,j=1,2,\ldots,5} = \begin{pmatrix} 0 & 2.5 & 3.5 & \mathit{10} & \mathit{11} \\ 2.5 & 0 & 6 & \mathit{7.5} & \mathit{8.5} \\ 3.5 & 6 & 0 & \mathit{13.5} & \mathit{14.5} \\ \mathit{10} & \mathit{7.5} & \mathit{13.5} & \mathit{0} & \mathbf{1} \\ \mathit{11} & \mathit{8.5} & \mathit{14.5} & \mathbf{1} & \mathit{0} \end{pmatrix},$$

we find that the *minimal non-diagonal entry is* $d_{4,5}^{(0)} = d_{5,4}^{(0)} = 1$ (displayed in bold-face).

Hence we unite the the groups $K_4^{(0)}$ *and* $K_5^{(0)}$.

We have to *delete the* 5*th row and* 5*th column* (displayed in italics) in $\mathbf{D}^{(0)}$ and *compute the new entries for the* 4*th row and* 4*th column* (displayed in italics).

Ex. 5.1: Classifying Digital Cameras

New partition after step 1: $\mathcal{P}^{(1)} = \{K_1^{(1)}, K_2^{(1)}, K_3^{(1)}, K_4^{(1)}\}$ with $K_1^{(1)} = \{e_1\}$, $K_2^{(1)} = \{e_2\}$, $K_3^{(1)} = \{e_3\}$ and $K_4^{(1)} = \{e_4, e_5\}$.

The new distance matrix $\mathbf{D}^{(1)}$ is given by

$$\mathbf{D}^{(1)} = (d_{i,j}^{(1)})_{i,j=1,2,\ldots,4} = \begin{pmatrix} 0 & 2.5 & 3.5 & \mathit{10} \\ 2.5 & 0 & 6 & \mathit{7.5} \\ 3.5 & 6 & 0 & \mathit{13.5} \\ \mathit{10} & \mathit{7.5} & \mathit{13.5} & \mathit{0} \end{pmatrix},$$

where the 4th row and 4th column anew (displayed in italics) were computed with the *nearest neighbor rule*: $d_{4,4}^{(1)} = 0$,

$$d_{4,1}^{(1)} = d_{1,4}^{(1)} = \min\{d_{4,1}^{(0)}, d_{5,1}^{(0)}\} = \min\{10, 11\} = 10,$$
$$d_{4,2}^{(1)} = d_{2,4}^{(1)} = \min\{d_{4,2}^{(0)}, d_{5,2}^{(0)}\} = \min\{7.5, 8.5\} = 7.5,$$
$$d_{4,3}^{(1)} = d_{3,4}^{(1)} = \min\{d_{4,3}^{(0)}, d_{5,3}^{(0)}\} = \min\{13.5, 14.5\} = 13.5.$$

Ex. 5.1: Classifying Digital Cameras

Step 2: The *minimal non-diagonal entry* in $\mathbf{D}^{(1)}$ is $d_{1,2}^{(1)} = d_{2,1}^{(1)} = 2.5$ (displayed in bold-face in the distance matrix $\mathbf{D}^{(1)}$ below).

Hence we *unite the two groups* $K_1^{(1)}$ *and* $K_2^{(1)}$.

New partition after step 2: $\mathcal{P}^{(2)} = \{K_1^{(2)}, K_2^{(2)}, K_3^{(2)}\}$ with $K_1^{(2)} = \{e_1, e_2\}$, $K_2^{(2)} = \{e_3\}$ and $K_3^{(2)} = \{e_4, e_5\}$

$$\mathbf{D}^{(1)} = (d_{i,j}^{(1)})_{i,j=1,2,\ldots,4} = \begin{pmatrix} \mathit{0} & \mathbf{2.5} & \mathit{3.5} & \mathit{10} \\ \mathbf{2.5} & \mathit{0} & \mathit{6} & \mathit{7.5} \\ \mathit{3.5} & \mathit{6} & 0 & 13.5 \\ \mathit{10} & \mathit{7.5} & 13.5 & 0 \end{pmatrix}.$$

We need to *delete the* 2*nd row and* 2*nd column* in $\mathbf{D}^{(1)}$ (displayed in italics) and *compute the new entries of the* 1*st row and* 1*st column* (displayed in italics).

Ex. 5.1: Classifying Digital Cameras

The *new distance matrix* $\mathbf{D}^{(2)}$ is given by

$$\mathbf{D}^{(2)} = (d_{i,j}^{(2)})_{i,j=1,2,3} = \begin{pmatrix} \mathit{0} & \mathit{3.5} & \mathit{7.5} \\ \mathit{3.5} & 0 & 13.5 \\ \mathit{7.5} & 13.5 & 0 \end{pmatrix},$$

where the new 1st row and 1st column (displayed in italics) were computed as follows:

$$\begin{aligned} d_{1,1}^{(2)} &= 0, \\ d_{1,2}^{(2)} = d_{2,1}^{(2)} &= \min\{d_{1,3}^{(1)}, d_{2,3}^{(1)}\} = \min\{3.5, 6\} = 3.5, \\ d_{1,3}^{(2)} = d_{3,1}^{(2)} &= \min\{d_{1,4}^{(1)}, d_{2,4}^{(1)}\} = \min\{10, 7.5\} = 7.5. \end{aligned}$$

Step 3: The *minimal entry in* $\mathbf{D}^{(2)}$ is given by $d_{1,2} = d_{2,1} = 3.5$ (displayed in bold-face in the matrix $\mathbf{D}^{(2)}$ on the next page).

Ex. 5.1: Classifying Digital Cameras

Hence we unite the groups $K_1^{(2)}$ *and* $K_2^{(2)}$.

New partition after step 3: $\mathcal{P}^{(3)} = \{K_1^{(3)}, K_2^{(3)}\}$ with
$K_1^{(3)} = \{e_1, e_2, e_3\}$, $K_2^{(3)} = \{e_4, e_5\}$

$$\mathbf{D}^{(2)} = (d_{i,j}^{(2)})_{i,j=1,2,3} = \begin{pmatrix} \mathit{0} & \mathbf{3.5} & \mathit{7.5} \\ \mathbf{3.5} & \mathit{0} & \mathit{13.5} \\ \mathit{7.5} & \mathit{13.5} & 0 \end{pmatrix}$$

We need to *delete the* 2*nd row and* 2*nd column of* $\mathbf{D}^{(2)}$ (displayed in italics) and *compute the new* 1*st row and* 1*st column* (displayed in italics). The new distance matrix is given by

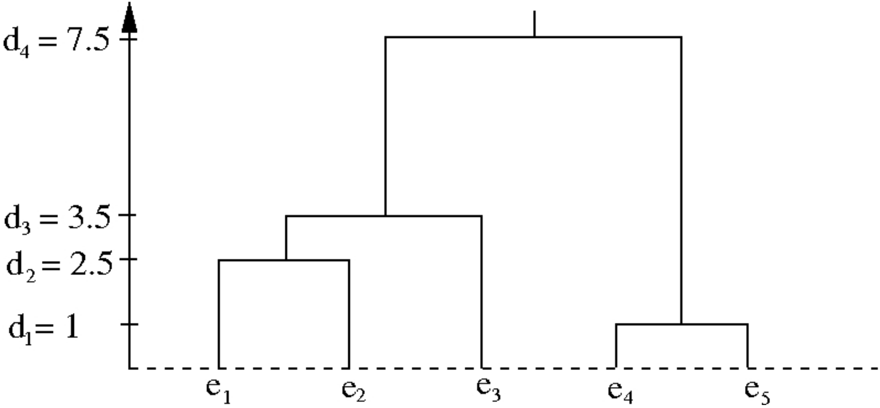
$$\mathbf{D}^{(3)} = (d_{i,j}^{(3)})_{i,j=1,2} = \begin{pmatrix} \mathit{0} & \mathit{7.5} \\ \mathit{7.5} & 0 \end{pmatrix},$$

where $d_{1,1}^{(3)} = 0$, $d_{1,2}^{(3)} = d_{2,1}^{(3)} = \min\{d_{1,3}^{(2)}, d_{2,3}^{(2)}\} = \min\{7.5, 13.5\} = 7.5$.

Ex. 5.1: Classifying Digital Cameras

Step 4: In the next step we finally *unite the remaining two groups* and obtain $\mathcal{P}^{(4)} = \{K_1^{(4)}\}$ with $K_1^{(4)} = \{e_1, e_2, e_3, e_4, e_5\}$.

The *minimal distance is* $d^3_{1,2} = d^{(3)}_{2,1} = 7.5$, but here we do not need to compute anything as $\mathbf{D}^{(0)} = (0)$.



Ex. 5.2: Classifying Digital Cameras

Determine the number of groups for the digital cameras from your results for Ex. 5.1.

Solution: For our digital camera example, we conclude from *inspecting the dendrogram* that we should have two groups:

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

since in the next (4th) step the distance increases drastically.

The *rule of thumb* provides

$$g \approx \sqrt{n/2} = \sqrt{5/2} \approx 1.58$$

which rounds to $g = 2$. This is also the number of groups that we determined from the dendrogram.

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

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

<u>Solution:</u> We have already computed the initial distance matrix in Ex. 5.1:

$$\mathbf{D} = (d_{i,j})_{i,j=1,2\dots,5} = \begin{pmatrix} \mathit{0} & \mathit{2.5} & \mathit{3.5} & 10 & 11 \\ \mathit{2.5} & \mathit{0} & \mathit{6} & 7.5 & 8.5 \\ \mathit{3.5} & \mathit{6} & \mathit{0} & 13.5 & 14.5 \\ 10 & 7.5 & 13.5 & \mathbf{0} & \mathbf{1} \\ 11 & 8.5 & 14.5 & \mathbf{1} & \mathbf{0} \end{pmatrix}$$

Numbers in italics are the distances between elements in $K_1 = \{e_1, e_2, e_3\}$, *numbers in bold-face* are the distances between elements in $K_2 = \{e_4, e_5\}$, and the *remaining numbers* are the distances between an element in $K_1 = \{e_1, e_2, e_3\}$ and an element in $K_2 = \{e_4, e_5\}$. Here $n_1 = 3$, $n_2 = 2$.

Ex. 5.3: Quality of the Classification of Digital Cameras

Average of the distances of the objects within a group:

$$g_1(K_1) = \frac{2}{3\cdot(3-1)} \cdot \left(d_{1,2} + d_{1,3} + d_{2,3}\right) = \frac{1}{3} \cdot \left(2.5 + 3.5 + 6\right) = \frac{12}{3} = 4,$$

$$g_1(K_2) = \frac{2}{2\cdot(2-1)} \cdot \left(d_{4,5}\right) = \frac{1}{1} = 1.$$

Distance of the least similar objects in a group:

$$g_2(K_1) = \max\{d_{1,2}, d_{1,3}, d_{2,3}\} = \max\{2.5, 3.5, 6\} = 6,$$

$$g_2(K_2) = \max\{d_{4,5}\} = \max\{1\} = 1.$$

Distance of the most similar objects in a group:

$$g_3(K_1) = \min\{d_{1,2}, d_{1,3}, d_{2,3}\} = \max\{2.5, 3.5, 6\} = 2.5,$$

$$g_3(K_2) = \min\{d_{4,5}\} = \min\{1\} = 1.$$

Ex. 5.3: Quality of the Classification of Digital Cameras

Complete linkage (furthest neighbor):

$$\begin{aligned} v_1(K_1, K_2) &= \max\{d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\} \\ &= \max\{10, 11, 7.5, 8.5, 13.5, 14.5\} = 14.5 \end{aligned}$$

Single linkage (nearest neighbor):

$$\begin{aligned} v_2(K_1, K_2) &= \min\{d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}, d_{3,4}, d_{3,5}\} \\ &= \min\{10, 11, 7.5, 8.5, 13.5, 14.5\} = 7.5 \end{aligned}$$

Average linkage: with $n_1 \cdot n_2 = 3 \cdot 2 = 6$,

$$\begin{aligned} v_3(K_1, K_2) &= \tfrac{1}{6}\left(d_{1,4} + d_{1,5} + d_{2,4} + d_{2,5} + d_{3,4} + d_{3,5}\right) \\ &= \tfrac{1}{6}\left(10 + 11 + 7.5 + 8.5 + 13.5 + 14.5\right) = \tfrac{65}{6} \approx 10.83 \end{aligned}$$

Ex. 5.3: Quality of the Classification of Digital Cameras

Squared Euclidean distance of the means:

With the data for the random variable X (see Table on page 70), we first compute the means in each group

$$\bar{\mathbf{x}}_1 = \frac{1}{3}\left[\begin{pmatrix} 1 \\ 6 \end{pmatrix} + \begin{pmatrix} 1.5 \\ 8 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 3 \end{pmatrix}\right] = \frac{1}{3}\begin{pmatrix} 3 \\ 17 \end{pmatrix} = \begin{pmatrix} 1 \\ 17/3 \end{pmatrix},$$

$$\bar{\mathbf{x}}_2 = \frac{1}{2}\left[\begin{pmatrix} 5 \\ 12 \end{pmatrix} + \begin{pmatrix} 6 \\ 12 \end{pmatrix}\right] = \frac{1}{2}\begin{pmatrix} 11 \\ 24 \end{pmatrix} = \begin{pmatrix} 11/2 \\ 12 \end{pmatrix}.$$

Now we can compute the Euclidean distance of the means:

$$\begin{aligned} v_4(K_1, K_2) &= \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|_2^2 = \left\|\begin{pmatrix} 1 \\ 17/3 \end{pmatrix} - \begin{pmatrix} 11/2 \\ 12 \end{pmatrix}\right\|_2^2 \\ &= \left\|\begin{pmatrix} -9/2 \\ -19/3 \end{pmatrix}\right\|_2^2 = \left(-\frac{9}{2}\right)^2 + \left(-\frac{19}{3}\right)^2 \approx 60.36. \end{aligned}$$