Stockholm
University

STOCKHOLM UNIVERSITY
Department of Statistics
Autumn 2018, period C-D
Michael Carlson (examiner)
Ulf Högnäs

EXAM – BASIC STATISTICS FOR ECONOMISTS

2019-01-14

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 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 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

An NGO (non-government agency) works in hazardous regions around the world. A survey was conducted to investigate the living conditions of adult women in a refugee camps. Data were collected on age, number of children, age and gender of the children, how long they had to queue for e.g. water and more. Each respondent also indicated on a 10-degree scale how they experienced their current situation where 1 was worst possible and 10 best possible.

a) Which of the following statements about the variables is **true**? (5p)

A. Age is a continuous numerical variable on an ordinal scale.

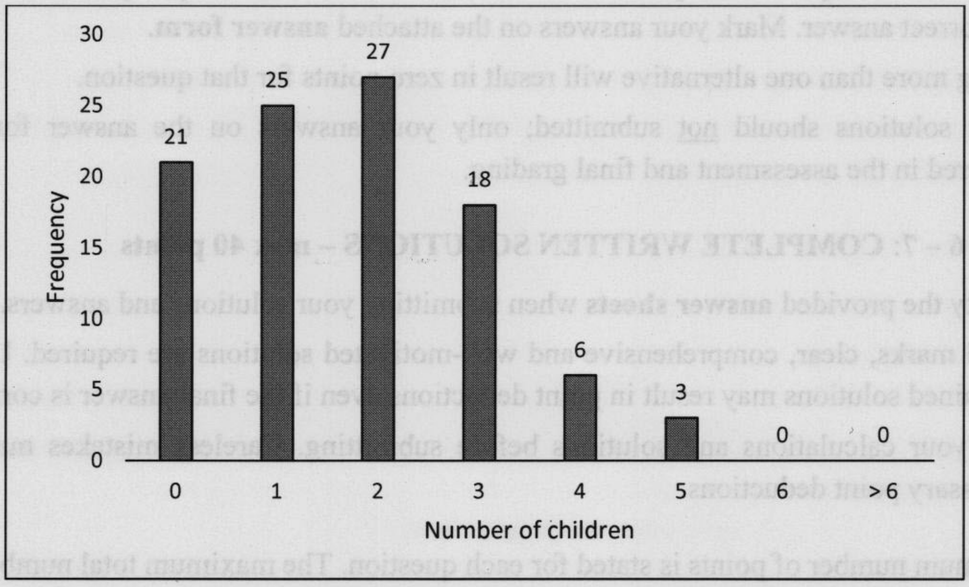
B. Number of children is a discrete numerical variable on a ratio scale (*kvotskala*)

C. Waiting time is continuous categorical variable on a nominal scale.

D. Gender is a categorical variable on an ordinal scale.

E. Current situation is a discrete numerical variable on an interval scale.

When compiling statistics from the survey of the number of children (per woman), the following frequencies were obtained for $n = 100$ women:



b) What is the sample mean and median for the number of children per women? (5p)

A. $\bar{x} = 1.72$ median $= 2.5$

B. $\bar{x} = 1.69$ median $= 1.5$

C. $\bar{x} = 2.25$ median $= 2{,}0$

D. $\bar{x} = 1.50$ median $= 1.5$

E. $\bar{x} = 1.72$ median $= 2{,}0$

Problem 2

A software company has analyzed its customer database and the sales figures for a particular software. The software is available in two versions: the latest and slightly more expensive "Ver. 2.0" and the older and slightly cheaper "Ver. 1.4". Based on sales records, 50% of the customers were classified as "faithful", 30% as "less faithful" and the remaining 20% as "new" (first time purchasers). The relative frequencies of the two software versions for each of the customer categories is as follows:

	Faithful (50%)	Less faithful (30%)	New (20%)	
Ver. 1.4	60%	80%	30%	70%
Ver. 2.0	40%	20%	70%	30%
	100%	100%	100%	

a) What is the probability that a randomly chosen customer purchased Ver. 2.0? (5p)

A. 0,350

B. 0,400

C. 0,433

D. 0,500

E. 0,650

b) Given that a customer purchased Ver. 2.0, what is the probability that the customer belongs to the category "New"? (5p)

A. 0,140

B. 0,200

C. 0,286

D. 0,350

E. 0,700

The probability distribution for a random variable Y is given in the following table:

y	0	1	2	3
$P(y)$	0.3	0.3	0.2	0.2

c) What is the mean and standard deviation of Y? (5p)

A. $\mu_Y = 1{,}3 \quad \sigma_Y = 1.1$

B. $\mu_Y = 1{,}5 \quad \sigma_Y = 1.118$

C. $\mu_Y = 1{,}3 \quad \sigma_Y = 1.21$

D. $\mu_Y = 1{,}5 \quad \sigma_Y = 1.25$

E. $\mu_Y = 1{,}3 \quad \sigma_Y = 1.118$

Problem 3

Assume that the probability that a Swedish CEO can conduct business in English is $P = 0.80$. A sample of Swedish CEOs of size $n = 8$ are selected at random. The observations are assumed to be independent of each other.

a) What is the probability that less than three of the CEOs in the sample cannot transact business in English? (5p)

A. 0.056

B. 0.203

C. 0.797

D. 0.944

E. 0.990

Jim is the CEO of a mid-sized Swedish company. Jim's systolic blood pressure X is a random variable with mean $\mu = 150$ mmHG and standard deviation of $\sigma = 20$ mmHg. It is assumed that $X \sim N(\mu, \sigma^2)$.

For Jim's age group, a mean value of 140 mmHg is the threshold for high blood pressure, above which treatment is required or at least recommended. A sample of $n = 9$ measurements are taken at randomly chosen moments over a longer period of time. The sample mean $\bar{X}$ of these measurements is calculated and compared to the threshold value. The measurements are assumed to be independent or each other.

b) What is the probability that Jim will not receive the sample mean is less than 140 mmHg? (5p)

A. 0.000

B. 0.050

C. 0.067

D. 0.409

E. 0.933

NOTE: The numbers in a) and b) above have been rounded to three decimals.

Problem 4

A sample of size was $n = 360$ consisting of randomly chosen of small companies was drawn and on the question whether they thought it would be a rise for their business the following year, $x = 234$ answered "Yes". Assume that the observations are *iid*.

a) Which of the following alternatives is a 99% confidence interval for $P =$ the true proportion of companies that are positive about the future and would have answered "Yes"? (5p)

A. (0.609; 0.691)

B. (0.601; 0.699)

C. (0.592; 0.708)

D. (0.585; 0.715)

E. (0.567; 0.733)

In a similar study in another country, a 95% confidence interval for the same proportion was reported to be (0.488; 0.612).

b) What was the sample size for that country's survey? Note that the calculations are subject to rounding errors and that the alternatives below are approximate. TIP: The midpoint of the interval is equal to the estimated proportion $\hat{p}$ for this other country. (5p)

A. $n = 8$

B. $n = 126$

C. $n = 247$

D. $n = 782$

E. $n = 1262$

Statistical inference relies on results from probability theory such as the central limit theorem, and we use different concepts such as bias, margin of error, confidence level etc.

c) Which of these statements is **correct**, i.e. true? (5p)

A. The expected value of a biased estimator is equal to the parameter it is estimating.

B. A statistically significant result is synonymous with a p-value larger than α.

C. The margin of error does not depend on the confidence level of the interval.

D. It is always necessary to apply the central limit theorem in estimation, even when the observations are drawn from a normal distribution.

E. In order to reduce the size of the standard error of an estimator by 50%, a sample size four times larger is typically needed.

Problem 5

Returning to the scenario in Problem 2a-b), you now have the actual sales numbers (frequencies) for each combination of version and customer category. In total there were $n = 100$ sales distributed across software versions and customer categories as follows:

	Faithful	Less faithful	New
Ver. 1.4	30	24	6
Ver. 2.0	20	6	14
	50	30	20

You are asked to do a test for independence between version purchase and customer category (*oberoendetest*), i.e. to test if the purchased version is affected by (depends on) customer category or not.

a) With a 1% significance level, which of the following is a correct specification of the test? (5p)

A. H_0: independent vs H_1: dependent; reject H_0 if $|\chi^2_{\text{obs}}| < 12.838$

B. H_0: dependent vs H_1: independent; reject H_0 if $\chi^2_{\text{obs}} > 9.210$

C. H_0: independent vs H_1: dependent; reject H_0 if $\chi^2_{\text{obs}} > 15.086$

D. H_0: dependent vs H_1: independent; reject H_0 if $\chi^2_{\text{obs}} > 12.838$

E. H_0: independent vs H_1: dependent; reject H_0 if $\chi^2_{\text{obs}} > 9.210$

Calculate the observed value of the test statistic.

b) What is the correct value and conclusion? (5p)

A. $\chi^2_{\text{obs}} = 12.5$ and we reject H_0; the p-value is smaller than 0.01

B. $\chi^2_{\text{obs}} = 12.5$ and we do not reject H_0; the p-value is larger than 0.01

C. $\chi^2_{\text{obs}} = 12.5$ and we reject H_0; the p-value is larger than 0.01

D. $\chi^2_{\text{obs}} = 16.07$ and we reject H_0; the p-vaule is smaller than 0.001

E. $\chi^2_{\text{obs}} = 16.07$ and we do not reject H_0; the p-vaule is larger 0.001

Complete written solutions are required for Problems 6 and 7.

Use separate answer sheets for 6 and 7 respectively.

Problem 6

Management for a chain of take-away restaurants is analyzing variations in order sizes (in Euros) that customers purchase at different times during a normal week. Statistics for two small samples, one for Friday nights and one for Saturday nights, is given below:

	Sample mean	Sample variance	Sample size
Friday nights (X)	$\bar{x} = 22$	$s_x^2 = 25$	$n_x = 12$
Saturday nights (Y)	$\bar{y} = 27$	$s_y^2 = 36$	$n_y = 10$

You are asked to test at $\alpha = 0.05$ if there is a significant difference between Friday and Saturday nights with respect to average order sizes.

a) State the assumptions that you need in order to solve the problem and perform the test. Based on your assumptions, what test statistic will you use and what is its distribution? Are any of the assumptions in doubt in this case? (5p).

b) State the hypotheses, the decision rule and critical value. (5p)

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

d) Explain briefly what a Type II error is. Only a couple of sentences is required. (4p)

Problem 7

In order to determine if the previous year's percent return on mutual funds could be a good predictor of the following year's percent return, an investment manager gathered data on $n = 10$ funds different for two consecutive years and estimated a simple regression model where Y = last year's percent return and X = this year's percent return, i.e. the model:

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

The complete dataset and some calculations are provided on the following page.

a) Estimate the model parameters and interpret them (i.e. explain the numerical values), state the estimated model. (5p)

b) Illustrate the data in a suitable graph and include the estimated regression line in your graph. (4p)

c) Calculate the residual variance and the coefficient of determination and interpret the result for the latter. (5p)

d) Construct a 95% confidence interval for the slope, and from the result draw a conclusion about the estimated model. Would you say that last year's return is a good predictor for the following year's return or not? Why? (6p)

DATA for Problem 7

i	x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	e_i	e_1^2
1	12	16	1	0	0	– 0.62432	0.38978
2	19	26	64	100	80	5.00541	25.05408
3	10	18	1	4	– 2	2.62432	6.88708
4	14	19	9	9	9	1.12703	1.27019
5	14	13	9	9	– 9	– 4.87297	23.74587
6	6	16	25	0	0	3.12162	9.74452
7	20	21	81	25	45	– 0.61892	0.38306
8	11	12	0	16	0	– 4.00000	16.00000
9	-1	11	144	25	60	2.49189	6.20953
10	5	8	36	64	48	– 4.25405	18.09698
SUM	110	160	370	252	231	0.0000	107.7811

Stockholms universitet

Department of Statistics

Correction sheet

Sic itur ad astra! /ML

Date: 14/01/2019

Room: Värtasalen

Course: Basic statistics for economists (eng)

Exam: Statistics for economists (eng)

Anonymous code: 0010-ZDE

☒ 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	~~X~~	~~X~~	5
Teacher's notes 10	15	10	15	10	20	20	\	\	

Points	Grade	Teacher's sign.
100	A	ML

STOCKHOLM UNIVERSITY
Department of Statistics

Autumn 2018 C-D

ANSWER FORM Exam – Basic statistics for economists
2019-01-14

Room: Vårtasalen

Anonymous code: 0010-2DE *(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)	□	□	□	☒	□
	c)	☒	□	□	□	□
Problem 3	a)	□	□	☒	□	□
	b)	□	□	□	□	☒
Problem 4	a)	□	□	□	☒	□
	b)	□	□	☒	□	□
	c)	□	□	□	□	☒
Problem 5	a)	□	□	□	□	☒
	b)	☒	□	□	□	□

See solutions!

← i had problem understand the question but i assumed that you meant not less than 140, or in other words greater than 140.

VR

60/60

SU, DEPARTMENT OF STATISTICS

Room: Vårtasalen **Anonymous code:** 0010-ZDE **Sheet number:** 7

Problem 6

a) Assumptions: We assume that the observations of X and Y are independent and identically distributed (iid). We also assume that the samples are independent of eachother. Also since the sample size is small we can not apply CLT so we assume that $X \sim N(\mu, 6^2)$ and $Y \sim N(\mu, 6^2)$ which means that $\bar{X} \sim N(\mu, \frac{6^2}{n})$ and $\bar{Y} \sim N(\mu, \frac{6^2}{n})$. We are also not provided

with the population variance so we assume that the sample variance for X any Y are good estimates proxy. We also assume that 6^2_x and 6^2_y are assumed equal.

The assumptions that X and Y are normally distributed are actually in doubt since the information was not handed to us anywhere. The same holds for the assumptions we had to make for equal variance. NICE!

Test statistics: $t_{n_x+n_y-2} = \dfrac{\bar{X} - \bar{Y} - 0}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$ which is t-distributed with df = ? /5

b) Hypothesis: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$

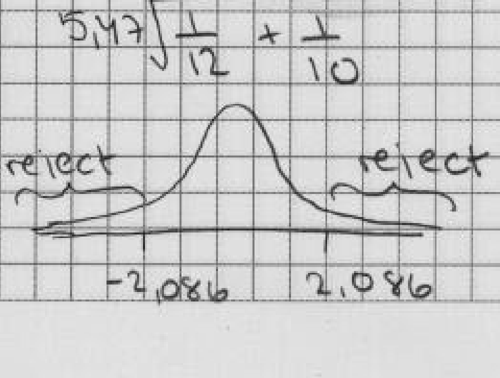
Decision rule: We reject H_0 if $|t_{obs}| > t_{crit}$

Critical value: $t_{(12+10-2);(0{,}025)} = t_{20;0{,}025} = 2086$ /5

c) Calculations: First we need to solve for the pooled variance: $S_p^2 = \dfrac{(12-1)\cdot 25 + (10-1)\cdot 36}{10+12-2}$

$S_p^2 = 29{,}95$ $S_p = \sqrt{29{,}95} \approx 5{,}47$

$t_{obs} = \dfrac{22 - 27 - 0}{5{,}47\sqrt{\frac{1}{12} + \frac{1}{10}}} \approx -2{,}134$ ☈



We reject Ho since $|t_{obs}| > t_{crit}$ which means that there are a (significant difference between X and Y at a 5% significant level.) /6

order size on friday

order size on saturdays

SU, DEPARTMENT OF STATISTICS

Room: Vörössalen Anonymous code: 0010-ZDE Sheet number: 2

Problem 6

d) Type II error is when you accept H_0 when H_0 is false, or in other words when H_1 is the correct decision. The probability of a Type II error is β. ½

20

Problem 7

$n = 10$

a) To estimate b_1 we use the formula $\frac{cov(x,y)}{s_x^2}$ we start of by calculating the covariance which we do by: $cov \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$ in our case: $\frac{231}{9} \approx 25{,}667$

After that we calculate the sample variance which is: $\frac{(x_i-\bar{x})^2}{n-1}$ which in our case is $\frac{370}{9} \approx 41{,}11$

b_1 is therefore : $\frac{25{,}667}{41{,}11} \approx \underline{0{,}624}$ R

now we can calculate b_0 which is: $\bar{y} + (b_1 \cdot \bar{x})$

to calculate $\bar{y}$ we simply divide the sum of y_i with the sample size, and the same goes for $\bar{x}$. Therefor $\bar{y} = \frac{160}{10} = 16$ and $\bar{x} = \frac{110}{10} = 11$.

$b_0 = 16 - (11 \cdot 0{,}624) \approx \underline{9{,}13}$ R

Estimated model: $\hat{y} = 9{,}13 + 0{,}624x_i$ R

expected to be / BRA!

this year

This means that if there is no percentage return the last years percentage return is 9,13 %. And if this years percentage return increase by one unit our last years percentage return is expected to increase by 0,624 %.

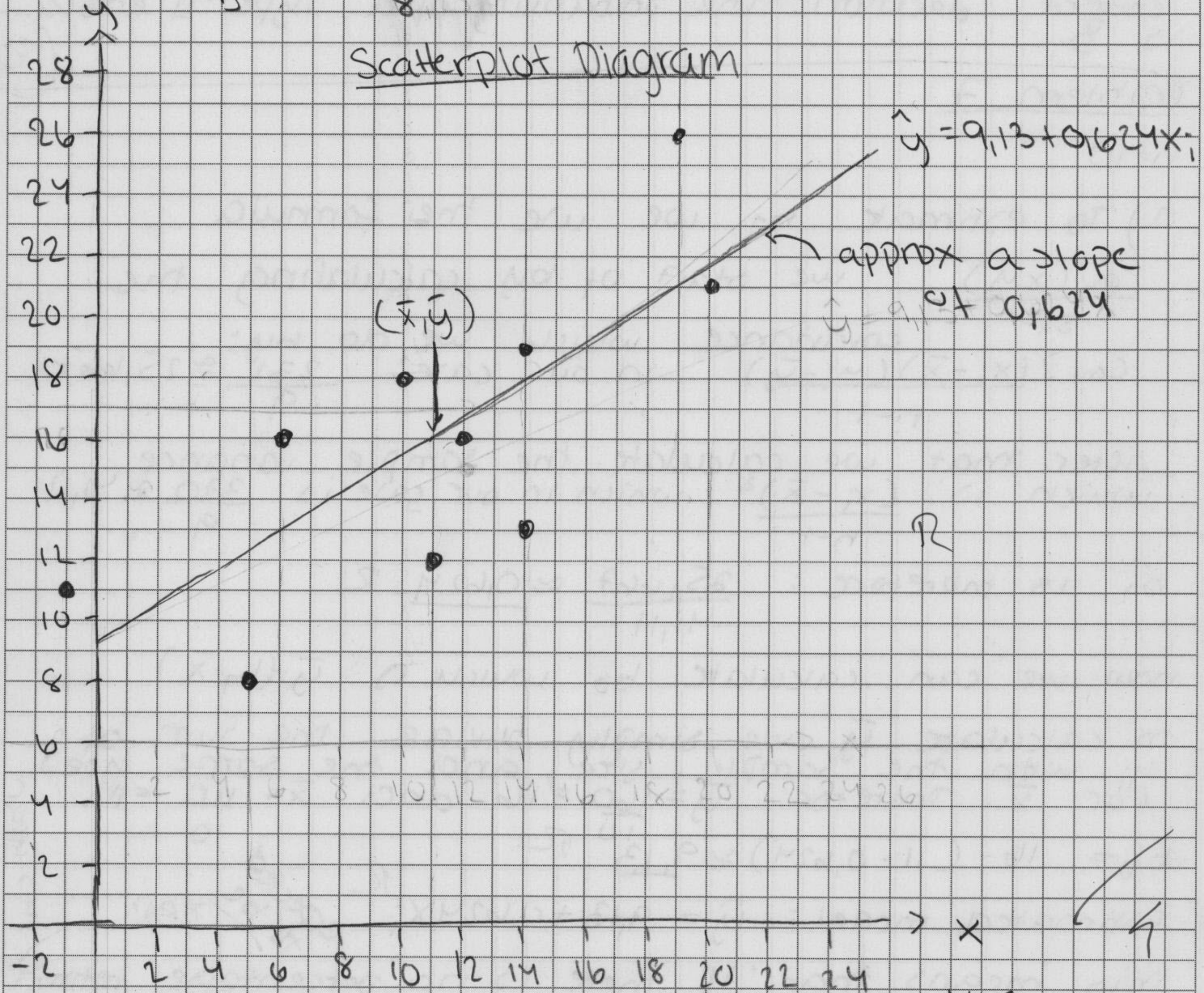
/5

→

b)

X	Y
12	16
19	26
10	18
14	19
14	13
6	16
20	21
11	12
-1	11
5	8

$\hat{y} = 9{,}13 + 0{,}624x$



K=1

c) Residual variance $\frac{e_i^2}{n-k-1} = \frac{107{,}7811}{10-1-1} \approx 13{,}47$

2

coefficient of determination: $1 - \frac{SSE}{SST}$

$SSE = e_i^2 = 107{,}7811$ $SST: (y_i - \bar{y})^2 = 252$

$R^2 = 1 - \frac{107{,}7811}{252} \approx 0{,}5723$

R

Explanation next page →

SU, DEPARTMENT OF STATISTICS

Room: Värasalen Anonymous code: 0010-ZDE Sheet number: 3

Problem 7

Forts c) A R^2 of 57,23% means that 57,23% of the variation in last years percent return can be explained by this years percent return. R

/5

a) 95% CI: $b_1 \pm t_{8;0,025} \cdot s_{b_1}$

$b_1 = 0,624$ $s_{b_1}^2 = \frac{se^2}{9 \cdot s_x^2} = \frac{13,47}{9 \cdot 41,11} \approx 0,03641$

$s_{b_1} = \sqrt{0,03641} = 0,1908$

$t_{8;0,025} = 2,306$ 95% CI $0,624 \pm 2,306 \cdot 0,1908 =$

$[0,1839 ; 1,064]$ R

Since our CI does not include 0 this is almost the same as doing a test and see that the slope is significant from 0. We can therefore from this CI conclude that the slope is significant from 0 And therefor the last year's return is a good prediction for the following years return.

/6

20