Exam Empirische Methoden

VU University Amsterdam, Faculty of Exact Sciences

March 25, 2011

NB. Use of a basic calculator is allowed; use of graphical/programmable calculators, mobile phones, etc. is not allowed.

Addendum: Formulas and Tables

NB. The exam can be made in the language of your preference: English or Dutch.

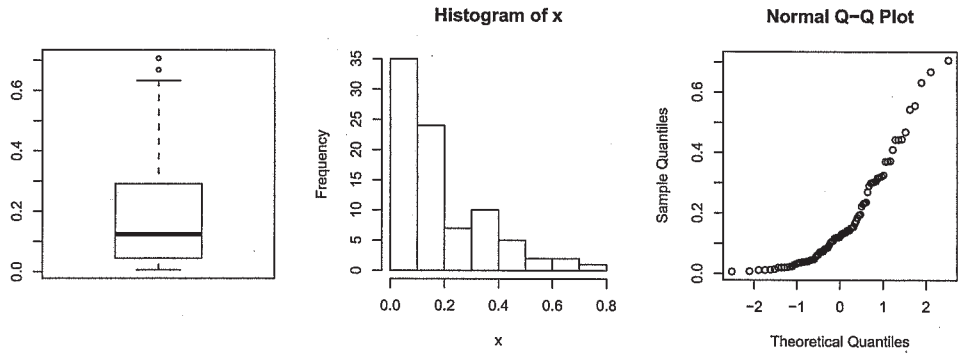
The 9 questions below all have the same weight.

1. a) Figure 1 shows a boxplot, a histogram and a normal QQ-plot for a representative sample x. Describe briefly what the boxplot tells you about the location, scale and shape of the underlying distribution of the data.

   b) Which aspects of the underlying distribution of the data can you see from the histogram that you cannot not infer from the boxplot?

   c) Which aspects of the underlying distribution of the data can you infer from the QQ-plot that you cannot not infer from the boxplot?

   d) Which of the two location measures will be larger for these data, the median or the mean? Why?



Figuur 1: Boxplot, histogram and normal QQ-plot of a sample x. (The whiskers in the boxplot connect the box with the most extreme data points that lie at most 1.5 times the interquartile range from the edges of the box.)

2. The (ordered) IQ-scores of a group of 36 people who did an IQ-test were

72	82	85	86	86	87
90	91	94	95	95	96
96	97	98	98	99	100
102	102	103	103	109	109
110	111	113	115	116	116
117	125	132	136	141	156.

The mean of these 36 numbers is $\bar{x} = 105$.

a) Give, based on the relative frequency, an estimate for the probability that a randomly chosen person from the population would have an IQ-score larger than 115.

IQ-tests are constructed such that IQ-scores in the population are normally distributed with expectation (population mean) $\mu = 100$ and standard deviation $\sigma = 15$.

b) What is the probability that a randomly chosen person would have an IQ-score larger than 115 for an IQ-test under the assumption that the IQ-scores in the population are normally distributed with expectation $\mu = 100$ and standard deviation $\sigma = 15$?

c) The estimate of part a) is different from the probability computed in part b). What could be the cause for this?

d) Someone said that such group with mean IQ-score 105 must be a very bright group. To check this, compute, under the same normality assumption as in part b), the probability that the mean IQ-score of 36 randomly chosen people is larger than or equal to 105. Is the person right?

3. A deck of cards has 4 red, 4 yellow, 4 blue cards. The red cards are all numbered 1, and the cards of the other two colors are numbered 1 to 4. Someone draws at random a card from the deck. Let A be the event that the card is red or yellow, and B the event that the card number is 1.

a) What is the probability that A and B occur? What is the probability that A or B occurs?

b) Are A and B independent? Why (not)?

Next, the person tosses a fair coin. If *head* comes up, he receives an amount of euros equal to the number of the drawn card, if *tails* comes up he receives nothing.

c) Construct a probability space for the *combined* experiment of drawing one card from the deck *and* tossing a fair coin.

d) Consider the random variable X which is the amount of euros received in the combined experiment. Construct the probability function $p(x) = P(X = x)$ of X based on the formal definition; you may present the results in a table.

e) Compute the expectation EX of X; do not only give the result, but also show how you obtained it.

4. In a study of changes in professional status of fathers and sons, professions were classified into levels 'high' and 'low'. In the generation of the fathers 60% had a low level job. A son of a father with a low level job has 70% chance to have a low level job too; a son of a father with a high level job has a chance of 80% to have a high level job too.
In the items below, do not only give your answer, but also show how you obtained it and name the rule(s) or property(ies) of probabilities that you have used for its computation.

a) What is the probability that two randomly chosen fathers both had a high level job?

b) What is the probability that at least one of three randomly chosen fathers had a high level job?

c) What is the probability that a randomly selected son has a high level job?

d) A randomly selected son has a high level job. What is the probability that his father had a high level job too?

5. In Question 2 the IQ-scores were given for $n = 36$ people. The mean and standard deviation of these 36 numbers were $\bar{x} = 105$ and $s_{36} = 18$. We now assume that we do not know anything about the distribution of IQ-scores in the population, and determine with these data a confidence interval for the mean μ of the IQ-scores in the population. From the Central Limit Theorem and because the standard deviation of the variable in the population is unknown and estimated by the sample standard deviation, we know that for large n the mean $\bar{X}$ of n IQ-scores lies with a probability of approximately 90% between the values $\mu - t_{n-1,0.95}s_n/\sqrt{n}$ and $\mu + t_{n-1,0.95}s_n/\sqrt{n}$, where $t_{n-1,0.95}$ denotes the 0.95-quantile of the t-distribution with $n - 1$ degrees of freedom.

a) Derive from this probability statement the *general expression* for a 90% confidence interval for the unknown value of μ.

b) What is the interpretation of the interval of part a)?

c) Compute corresponding to the general expression of part a) the 90% confidence interval for the unknown value of μ based on the given data set of 36 numbers.

d) Would the interval of part c) have been smaller or larger if the 0.95-quantile of the standard normal distribution would have been used, instead of the 0.95-quantile of the t-distribution? Explain shortly.

6. A large IT company needs to update its contract with a telecom provider. It has to choose between iPhones and HTC Androids for its employees. The management thinks that the majority of the employees would be happier with an HTC Android, and that only 20% of the employees would prefer an iPhone. To make sure that they make the right deal, the company conducts a poll among 36 employees to find out which proportion of the employees would prefer an iPhone above an HTC Android. Of these 36 employees 9 people said they would prefer an iPhone, and the other 27 said they would rather have an HTC Android.
Let p be the fraction of employees of the company that prefers an iPhone.

   a) Determine the usual point estimate $\hat{p}$ for p, and the margin of error for the 95% confidence interval for p.

   b) Do you think based on the answers of part a) that the result of the poll confirms the assumption of the management that 20% of the employees prefers an iPhone? Motivate your answer.

   c) What should have been the size of the poll if the management would have wanted the margin of error for the 95% confidence interval for p to be maximally 0.1?

   d) The management also wants to investigate its conjecture that 20% of the employees would prefer an iPhone with a statistical test. Formulate the null hypothesis H_0 and the alternative hypothesis H_a with respect to the fraction p, and give the expression for the test statistic, i.e. the standard score, for this situation.
   NB. You do not need to perform the test.

7. From a study on long-term effects of drinking alcohol on mental abilities two data sets of 40 test scores are available. For the first data set the mean and standard deviation of the test scores are $\bar{x}_1 = 83.3$ and $s_1 = 5.5$, for the second set $\bar{x}_2 = 73.3$ and $s_2 = 7.5$. Some other characteristics of the two samples are: the standard deviation of the differences between the score from the first sample and the corresponding second score from the second sample is $s_d = 7.05$; the square root of the pooled sample variance is $\bar{s}_{n_1,n_2} = 6.6$; $\bar{s}_{n_1,n_2}\sqrt{1/n_1 + 1/n_2} = 1.48$; $\sqrt{s_1^2/n_1 + s_2^2/n_2} = 1.47$; $df_{adjust} = 72$.

   Perform the parts a) and b) below with the appropriate characteristics.
   State in each of the two parts which requirements and/or assumptions need to be fulfilled for the methods that you use.

	gamma study	other study	total
Amsterdam	34 (30)	16 (20)	50
Den Haag	26 (30)	24 (20)	50
total	60	40	100

Tabel 1: Types of students present at demonstrations in Amsterdam and Den Haag.

a) First assume that the first sample is from 40 heavy drinkers just before they stopped drinking alcohol, and the second from the same 40 people after 2 years without alcohol. Test with a suitable t-test the claim that the alcohol has a permanent negative influence on mental abilities. Take significance level $\alpha = 0.05$.

b) Now assume that the data sets originate from two groups of size 40, one of now abstinent, but formerly moderate drinkers, and one of now abstinent, but formerly heavy drinkers. Give a (point) estimate of and a 95% confidence interval for the difference between the mean test scores for formerly moderate and formerly heavy drinkers.

8. During recent student demonstrations against the government's plans to have students pay 3000 euros if they study too long, polls of size 50 were held in Amsterdam and Den Haag to see what type of studies the students who took part in the demonstrations, were doing. Someone claimed that the number of "gamma study"students in the Amsterdam demonstration was larger than in Den Haag. This claim is investigated by means of a statistical test. The results of the polls are given in Table 1. The numbers between brackets in the table are the expected frequencies if there was no difference between the numbers of the two types of studies between the two cities.

   a) Does this concern a test for homogeneity for two samples or a test for independence of variables for one sample?

   b) Formulate the null and alternative hypothesis.

   c) Test the hypotheses with a chi-square test. Take significance level 0.05.

   d) Describe how one could investigate the claim with Fisher's exact test: formulate the null and alternative hypothesis, give the test statistic and its distribution under the null hypothesis (no need to specify parameters), and indicate when the null hypothesis will be rejected, for large values, for small values, or for both large and small values of the test statistic.
   NB. You do not need to perform this investigation yourself.

9. In Figure 2 a scatter plot is presented of the life expectancy of men in 40 different countries against the Gross National Product (GNP) of the country.

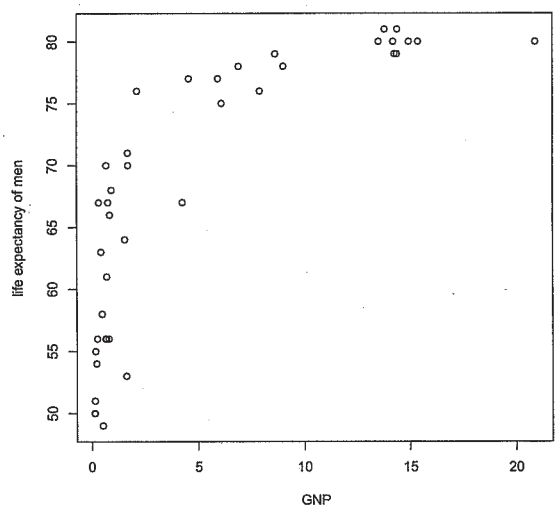
   a) Draw the best-fit line and estimate the sample correlation coefficient by eye.

   b) How much of the variation in the y-variable can be approximately accounted for by the best-fit line?

   Let b_1 be the unknown slope of the linear regression model with response variable *life expectancy of men* and explanatory variable *GNP*, $\hat{b}_1 = 1.4$ the corresponding estimate and $s_{\hat{b}_1} = 0.18$ the estimated standard deviation of $\hat{b}_1$.

   c) Test the claim that $b_1 = 0$, or, equivalently, that there is no linear relationship between the variables *life expectancy of men* and *GNP*. Take significance level $\alpha = 0.05$.

   d) Do you judge that the linear regression model is an appropriate model? Why (not)?



Figuur 2: Life expectancy of men against Gross National Product.

Formulas and Tables for Exam Empirische Methoden

Probability

We use the following notation:
$(\Omega, \mathcal{A}, P)$ probability space,
$A, B_1, B_2, \ldots, B_m \in \mathcal{A}$ events,
$B_1, B_2, \ldots, B_m$ a partition of Ω with $P(B_i) > 0$ for all $i \in \{1, 2, \ldots, m\}$; $r \in \{1, 2, \ldots, m\}$.

Rule of Total Probability:

$$P(A) = \sum_{i=1}^{m} P(A \cap B_i) = \sum_{i=1}^{m} P(A|B_i)P(B_i).$$

Bayes' Rule:

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{m} P(A|B_i)P(B_i)} = \frac{P(A|B_r)P(B_r)}{\sum_{i=1}^{m} P(A|B_i)P(B_i)}.$$

Two *independent* samples

(The formulas below hold under certain conditions.)

(i) If, for two *independent* samples, the population variances of the corresponding populations satisfy $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the statistic

$$T^{(2)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\bar{s}_{n_1,n_2}\sqrt{1/n_1 + 1/n_2}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Here $\bar{s}_{n_1,n_2}$ is the square root of the 'pooled' sample variance $\bar{s}^2_{n_1,n_2}$ given by

$$\bar{s}^2_{n_1,n_2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

(ii) If, for two *independent* samples, the population variances of the corresponding populations satisfy $\sigma_1^2 \neq \sigma_2^2$, then the denominator of $T^{(2)}$ is replaced by

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and the distribution of $T^{(2)}$ approximately is a t-distribution with *adjusted* number of degrees of freedom the following number rounded towards the nearest integer:

$$df_{adjust} = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

(iii) For two *independent* samples a confidence interval for $(p_1 - p_2)$ with confidence of approximately 95% is

$$[(\hat{p}_1 - \hat{p}_2) - E, (\hat{p}_1 - \hat{p}_2) + E],$$

with

$$E = 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

(iv) Moreover, if $p_1 = p_2$, then the statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2}}$$

is approximately standard normally distributed. Here $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$ is the 'pooled' sample fraction.

Correlation

Under certain conditions the statistic

$$T_{cor} = \frac{r - \rho}{\sqrt{(1-r^2)/(n-2)}}$$

has a t-distribution with $n-2$ degrees of freedom. Here r is the sample correlation coefficient given by

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}.$$

Linear regression

Let b_0 be the unknown intercept and b_1 the unknown slope of a linear regression model with one explanatory variable, and let $\hat{b}_0$ and $\hat{b}_1$ be the corresponding estimators, i.e. the intercept and slope of the 'best' line. Then $\hat{b}_0$ and $\hat{b}_1$ are given by

$$\hat{b}_1 = r\frac{s_y}{s_x}$$

and

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

Let b_k be the unknown coefficient of the k-th explanatory variable in a linear regression model with p explanatory variables ($p \geq 1$). Let $\hat{b}_k$ be its estimator, and $s_{\hat{b}_k}$ the estimated standard deviation of this estimator. If the measurement errors are independent and normally distributed, then the statistic

$$T_k = \frac{\hat{b}_k - b_k}{s_{\hat{b}_k}}$$

has a t-distribution with $n - p - 1$ degrees of freedom.

Tables standard normal, t- and chisquare distribution
for exam Empirische Methoden

Table 5.1 Standard Scores and Percentiles for a Normal Distribution (cumulative values from the left)

Standard score	%	Standard score	%	Standard score	%	Standard score	%
−3.5	0.02	−1.0	15.87	0.0	50.00	1.1	86.43
−3.0	0.13	−0.95	17.11	0.05	51.99	1.2	88.49
−2.9	0.19	−0.90	18.41	0.10	53.98	1.3	90.32
−2.8	0.26	−0.85	19.77	0.15	55.96	1.4	91.92
−2.7	0.35	−0.80	21.19	0.20	57.93	1.5	93.32
−2.6	0.47	−0.75	22.66	0.25	59.87	1.6	94.52
−2.5	0.62	−0.70	24.20	0.30	61.79	1.7	95.54
−2.4	0.82	−0.65	25.78	0.35	63.68	1.8	96.41
−2.3	1.07	−0.60	27.43	0.40	65.54	1.9	97.13
−2.2	1.39	−0.55	29.12	0.45	67.36	2.0	97.72
−2.1	1.79	−0.50	30.85	0.50	69.15	2.1	98.21
−2.0	2.28	−0.45	32.64	0.55	70.88	2.2	98.61
−1.9	2.87	−0.40	34.46	0.60	72.57	2.3	98.93
−1.8	3.59	−0.35	36.32	0.65	74.22	2.4	99.18
−1.7	4.46	−0.30	38.21	0.70	75.80	2.5	99.38
−1.6	5.48	−0.25	40.13	0.75	77.34	2.6	99.53
−1.5	6.68	−0.20	42.07	0.80	78.81	2.7	99.65
−1.4	8.08	−0.15	44.04	0.85	80.23	2.8	99.74
−1.3	9.68	−0.10	46.02	0.90	81.59	2.9	99.81
−1.2	11.51	−0.05	48.01	0.95	82.89	3.0	99.87
−1.1	13.57	0.0	50.00	1.0	84.13	3.5	99.98

NB: these are percentages!

Table 10.1 Critical Values of t

Degrees of freedom ($n-1$)	Area in one tail 0.025	Area in one tail 0.05
	Area in two tails 0.05	Area in two tails 0.10
1	12.706	6.314
2	4.303	2.920
3	3.182	2.353
4	2.776	2.132
5	2.571	2.015
6	2.447	1.943
7	2.365	1.895
8	2.306	1.860
9	2.262	1.833
10	2.228	1.812
11	2.201	1.796
12	2.179	1.782
13	2.160	1.771
14	2.145	1.761
15	2.131	1.753
16	2.120	1.746
17	2.110	1.740
18	2.101	1.734
19	2.093	1.729
20	2.086	1.725
21	2.080	1.721
22	2.074	1.717
23	2.069	1.714
24	2.064	1.711
25	2.060	1.708
26	2.056	1.706
27	2.052	1.703
28	2.048	1.701
29	2.045	1.699
30	2.042	1.697
31	2.040	1.696
32	2.037	1.694
34	2.032	1.691
36	2.028	1.688
38	2.024	1.686
40	2.021	1.684
50	2.009	1.676
100	1.984	1.660
Large	1.960	1.645

df **$t_{df;\,0.975}$ quantile** **$t_{df;\,0.95}$ quantile**

Table 10.7 Critical Values of χ^2: Reject H_0 Only If $\chi^2 >$ Critical Value

Table size (rows × columns)	Significance level 0.05	0.01
2 × 2	3.841	6.635
2 × 3 or 3 × 2	5.991	9.210
3 × 3	9.488	13.277
2 × 4 or 4 × 2	7.815	11.345
2 × 5 or 5 × 2	9.488	13.277

0.95- and 0.99-quantiles chisquare distribution (the critical values)