Exam Empirische Methoden

VU University Amsterdam, Faculty of Exact Sciences

February 4, 2014

NB. Only the use of a basic calculator is allowed; use of graphical/programmable calculators, mobile phones, smart watches, etc. is not allowed.

Addendum: Formulas and Tables

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

Division of points: (1) a,b,c:1; d:3; e:2. (2) a,b,c:3; d:2. (3) a:2; b,d,e,:1; c:4. (4) a:1; b:5; c:2. (5) a:1; b:5; c:2. (6) a:1; b:2; c:5; d:2. The exam grade will be 1 + (total points)/6.

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

a) In a box plot the five number summary of the data is visualized: the minimum, the first quartile, the mean, the third quartile and the maximum.

b) For a sample from a right-skewed distribution the sample mean will generally be larger than the median.

c) For a test with significance level 0.10, we cannot say anything about the probability of a type I error if we do not know the sample size.

d) The probability that a normally distributed random variable with mean 5 and standard deviation 2 is larger than 8.4 is equal to 4.46%.

e) For the data in the following contingency table it is given that the value of the chi-square statistic is 4.18.

43	15
35	29

The statement to evaluate is the following.
For these data the chi-square test for testing whether or not there is a relationship between the row- and column-variable, rejects the null hypothesis of no relationship for significance level 5%, but not for significance level 1%.

2. Urn M and urn N each contain 3 white chips, 2 blue chips and 1 red chip; for each urn the white chips are numbered $1, 2, 3$, the blue chips $1, 2$, and the red chip has number 1. Tim randomly draws two chips, one from each urn.

   a) Consider the experiment of drawing the two chips. Give the outcome space Ω and the probability measure P for this experiment.

   b) Let A be the event that Tim draws 1 blue and 1 red chip and B the event that the chip that is drawn from urn M is red. Are A and B independent? Motivate your answer.

   If the chip that Tim draws from urn M is white, Tim receives 1 euro, if it is blue he receives 2 euro, and if it is red, he has to pay 1 euro; if the chip that Tim draws from urn N is red he also has to pay 1 euro.

   c) Consider the random variable X which is the amount (in euros) Tim earns. Make a table with two columns: one with all possible values x of X, and one with the corresponding probabilities $P(X = x)$ that X takes the value x. Do not only give the table, but also show how the probabilities were computed.

   d) Compute, using the results of part c, the expectation EX of X. Do not only give the result, but also show how it was obtained.

3. One of the tasks of the public service provider RWD is to monitor the technical conditions of vehicles. RDW inspects 3% of the cars that have passed the general periodical inspection (APK) to check whether they were correctly APK-approved. In what follows p denotes the proportion of all APK-approved cars that are wrongly APK-approved.

   a) What is the interpretation of a 95% confidence interval for an unknown population proportion p?

   b) On one day RDW inspected 324 cars and found 29 of them to be wrongly APK-approved. Give, based on these data, a point estimate of p.

   c) Compute, based on the same data, the margin of error for the 95% confidence interval for the unknown proportion p of wrongly APK-approved cars, and compute the corresponding 95% confidence interval for p.

   d) To determine how many APK-approved cars need to be inspected in order to have a margin of error for the 95% confidence interval for p based on that day's data to be below a certain value, one could use the formula $n \approx 1/E^2$, where n denotes the number of inspected APK-approved cars and E denotes the margin of error. Compute which margin of error one would approximately obtain for $n = 324$ according to this formula.

   e) Compare the values that you found for the margin of error in parts c and d, and explain their difference/similarity.

4. Listed below are two sets of body temperatures (in ^{o}C):
sample 1: 36.1 35.7 36.4 35.8 36.6 37.3;
sample 2: 36.7 37.0 37.1 36.7 37.0 36.4.
The sample means and sample standard deviations for these data are for sample 1: $\bar{x}_1 = 36.3$, $s_1 = 0.59$; for sample 2: $\bar{x}_2 = 36.8$, $s_2 = 0.26$; for the pairwise differences: $\bar{x}_d = -0.5$, $s_d = 0.75$. Some other characteristics of the data that you may or may not use are $\bar{s}\sqrt{1/n_1 + 1/n_2} = \sqrt{s_1^2/n_1 + s_2^2/n_2} = 0.26$, $df_{adjust} = 7$.
Assume that the data are temperatures of 6 subjects, measured at 8:00 AM (sample 1) and 12:00 AM (sample 2).

   a) Give a point estimate of the population parameter $\mu_1 - \mu_2$, the difference between the mean body temperature at 8:00 AM and the mean body temperature at 12:00 AM.

   b) Investigate with an appropriate test the claim that the mean body temperature at 8:00 AM is lower than the mean body temperature at 12:00 AM. Take significance level 5%. As always, formulate the relevant H_0 and H_a, give a formula for the test statistic and specify its distribution under H_0, compute the observed value of the test statistic, and perform the test.

   c) The test that you performed in part b should only be used under some requirements for the two samples. Which are these requirements and is it reasonable to assume that they are fulfilled in this case?

5. Consider the same data as in Question 4, but now assume that all measurements were taken at 8:00 AM: sample 1 of six men and sample 2 of six women.

   a) Give a point estimate of the population parameter $\mu_1 - \mu_2$, the difference between the mean body temperature at 8:00 AM of men and the mean body temperature at 8:00 AM of women.

   b) Investigate with an appropriate test the claim that the mean body temperature at 8:00 AM of men and the mean body temperature at 8:00 AM of women are different. Take significance level 5%. Again, formulate the relevant H_0 and H_a, give a formula for the test statistic and specify its distribution under H_0, compute the observed value of the test statistic, and perform the test.

   c) The test that you performed in part b should only be used under some requirements for the two samples. Which are these requirements and is it reasonable to assume that they are fulfilled in this case?

6. In Figure 1 a scatter plot of 36 points corresponding to the data sets x and y for two variables is presented, as well as a normal QQ-plot of the residuals of a linear regression of y on x. Some characteristics of the data that you may or may not use are: $\bar{x} = 10.24$, $\bar{y} = -35.39$, $s_x = 6.97$, $s_y = 31.67$, $r = -0.78$, $\sqrt{(1-r^2)/(n-2)} = 0.11$, $\hat{b}_0 = 0.86$, $\hat{b}_1 = -3.54$, $s_{\hat{b}_1} = 0.49$.

   a) How much of the variation in the y-variable can approximately be accounted for by the x-variable using a linear regression of y on x, i.e. by the 'best-fit' line?

   b) Do you identify any outliers in the plot of Figure 1? If so, briefly discuss the effect of the presence of the outlier(s) on the strength of the correlation between x and y and on the position of 'best-fit' line.

   c) Using significance level 5% test the claim that there is no linear relationship between the explanatory variable x and the response variable y. (Formulate the relevant H_0 and H_a, give a formula for the test statistic and specify its distribution under H_0, compute the observed value of the test statistic, and perform the test.)

   d) In view of the plots, the data characteristics and your answers to parts (a)–(c): do you judge that the linear regression model is an appropriate model for these data? Motivate your answer.

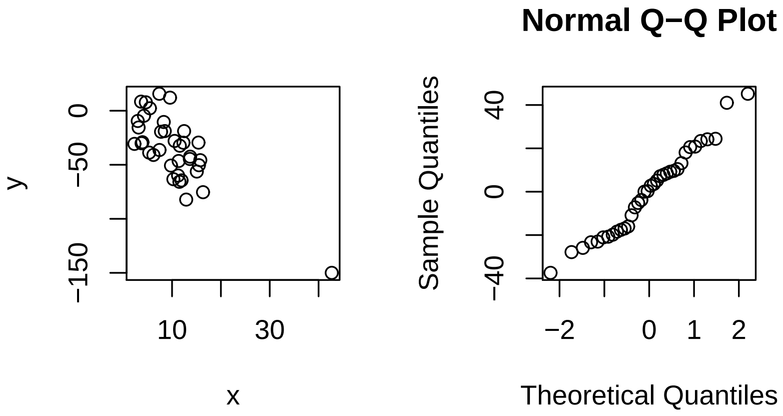


Figure 1: Scatter plot of x and y and normal QQ-plot of the residuals of linear regression of y on x.

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 $i \in \{1, 2, \ldots, m\}$.

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.)

For two *independent* samples,
(i) if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the statistic

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\bar{s}\sqrt{1/n_1 + 1/n_2}}$$

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

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

(ii) if $\sigma_1^2 \neq \sigma_2^2$, we use the general result that the statistic

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

approximately has a t-distribution with $\tilde{n}$ degrees of freedom. Here $\tilde{n}$ equals the following number rounded towards the nearest integer:

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

(iii) the statistic

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

approximately has a standard normal distribution.

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

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

approximately has a standard normal distribution. 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 ρ is the population correlation coefficient and r is the sample correlation coefficient given by

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

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 regression line (the 'best' line). Then $\hat{b}_0$ and $\hat{b}_1$ are given by

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

and

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

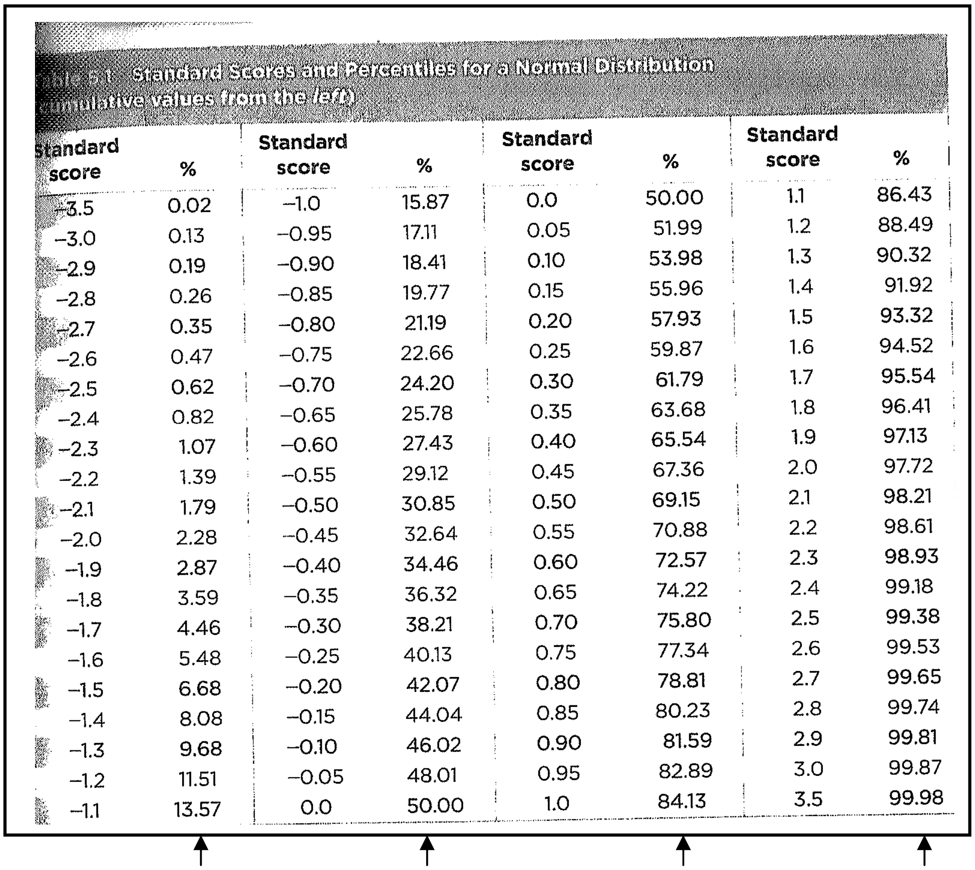
If the measurement errors are independent and normally distributed, then the statistic

$$t_1 = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

has a t-distribution with $n-2$ degrees of freedom. Here $s_{\hat{b}_1}$ is the estimated standard deviation of the estimator $\hat{b}_1$.

Tables standard normal, t- and chi-square distributions

for exam `Empirische Methoden’



ble 5.1 Standard Scores and Percentiles for a Normal Distribution
umulative values from the *left*)

tandard 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

note: these are percentages!

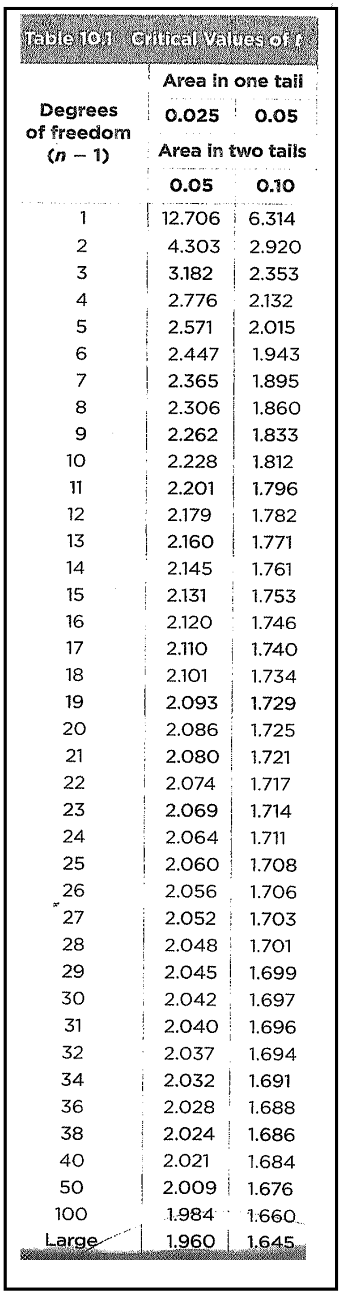


Table 10.1 Critical Values of *t*

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

number of degrees graden of freedom, df

$t_{df;\,0.975}$ quantile

$t_{df;\,0.95}$ quantile

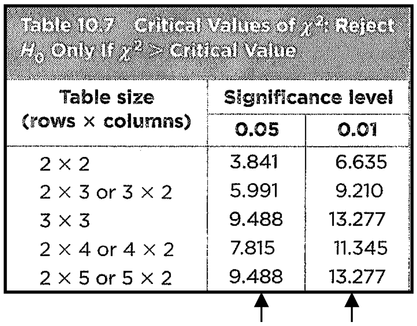


Table 10.7 Critical Values of χ^2: Reject H_0 Only If χ^2 > Critical Value

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 chi-square distribution (the critical values)