Stockholm University

STOCKHOLM UNIVERSITY
Department of Statistics
Spring 2019, period C-D

Michael Carlson (examiner)
Ulf Högnäs

EXAM – BASIC STATISTICS FOR ECONOMISTS

2019-08-15

Time:	19.00 - 15.00 (10AM – 3PM)
Approved aid:	Hand-held calculator with no stored text, data or formulas
Provided aid:	*Formula Sheet and Probability Distribution Tables*, returned after the exam, English-Swedish dictionaries available on-site

- **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 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.
  - 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

Transport Analysis (*Trafikanalys*) is a Swedish government agency charged with providing policy advice in the general sphere of transportation. The agency is also responsible for the production of official statistics in the transport and communication sectors, including commodity flows (*varuflöden*). The following table is a simplified version of an official table describing commodity flows of shipments in and out of Sweden (i.e. import and export) for the year 2016 by mode of transport in terms of weight and value:

	Incoming (import)		***Outgoing (export)***	
Mode of transport	*Weight 1 000 tons*	*Value SEK million*	*Weight 1 000 tons*	*Value SEK million*
Road	5 092	114 305	11 512	222 049
Railway	701	6 310	1 708	14 055
Sea	39 609	140 778	26 248	125 752
Air	12	5 021	14	6 390
Combination or unknown	11 827	275 691	44 248	468 318
Total	**57 241**	**542 106**	**83 731**	**836 564**

Source: https://www.trafa.se/en/travel-survey/commodity-flows/; date 2019-08-13.

a) Which of the following statements is not correct or cannot be derived from the table? (5p)

A. Weight is a numerical variable measured on a ratio scale.

B. In total, the value of the outgoing flow is approximately 54.3% larger than the incoming.

C. The average value of incoming per ton by air is larger than by sea.

D. The average weight of outgoing per shipment is larger by road than by railway.

E. Mode of transport is a categorical variable measured on a nominal scale.

For a certain company, a sample of outgoing shipments by air is collected. Define x = weight in kilograms and y = value in 1000 SEK for each shipment. You obtain the following $n = 5$ values on x and y:

x:	30	40	60	110	160
y:	3	4	5	6	7

b) Calculate the sample covariance s_{xy} and sample correlation coefficient r_{xy} between x and y. (5p)

A. $s_{xy} = 82.5$ $\quad r_{xy} = 0.961$

B. $s_{xy} = 82.5$ $\quad r_{xy} = 1.04$

C. $s_{xy} = 0$ $\quad r_{xy} = 0$

D. $s_{xy} = 582.5$ $\quad r_{xy} = 0.958$

E. $s_{xy} = 582.5$ $r_{xy} = -0.958$

Problem 2

A copier service company has identified three types of errors that cause some kind of action on their behalf and have also determined the probabilities of these events occurring for a randomly selected customer during a given month:

A = "software related"	$P(A) = 0{,}05$
B = "mechanical failure"	$P(B) = 0{,}10$
C = "user error (no real error)"	$P(C) = 0{,}20$

It is assumed that A is independent of both B and C. It is also assumed that B and C are disjoint events.

a) Which of the following statements is **false**? (5p)

A. $P(A \cap B) = 0.005$

B. $P(A \cup C) = 0.24$

C. $P(C|A) = 0.20$

D. $P(B \cap C) = 0$

E. B and C are independent

The service company also collects data on the number of errors that are fixed per customer visit and the following probability distribution has been compiled for the random variable X = number of errors:

x	0	1	2	3
$P(X = x)$	0,1	0,5	0,3	0,1

b) Calculate the expected value and the standard deviation of X. (5p)

A. $\mu = 1{,}5$ $\sigma = 0.64$

B. $\mu = 1{,}4$ $\sigma = 0{,}80$

C. $\mu = 1{,}5$ $\sigma = 1.29$

D. $\mu = 1{,}4$ $\sigma = 0.64$

E. $\mu = 1{,}0$ $\sigma = 0.80$

Problem 3

The weight of a randomly chosen apple from a large batch of Gala apples is assumed to be normally distributed with a mean of 200 grams and a standard deviation of 40 grams. Anna randomly chooses $n = 9$ apples from this batch at a grocery store. We assume that the individual weight of each apple is independent of the others.

a) What is the probability that the **total weight** of her 9 apples is **less** than 2 kg? (5p)

A. 0.500
B. 0.506
C. 0.711
D. 0.952
E. 1.000

b) What is the probability that **at most** 5 of her $n = 9$ apples **individually** weigh more than 200 grams? HINT: First you need to determine the probability P that an individual apple weighs more than 200 grams. (5p)

A. 0.254
B. 0.500
C. 0.746
D. 0.914
E. 0.246

Anna collects an even larger batch of apples and increases the number of apples to $n = 100$.

c) Using the approximation technique described in the course literature, what is the probability that **more than** half of the apples weigh more than 200 grams? (5p)

A. 0.254
B. 0.500
C. 0.691
D. 0.994
E. 0.520

NOTE: The numbers in a) – c) above have been rounded to three decimals, choose the closest value.

Problem 4

In a sample of $n = 162$ companies, 100 companies reported profits that were better than they had predicted, 28 companies matched their predictions and 34 fell short of (i.e. were less than) their prior predictions. Let P denote the proportion of companies in the entire population that fell short.

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

A. $(14.7\% ; 27.3\%)$

B. $(20.8\% ; 21.2\%)$

C. $(27.7\% ; 40.3\%)$

D. $(15.7\% ; 26.3\%)$

E. $(12.3\% ; 29.7\%)$

b) Given a confidence level of 95% and assuming $P = 0.25$, what is the smallest sample size required if the margin of error is to be less than 5%? (5p)

A. $n = 12$

B. $n = 147$

C. $n = 289$

D. $n = 666$

E. $n = 1153$

Problem 5

In 1998, the medical academic journal The Lancet published a study of alcohol consumption among English medical students in residency. The table below presents consumption of alcohol, measured in grams per week, for a sample of men and women respectively. You are asked to test at the 5% level whether there was a difference in consumption between the average for men (μ_M) and the average for women (μ_F), i.e. use the following hypotheses:

H_0: there is no difference between male and female students

H_1: there is a difference between male and female students

Sex	*n*	*Mean (grams)*	*Standard deviation (grams)*
Men	39	230	160
Women	51	160	108

a) Which of the following is the correct decision rule for this test? (5p)

A. Reject the null if $z_{obs} > 1{,}96$

B. Reject the null if $|z_{obs}| > 1{,}96$

C. Reject the alternative if $z_{obs} > 1{,}96$

D. Reject the alternative if $|z_{obs}| > 1{,}96$

E. Since the distribution of alcohol consumption is unknown, we cannot conduct this test.

b) What is the observed value of the test variable and conclusion of the test?

A. 0.012; drinking habits differ between men and women.

B. 1.720; drinking habits do not differ between men and women.

C. 1.720; drinking habits differ between men and women.

D. 2.353; drinking habits do not differ between men and women.

E. 2.353; drinking habits differ between men and women.

c) Relating to the problem above, which of the following is a <u>false</u> or irrelevant assumption? (5p)

A. The selection of students are assumed to be mutually independent of each other within each sample.

B. The population variances of drinking habits for both men and women is unknown but we approximate our calculations by using the sample variances.

C. The two samples are assumed to independent of each other.

D. We rely on the Central Limit Theorem (CLT) that states that a sample mean is approximately normal distributed if the sample size is sufficiently large.

E. The variances for men and women are assumed to be equal and we estimate this variance using a pooled variance estimate.

Complete written solutions are required for Problems 6 and 7.

Use separate answer sheets for 6 and 7 respectively.

Problem 6

The price-earnings ratio, or P/E ratio, is the ratio of a company's share (stock) price to the company's earnings per share. A sample of companies in the consumer products industry and a sample of companies in the banking industry were collected. The respective P/E ratios were categorized into size classes. The number of companies (i.e. the frequencies) for each combination of industry type and P/E ratio are shown in the cross-tabulation on the following page. You are asked to determine if there is a relationship between industry type and P/E ratio.

a) Test on the 5% level whether industry type and P/E ratio are independent or not. State your assumptions, your hypotheses, the test variable, the decision rule and critical value. (8p).

b) Calculate the test statistic, draw conclusions and interpret your results in words. (6p)

c) Briefly explain the terms Type I and Type II errors and how they relate to the significance level of a test. Your answer doesn't need to be very long, ½ page should suffice. (6p)

Problem 7

The Beta coefficient for a stock is defined as the slope coefficient obtained from a simple linear regression model with the monthly percentage return of a *stock index* as independent variable (X) and the monthly percentage return of the stock itself as dependent variable (Y):

$$\text{Model:} \qquad Y = \beta_0 + \beta_1 X + \varepsilon$$

The Beta coefficient β_1 is a measure of how responsive the return of the stock is to the movements of the market as a whole (see e.g. NCT p. 456). Based on $n = 36$ months of data, a finance professor calculated the Beta coefficient for The Coca Cola Company against the S&P 500 and the results of the analysis is provided on the following page.

a) Test at the 5% level of significance whether the Beta coefficient is **greater than** zero. Formulate your hypotheses, test statistic, critical value and decision rule, calculations and conclusions. You do not need to state any assumptions. (8p)

b) Calculate a 95% confidence interval for the **mean** monthly percentage return of Coca Cola (Y) given that the monthly percentage return of S&P 500 (X) is 1% that same month. (8p)

c) Briefly explain the term heteroscedasticity. Is heteroscedasticity one of the assumptions of the linear regression model? (4p)

DATA for Problem 6

	P/E ratio					
Industry	5-9	10-14	15-19	20-24	25-29	***Total***
Consumer	4	10	18	10	8	50
Banking	14	14	12	6	4	50
Total	18	24	30	16	12	100

DATA for Problem 7

	Sample mean	Sample variance
Monthly percentage return S&P 500 (X)	$\bar{x} = 0.73\%$	$s_x^2 = 0.13\%$
Monthly percentage return Coca Cola (Y)	$\bar{y} = 0.57\%$	$s_y^2 = 0.12\%$

Regression Statistics	
Multiple-R	0.31932
R Square	0.10196
Adjusted R Square	0.07555
Standard Error	0.03399
Observations	36

← NOTE: this is the standard deviation of the residuals

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	0.00446	0.00446	3.86042
Residual	34	0.03928	0.00116	
Total	35	0.04373		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.00349	0.00578	0.60332	0.55030
S&P 500 (X)	0.31048	0.15802	1.96480	0.05765

Department of Statistics

Stockholms universitet

Correction sheet

Date: 2019-08-15

Room: Värtasalen

Exam: Statistics for Economists

Course: Basic Statistics for Economists

Anonymous code: 0035-F2P

☒ 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			3
Teacher's notes 10	5	15	10	15	16	16			

Points	Grade	Teacher's sign.
87	B	ME

STOCKHOLM UNIVERSITY
Department of Statistics

Spring 2019 C-D

ANSWER FORM Exam – Basic statistics for economists
2019-08-15

Room: VÄ

Anonymous code: 0035-FZP *(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, write your answer in the margin to the right and explain you reasoning on the back.

		A	B	C	D	E	
Problem 1	a)	☐	☐	☐	☒	☐	R
	b)	☒	☐	☐	☐	☐	R
Problem 2	a)	☐	☐	☐	☐	☒	R
	b)	☐	☐	☐	☒	☐	–
Problem 3	a)	☐	☐	☐	☒	☐	R
	b)	☐	☐	☒	☐	☐	R
	c)	☐	☒	☐	☐	☐	R
Problem 4	a)	☒	☐	☐	☐	☐	R
	b)	☐	☐	☒	☐	☐	R
Problem 5	a)	☐	☒	☐	☐	☐	R
	b)	☐	☐	☐	☐	☒	R
	c)	☐	☐	☐	☐	☒	R

55

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: 0035-FZP Sheet number: _____

Problem 6

(independent and identically distributed)

a) Assumptions: The samples are randomly collected and iid ✓

Hypotheses: $H_0: P(C) = P(B) = 0,5$ vs $H_1: P(C) \neq P(B)$ ✓

Test variable: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ where $E_{ij} = \frac{R_i C_j}{n}$ ✓

Critical value: 5% significance level; $v = (R-1)\cdot(C-1) = 4$

$\rightarrow \chi^2_{4;0,05} \rightarrow \chi^2_{crit} = 9,488$ ✓

Decision rule: Reject H_0 if $\chi^2_{obs} > \chi^2_{crit} = 9,488$ ✓

5/6

b)

Calculating E_{ij}:

	P/E ratios					
Industry	5-9	10-14	15-19	20-24	25-29	Total
Consumer	$50 \cdot 18/100 = 9$	$50 \cdot 24/100 = 12$	$50 \cdot 30/100 = 15$	$50 \cdot 16/100 = 8$	$50 \cdot 12/100 = 6$	50
Banking	$50 \cdot 18/100 = 9$	$50 \cdot 24/100 = 12$	$50 \cdot 30/100 = 15$	$50 \cdot 16/100 = 8$	$50 \cdot 12/100 = 6$	50
Total	18	24	30	16	12	100

Calculating χ^2_{obs}:

$$\frac{(4-9)^2}{9} + \frac{(10-12)^2}{12} + \frac{(18-15)^2}{15} + \frac{(10-8)^2}{8} + \frac{(8-6)^2}{6} + \frac{(14-9)^2}{9} + \frac{(14-12)^2}{12}$$

$$+ \frac{(12-15)^2}{15} + \frac{(6-8)^2}{8} + \frac{(4-6)^2}{6} \approx 9,7556$$ ✓

Conclution: $\chi^2_{obs} = 9,7556 > \chi^2_{crit} = 9,488 \rightarrow H_0$ is rejected

On a 5% level, The P/E ratio and the industry type are significally dependent on each other. ✓

(However, conclusion doesn't relate to your hypotheses in a)

6/6

g) Type I and Type II errors are when you incorrectly reject / do not reject the Ho Hypotheses.

	Rejected	Not rejected
Shoud be Rejected	:)	Type II error
Shoud not be rejected	Type I error	:)

R

If the significance level is very big, we get a small critical value and we risk rejecting a Ho That should not have been rejected = Type II

If the significance level is very small, we get a big critical value and we risk not being able to reject a Ho that should have been rejected = Type I

$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$

4

$6 + 6 + 4 = 16$

SU, DEPARTMENT OF STATISTICS

Room: VÄ Anonymous code: 0035-FZP Sheet number: _____

Problem 7

a) Hypotheses: $H_0: \beta_1 = 0$ vs $H_1: \beta_1 > 0$ R

Test statistic: $t_{n-k-1} = \frac{b_j - \beta_j^*}{s_{bj}}$ and $S^2_{b_1} = \frac{S^2_E}{(n-1)S^2_X}$ R

Critical value: $\nu = 36-1-1 = 34$; $\alpha = 0,05 \rightarrow t_{34;0,05} \approx 1,69$ R

Decision rule: Reject H_0 if $t_{obs} > t_{crit} \approx 1,69$ R

Calculations:

$S^2_{b_1} = \frac{0,033992^2}{35 \cdot 0,0013} \approx 0,025392$ $S_{b_1} \approx 0,15935$ (see output)

ok

$t_{34} = \frac{0,31048 - 0}{0,15935} \approx 1,9484$ ok

Conclution: $t_{obs} = 1,9484 > t_{crit} \approx 1,69$

H_0 is rejected. The Beta coefficient β_1 is greater than zero on a 5% significance level. R

/8

Double check since $\nu = 34$ is not in table 3:

$t_{obs} = 1,9484 > t_{30;0,05} = 1,697 \rightarrow$ Reject H_0

$t_{obs} = 1,9484 > t_{35;0,05} = 1,690 \rightarrow$ Reject H_0 :)

b)

$$(b_0 + b_1 X) \pm t_{n-2;\alpha/2} \sqrt{s_\varepsilon^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2} \right)}$$ R

$t_{34,0,025} \approx 2,03$

$(b_0 + b_1 X) = 0,00349 + 0,31048 \cdot 0,01 = 0,00659948$ R

$$\sqrt{s_\varepsilon^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2} \right)} = \sqrt{0,033992 \left(\frac{1}{36} + \frac{(0,01-0,0073)^2}{(36-1)\,0,0013} \right)} \approx 0,0056813$$ R

$2,03 \cdot 0,0056813 = 0,011533$

$0,00659948 \pm 0,011533$

$-0,0049 \; ; \; 0,0181$

or

$-0,49\% \; ; \; 1,81\%$ /8

c) —

8 8 —

16