Stockholm University. Department of Statistics
Basic Statistics for Economists
Solution to the exam 220211

Problem 1. The table below shows the total number of eligible voters (in thousands) and the percentage of them who voted in the election for Swedish parliament 2014, by age category and sex.

	Men		Women	
age	eligible voters	voted, %	eligible voters	voted, %
18—24	420	79,3	393	83,3
25—29	290	78,9	286	84
30—34	269	82,5	265	85,4
35—39	276	85,3	267	86,8
40—44	307	87,0	302	88,2
45—49	321	85,5	320	88,2
50—54	300	86,7	292	88,8
55—59	281	86,9	279	91,0
60—64	277	88,2	277	90,8
65—69	272	91,6	281	92,2
70—74	250	91,3	259	90,5
75—79	152	87,4	178	86,7
80+	199	80,8	316	69,5
total	3614	85,2	3715	86,4

a. What percentage of eligible female voters older than 69 voted in the election? Choose the alternative closest to your own answer. (5p.)

(A) 20.3%

(B) 21.9%

(C) $\boxed{80.8\%}$

(D) 82.2%

(E) 83.9%

b. Find the interval that contains the first quartile for age among men who voted. Tip: First calculate the number of voters in each category. (5p.)

(A) 18-24

(B) 25-29

(C) $\boxed{30\text{-}34}$

(D) 35-39

(E) 40-44

Solution 1. Let x_k be the number of eligible female voters in the kth age category and y_k be the percentage of females in the kth age category who voted in the election. Let also $z_k = x_k \times y_k$ be the total number of females in the kth age category who voted and Z_k the cumulative frequency of z_k, i.e. $Z_k = \sum_{i=1}^{k} z_i$. The observed values are shown in the following table (**Note:** Calculations are carried out by keeping all decimals, although they are not shown for improving readability.)

	Men				Women			
age k	Eligible x_k	voted, % y_k	z_k	Z_k	Eligible x_k	voted, % y_k	z_k	Z_k
18—24	420	79,3	333	333	393	83,3	327	327
25—29	290	78,9	229	562	286	84,0	240	568
30—34	269	82,5	222	784	265	85,4	226	794
35—39	276	85,3	235	1019	267	86,8	232	1026
40—44	307	87,0	267	1286	302	89,0	266	1294
45—49	321	85,5	274	1561	320	88,2	282	1577
50—54	300	86,7	260	1821	292	88,8	259	1836
55—59	281	86,9	244	2065	279	91,0	254	2090
60—64	277	88,2	244	2309	277	90,8	252	2341
65—69	272	91,6	249	2559	281	92,2	259	2600
70—74	250	91,3	228	2787	259	90,5	234	2835
75—79	152	87,4	133	2920	178	86,7	154	2989
80+	199	80,8	161	3079	316	69,5	220	3210
total	3614	85,2	3079		3715	86,4	3210	

a. The desired percentage is of the form t_y/t_z where t_y is the total of females older than 69 who voted in the election and t_z is the total of eligible female voters older than 69. This is obtained as

$$\frac{t_y}{t_z} = \frac{234 + 154 + 220}{259 + 178 + 316} = 80.8\%$$

b. The first quartile is the 25th percentile. We have $n = 3\,079\,000$, so $(n+1)p/100 \approx 770\,000$, therefore $a = 770\,000$ and $b = 0$. So the 25th percentile is the observation 770 000 from smallest to largest. As the values are in thousands, this will be the observation 770 which falls in the third interval, i.e. the first quartile is contained in the interval 30—34.

Problem 2.

a. The following table describes a random variable X.

x	-1	0	1
$P(X = x)$	0.4	0.2	0.4

Find the variance of X. Choose the alternative closest to your answer (5p.)

(A) 0

(B) 0.25

(C) 0.40

(D) 0.50

(E) 0.80

b. A studio software company has analyzed its customer database and the sales figures for a particular software. The software is available in two versions: the "Lite" version and the more expensive "Pro" version. Based on sales records, 50% of the customers were in the age group 16-24, 30% were in the age group 25-34 and the remaining 20% were 35 and older. The relative frequencies of the two software versions for each of the customer categories is as follows:

	16—24	25—34	35+
Lite	80%	60%	25%
Pro	20%	40%	75%

What is the probability that a randomly chosen customer purchased the Pro version? Choose the alternative closest to your answer.(5p)

(A) 0.22

(B) 0.27

(C) 0.37

(D) 0.45

(E) 0.48

c. A test for celiac disease (an autoimmune disorder that causes gluten intolerance) is 93% accurate when the person does have the disease, and 97% accurate when the person does not have the disease. Suppose that 1% of the population of some country has this disease. We test a randomly selected person from the population - and the test is positive (so the test indicates celiac disease). Find the probability that the test shows the correct result. Choose the alternative closest to your answer.(5p)

(A) 0.12

(B) $\boxed{0.25}$

(C) 0.53

(D) 0.77

(E) 0.93

(Note: In this problem, a patient is tested at random. In practice, patients are tested because the show symptoms that are typical of the disease. When this is the case, the test will be much more accurate. You should not administer this test at random!)

Solution 2.

a.

$$E(X) = \sum xp(x) = (-1) \cdot 0.4 + 0 \cdot 0.2 + 1 \cdot 0.4 = 0$$

$$Var(X) = \sum x^2 p(x) - (E(X))^2 = (-1)^2 \cdot 0.4 + 0 \cdot 0.2 + 1^2 \cdot 0.4 - 0 = 0.4 + 0.4 = 0.8$$

b. First, let us define some notation:

- let E_{16-24} be the event "a randomly chosen customer is between 16 and 24 years old";
- let E_{25-34} be the event "a randomly chosen customer is between 25 and 34 years old";
- let E_{35+} be the event "a randomly chosen customer is at least 35 years old";
- let A_l be the event "a randomly chosen customer purchased the Lite version; and,
- let A_p be the event "a randomly chosen customer purchased the Pro version.

We are looking for $P(A_p)$ and it is known that

$$\begin{array}{rrr} P(A_l|E_{16-24}) = 0.8 & P(A_l|E_{25-34}) = 0.6 & P(A_l|E_{35+}) = 0.25 \\ P(A_p|E_{16-24}) = 0.2 & P(A_p|E_{25-34}) = 0.4 & P(A_p|E_{35+}) = 0.75 \\ P(E_{16-24}) = 0.5 & P(E_{25-34}) = 0.3 & P(E_{35+}) = 0.2 \end{array}$$

Taking into account that E_{16-24}, E_{25-34} and E_{35+} form a partition of the sample space, we can use the law of total probability, therefore we have

$$P(A_p) = P(A_p|E_{16-24})P(E_{16-24}) + P(A_p|E_{25-34})P(E_{25-34}) + P(A_p|E_{25-34})P(E_{25-34}) = 0.37$$

c. Let A be the event "a randomly chosen person has the celiac disease" and B be the event "a randomly chosen person tests positive". We are looking for $P(A|B)$ and it is known that

$$P(B|A) = 0.93 \qquad P(\bar{B}|\bar{A}) = 0.97 \qquad P(A) = 0.01.$$

It follows that $P(\bar{A}) = 0.99$. Also, note that

$$0.93 = P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B \cap A)}{0.01} \longrightarrow P(A \cap B) = 0.0093.$$

Therefore $P(A \cap \bar{B}) = P(A) - P(A \cap B) = 0.0093$. In the same way,

$$0.97 = P(\bar{B}|\bar{A}) = \frac{P(\bar{B} \cap \bar{A})}{P(\bar{A})} = \frac{P(\bar{B} \cap \bar{A})}{0.99} \longrightarrow P(\bar{A} \cap \bar{B}) = 0.9603.$$

Therefore $P(\bar{A} \cap B) = P(\bar{A}) - P(\bar{A} \cap \bar{B}) = 0.0297$ and $P(B) = P(A \cap B) + P(\bar{A} \cap B) = 0.039$.

Finally, we get

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = 0.2385.$$

Problem 3.

a. A pizzeria sells pizzas for €10 each. Suppose that the number of pizzas sold during a particular week is approximately normally distributed with mean 1500 and standard deviation 180. Find the probability that the revenue (total sales) from pizza is greater than €18000 that week. Choose the alternative closest to your answer. (5p.)

(A) 0%
(B) 5%
(C) 50%
(D) 95%
(E) 100%

b. The pizzeria sells only two variations of pizza: Margarita and Marinara. Past sales data shows that 25% of customers prefer Marinara, while 75% of customers prefer Margarita. Suppose that we randomly select 16 customers. What is the probability that at least 4 out of these 16 customers prefer the Marinara pizza? Choose the alternative closest to your (5p.)

(A) 37%
(B) 40%
(C) 60%
(D) 63%
(E) 66%

c. Suppose that we randomly select 100 customers. Find the approximate probability that at least 30 of these 100 customers prefer the Marinara pizza. Use approximation method taught in the course. Choose the alternative closest to your answer. (5p)

(A) 0%
(B) 15%
(C) 41%
(D) 85%
(E) 100%

Solution 3.

a. Let X = number of pizzas sold and $Y = 10X$ total revenue from pizzas. We know that X follows a normal distributions with mean 1500 and standard deviation 180, $X \sim N(1500, 180^2)$, therefore $Y \sim N(15000, 100 \cdot 180^2)$ We are interested in $P(Y > 18000)$.

$$P\left(Y > 18000\right) = P\left(Z > \frac{18000 - 15000}{10 \cdot 180}\right) = P(Z > 1.66) = 1 - P(Z < 1.66) = 1 - 0.95154.$$

b. Let Y = number of customers that like Marinara. Y follows a binomial distributions with parameters 16 and 0.25, $Y \sim Bin(16, 0.25)$. We are interested in $P(Y \geq 4)$.

$$P(Y \geq 4) = 1 - P(Y < 4) = 1 - P(Y \leq 3) = 1 - 0.40499 = 0.59501$$

c. Let Y = number of customers that like Marinara. Y follows a binomial distributions with parameters 100 and 0.25, $Y \sim Bin(100, 0.25)$. We are interested in $P(Y \geq 30)$.

$$E(Y) = n \cdot p = 100 \cdot 0.25 = 25$$

$$Var(Y) = n \cdot p \cdot (1 - p) = 100 \cdot 0.25 \cdot 0.75 = 18.75$$

As n is large and $n \cdot p \cdot (1 - p) > 5$, the approximation $Y \sim N(25\,, 18.75)$ holds, therefore

$$P(Y \geq 30) = P\left(Z \geq \frac{30 - 25}{\sqrt{18.75}}\right) = P(Z \geq 1.15) = 1 - P(Z < 1.15) = 1 - 0.87493$$

Problem 4.

a. A statistics student wants to estimate the mean age of first-year business students. She collects a simple random sample of 30 business students from Stockholm University and a sample of 40 business students from Uppsala University.

	Sample mean	Sample Standard Deviations	n
Stockholm	21.93	4.95	30
Uppsala	22.52	4.56	40

Find a 90% confidence interval for the difference in mean age between the two populations of students (Stockholm minus Uppsala). Choose the alternative closest to your answer. (5p)

(A) (-1.46, 0.28)

(B) (-2.29, 1.11)

(C) (-2.49, 1.31)

(D) (-2.86, 1.68)

(E) (-3.26, 2.08)

b. As part of a scientific study, a random sample of 10 overweight volunteers are given a course in nutrition by a nutritionist. The body weight of each participant is measured at the beginning of the study, and then again six months later. Assume that the body weights are normally distributed. You can find the weights in kilograms below:

Volunteer	1	2	3	4	5	6	7	8	9	10
Weight before	80	95	78	105	90	90	91	79	115	85
Weight after	80	92	77	100	91	89	93	77	112	85

Find a 95% confidence interval for the average weight change, after minus before, of a (future) participant in this type of study. Choose the alternative closest to your answer. (5p.)

(A) (-2.24, -0.15)

(B) (-2.37, -0.03)

(C) (-2.50, -0.10)

(D) (-2.70, 0.30)

(E) (-11.9, 9.5)

Solution 4.

a. We use the formula

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

with

$$z_{\alpha/2} = z_{0.1/2} = z_{0.05} = 1.6449$$

$$\bar{x} - \bar{y} = 21.93 - 22.52 = -0.59$$

$$\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} = \frac{4.95^2}{30} + \frac{4.56^2}{40} = 0.51984 + 0.81675 = 1.33659$$

$$\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = \sqrt{1.33659} = 1.15611$$

which yields

$$\bar{x} - \bar{y} \pm Z_{\alpha/2}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = -0.59 \pm 1.6449 \cdot 1.15611 = -0.59 \pm 1.9016 = (-2.49\,, 1.31)$$

b. We have

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-12}{10} = -1.2$$
$$s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{39.6}{10-1} = 4.4$$
$$s_d = \sqrt{s_d^2} = \sqrt{4.4} = 2.097$$
$$t_{n-1;\alpha/2} = t_{10-1;0.05/2} = t_{9;0.025=2.262}.$$

Therefore, we get

$$\bar{d} \pm t_{n-1;\alpha/2}\frac{s_d}{\sqrt{n}} = -1.2 \pm 2.262\frac{2.097}{3.16} = -1.2 \pm 1.501 = (-2.701\,, 0.301)$$

Some calculations are shown in the following table:

Volunteer	Weight before	Weight after	d_i = after -before	$(d_i - \bar{d})^2$
1	80	80	80-80=0	$[0-(-1.2)]^2 = 1.44$
2	95	92	92-95=-3	$[-3-(-1.2)]^2 = 3.24$
3	78	77	77-78=-1	$[-1-(-1.2)]^2 = 0.04$
4	105	100	100-105=-5	$[-5-(-1.2)]^2 = 14.44$
5	90	91	91-90=-1	$[1-(-1.2)]^2 = 4.84$
6	90	89	89-90=-1	$[-1-(-1.2)]^2 = 0.04$
7	91	93	93-91=2	$[2-(-1.2)]^2 = 10.24$
8	79	77	77-79=-2	$[-2-(-1.2)]^2 = 0.64$
9	115	112	112-115=-3	$[-3-(-1.2)]^2 = 3.24$
10	85	85	85-85=0	$[0-(-1.2)]^2 = 1.44$
sum			-12	39.6

Problem 5. The scientists in problem 4b also selected a random sample of 10 overweight volunteers to use as control group. They suspected that just being part of a study at all might have an effect on weight loss. The body weights of the participants in the control group were also measured in the beginning of the study and then again after six months. Assume that the body weights are normally distributed in both groups and that the variances are equal. The weight changes (where a negative means a weight loss) and sample standard deviations can be found in the table below:

	Mean weight change	Standard Deviation	n
Treatment Group	-1.2	2.01	10
Control Group	-0.5	2.00	10

Test whether the treatment group loses more weight (more negative weight change, use treatment minus control) than the control group. Use 5% level of significance.

a. Find the decision rule of the test. (5p)

(A) $\boxed{\text{Reject } H_0 \text{ if } t_{obs} < -1.73}$

(B) Reject H_0 if $t_{obs} < -1.64$

(C) Reject H_0 if $|t_{obs}| > 1.64$

(D) Reject H_0 if $|t_{obs}| > 1.73$

(E) Reject H_0 if $|t_{obs}| > 1.96$

b. Find the value of the test variable. Choose the alternative closest to your answer. (5p)

(A) -0.55

(B) $\boxed{\text{-0.78}}$

(C) -1.56

(D) -2.14

(E) -2.28

Solution 5. Let x denote the treatment group and y, the control group. We have the following hypothesis

$$H_0 : \mu_x - \mu_y \geq 0 \qquad \text{vs.} \qquad H_1 : \mu_x - \mu_y < 0$$

Test Statistic:

$$t_{obs} = \frac{\bar{x} - \bar{y} - D_0}{s_p\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \qquad \text{where} \qquad S_p = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

a. The decision rule is to reject H_0 if $t_{obs} < -t_{crit}$ where $t_{crit} = t_{n_x+n_y-2;\alpha} = t_{18;0.05} = 1.734$.

b. First we need to find s_p^2,

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{9 \cdot 2.01^2 + 9 \cdot 2^2}{18} = 4.02.$$

Therefore, $s_p = 2$, and

$$t_{obs} = \frac{\bar{x} - \bar{y} - D_0}{s_p\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} = \frac{-1.2 - (-0.5) - 0}{2\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-0.7}{2\sqrt{\frac{2}{10}}} = -0.78$$

Problem 6. A candy manufacturer produces multi-colored button-shaped chocolates. There are five different colors: red, green, blue, yellow, and brown. The manufacturer claims that each color is equally frequent in production, but that the distribution of colors in each individual bag is random. A student collects a random sample of 300 candies and counts the colors. You can find the counts in the table below:

Red	Green	Blue	Yellow	Brown
72	65	67	52	44

Test at the 5% level of significance whether the population of candies is equally distributed between the five colors.

a. State your hypotheses, test statistic, critical value and decision rule. (5p)

b. Calculate the test variable. (5p)

c. State your conclusions and give a verbal interpretation. (5p)

d. Explain briefly what a type-I error is. Illustrate a situation in which this test would result in a type-I error. (5p.)

Solution 6.

a. Hypotheses: H_0: $P_{Red} = P_{Green} = P_{Blue} = P_{Yellow} = P_{Brown} = 0.2$
H_1: at least one $P_i \neq 0.2$, i=Red, Green, Blue, Yellow, Brown
Test variable:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1} = \chi^2_4$$

where O_i are the observed frequencies, $E_i = n \cdot P_i$ are the expected frequencies under H_0 (i.e. under the assumption that H_0 is true). The test variable is χ^2distributed with $K - 1 = 5 - 1 = 4$ degrees of freedom.
Critical value:

$$\chi^2_{crit} = \chi^2_{K-1;\alpha} = \chi^2_{5-1;0.05} = \chi^2_{4;0.05} = [Table4] = 9.488$$

Decision rule: We reject the null hypothesis at the 1% significance level if

$$\chi^2_{obs} > \chi^2_{crit}$$

b. Calculations:

	Red	Green	Blue	Yellow	Brown
O_i	72	65	67	52	44
E_i	60	60	60	60	60
$O_i - E_i$	12	5	7	-8	16
$(O_i - E_i)^2$	144	25	49	64	256
$\frac{(O_i - E_i)^2}{E_i}$	144/60=2.4	25/60=0.417	49/60=0.816	64/60=1.07	256/60=4.26

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = 2.4 + 0.417 + 0.816 + 1.07 + 4.26 = 8.963$$

c. Conclusion: Since $\chi^2_{obs} = 8.963 < \chi^2_{crit} = 9.488$ we cannot reject the H_0 at the 5% significance level. The population of candies is equally distributed.

d. Type I error is the probability of rejecting the null hypothesis when it is true. In this case this type of error would occur if, in fact, the different colors are produced in the exact same proportion but our sample was "unfortunate" and it had too many candies of the same color, e.g. 200 red candies.

Problem 7. An American businessman owns a chain of airport liquor stores. He wants to study the relationship between price and sales for a new brand of bourbon (a type of distilled alcoholic drink). The price varies between his airport stores. In some airports, he has paid for advertising billboards, in other airports, he has no advertising. He uses a random sample of 12 airports to estimate three models:
Model 1: $sales = \beta_0 + \beta_1 \cdot price + \epsilon$
Model 2: $sales = \beta_0 + \beta_1 \cdot price + \beta_2 \cdot ad + \epsilon$
Model 3: $sales = \beta_0 + \beta_1 \cdot price + \beta_2 \cdot ad + \beta_3 \cdot (price \cdot ad) + \epsilon$.

Model 1:

Multiple R	0.252
R square	0.064
Adjusted R square	-0.030
Standard error	6.868
Observations	12

	df	SS
Regression	1	32.029
Residual	10	471.637
Total	11	503.667

	Coefficients	Standard error
Intercept	44.873	11.943
Price	-0.971	1.178

Model 2:

Multiple R	
R square	
Adjusted R square	
Standard error	5.382
Observations	12

	df	SS
Regression	2	242.980
Residual	9	260.686
Total	11	503.667

	Coefficients	Standard error
Intercept	50.381	9.580
Price	-2.063	1.008
Ad	9.287	3.441

Model 3:

Multiple R	0.748
R square	0.559
Adjusted R square	0.394
Standard error	5.267
Observations	12

	df	SS	MS	F	p-value
Regression	3	281.702	93.091	3.384	0.075
Residual	8	221.965	27.746		
Total	11	503.667			

	Coefficients	Standard error
Intercept	64.111	14.933
Price	-3.556	1.603
Ad	-14.063	20.050
Price×Ad	2.402	2.034

a. Use Model 2 to find a point estimate for the number of bottles sold, given that the price of one bottle is \$20 and that the businessman is paying for advertising at that airport. Round to the nearest integer. (5p)

b. Formally test whether Model 3 is better than Model 2 (i.e. test whether or not the interaction term should be included, given that price and ad are included in the model). Use 5% level of significance. (5p)

c. Calculate and interpret the coefficient of determination for Model 2. (5p)

d. Perform an F-test of Model 3. State the hypotheses and use the p-value to reach a conclusion. Use 5% level of significance. (5p)

Solution 7.

a.

$$\widehat{sales} = b_0 + b_1 \cdot 20 + b_2 \cdot 1 = 50.381 - 2.063 \cdot 20 + 9.8287 = 18.408$$

b. Hypotheses: $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$:
Test variable:

$$t = \frac{b_3 - \beta_3^*}{s_{b_3}}$$

Critical value:

$$t_{crit} = t_{n-K-1;\alpha/2} = t_{12-3-1;0.05/2} = t_{8;0.025} = 2.306$$

Decision rule: We reject the null hypothesis at the 5% significance level if

$$|t_{obs}| \geq t_{crit}$$

Calculations:

$$t_{obs} = \frac{2.402 - 0}{2.034} = 1.18$$

Conclusions: Since $|t_{obs}| = 1.18 < t_{crit} = 2.306$ we fail to reject the H_0 at the 5% significance level and conclude that the interaction term between price and advertising should not be included in the model. Thus, model 2 is better than model 3.

c.

$$R^2 = \frac{SSR}{SST} = \frac{242.98}{503.667} = 0.4824$$

48.2% of the variation of y can be explained by the explanatory variables (price and advertising).

d. Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs H_1 : at least one $\beta_i \neq 0, i = 1, 2, 3$
Test variable:

$$F = \frac{\frac{SSR}{K}}{\frac{SSE}{n-K-1}}$$

$$p - value = 0.075 > \alpha = 0.05$$

Since p-value is higher than our significance level, we fail to reject H_0. Thus, we cannot reject the hypothesis that all the βs are not statistically significant from zero.