Stockholm University

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

Michael Carlson (examiner)
Ulf Högnäs

EXAM – BASIC STATISTICS FOR ECONOMISTS

2019-06-05

Time:	09.00 - 14.00 (9AM – 2PM)
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 sight

- **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 <u>not</u> 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

The U.S. Census Bureau publishes various statistics about the population in the U.S. and one example is the population's poverty status in the past 12 months. A person is characterized as being below the poverty level if he or she belongs to a household were income is below a pre-defined level. Below is an excerpt from this study for the year 2017. The numbers are in thousands and are estimates from an annual community survey.

		Total	Below poverty level	Percent below poverty level
Total	population	313 049	45 650	14.6% (± 0.1)
AGE	< 5 years	19 533	4 390	22.5% (± 0.2)
	5 - 17 years	52 897	10 320	19.5% (± 0.2)
	18 - 34 years	71 085	12 854	18.1% (± 0.1)
	35 - 64 years	123 108	13 769	11.2% (± 0.1)
	> 65 years	46 425	4 317	9.3% (± 0.1)
SEX	Female	159 760	25 242	15.8% (± 0.1)
	Male	153 289	20 409	13.3% (± 0.1)

a) Which of the following statements is false or cannot be deduced from the table? (5p)

A. The combined effect of being young and female strongly increases the risk of poverty.

B. Age is normally treated as a numeric variable on a ratio scale but not so in this table.

C. The number of people below the poverty level is a numerical variable on a ratio scale.

D. Age affects the probability of being below the poverty level.

E. Of those below poverty level, more than 30% are younger than 18 years old.

The Swedish Arts Council allocates grants to various cultural center organizations. The following data are the grants in thousands of SEK awarded to $n = 11$ organizations for 2019 ordered by size:

420 650 1190 1265 1680 2280 2570 3075 3485 3650 4200

b) Using the method for calculating percentiles given in the course literature, what is the interquartile range for these data? (5p)

A. IQR = 2280

B. IQR = 2295

C. IQR = 3000

D. IQR = 1810

E. IQR = 3780

Problem 2

A book publisher analyses sales of new first edition publications and whether or not these new books were revised to a second edition or not. The findings are that 60% of all new books sell less than projected and a 10% sell more than projected, the rest sell close to the projected number. If a new book was revised or not depends on how it sold. The publisher compiles the following table with relative frequencies on how well new books have sold and the proportions within each category that were revised or not:

	Sales		
	Less than projected	Close to projected	More than projected
	0.60	0.30	0.10
Revised	0.20	0.50	0.70
Not revised	0.80	0.50	0.30

a) What is the probability that a randomly chosen new book sold more than projected given that it was revised to a second edition, i.e. what is $P(\text{more than projected}|\text{revised})$? (5p)

A. 0.045

B. 0.238

C. 0.467

D. 0.000

E. 0.206

NOTE: The numbers above have been rounded to three decimals, choose the closest value.

Among the more successful books published, the total number of revisions over the years was counted. In the table below the relative frequencies (probabilities) of books and number of revisions (x) is displayed:

x	0	1	2	3
$P(X = x)$	0.30	0.25	0.30	0.15

b) What is the mean and variance of X = the number of revisions? (5p)

A. $\mu_X = 1.3 \quad \sigma_X^2 = 1.25$

B. $\mu_X = 1.3 \quad \sigma_X^2 = 1.11$

C. $\mu_X = 1.0 \quad \sigma_X^2 = 1.11$

D. $\mu_X = 1.5 \quad \sigma_X^2 = 1.25$

E. $\mu_X = 1.5 \quad \sigma_X^2 = 1.08$

Problem 3

An expert analyst specializing on the automobile market says that 70% of car buyers nowadays use the Internet for research and price comparisons. A sample of $n = 14$ recent car buyers was drawn and the respondents were asked if they used the Internet before purchasing their cars.

a) Assuming that the analyst's statement is true and that the car buyers in the sample are independent of each other, what is the probability that more than half of the 14 respondents used the Internet for research and price comparisons? (5p)

A. 0.969

B. 0.093

C. 0.907

D. 0.781

E. 0.700

The credit scores of 35-64 year-olds applying for a loan at a given bank to purchase a new car is assumed to be (approximately) normally distributed with mean 600 and standard deviation 100 and they are also assumed to be independent of each other.

b) Find the score that defines the upper 5% of the applicants, i.e. determine the value x such that $P(X > x) = 0.05$ (5p)

A. 436

B. 764

C. 796

D. 404

E. 700

c) During a working day, the bank receives applications from ten 35-64 year-olds. What is the probability that the average credit score of the ten applicants is larger than 625? (5p)

A. 0.785

B. 0.096

C. 0.994

D. 0.215

E. 0.500

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

Problem 4

A mail order company uses the mail to distribute a particular popular product. As a basis for calculating the postage cost, $n = 4$ four already packaged copies of the product were weighed whereby the following weights in grams were obtained: 522, 534, 538 and 522.

a) Assuming that the required assumptions are fulfilled, see c. below, which of the following is a 90% confidence interval for the average weight? (5p)

A. (522.2 ; 535.8)

B. (515.9 ; 542.1)

C. (517.1 ; 540.4)

D. (519.3 ; 538.7)

E. (524.1 ; 533.9)

Using a different kind of packaging that is both lighter in weight and more secure, the average weight of a package can be significantly reduced. A much larger sample of size $n = 50$ was obtained using the new package and you determined the 95% confidence interval for the average weight to be (446 ; 450), using the normal distribution as an approximation. However, when asked, you can't remember what the standard deviation of the weights was and you need to quickly calculate it from the given information.

b) What is the standard deviation of the weights in this sample? (5p)

A. 7.22

B. 51.0

C. 8.60

D. 3.61

E. 9.45

Statistical inference is the formal process of analyzing limited data to infer properties of an underlying probability distribution or that of a greater population, e.g. by providing estimates and confidence intervals or testing hypotheses. Statistical inference requires some assumptions concerning the generation of the observed and similar (unobserved) data.

c) Relating to the problems above, which of the following is a false or irrelevant assumption? (5p)

A. The weights in a) are assumed to be mutually independent random variables.

B. The weights in b) are assumed to be realizations drawn from the same distribution.

C. In b) we assume that the weights have the same mean and variance but we do not assume anything about the underlying distribution.

D. In a) we rely on the Central Limit Theorem (CLT) that states that the sample mean is normally distributed.

E. In a) we assume that the weights are normally distributed $N(\mu, \sigma^2)$.

Problem 5

A project exploring the bottled water phenomena conducted an experiment where 100 students participated in a double-blind study where they tasted three different commercial brands of bottle water (A, B and C) and common tap water (T). Each student were asked to indicate which of the four types they preferred. The results of the experiment are displayed in the table below:

	A	B	C	T
Observed frequency	27	34	26	13

You are tasked with doing a formal hypothesis test, at the 5% significance level, to determine if the four types are equally likely in preference or if they differ, i.e. if some types are more preferred than others.

a) Given the data, what is the result and conclusion of the test? (5p)

A. $\chi^2_{\text{obs}} = 9.20$; H_0 is rejected; all four types are equally likely in preference.

B. $\chi^2_{\text{obs}} = 13.65$; H_0 is not rejected; all four types are equally likely in preference

C. $\chi^2_{\text{obs}} = 9.20$; H_0 is not rejected; some types are more preferred than others

D. $\chi^2_{\text{obs}} = 13.65$; H_0 is rejected; some types are more preferred than others

E. $\chi^2_{\text{obs}} = 9.20$; H_0 is rejected; some types are more preferred than others

b) The p-value for the test above lies between which two values? (5p)

A. Larger than 0.10

B. Between 0.05 and 0.10

C. Between 0.025 and 0.05

D. Between 0.01 and 0.025

E. None of the above

Complete written solutions are required for Problems 6 and 7.

Use separate answer sheets for 6 and 7 respectively.

Problem 6

An experimental time-saving surgical procedure is being tested as an alternative to the old method. A small scale study was done where five surgeons performed the operation on two patients each, one using the old method and one with the new. The patients were matched pairwise by age, sex and other relevant factors so that they would resemble each other as much as possible. The times in minutes to complete the surgeries were recorded and are displayed in the table on the following page. At the 5% significance level, can it be concluded that the new method on average is faster compared to the old method?

a) State the hypotheses and the assumptions that you need in order to solve the problem. State the test statistic and its distribution, the decision rule and critical value. (8p).

b) Finish your calculations, state your conclusions and give a verbal interpretation. (6p)

c) It was later explained to you that the new method is more expensive than the older method and could be justified only if it was on average more than 15 minutes faster. How would you adjust your test above to test this and what would your conclusion be with the given data? Note that you do not need to reiterate the entire test, you need only change the hypotheses, the test statistic and your final conclusion. (6p)

Problem 7

The human resources department of a large corporation conducted a study of the sleeping habits of their employees. They suspected that the average hours worked per week affects the average number of hours the employees sleep each night. The following two models were estimated:

Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

where Y is the average **hours slept** each night, X_1 is the average **hours worked** per week, and X_2 is the **age** of the employee and ε is the error term such that $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. On the next page, you can find output for the two estimated models.

a) Use Model 1, $\bar{x}_1 = 41$ and $s_{x_1} = 160$ to find a 90% prediction interval for the number of hours sleep, given that a person works 41 hours. Interpret the result. (6p)

b) Calculate 95% confidence intervals for β_1 in both Model 1 and Model 2. Comment briefly on the results. Would you conclude that there is a linear relationship between X_1 and Y? How does the relationship between X_1 and Y change when you control for $X_2 =$ age? Explain briefly. (8p)

c) You estimate a model without X_1 but instead include gender as an explanatory variable, i.e. $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ where $X_3 = 0$ for males and $X_3 = 1$ for females. How would you interpret the estimate of β_3? Illustrate the regression model in a suitable way in a graph. (6p)

DATA for Problem 6

	Surgeon 1	Surgeon 2	Surgeon 3	Surgeon 4	Surgeon 5
Old method	36	55	28	40	62
New method	29	42	30	32	56

DATA for Problem 7

MODEL 1:

$R^2 = 0.24723$ $R^2_{adj} = 0.22034$ $s_e = 1.45428$ $n = 30$

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	19.4485	19.4485	9.1958
Residual	28	59.2181	2.11493	
Total	29	78.6667		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	13.7672	2.30190	5.98078	1.93E-06
X1 (hours work)	-0.16912	0.05577	-3,03246	0.005183

MODEL 2:

$R^2 = 0,47414$ $R^2_{adj} = 0,43519$ $s_e = 1,23779$ $n = 30$

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	37.2992	18.6496	12.1724
Residual	27	41.3674	1.5321	
Total	29	78.6667		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	12.9526	1.97371	6.56254	4.87E-07
X1 (hours work)	-0.08808	0.05307	-1.65955	0.10858
X2 (age)	-0.06967	0.02041	-3.41335	0.00204

Stockholms
universitet

Department of Statistics

Correction sheet

Date: 05/06/2019

Room: Uggleviksssalen

Exam: Statistics for Economists

Course: Basic Statistics for Economists

Anonymous code: 0005-KWS

☒ 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			4
Teacher's notes 5	10	15	15	10	20	16			

Points	Grade	Teacher's sign.
91	A	

STOCKHOLM UNIVERSITY
Department of Statistics

Spring 2019 C-D

ANSWER FORM Exam – Basic statistics for economists 2019-06-05

Room: Ugglevikssalen

Anonymous code: 0005-KWS *(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)	☐	☐	☐	☒	☐	✓
	b)	☐	☒	☐	☐	☐	
Problem 2	a)	☐	☐	☐	☐	☒	
	b)	☐	☒	☐	☐	☐	
Problem 3	a)	☐	☐	☒	☐	☐	
	b)	☐	☒	☐	☐	☐	
	c)	☐	☐	☐	☒	☐	
Problem 4	a)	☐	☐	☐	☒	☐	
	b)	☒	☐	☐	☐	☐	
	c)	☐	☐	☐	☒	☐	
Problem 5	a)	☐	☐	☐	☐	☒	55/60
	b)	☐	☐	☒	☐	☐	

SU, DEPARTMENT OF STATISTICS

Room: Vagaleviksssalen Anonymous code: 0005-KWS Sheet number: 1

6.

(Observed time)	1	2	3	4	5	Σ
Old method	36	55	28	40	62	221
New method	29	42	30	32	56	189
Σ	65	97	58	72	118	410

Can it be concluded that the new method on average is faster compared to the old method?

X = the time in minutes with old method
Y = the time in minutes with new method.

X_i	36	55	28	40	62
y_i	29	42	30	32	56
d_i	7	13	-2	8	6

R

$n_x = 5$ ~~$n_y = 5$~~ $\bar{x} = \frac{221}{5} = 44,2$ $\bar{y} = \frac{189}{5} = 37,8$

$\bar{d} = \frac{32}{5} = 6,4$ or $(\bar{x} - \bar{y} = \bar{d} \Rightarrow 44,2 - 37,8 = 6,4)$ $s_d^2 = \frac{322 - 5 \cdot 6,4^2}{5-1} = 29,3$ R

$S_d = \sqrt{29,3} = 5,412947441$

a) Assumptions:

- The two samples, x and y, are not independent of each other since it is the same surgeon that perform the operations x_i and y_i. The samples contains paired observations, and the patients are matched to resemble each other as much as possible.
- The difference, $D_i = X_i - Y_i$ is a random variable with observations that are iid, independent of each other and identically distributed ~~+ normal distr~~ R
- The mean and variance are unknown, instead we use sample mean, $\bar{d}$ and s_d^2, sample variance.
- The sample sizes, n_x ~~and n_y~~ are both small, < 30, so we assume a normal distribution.

fort. >

Hypothesis: $H_0: \mu_D = 0$ $(\mu_x - \mu_y = 0)$

$H_1: \mu_D > 0$ $(\mu_x - \mu_y > 0)$ R

This is a one-sided test with $\alpha = 0{,}05$

Test variable: $t_{n-1} = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}} \sim t_{5-1=4 \text{ degrees of freedom}}$ (table 3)

R

Decision rule & critical value: $t_{crit} = t_{4;\alpha} = t_{4;0{,}05} = 2{,}132$ R

Reject H_0 if $t_{obs} > 2{,}132$

8

b) Calculations: $t_{obs} = \frac{\bar{d} - 0}{S_d/\sqrt{n}} = \frac{6{,}4}{\sqrt{\frac{29{,}3}{5}}} = \frac{6{,}4}{2{,}420743688} = 2{,}6438$

Conclusion: $2{,}6438 > 2{,}132$, so we reject the null hypothesis at 5% significance level. The new method is on average faster than the old method, (but we don't know how much faster, just that it is on average, faster)

6

c) $H_0: \mu_D = 15$ $(\mu_x - \mu_y = 15)$ R

$H_1: \mu_D > 15$ $(\mu_x - \mu_y > 15)$ (the old method needs an average time of +15 or more, than the new method).

Test variable: $t_{obs} = \frac{\bar{d} - 15}{S_d/\sqrt{n}} = \frac{6{,}4 - 15}{\sqrt{29{,}3/5}} = -3{,}5526$ R

Conclusion: Now, $-3{,}5526 < 2{,}132$ and we would not reject the null hypothesis at 5% significance level. ~~This~~ First test shows that the new method is on average faster than the old method but this test shows not as much as on average 15 minutes faster.

So, in this case the new method would not be justified since it is not more than 15 minutes faster on average.

6

20

SU, DEPARTMENT OF STATISTICS

Room: Vaglevikssalen   Anonymous code: 0005-KWS   Sheet number: 2

7. Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Y = average hours slept each night

X_1 = average hours worked / week

X_2 = age of the employee

a) Model 1, $\bar{X}_1 = 41$, $S_{x_1} = 160$ $\alpha = 0{,}10$, $X_1 = 41$, $S_x^2 = 160^2 = 25600$

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

From the output:

$b_0 = 13{,}7672$ $b_1 = -0{,}16912$ $n = 30$

$s_e^2 = 1{,}45428^2 = 2{,}114930318$

$t_{n-2;\alpha/2} = t_{30-2;0{,}10/2} = t_{28;0{,}05} = 1{,}701$ (table 3) /

Calculations:

$$(13{,}7672 + (-0{,}16912 \cdot 41)) \pm 1{,}701 \cdot \sqrt{2{,}114930318\left(1 + \frac{1}{30} + \frac{(41-41)^2}{(30-1)\cdot 25600}\right)}$$

$$6{,}83328 \pm 1{,}701 \cdot \sqrt{2{,}114930318\left(1{,}0333 + \frac{0}{742400}\right)}$$

$$6{,}83328 \pm 1{,}701 \cdot \sqrt{2{,}185427995}$$

$$6{,}83328 \pm 2{,}514621154$$

$(4{,}319 ; 9{,}515)$ /   2

In 90% of times, when an employee works 41 hours / week he or she will sleep on average between 4 and 9 hours each night.

6

b)

95% confidence interval for β_1 = $b_1 \pm t_{n-k-1;\alpha/2} \cdot Sb_1$

Model 1:

$b_1 = -0,16912$ $Sb_1 = 0,05577$ (from output)

$t_{n-k-1;\alpha/2} = t_{30-1-1;0,05/2} = t_{28;0,025} = 2,048$ (table 3)

$-0,16912 \pm 2,048 \cdot 0,05577$

$-0,16912 \pm 0,11421696$

$(-0,283; -0,0549)$ 2

Model 2:

$b_1 = -0,08808$ $Sb_1 = 0,05307$

$t_{n-k-1;\alpha/2} = t_{30-2-1;0,05/2} = t_{27;0,025} = 2,052$ (table 3)

$-0,08808 \pm 2,052 \cdot 0,05307$

$-0,08808 \pm 0,10889964$

$(-0,197; 0,021)$ 2

Since the value zero is not included in the confidence interval for Model 1 at 5% significance level we would reject $H_0: \beta_1 = 0$ against $H_1: \beta \neq 0$ at 5% level. This shows that the model holds and that hours worked/week (X_1) is a good predictor. 2

In model 2 we can see that the value zero is included at 5% significance level, hence we would fail to reject $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$.* The slope coefficient in this model could be equal to zero and is not as good predictor in Model 2 as in model 1

Yes there is a linear relationship between X_1 and Y in model 1, but not strongly since it is close to zero. In model 2 there are almost none linear relationship between X_1 and Y, X_2 is a better predictor and affects X_1 to the worse predictor. 1

/7

* At 5% significance level.

SU, DEPARTMENT OF STATISTICS

Room: Ugglevikssalen Anonymous code: 0005-KWS Sheet number: 3

7. c) $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

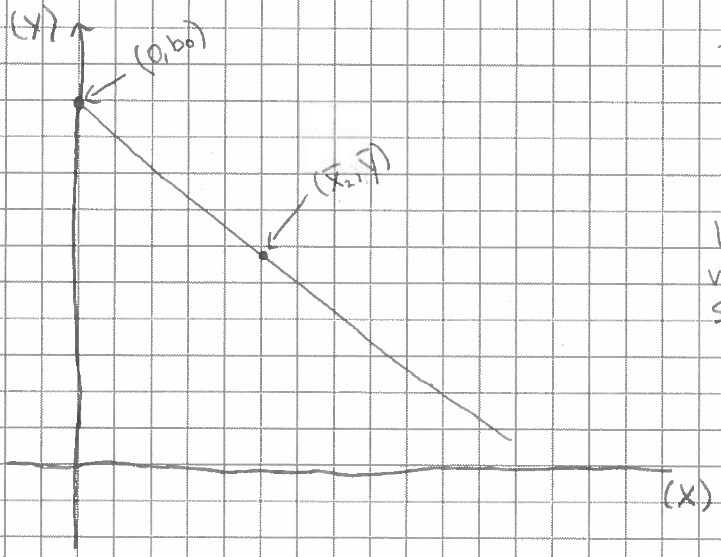
Y = average hours slept / night

X_2 = age

X_3 = gender, 0 = male and 1 = female (dummy variable)

$\hat{Y} = b_0 + b_2 X_i + b_3 X_i$

b_3 = a slope coefficient, the cange in Y when X change. if the estimate b_3 would be equal to 0,05, then the average hours slept each night (Y) would increas 0,05 hours more if the employee was a female than if it was a male. If the employee was a male the average hours slept each night would not change, it would only change if it was a female.



X_3

the dummy variable is either 0, and is not on the graph or, 1 and is in that case equal to the estimate of β_3, $b_3 \cdot 1$.

If I would have had X_i and Y_i values, I would have done a scatter plot and then drawn the regression line in between

3