Exam Empirische Methoden

VU University Amsterdam, Faculty of Exact Sciences

December 17, 2013

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

1. For the following situations identify which of the following applies: simple random sample, systematic sample, convenience sample, stratified sample, or cluster sample. In each case, state whether you think the procedure is likely to yield a representative sample or a biased sample, and briefly explain why.

   a) *People* magazine chooses its "best dressed celebrities" by compiling responses from readers who mailed the magazine their answers to the questions in a survey that was printed in the magazine.

   b) A marketing expert for MTV is planning a survey in which 500 people will be randomly selected from each age group of 10-19, 20-29, and so on.

   Determine whether the data described in parts c and d are qualitative or quantitative and give their level of measurement. Indicate also which type of visualization is most suited for these data and why.

   c) A question in a survey has five possible answers, 1, 2, 3, 4, and 5, which stand for very unhappy, unhappy, neutral, happy, and very happy, respectively. The data consist of the answers to this question of 150 people.

   d) With carbon dating, the ages (in years) of 78 specimens of wood were determined.

2. *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.*

   An allergy drug is tested by giving 120 people the drug, 100 people a placebo, and 80 people no treatment. Of the three groups 65, 42 and 31 people, respectively, showed improvements. What is the probability that

   a) a randomly selected person in the study was given the drug or improved?

   b) a randomly selected person in the study either improved or did not improve?

   c) at least one of three randomly selected people in the study was given the drug and improved?

   d) given that a randomly selected person in the study improved, he/she was given the drug?

3. In a photographic process, the developing time of prints (in seconds) may be assumed to be a normally distributed random variable with mean $\mu = 16.28$ and standard deviation $\sigma = 0.12$.

   a) What is the probability that it will take anywhere from 16.00 to 16.50 seconds to develop one of the prints?

   b) What is the probability that the mean developing time of 16 randomly selected prints is smaller than 16.25 seconds?

   For a second photographic process, the developing time of prints is a normally distributed random variable with unknown mean μ and unknown standard deviation σ. Suppose that the mean developing time of a sample of 16 randomly selected prints with this process is $\bar{x} = 16.50$ seconds, and the sample standard deviation $s = 0.10$ seconds.

   c) What is the interpretation of a 95% confidence interval for an unknown population mean μ?

   d) Give a 95% confidence interval for the unknown population mean μ of the second photographic process based on the sample of 16 developing times.

   e) Based on your result of part d, do you think that the mean developing time of prints in the second photographic process equals 16.28 seconds? Why (not)?

4. The Organization for Economic Cooperation and Development (OECD) summarizes data on labor-force participation rates. Independent samples were taken of 300 U.S. women and 250 Canadian women. Of the U.S. women, 215 were found to be in the labor force; of the Canadian women, 186 were found to be in the labor force. Let p_1 and p_2 denote the proportion of women who participate in the labor force in the U.S. and Canada, respectively. Some characteristics of the two samples, that you may or may not use, are: the pooled sample fraction is $\bar{p} = 0.729$; $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = 0.0366$; $\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2} = 0.0368$.

   a) Give based on the data a (point) estimate for the difference between the proportions of women who participate in the labor force in the U.S. and of those in Canada.

   b) Investigate the claim that the labor-force participation proportion of U.S. women is smaller than that of Canadian women with a suitable test: formulate H_0 and H_A in terms of the population parameters of interest, give the expression of the test statistic and its distribution under H_0, compute the observed value of the test statistic, and perform the test. Take significance level $\alpha = 10\%$.

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

5. In each province a number of randomly selected people were asked whether or not they think that the appearance of Zwarte Piet should change. The results for Limburg, Groningen, and Noord-Holland are given in the following table.

	change	no change	total
Limburg	3	147	150
Groningen	6	194	200
Noord-Holland	22	228	250
total	31	569	600

   a) Use the table to give, for each of the three provinces separately and under the assumption that there is no relationship between the variables 'province' and 'change', the expected number of people in the sample from that province who think that the appearance of Zwarte Piet should change.

   b) Suppose that we wish to investigate with a chi-square test whether or not there is a relationship between the variables 'province' and 'change'. Formulate suitable H_0 and H_a, specify the test statistic (also tell what the symbols that you use in the formula that you give for the test statistic, stand for), and its distribution under H_0. *(You do not need to compute the observed value of the test statistic.)*

   c) The observed value of the test statistic for these data is 11.72. What would be the conclusion of the test that you described in part b for significance level 1%? Motivate your answer.

   d) The test that you described in part b should only be used under a condition on the sample. What is this condition and is it satisfied in this case?

6. In Figure 1 a scatter plot and the 'best-fit' line (the regression line) of 30 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:
$\bar{x} = 78.00$, $\bar{y} = 60.57$, $s_x = 5.83$, $s_y = 10.17$, $r = -0.54$, $\sqrt{(1-r^2)/(n-2)} = 0.16$, $\hat{b}_0 = 134.53$, $\hat{b}_1 = -0.95$, $s_{\hat{b}_0} = 21.65$, $s_{\hat{b}_1} = 0.28$.

   a) Using significance level 5%, test the claim that the population correlation coefficient ρ equals 0. (As always, formulate the relevant H_0, H_a, give a formula for the test statistic and specify its distribution under H_0, and perform the test.)

   b) In view of the scatter plot, the data characteristics and your conclusion in part a: do you judge that the linear regression model is an appropriate model for these data? Motivate your answer.

   c) What is a normal QQ-plot and what can it tell us?

   d) What does the QQ-plot in Figure 1 tell us, and in which sense is this relevant for the conclusions of a regression analysis?

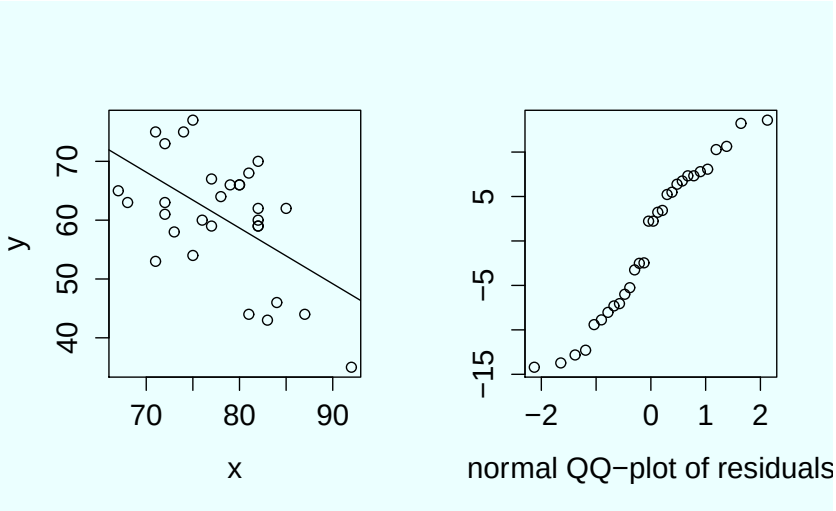

Figure 1: Scatter plot of x and y with linear regression line and normal QQ-plot of the residuals of linear regression of y on x.

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:

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

(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'

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

note: these are percentages!

Table 10.1 Critical Values of *t*

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

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 $\chi^2 >$ 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 chi-square distribution (the critical values)