



EXAM – BASIC STATISTICS FOR ECONOMISTS
2018-05-29

Time: 15.00 - 20.00 (3PM – 8PM)
Approved aid: Hand-held calculator with no stored text, data or formulas
Provided aid: *Formula Sheet and Probability Distribution Tables*, returned after the exam

• **Problems 1 – 5: MULTIPLE CHOICE QUESTIONS – max 60 points**

- A total of 12 multiple choice questions with five alternative answers per question one of which is the correct answer. Mark your answers on the attached **answer form**.
- Marking more than one alternative will result in zero points for that question.
- Written solutions should not be submitted; only your answers on the answer form will be considered in the assessment and final grading.

• **Problems 6 – 7: COMPLETE WRITTEN SOLUTIONS – max 40 points**

- Use only the provided **answer sheets** when submitting your solutions and answers.
- Answers in English, Swedish or a mix of both are allowed.
- For full marks, clear, comprehensive and well-motivated solutions are required. Unclear and unexplained solutions may result in point deductions even if the final answer is correct.
- Check your calculations and solutions before submitting. Careless mistakes may result in unnecessary point deductions.

- The maximum number of points is stated for each question. The maximum total number of points is $60 + 40 = 100$. At least 50 points is required to pass (grades A-E). The grading scale is as follows:

- A: 90 – 100 points
- B: 80 – 89 points
- C: 70 – 79 points
- D: 60 – 69 points
- E: 50 – 59 points
- Fx: 40 – 49 points
- F: 0 – 40 points

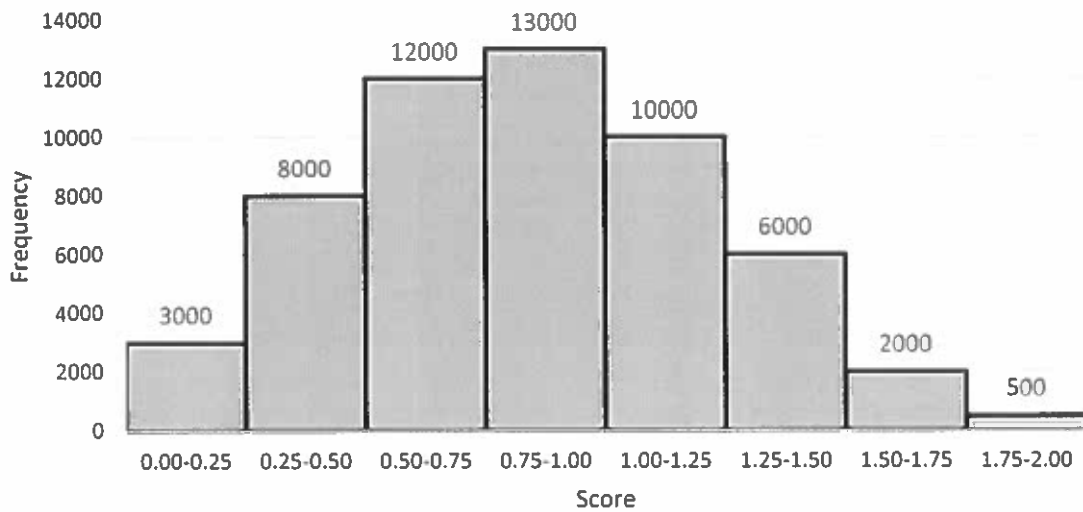
NOTE! Fx and F are failing grades that require re-examination. Students who receive the grade Fx or F cannot supplement for a higher grade.

- Solutions will be posted on Mondo shortly after the exam.

GOOD LUCK!

Problem 1

The histogram below shows the distribution of scores for the Swedish college entry exam (*Högskoleprovet*), autumn 2016. The frequencies have been rounded to make calculations easier, and the frequency of each interval is noted in the diagram.



a) Based on the histogram and the frequencies, determine the intervals which contain Q_1 and Q_3 respectively, i.e. the first and third quartiles. (5p)

- A. Q_1 lies in 0.25-0.50; Q_3 lies in 0.75-1.00
- B. Q_1 lies in 0.25-0.50; Q_3 lies in 1.00-1.25
- C. Q_1 lies in 0.25-0.50; Q_3 lies in 1.25-1.50
- D. Q_1 lies in 0.50-0.75; Q_3 lies in 1.00-1.25
- E. Q_1 lies in 0.50-0.75; Q_3 lies in 1.25-1.50

A sample of size $n=10$ has been drawn and in the table below we have the joint frequency distribution for two variables x and y :

Frequency	$y = 0$	$y = 1$
$x = 0$	5	1
$x = 1$	0	4

b) What is the covariance between x and y ? (5p)

- A. 0
- B. -0.222
- C. 0.200
- D. 2.000
- E. 0.222

Problem 2

A small business has $n = 24$ employees. The manager is sending a pair of employees, i.e. $k = 2$, to represent the company at a large business conference.

a) If selection order does not matter, how many possible pairs of employees can the manager choose? Note that a pair consists of two different employees. (4p)

- A. 132
- B. 276
- C. 521
- D. 552
- E. 576

The joint probability distribution of two random variables X and Y is given in the table below:

$P(x, y)$	$Y = 0$	$Y = 100$	$Y = 200$
$X = 0$	0.2	0.2	0.1
$X = 1$	0.1	0.3	0.1

b) Calculate the conditional mean of Y given that $X = 0$, i.e. calculate $\mu_{Y|X=0} = E[Y | X = 0]$. (6p)

- A. $\mu_{Y|X=0} = 40$
- B. $\mu_{Y|X=0} = 80$
- C. $\mu_{Y|X=0} = 90$
- D. $\mu_{Y|X=0} = 100$
- E. $\mu_{Y|X=0} = 140$

Hint: Determine the conditional distribution of Y given $X = 0$.

Problem 3

Suppose that 5% of adult residents of Malmö were victims of theft during 2016. The Council of Crime Prevention (*BRÅ*) decides to ask $n = 17$ randomly chosen adult residents of Malmö the question “were you the victim of theft during 2016?” Assume that their answers are accurate and independent of each other.

a) What is the probability that exactly 3 of the 17 selected residents answer “yes”? (5p)

- A. 0,850
- B. 0,008
- C. 0,041
- D. 0,158
- E. 0,991

Caligula and Nero work at the same office. They are often late, so the HR department has developed a stochastic model to describe their tardiness, i.e. the number of minutes each of them are late are considered to be random variables. Let X denote the number of minutes Caligula is late, and Y the number of minutes Nero is late, on any given day. For simplicity it is assumed that

$$X \sim N(5; 5^2) \quad \text{and} \quad Y \sim N(10; 5^2)$$

Since both of them walk to work, and from opposite parts of town, their arrival times are assumed to be independent of each other. A negative number means the corresponding person is early.

b) What is the probability that both Caligula and Nero are late on a given day? (5p)

- A. 0
- B. 0,038
- C. 0,380
- D. 0,708
- E. 0,822

Hint: How do you calculate $P(X > 0 \cap Y > 0)$?

NOTE: The alternatives for a) – b) have been rounded to 3 digits.

Problem 4

An opinion poll is being planned and one of the questions concerns a proposed change of unemployment benefits. Respondents may respond “yes” or “no” to the question whether they think the new proposal is good or not.

Your task is to estimate P = the proportion of “yes” answers for the whole population with \hat{p} = the sample proportion of “yes” answers and a 95% confidence interval for P which is calculated as

$$\hat{p} \pm \text{margin of error}$$

a) How large should the sample size n be if the margin of error should be smaller than 0,05? (6p)

- A. 196
- B. 335
- C. 385
- D. 769
- E. 1537

HINT: Since you do not know the true proportion P in the population nor what the sample proportion \hat{p} will be you'll have to choose some value for the proportion in order to determine the required sample size. Which value for \hat{p} should you use in order to guarantee a margin of error smaller than 0,05?

The researchers eventually settle for a sample size of $n = 400$ and it is determined that 132 answer “yes” to the question.

b) Which of the following is a 95% confidence interval for P ? (5p)

- A. 0.33 ± 0.046
- B. 0.33 ± 0.080
- C. 0.33 ± 0.039
- D. 0.33 ± 0.151
- E. 0.67 ± 0.046

In statistical inference we use different terms such as bias, Type I and II errors, p -value, confidence and significance, power and so on.

c) One of the following statements is false, which? (4p)

- A. A 95 % confidence interval contains 95 % of the observed values in a given sample.
- B. Type II error means to wrongfully accept H_0 when H_0 is false.
- C. The power of a test is the probability to reject a false null hypothesis.
- D. If the bias of an estimator is zero, then the expected value of the estimator is equal to the true value of the parameter being estimated.
- E. The p -value is the lowest significance level at which you reject H_0 for a given sample.

Problem 5.

A survey of $n = 200$ statistics students at the University of Adelaide included a question about the respondents' smoking habits. The answers were divided into 3 categories of smoking habits and across sexes, female and male. The results of the survey are presented in the table below.

	Daily Smoker	Occasional Smoker	Never Smokes	Sum
Female	10	5	85	100
Male	15	10	75	100
Sum	25	15	160	200

Suppose that you want to perform a χ^2 -test to determine whether a person's smoking habits is dependent or independent of the person's sex using a significance level of $\alpha = 5\%$.

a) What is the critical value for this test? (4p)

- A. $\chi_{\text{crit}}^2 = 1.960$
- B. $\chi_{\text{crit}}^2 = 3.841$
- C. $\chi_{\text{crit}}^2 = 5.991$
- D. $\chi_{\text{crit}}^2 = 11.070$
- E. $\chi_{\text{crit}}^2 = 233.99$

b) What is the observed value of the test statistic? (6p)

- A. $\chi_{\text{obs}}^2 = 3.29$
- F. $\chi_{\text{obs}}^2 = 3.54$
- G. $\chi_{\text{obs}}^2 = 3.75$
- H. $\chi_{\text{obs}}^2 = 4.02$
- I. $\chi_{\text{obs}}^2 = 4.19$

c) Which of the following conclusions is correct? (5p)

- A. Reject H_0 , smoking habits and sex are dependent.
- B. Reject H_0 , smoking habits and sex are independent.
- C. Do not reject H_0 , smoking habits and sex are dependent.
- D. Do not reject H_0 , smoking habits and sex are independent.
- E. We can neither accept nor reject H_0 , there is no evidence for either decision.

Complete written solutions are required for Problems 6 and 7.

Use separate answer sheets for 6 and 7 respectively.

Problem 6

In 2015, Statistics Sweden (SCB) conducted a survey of the yearly housing costs of people living alone without children in their own condominiums (*bostadsrätt*). The results for Stockholm and Gothenburg are given in the table below where the cost per year is given in thousands of SEK:

	Average cost/year	Sample variance	Sample size
Stockholm	62,4	7569	1745
Gothenburg	57,0	3364	575

- Calculate a 95% confidence interval for the average cost/year in Stockholm. State your assumptions. (6p)
- Test at a 5% significance level whether the average cost of housing was higher in Stockholm than in Gothenburg. State your assumptions, hypotheses, test statistics, decision rule and critical value, calculations and your conclusions. (10p)
- Briefly explain what the significance level α (alpha) of a test is. Answer briefly, you should be able to explain this on half a page at most. (4p)

Problem 7

A chain of Italian themed restaurants called “Alba Longa” has many locations across the United States. The owners carry out an experiment where they change advertising expenditures in $n = 8$ randomly selected regions. The table on the following page shows the percentage increase in advertising and the percentage increase in sales, for the eight sampled regions.

A business analyst proposes the following simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

where Y_i denotes the percentage increase in sales, and X_i denotes the percentage increase in advertising for region i . On the following page you'll find an Excel-output for the estimated regression model with some numbers missing.

- Is there a linear relationship between X and Y , i.e. is the slope equal to 0 or not? Test this at the 5% significance level. Formulate your hypotheses, test statistic, critical value and decision rule, calculations and conclusions. You do not need to state any assumptions. (8p)
- The owners want to estimate the expected change in sales given an increase in advertising expenditure by 10 percent. Calculate a 95% confidence interval for the mean percentage change in sales Y given $X = 10$. (8p)
- Explain briefly what multicollinearity is. What are the potential effects of multicollinearity? Answer briefly, you should be able to explain this on half a page at most. (4p)

DATA

Region, <i>i</i>	1	2	3	4	5	6	7	8
Increase, ad expenditure (%), <i>x</i>	0	2	7	5	5	4	9	8
Increase, sales (%), <i>y</i>	2.4	7.5	11.4	9.2	11.0	4.4	7.8	11.5

ESTIMATED REGRESSION MODEL

Regression Statistics

Multipel-R	0,69897	
R Square	0,48856	
Adjusted R Square	0,40332	
Standard Error	2,58972	← NOTE: this is the standard deviation of the residuals
Observations	8	

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>
Regression	1	38,44			
Residual	6	40,24			
Total	7	78,68			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	4,275	1,85960	2,29888	0,06119
X	0,775			



Correction sheet

Date: 29/05/2018

Room: Brunnsvikssalen

Course: Basic statistics for economists (eng)

Exam: Statistics for economists (eng)

*Sic itur
ad astra.
/me*

Anonymous code: 0061-SZK

I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

	1	2	3	4	5	6	7	8	9	Total number of pages
	X	X	X	X	X	X	X			5 IF
Teacher's notes	10	10	10	15	15	20	20			

Points	Grade	Teacher's sign.
100	A	ME

ANSWER FORM Exam – Basic statistics for economists
2018-05-29

Room: Brunnsvikssalen

Anonymous code: 0061-SZK (write clearly!)

Mark your answers with a clear cross (X) in the corresponding boxes below.

NOTE! Only one cross per question. If more than one alternative has been marked, zero points will be awarded for that question.

NOTE! If, after checking your calculations properly, you are convinced that the correct answer is not included among the given alternatives, type your answer in the margin to the right.

		A	B	C	D	E
Problem 1	a)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	b)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Problem 2	a)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	b)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Problem 3	a)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	c)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Problem 4	a)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	b)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	c)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Problem 5	a)	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	b)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	c)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

60/60



Problem 6

We denote Stockholm to X and Gothenburg to Y } This stands through the whole problem

What we know [From the problem]

$n_x = 1745$ $\bar{x} = 62,4$ (cost/year) $S_x^2 = 7569$ $S_x = \sqrt{S_x^2} = 87$
 $n_y = 575$ $\bar{y} = 57,0$ (cost/year) $S_y^2 = 3364$ $S_y = \sqrt{S_y^2} = 58$

a) Calculate a 95% CI for \bar{X} .

Assumptions

All observations are independently and identically distributed (iid). The population variance is unknown \Rightarrow We use sample variance. The sample is large ($n = 1745 \geq 30$) which means that CLT applies $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ☺

The formula we use for the 95% CI is

$$\bar{X} \pm Z_{\alpha/2} \frac{S_x}{\sqrt{n}}$$
 R

with $\alpha/2 = 0,05/2 = 0,025$

95% CI $62,4 \pm Z_{0,025} \cdot \frac{87}{\sqrt{1745}}$ [table 2]

$\Rightarrow 62,4 \pm 1,96 \cdot 2,082675159$


$\Rightarrow 62,4 \pm 4,082043311$ R

The 95% CI is $[58,32; 66,48]$ ← Answer R

6

b) We are testing at a 5% sign level whether the average cost of housing was higher in Stockholm than in Gothenburg.

Assumptions

All observations are iid. We assume it's two independent samples with unknown variances. Large samples
→ CLT 

① Hypothesis:

$$H_0: \mu_x - \mu_y = 0$$

$$H_1: \mu_x - \mu_y > 0 \quad \mathcal{R}$$

Stockholm: μ_x

Gothenburg: μ_y

② Test statistic:

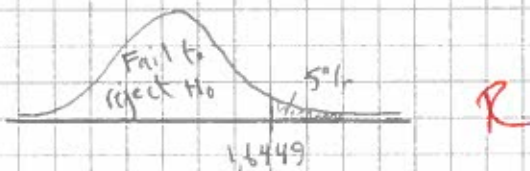
$$Z_{obs} = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{S_x^2/n_x + S_y^2/n_y}} \sim N(0,1) \quad \mathcal{R}$$

③ Critical value:

one-sided with $\alpha = 0,05$

$$Z_{crit} = Z_\alpha = Z_{0,05} = 1,6449 \quad [\text{table 2}] \quad \mathcal{R}$$

④ Draw/Decision rule:



Decision rule:

$$\text{Reject } H_0 \text{ if } Z_{obs} > Z_{crit} \Rightarrow \text{if } Z_{obs} > 1,6449 \quad \mathcal{R}$$

(5) Calculate:

$$Z_{obs} = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{S_x^2/n_x + S_y^2/n_y}} = \frac{624 - 570 - 0}{\sqrt{\frac{7569}{1745} + \frac{3364}{575}}} = 1.691803497 \approx 1.69$$

(6) Conclusion:

We reject H_0 at a 5% significance level since $Z_{obs} = 1.69 > 1.6449$ and conclude that the average cost of housing was higher in Stockholm. 😊

(7) The significance level α (alpha) is the "error term" of a test. Since we don't know the true parameter we need to have a small % of uncertainty to our tests and this is where α comes in. It's the chance that we reject H_0 when H_0 actually is true. This is also known as the type I error.

	H_0 True	H_0 False
Accept	OK!	Type II error
Reject	Type I error (α)	OK! (error)



4

10

Problem 7

Regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

Y_i : the percentage increase in sales (Dependent variable)

X_i : the percentage increase in advertising for region i (Independent variable)

Some things we know [from the table and regression output]

$b_0 = 4,275$ (the estimate of β_0 , intercept)

$b_1 = 0,775$ (the estimate of β_1 , the slope)

$n = 8$

$SSR = 38,44$

$SSE = 40,24$

$SST = 78,68$

a) We want to test if the slope is 0 or not

We conduct a test at the 5% significance level for this

(1) Hypothesis

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

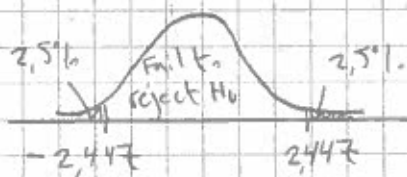
(2) Test statistic $\beta_1^0 =$ under the null hypothesis $= 0$

$$|t_{obs}| = \frac{b_1 - \beta_1^0}{s_{b_1}} \sim t_{n-k-1}$$

- (3) Critical value
 double-sided with $\alpha/2 = 0,05/2 = 0,025$
 and $k=1$ (independent variables)

$$t_{crit} = t_{n-k-1; \alpha/2} = t_{8-1-1; 0,025} = t_{6; 0,025} = 2,447 \text{ [table 3]}$$

- (4) Draw/Decision rule



Decision rule:

$$\text{Reject } H_0 \text{ if } |t_{obs}| > t_{crit} \Rightarrow \text{if } |t_{obs}| > 2,447$$

- (5) Calculate

The formula is

$$|t_{obs}| = \frac{b_1 - \beta_1}{s_{b_1}}$$

So we need s_{b_1}

$$s_{b_1}^2 = \frac{s_e^2}{(n-1) s_x^2}$$

We get s_e^2 and s_x^2 with the formulas.

$$s_e^2 = \frac{SSE}{n-k-1} = \frac{40,24}{8-1-1} = 6,70666667$$

[Data table]

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{40}{8} = 5$$

$$s_x^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{0^2 + 2^2 + 7^2 + 5^2 + 5^2 + 4^2 + 7^2 + 9^2 - 8 \cdot 5^2}{8-1} = 9,142857143$$

Now we can calculate $S_{b_1}^2$

$$S_{b_1}^2 = \frac{S_e^2}{(n-1)S_x^2} = \frac{6,706666667}{(8-1) \cdot 9,142857143} = 0,1047916667$$

$$S_{b_1} = \sqrt{S_{b_1}^2} = \sqrt{0,1047916667} = 0,32371541$$

Now we have everything for $|t_{obs}|$

$$|t_{obs}| = \frac{b_1 - \beta_0^*}{S_{b_1}} = \frac{0,775 - 0}{0,32371541} = 2,394078181$$

⑥ Conclusions.

Since $|t_{obs}| = 2,394 < 2,447$ we fail to reject the null hypothesis and conclude that the slope is not significant from 0 on a 5% significance level. i.e. there is no linear relationship between X and Y

b) Calculate a 95% confidence interval for the mean percentage change in sales Y given X=10, $\alpha=0,05$

The formula we are going to use is

$$(b_0 + b_1 X) \pm t_{n-2, \alpha/2} \sqrt{S_e^2 \left(\frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)S_x^2} \right)}$$

We have everything we need from the problem above. Now we just plug it in the formula:

$$(4,275 + 0,775 \cdot 10) \pm t_{8-2, 0,05/2} \sqrt{6,706666667 \left(\frac{1}{8} + \frac{(10-5)^2}{(8-1) \cdot 9,142857143} \right)}$$

[table 3]

$$12,025 \pm 2,447 \cdot 1,859603452$$

$$12,025 \pm 4,556449648$$

95% confidence interval is.

$$[7,475; 16,575]$$

C) Multicollinearity is when you have two or more independent variables that have a strong correlation between each other in your model. This can lead to a very misleading model even if your R^2 gives you a high number close to 1 and tells you that the variability in Y can be explained by the variation in X . In other words, you can't trust your model to approximate future \hat{y} very well.