Stockholm University

**STOCKHOLM UNIVERSITY**

**Department of Statistics**

**Autumn 2019, period C-D**

Ulf Högnäs (examiner)

# EXAM – BASIC STATISTICS FOR ECONOMISTS

**2020-01-15**

| | |
|---|---|
| **Time:** | 10.00 - 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.

- **GOOD LUCK!**

**Problem 1**

a) A town in Sweden has two highschools, *Katedralskolan* and *Polhemskolan*. Each school has two kinds of students: students in college preparation programs (category C) and students in trade preparation programs (category T). The table below shows the absolute frequencies of students in the two categories C and T, for the two schools. **Use the second table to fill in the relative frequencies, conditional on school. What percentage should be in the cell marked (X)?** Choose the value closest to your answer. (5 p)

A) 17%

B) 24%

C) 42%

D) 71%

E) 83%

| Absolute Frequency | C | T | Sum |
|---|---|---|---|
| Katedralskolan | 1600 | 200 | 1800 |
| Polhemskolan | 1400 | 1000 | 2400 |
| Sum | 3000 | 1200 | 4200 |

| Relative frequency, Conditional on School | C | T | Sum |
|---|---|---|---|
| Katedralskolan | | | |
| Polhemskolan | | (X) | |
| Sum | | | |

b) Figure 1 shows a histogram of some numeric data $x$. All values are between 0.0 and 8.0, but the exact values are not known. **Which of these intervals contains the median value?** (5p)

A) 0.0-2.0

B) 2.0-3.0

C) 3.0-4.0

D) 4.0-5.0

E) 5.0-8.0

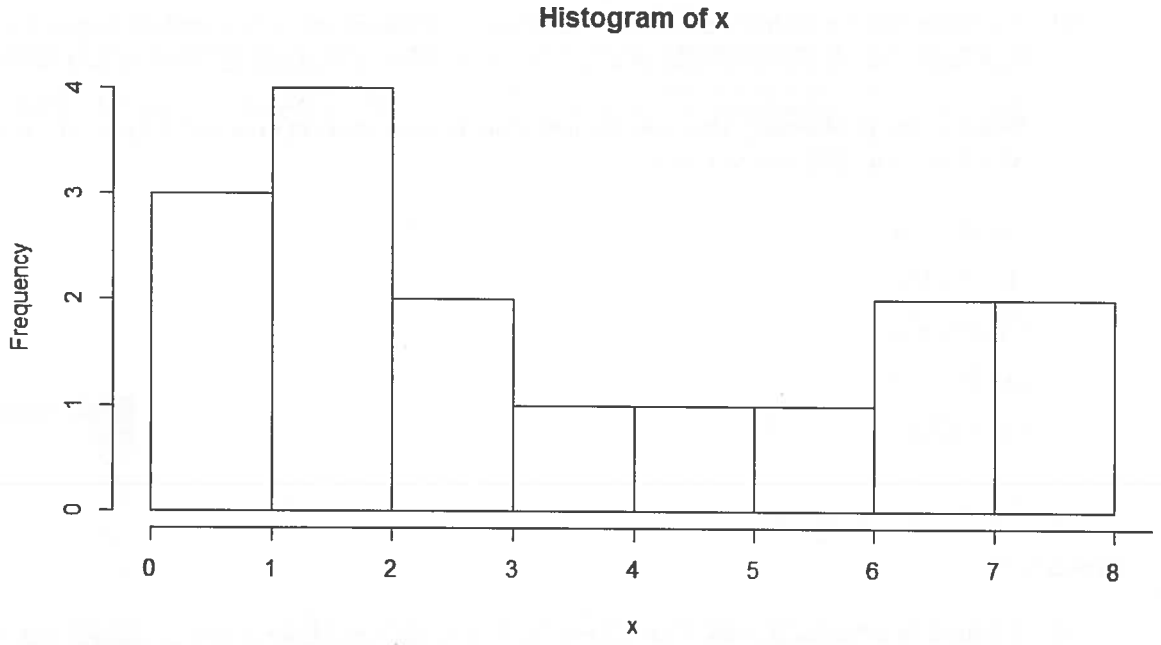


Figure 1

**Problem 2**

a) The best-selling product at particular clothes boutique in Stockholm is a t-shirt. The t-shirt comes in two colors, black and white, and three sizes, Small, Medium, and Large. The colors are equally popular; a randomly chosen customer who buys the t-shirt will choose size with the following probabilities: 30% choose size Small, 50% choose size Medium, and 20% choose size Large. Assume that color is independent of size.

$B$ = the customer buys a black t-shirt
$W$ = the customer buys a white t-shirt
$S$ = the customer buys a size Small.
$M$ = the customer buys a size Medium.
$L$ = the customer buys a size Large.

Consider the next t-shirt sold (to a randomly chosen customer). **Which of the following statements is false?**

(A) $P(S \cup M) = 0.7$

(B) $P(W \cap M) = 0.25$

(C) $P(W \cap \bar{L}) = 0.30$

(D) $P(S \mid B) = 0.20$

(E) $P(M \mid B) = P(B \mid M)$

b) A student takes a multiple choice test in Japanese. Unfortunately, the student does not speak or read Japanese, so she randomly guesses the answer for each question. The test has twelve questions and each question has five alternatives. Assume independence between questions. **What is the probability that the student makes zero correct guesses?** Choose the interval which contains the correct answer.

A) 0%-3%

B) 3%-6%

C) 6%-9%

D) 9%-12%

E) >12%

**Problem 3**

a) A course at a business school has 10 enrolled students. On Monday morning, each individual student will attend class with probability 0.55. The probability that anyone of the 10 students will attend is independent from the others. **Find the probability that <u>four or fewer</u> students attend class on Monday morning.** Choose the interval which contains the correct answer. (5p)

A) 0%-5%

B) 5%-6%

C) 6%-7%

D) 7%-8%

E) >8%

b) Moa sells ice-cream and soda during the summer. After years of careful data-collection, she has developed a mathematical model for her total sales over a whole summer. If $X$ is her revenue from ice-cream (in thousands of crowns) and $Y$ is her revenue from soda (in thousands of crowns) next summer, then

$X \sim N(30, 5^2)$
$Y \sim N(20, 3^2)$

according to her model. Furthermore, the correlation between $X$ and $Y$ is 0.9. **Find the probability that Moa's total sales for the summer is above 60 (thousand crowns), according to her model.** Choose the interval which contains the correct answer. (5p)

A) 0%-5%

B) 5%-15%

C) 15%-30%

D) 30%-45%

E) >45%

c) A scientist conducts scientific study on laboratory rats. Every rat in his treatment group of 500 rats is injected with 1 ml of some drug. Sadly, each rat has a 17% probability of dying within one month of being injected. Assume that each rat is independent of every other rat. **What is the approximate probability that more than 100 out of the 500 rats die within a month?** Choose the interval which contains the correct answer. (5p)

A) 0%-5%

B) 5%-6%

C) 6%-7%

D) 7%-8%

E) >8%

**Problem 4**

a) A group of researchers collects data on baby elephants born in zoos. They want to estimate the mean weight of newborn singleton (not twin) baby elephants. Their sample consisted of four newborn elephants and the babies' weights were as follows:

98, 105, 111, 101

The weights are in kilograms. Assume that the sample is independent, identically distributed, and that the weights are normally distributed. **Based on this information, which of the following is a 95% confidence interval for the mean weight of a baby elephant (singletons, born in a zoo)**. Choose the alternative closest to your answer. (5p)

A) (92.7, 114.8)

B) (94.8, 112.7)

C) (95.9, 111.6)

D) (98.0, 101.0)

E) (98.2, 109.3)

b) A political polling agency in United States periodically asked the question:

"Do you approve or disapprove of the way Donald Trump is handling his job as president?"

The question was asked two times, in June and then in December, each time to a separate, representative sample of 1000 people. In June 46% answered "Yes" and in December 44% answered "Yes."

Let $P_{Dec}$ be the proportion of the whole population who would answer "Yes" in December and $P_{June}$ be the proportion of the whole population who would answer "Yes" in June. **Create a 95% confidence interval for the difference in proportion $P_{Dec}$-$P_{June}$.** Choose the alternative closest to your answer. (5p)

A) (-0.021, -0.019)

B) (-0.024, -0.016)

C) (-0.064, 0.024)

D) (-0.057, 0.017)

E) (-0.051, 0.011)

c) Which of the following statement is **true** of a confidence interval? (5p)

A) Before the sample is drawn, a 95% confidence interval is a random interval that will contain the true parameter with 95% probability.

B) A 95% confidence interval covers 95% of the population.

C) The true parameter is a random value that has 95% chance of ending up in the 95% confidence interval.

D) A larger sample size will increase the size of the confidence interval.

E) The width of the confidence interval is also called the margin of error.

**Problem 5**

A music industry analyst wanted to investigate differences in age among fans of two popular artists, artist A and artist B. The analyst obtained two representative, random samples from each fan base.

| | Artist A | Artist B |
|---|---|---|
| Sample mean | 24.3 | 21.0 |
| Sample variance | 11.1 | 8.9 |
| Sample size | 40 | 40 |

Let $\mu_A$ and $\mu_B$ denote the true mean ages of the two artists' fans. Preform the analyst's hypothesis test at the **5%** level with the following hypotheses:

$$H_0 : \mu_A - \mu_B = 2$$
$$H_1 : \mu_A - \mu_B > 2$$

a) **What is the decision rule?** (5p)

A) Reject $H_0$ if $z_{obs} > 1.6649$

B) Reject $H_0$ if $|z_{obs}| > 1.6649$

C) Reject $H_0$ if $z_{obs} > 1.96$

D) Reject $H_0$ if $z_{obs} > 1.96$

E) Reject $H_0$ if $|z_{obs}| < 1.96$

b) **What is the value of the test variable?** Choose the interval which contains the correct answer. (5p)

A) < 0.5

B) 0.5-1.0

C) 1.0-1.5

D) 1.5-2.0

E) > 2.0

**Problem 6**

To investigate how users of different phones experienced a new app, a sample of n = 100 customers was selected. Respondents were asked to answer the following two questions:

1) Do you like the app (Yes/No)?

2) Do you use an iPhone?

The following results were found:

- 40 answered "Yes" to both questions
- 20 answered "No" to both questions.
- 10 answered "No" to the first question and "Yes" to the second.
- 30 answered "Yes" to the first question and "No" to the second.

We want to see if the answer to the first question depends on whether a person is an iPhone user or not. Test the hypothesis using the 5% significance level.

a) State your assumptions, hypotheses, test statistic, critical value and decision rule. (8p)
b) Finish your calculations, state your conclusions and give a verbal interpretation. (6p)
c) Explain briefly how the p-value can be used to determine the outcome of the test, no more than 2-4 sentences is required. Then use the $\chi^2$-table to approximately determine the p-value of the observed value of the test statistic in b). (6p)

**Problem 7**

A biology student wanted to study human height among the local population for her undergraduate thesis. She collected data from 100 adults using simple random sampling and estimated two regression models:

**Model 1:** $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$

**Model 2:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$

Where $Y$ is the adult's height, $X_1$ is their mother's height, $X_2$ is their father's height, and $X_3$ is a dummy variable describing the adult's biological sex, where $X_3 = 1$ if the adult is female and $X_3 = 0$ if the adult is male. All heights are in centimeters. You can find output from the two models on the next page.

(a) Test whether mother's height explains the adult's height, given that that the adult's father's height and the adult's biological sex is included in the model. State the hypotheses, the test statistic and its distribution, the decision rule and critical value, as well as the outcome and interpretation of the test. If you cannot find the correct degrees of freedom, use the nearest value that you can find in the table. (8p)
(b) Calculate $R^2_{adj}$ for both model 1 and model 2. Based on this, which model is better? (4p)
(c) Interpret the estimated coefficient for the dummy variable $X_3$ in model 2. (2p)
(d) Interpret the estimated coefficient for mother's height in model 1 and then interpret the estimated coefficient for mother's height in model 2. Look at the plot in figure 2. Based on this plot, try to explain why the estimated coefficient is higher in model 1. (6p)

**MODEL 1**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,768485977 |
| R Square | [redacted] |
| Adjusted R Square | |
| Standard Error | 8,832896306 |
| Observations | 100 |

ANOVA

| | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | 2 | 10916,18707 | 5458,093537 | 69,95756907 |
| Residual | 97 | 7567,945543 | 78,02005715 | |
| Total | 99 | 18484,13262 | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 37,17526876 | 17,0142677 | 2,184946741 | 0,031301196 |
| mother's height | 0,83918379 | 0,099485278 | 8,435256029 | 3,1629E-13 |
| daughter | -12,12312529 | 1,795445914 | -6,752152876 | 1,07406E-09 |

**MODEL 2**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,806470043 |
| R Square | [redacted] |
| Adjusted R Square | |
| Standard Error | 8,20452426 |
| Observations | 100 |

ANOVA

| | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | 3 | 12021,96766 | 4007,322552 | 59,53159156 |
| Residual | 96 | 6462,16496 | 67,31421834 | |
| Total | 99 | 18484,13262 | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | -2,553580138 | 18,59693859 | -0,137311855 | 0,891071877 |
| mother's height | 0,597724343 | 0,109947192 | [redacted] | |
| father's height | 0,451234528 | 0,111332317 | | |
| daughter | -12,43360621 | 1,669476366 | | |


Parent's heights
mother's height
father's height
150
160
170
180
190
160
170
180
190
200

Figure 2

Department of Statistics

Stockholms universitet

# Correction sheet

**Date:** 15/1 - 2020

**Room:** Värtasalen

**Exam:** Statistics for Economists

**Course:** Basic Statistics for Economists

**Anonymous code:** 0004-PAF

☒ 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 | - | - | 03 |
| Teacher's notes | | | | | | | | | |

| Points | Grade | Teacher's sign. |
|---|---|---|
| 97 | A | |

Highest score out of 178 exams. Sic ltur Ad Astra

STOCKHOLM UNIVERSITY
Department of Statistics

Autumn 2019 C-D

# ANSWER FORM Exam – Basic Statistics for Economists
## 2020-01-15

Room: VÄ

Anonymous code: 0004 - PAF *(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 |
|---|---|---|---|---|---|
| 1a | ☐ | ☐ | ☒ | ☐ | ☐ |
| 1b | ☐ | ☒ | ☐ | ☐ | ☐ |
| 2a | ☐ | ☐ | ☐ | ☒ | ☐ |
| 2b | ☐ | ☐ | ☒ | ☐ | ☐ |
| 3a | ☐ | ☐ | ☐ | ☐ | ☒ |
| 3b | ☐ | ☒ | ☐ | ☐ | ☐ |
| 3c | ☒ | ☐ | ☐ | ☐ | ☐ |
| 4a | ☐ | ☒ | ☐ | ☐ | ☐ |
| 4b | ☐ | ☐ | ☒ | ☐ | ☐ |
| 4c | ☒ | ☐ | ☐ | ☐ | ☐ |
| 5a | ☒ | ☐ | ☐ | ☐ | ☐ |
| 5b | ☐ | ☐ | ☐ | ☒ | ☐ |

(12/12)

60/60

**SU, DEPARTMENT OF STATISTICS**

Room: VÄ Anonymous code: 0004-PAF Sheet number: 02

6)

a)

| | Do you like the app | Do you use an iphone | |
|---|---|---|---|
| yes | 70 | 50 | 120 |
| no | 30 | 50 | 80 |
| | 100 | 100 | 200 |

$1-\alpha = 0,95 \Rightarrow \alpha = 0,05$

assumptions: iid independent samples, $n > 30$ → approx. normally distributed according to CLT. 1

Hypotheses:

$H_0$: answer to question independent of iphone ownership vs

$H_1$: answer depends on iphone ownership 2

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$

where $E_{ij} = \frac{R_i C_j}{n}$ 2

Critical value: $\chi^2_{crit} = \chi^2_{(2-1)(2-1)0.05} = \chi^2_{1,0.05} =$

$= [\text{Table 4}] = 3,841$ 2

Decision rule: Reject $H_0$ if $\chi^2_{obs} > 3,841$ 1

8

6 b)

| | Do you like the app | Do you use an iphone | |
|---|---|---|---|
| Yes | 70 | 50 | 120 |
| no | 30 | 50 | 80 |
| | 100 | 100 | 200 |

$O_{ij} - E_{ij}$:

| | | |
|---|---|---|
| $70 - \frac{100 \cdot 120}{200} = 10$ | $50 - \frac{100 \cdot 120}{200} = -10$ | 0 |
| $50 - \frac{100 \cdot 80}{200} = -10$ | $50 - \frac{100 \cdot 80}{200} = 10$ | 0 |
| 0 | 0 | 0 |

$O_{ij} - E_{ij}$ sums to 0

$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$:

| | | |
|---|---|---|
| $\frac{10^2}{60} = 1,66...$ | $\frac{(-10)^2}{60} = 1,66...$ | 3,33... |
| $\frac{(-10)^2}{40} = 2,5$ | $\frac{10^2}{40} = 2,5$ | 5 |
| 4,166... | 4,166... | 8,33 |

$$\chi^2_{obs} = 8,33 > \chi^2_{crit} = 3,841$$

conclusion: we reject $H_0$ at the 5% level, the answers to the second question depends on iphone ownership. 4

c) The p-value is the probability to see a value larger then the one observed. If the p-value is lower then the significance level $\alpha$ we can reject the null hypotheses.

the p-value for this test is between 0,001 and 0.005. 5

**SU, DEPARTMENT OF STATISTICS**

Room: VÄ Anonymous code: 0004-PAF Sheet number: 03

7 a) Test on model 2

Test if mothers height explains adults height given that adult fathers height and the adults biological sex is included in the model → model 2.

Hypotheses:

$H_0: b_1 = 0$

vs

$H_1: b_1 \neq 0$

a double sided test, 5% significance level chosen, $\alpha/2 = 0,025$

Test statistic:

$$t = \frac{b_1 - 0}{s_{b_1}} \sim t_{n-k-1}$$

$k = 3$

Critical value:

$$t_{crit} = t_{100-4,\,0,025} = t_{96,\,0.025} = [\text{Table 3}] = 1,985$$

$t_{95,\,0.025}$ chosen

Decision rule:

Reject $H_0$ if $|t_{obs}| > t_{crit}$

$$t_{obs} = \frac{0,597724343}{0,109947192} = 5,4364\ldots \approx 5,437 > t_{crit}$$

Conclusion: $|t_{obs}| > t_{crit}$, we reject the null at the 5% level. $b_2$ is significantly different from zero at the 5% level and thus can explain the adult's height.

2

/9

b) $R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$

$R^2_{adj}\ model\ 1 = 1 - \frac{7567,945543/100-2-1}{18484,13262/99} = 0,5821... \approx 58\%$ 2

$R^2_{adj}\ model\ 2 = 1 - \frac{6462,16496/100-3-1}{18484,13262/99} = 0,6394... \approx 64\%$ 2

Based on the $R^2_{adj}$, model 2 is better. Here 64% of the variation of $y_i$ can be explained by the variation of $x_i$. 2

c) If the dummy variable in model 2 is equal to 1 (girl) the adult's height will decrease by 12,43 centimeters. If it is equal to zero (male) the dummy variable will not affect the height, in other words 12,43 cms will NOT be subtracted. ok 3

d) The coefficient for mother's height is larger in model 1 than in model 2. In model 1 each increase of one centimeters of the mother's height increases the adult's height by 0,839 centimeters. In model 2 each increase of one centimeter of the mother's height increases the adult's height by 0,598 centimeters, a smaller value. With its lower adjusted coefficient of determination and large residuals in figure 2, model 1 could have a higher coefficient for the mothers height due to being less accurate. /