Exam Empirische Methoden

VU University Amsterdam, Faculty of Exact Sciences

March 29, 2012

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 7 questions below all have the same weight.

1. Are the following statements sensible/correct? Motivate your answer.

   a) According to the "68-95-99.7 Rule", about 95% of the data points in a sample will fall within 2 standard deviations of the population mean.

   b) Whether or not two independent samples originate from the same distribution can be investigated with a QQ-plot.

   c) For a statistical test the following holds: when the sample size becomes larger, the probability of a type I error becomes smaller and the power becomes larger.

   d) In order to estimate a population mean with a specified margin of error of at most E, the sample size n should at least be $(2s/E)^2$, where s is the sample standard deviation.

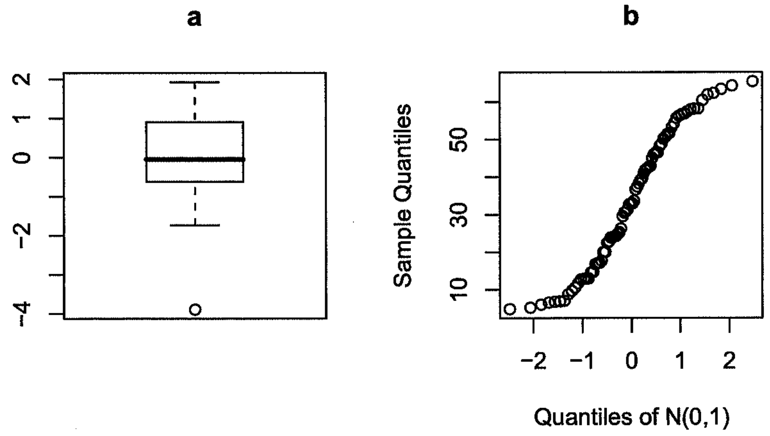
2. a) Briefly describe two methods to visualize a set of data that are measured on a nominal level.

   b) Figure 1.a shows a boxplot of a data set x. What can you deduce from this graph with respect to location, scale (= spread) and shape of the data distribution?

   c) In Figure 1.b a normal QQ-plot is shown for another data set, y. Based on this plot: to which of the following location-scale families could the underlying distribution of the data belong, that of the N(0,1), the $exp(1)$ distribution, the χ^2_{10} distribution, or the unif(0,1) distribution? Motivate your answer.

3. Consider an experiment consisting of two tosses with a biased coin. The probability of head is 0.7, the probability of tail is 0.3.
   (*In the items below, you may leave a ratio or product in your answer.*)

   a) What are the outcome space Ω and the probability measure P for this experiment?



Figuur 1: a. Boxplot of sample x; b. QQ-plot of sample y.

Let A be the event that in both tosses head comes up, and B the event that in both tosses tail comes up.

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

If head comes up the person who tosses receives 1 euro, if tail comes up he gets nothing.

c) Consider the random variable X which is the total amount of euros received in the experiment. Construct the probability function $p(x) = P(X = x)$ of X based on the formal definition.

d) Compute the expectation EX of X; do not only give the result, but also show how this was obtained.

4. Three years ago 80% of the adult population had "good" financial credit ratings, while the remaining 20% had "bad" financial credit ratings. At present 30% of the people who had bad ratings three years ago now have good ratings, and 15% of those with good ratings three years ago now have bad ratings.
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 people from the population both had bad credit ratings three years ago?

b) What is the probability that at least one of three randomly chosen people had good credit ratings three years ago?

c) What is the probability that a randomly selected person now has good ratings?

d) What is the probability that a person who has good ratings now, had bad ratings three years ago?

5. a) Suppose that the birth weight of male babies is normally distributed with expectation $\mu = 3.39$ and $\sigma = 0.67$. What is the probability that the average of 16 male babies is larger than 3.675 kg?

b) When birth weights were recorded for a sample of 16 male babies born to mothers taking a special vitamin supplement, the sample had a mean of $\bar{x} = 3.675$ kg and a standard deviation of $s = 0.657$ kg. Suppose that the distribution of the birth weights of all male babies of mothers given vitamins is normally distributed with unknown expectation μ and unknown σ. Test the claim that the mean birth weight for male babies of mothers given vitamins is different from 3.39 kg with a suitable test.

c) Based on these results, does the vitamin supplement appear to have an effect on birth weight?

6. To investigate the claim that ibuprofen, if taken six hours before climbing, reduces altitude sickness, one group of 44 climbers who had taken ibuprofen was compared to a control group of 42 climbers who had taken a placebo. In the ibuprofen-group 19 people suffered from altitude sickness; in the control group 29 people suffered from altitude sickness.
Let p_1 denote the proportion of people that will suffer from altitude sickness in the population of ibuprofin-taking climbers, and p_2 the proportion of people that will suffer from altitude sickness in the population of non-ibuprofen-taking climbers.
Some characteristics of the two samples, that you may or may not use, are:
the pooled sample fraction is $\bar{p} = 0.558$; $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = 0.103$; $\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2} = 0.107$.

a) Formulate the claim in terms of p_1 and p_2, and give based on the data a (point) estimate for the difference in proportion of people that will suffer from altitude sickness among ibuprofen-taking climbers and among non-ibuprofen-taking climbers.

b) Investigate the claim with a suitable test: formulate H_0 and H_A, give the expression of the test statistic and its distribution under H_0, and perform the test. Take significance level $\alpha = 5\%$.

Another way to investigate the claim would be to consider the same data but now given in the two-way table

	sick	not sick
ibuprofen	19	25
placebo	29	13

and to perform a test of homogeneity for the two samples.

c) Describe how it could be tested whether or not ibuprofen influences altitude sickness with this approach: formulate H_0 and H_A, give the expression of the test statistic and its distribution under H_0.
NB. You do not need to perform the test.

d) Which test do you prefer for this situation: the one of part b or the one of part c? Briefly motivate your answer.

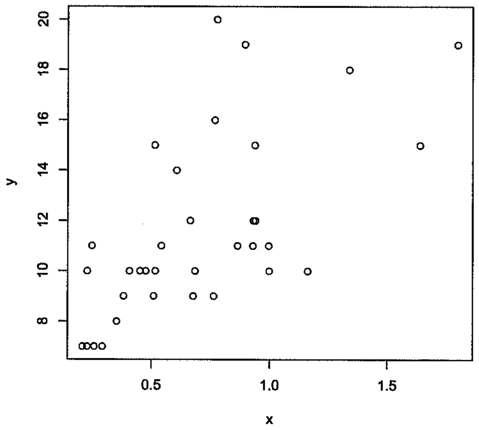
7. In Figure 2 a scatter plot of 34 points for two variables x and y is presented. Some characteristics of the data that you may or may not use are:
$\bar{x} = 0.70$, $\bar{y} = 11.59$, $s_x = 0.39$, $s_y = 3.60$,
$r = 0.64$, $\sqrt{(1-r^2)/(n-2)} = 0.14$, $s_{\hat{b}_1} = 1.25$.

a) Make a sketch of the best-fit line and give its intercept and its slope (as estimated by eye).

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

c) Give an estimate and a 95% confidence interval for the population correlation coefficient of the two variables.

d) Is there sufficient evidence in the data to conclude that there is a significant linear correlation? Motivate your answer.



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

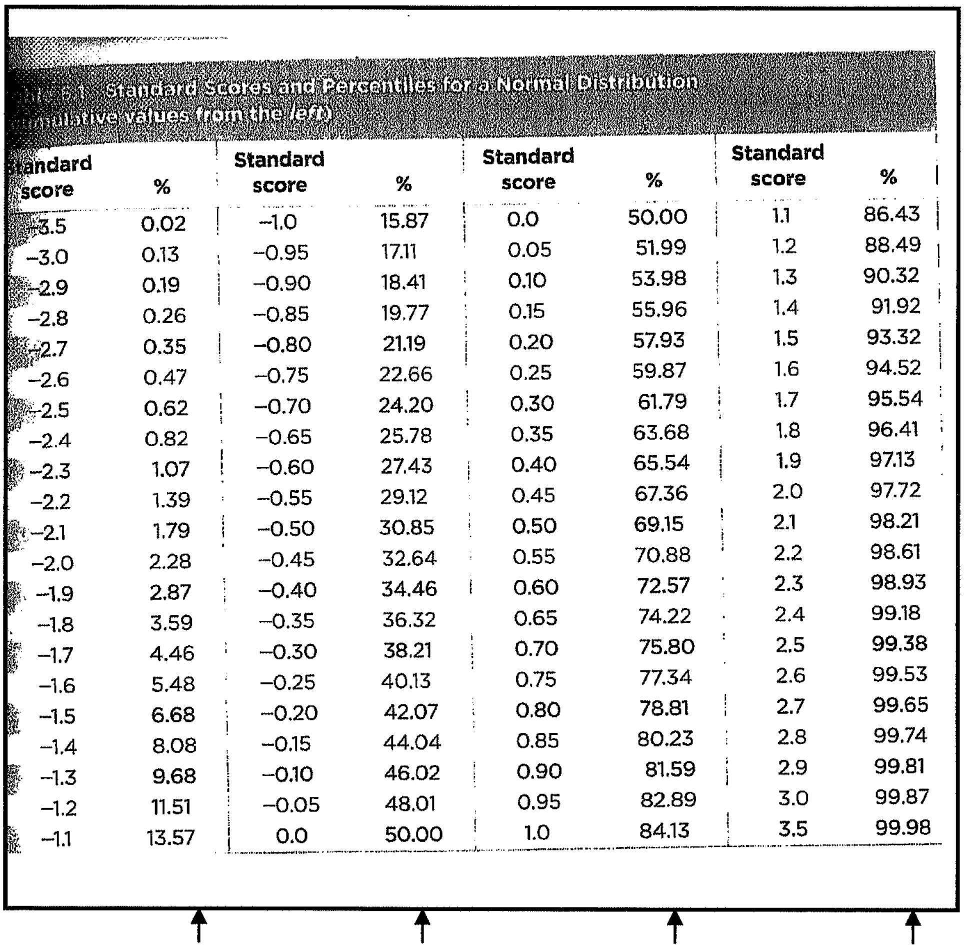
$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}.$$

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



1 Standard Scores and Percentiles for a Normal Distribution
mulative values from the *left*)

andard 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!

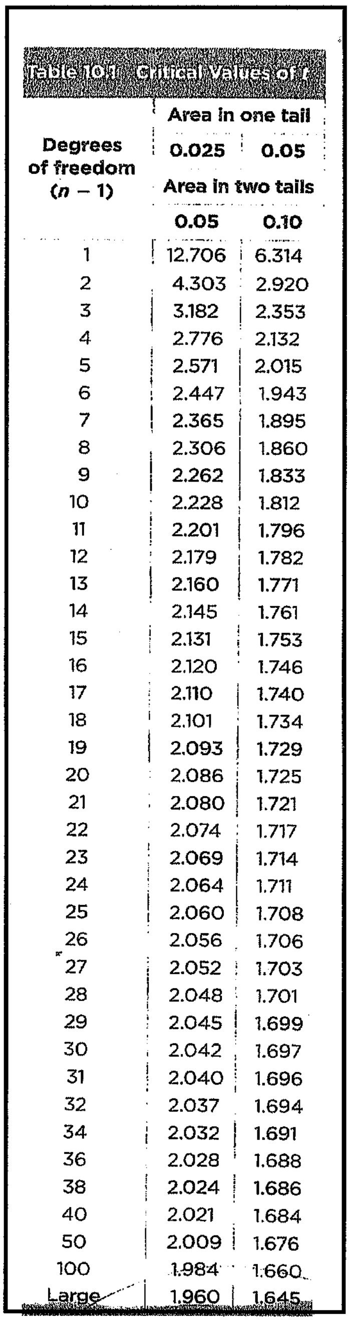


Table 10.1 Critical Values of *t*

Degrees of freedom $(n-1)$	Area in one tail	
	0.025	0.05
	Area in two tails	
	0.05	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**

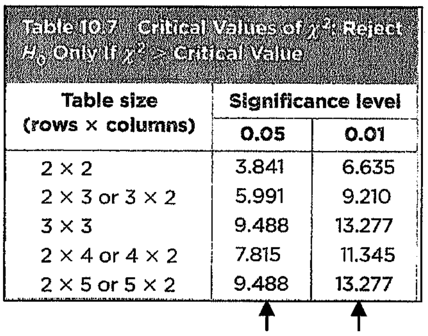


Table 10.7 Critical Values of χ^2: Reject H_0 Only If χ^2 > Critical Value

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)**