Answer form for multiple choice. You can make your own form, put please be clear and answer on one page. Do <u>not</u> submit solutions to the multiple-choice problems.

Number	Part	A	B	C	D	E
1	**a.**	☐	■	☐	☐	☐
1	**b.**	☐	☐	☐	■	☐
2	**a.**	☐	☐	☐	☐	■
2	**b.**	■	☐	☐	☐	☐
3	**a.**	☐	☐	☐	☐	■
3	**b.**	☐	■	☐	☐	☐
4	**a.**	☐	☐	■	☐	☐
4	**b.**	☐	☐	☐	■	☐
5	**a.**	☐	☐	☐	■	☐
5	**b.**	■	☐	☐	☐	☐

This is a draft of my suggested solutions. Please let me know if you find typos, errors, or if something is unclear: ulf.hognas@stat.su.se

PROBLEM 1

a. Find the inter quartile range (IQR) of the number of pets. *(5p)*

We need to locate the first and third quartiles. Then we will calculate the difference between the two. From the formula sheet:

Percentiles: Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the *ordered* sample, ordered by size from the smallest value $x_{(1)}$ to the largest $x_{(n)}$.

Let a = integer part of $(n+1)\frac{p}{100}$

Let b = decimal part of $(n+1)\frac{p}{100}$

p:te percentile $= x_{(a)} + b \cdot (x_{(a+1)} - x_{(a)})$

Ex.: 25:th percentile and $n = 9 \Rightarrow a = 2$ and $b = 0{,}5$

$\Rightarrow Q_1 = x_{(2)} + 0{,}5 \cdot (x_{(3)} - x_{(2)})$

First quartile

$(39+1)\frac{25}{100} = 10.0$

So, the first quartile will be the 10th ordered element.

Second quartile

$(39+1)\frac{75}{100} = 30.0$

The third quartile will be the 30th ordered element. We see that the 10th ordered element is a zero, since the thirteen smallest elements in the data are all zeros (13 out of 39 households have zero pets). Similarly, the 30th element is 2.

So,

$$IQR = 2 - 0 = 2$$

Answer: B

b. Find the sample correlation between the number cars and the number of pets. *Choose the alternative closest to your answer. (5p)*

Covariance: $$s_{xy} = Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Correlation: $$r_{xy} = Corr(x,y) = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$$

We need to find the covariance and then divide by the square root of the product of the variances, according to the formulas above.

Household	1	2	3	4	5	sum
Cars (x)	2	0	1	0	1	4
Pets (y)	4	2	1	0	2	9
x*y	8	0	1	0	2	11

I leave calculating the variances of x and y to you.

$s_x^2 = 0.7$
$s_y^2 = 2.2$

$\bar{x} = \frac{4}{5} = 0.8$
$\bar{y} = \frac{9}{5} = 1.8$
$\text{Cov}(x, y) = \frac{11 - 5 \cdot 0.8 \cdot 1.8}{5 - 1} = 0.95$
$r_{xy} = \frac{0.95}{\sqrt{0.7 \cdot 2.2}} \approx 0.77$

Answer: D

PROBLEM 2

A blood test for the disease Lycanthropy is not 100% reliable. A positive test is supposed to indicate that a patient is infected, while a negative test is supposed to indicate that the patient is not infected. The probability that the test is positive given that the patient is infected is 90%. The probability that the test is negative given that the patient is not infected is also 90%. Suppose that 5% of the population is infected.

a. ***Find the probability that a randomly selected ~~patient~~ member of the population tests negative (correctly or incorrectly).*** *Choose the alternative closest to your answer. (5p)*

Law of total probability:

$$P(A) = P(A \cap E_1) + \cdots + P(A \cap E_K)$$
$$= P(A|E_1)P(E_1) + \cdots + P(A|E_K)P(E_K) = \sum_{i=1}^{K} P(A|E_i)P(E_i)$$

The events "is infected" and "is not infected" is a partition of the sample space. So, we can calculate the probability that the randomly chosen member of the population tests negative as

$$\text{P}(Negative \mid Not\ infected) \cdot \text{P}(Not\ infected) + \text{P}(Negative \mid Infected) \cdot \text{P}(Infected)$$
$$= 0.9 \cdot 0.95 + 0.1 \cdot 0.05 = 0.86$$

Answer: E

A randomly selected patient is tested and the test is negative.

b. ***Find the probability that the patient is infected by the disease, despite the negative test result.*** *Choose the alternative closest to your answer. (5p)*

This is Baye's rule

Baye's rule: $$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)}$$

$$\mathrm{P}(Infected \mid Negative) = \frac{\mathrm{P}(Negative \mid Infected) \cdot \mathrm{P}(Infected)}{\mathrm{P}(Negative)}$$

$$= \frac{0.1 \cdot 0.05}{0.86} = 0.005813953$$

Answer: A

PROBLEM 3

A biologist estimates that the weight of a randomly chosen male red fox is normally distributed with mean 8 kg and standard deviation 2 kg.

a. ***Find the probability that a randomly chosen male fox weighs more than 11kg, according to the biologist's model.*** *Choose the alternative closest to your answer. (5p)*

The distribution of the weight X of a randomly chosen student is

$$X \sim N(8, 2)$$

We are asked to find $\mathrm{P}(X > 11)$. Standardize and calculate:

$$\mathrm{P}(X > 11) = \mathrm{P}\left(Z > \frac{11-8}{2}\right) = \mathrm{P}(Z > 1.5) = 1 - \mathrm{P}(Z < 1.5) = 1 - 0.93319 \approx 0.067$$

Answer: E

John and Jane invite 11 other couples to their New Year's Eve party. Assume that each of the 11 couples will answer "yes, we are coming to the party" with 80% probability and that each couple will answer independently of the other couples. If more than 8 of the couples answer "yes," there will not be enough space at the dinner table.

b. ***Find the probability that at most 8 of the 11 couples answer "yes."*** *Choose the alternative closest to your answer. (5p)*

Let X be the number of "yes" answers. Then $X \sim Bin(n = 11, P = 0.8)$. We seek $P(X \leq 8)$, so we need to use the table. But since $P > 0.5$, we need to use the transformation $Y = 11 - X$. Then $Y \sim Bin(n = 11, P = 0.2)$. From the table, we see that $P(X \leq 8)$ is the same as $1 - P(Y \leq 2)$. We can find $P(Y \leq 2)$ in the binomial table.

$$1 - P(Y \leq 2) = 1 - 0.61740 = 38\%$$

Answer: B

X	0	1	2	3	4	5	6	7	8	9	10	11
Y	11	10	9	8	7	6	5	4	3	2	1	0

PROBLEM 4.

a. ***Find a <u>90%</u> 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)*

This is difference, two independent samples where the sample sizes are both 30 or larger, so

– for the difference $\mu_X - \mu_Y$ (two independent samples)

$n_X, n_Y \geq 30$ regardless of distribution:	σ_X^2, σ_Y^2 known:	$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}$
	σ_X^2, σ_Y^2 unknown:	$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

It does not matter whether the variance is unknown or known, but since nothing is said about the variance being known, the problem text suggests that the standard deviation is calculated from the sample. Plug the numbers into one of these formulas to get:

$$(-2.49, 1.31)$$

Answer: C

An analyst at a marketing firm wants to survey Swedes about their opinions of the company Legendary Nuts, a big client of the marketing firm. The survey will include the question "do you have a favorable view of Legendary Nuts?" The analyst wants to find a <u>95%</u> confidence interval for the proportion of Swedes who would answer "yes."

b. ***What sample size is needed to guarantee a margin of error of at most 3% (3 percentage points)?*** *Assume 100% response rate. (5p)*

[…]

Hint: what proportion of "yes" answers should we assume when solving this problem?

Since we want to guarantee that the margin of error is at most 3%, we need to find a sample size such that the margin or error is 3% or less, no matter what $\hat{p}$ turns out to be in the sample. The formula for the confidence interval is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the margin of error is the term:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

As discussed in class, for a fixed n, this is maximized when $\hat{p} = 0.5$. We solve the equation:

$$0.03 \geq 1.96\sqrt{\frac{0.5(1-0.5)}{n}}$$

[…]

$$n \geq \frac{0.25}{\left(\frac{0.03}{1.96}\right)^2} = 1067.11$$

Answer: D

PROBLEM 5

Treat these sales figures as an independent identically distributed sample from the local population. Test at the 5% level of significance whether the distribution of sizes have changed compared to previous years.

a. ***Find the critical value.*** *(5p)*

We are asked to test whether the population proportions are 20% small, 50% medium, and 30% large, or not. This is a goodness-of-fit test:

STATISTICAL INFERENCE – TEST VARI

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad \text{where } E_i = nP_i$$

The test variable is Chi-square distributed with $K-1$ degrees of freedom where K is the number of categories, so $K = 3$ and $K - 1 = 2$.

$$\chi^2_{crit} = 5.991$$

Answer: D

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

We make a table

	Small	Medium	Large	Sum
Pi	0.2	0.5	0.3	1
Ei=nPi	40	100	60	200
Oi	39	108	53	200
Oi-Ei	-1	8	-7	0
(Oi-Ei)^2	1	64	49	114
(Oi-Ei)^2/Ei	0.025	0.64	0.816667	1.481667

Answer: A

PROBLEM 6

Assume that the score of each individual student is (approximately) normally distributed, both before and after taking the course. ***Test at 5% level of significance whether taking the course is associated with an increase (score after minus score before) in average score of at least 10.***

a. State the hypotheses, test variable, critical value and decision rule. *(5p)*

This is a t-test of paired differences

Hypotheses:
H_0: $\mu_d < 10$
H_1: $\mu_d \geq 10$

Test variable:

$$t_{n-1} = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}}$$

Critical value:
$n = 6$
$n - 1 = 5$
$\alpha = 0.05$
$t_{5;\,0.05} = 2.015$

Decision rule:
Reject the null if $t_{obs} > 2.015$.

b. Calculate the test variable and interpret the outcome of the test. (5p)
Find the mean difference

$$\bar{\boldsymbol{d}} = \frac{110}{6} = 18.333$$

We see that this is higher than 10, so this could be significant evidence in favor of the alternative.

Make a table:

Student	**1**	**2**	**3**	**4**	**5**	**6**	**Sum**
Score before	60	78	75	63	95	77	
Score after	79	105	87	77	117	93	
$\boldsymbol{d}$	19	27	12	14	22	16	110
$\boldsymbol{d - \bar{d}}$	0.666667	8.666667	-6.33333	-4.33333	3.666667	-2.33333	7.11E-15
$\boldsymbol{(d - \bar{d})^2}$	0.444444	75.11111	40.11111	18.77778	13.44444	5.444444	153.3333

$s^2 = \frac{153.3333}{6-1} = 30.667$

Now we are ready calculate the test variable:

$$t_{obs} = \frac{18.333 - 10}{\sqrt{30.667/6}} = 3.686$$

Outcome or conclusion: Since $t_{obs} = 3.686 > 2.015$, we reject the null.

Interpretation: We have found significant evidence for the hypothesis that taking the course is associated with an average score increase of at least 10.

c. Use the formula sheet to find an approximate p-value of the test. Briefly explain why we can use the p-value to determine to outcome of the test. (5p)

v	α = 0.1	0.05	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032

Looking at table 3, we see that the observed test variable is between $\alpha = 1\%$ and $\alpha = 0.5\%$. Since alphas (0.1, 0.05, 0.025, …) of table 3 gives us the right tail of the t-distribution, we know that the tail to the right of t_{obs} must be between these values. The tail to the right of t_{obs} is the same as the p-value for this one-sided test, so we know that the p-value is between 0.5% and 1%.
If that tail is smaller than the critical region, we know that t_{obs} is inside the critical region, so we can reject the null whenever the p-value is smaller than α.

For part d and part e, assume that the <u>true average (population) increase is 10 and that the population variance is 100</u>. The company decides to conduct another study with six new students; otherwise, the study is identical to the first.

d. Find the probability that a randomly chosen student does <u>not improve</u> their score after taking part in the study. *(5p)*

According to the problem, the distribution of the increase D of a randomly chose student is

$$D \sim N(10, 100)$$

We seek $\mathrm{P}(D > 0)$. Standardize and calculate:

$$\mathrm{P}(D > 0) = \mathrm{P}\left(Z > \frac{0 - 10}{10}\right) = \mathrm{P}(Z > -1) = 1 - \mathrm{P}(Z < 1) = 1 - 0.84134 \approx 0.1587$$

e. ***Find the probability that none of the six students improve their score after taking part in the study.*** *(5p)*

Use the probability from part d and independence to get

$$0.1587^6 \approx 0.000016$$

PROBLEM 7

a. ***Find the estimated price of a diamond according to model 2, given that the diamond weighs 1 carat, has a clarity rating that is <u>less good</u> than second best, and a "premium" cut.*** *(5p)*

$$\widehat{PRICE} = -2592.874 + 7253.921 \cdot 1 + 1086.982 \cdot 0 + 485.726 \cdot 1 = 5146.773$$

The estimated price of such a diamond is $5147, according to the model.

For part b and part c, you are asked to test at the <u>1% level of significance</u> whether $\beta_2 > 500$, given that CARAT and CUT is included in the model.

b. ***Clearly state hypotheses, test variable, critical value and decision rule.*** *(5p)*

Hypotheses:
H_0: $\mu_d < 500$
H_1: $\mu_d \geq 500$

Test variable:

$$t_{n-K-1} = \frac{b_j - \beta_j^*}{s_{b_j}}$$

Critical value:
$n = 100$
$K = 3$
$n - K - 1 = 96$
$\alpha = 0.01$
$t_{96;\,0.01} \approx t_{95;\,0.01} = 2.366$

Decision rule:
Reject the null if $t_{obs} > 2.366$.

c. ***Use the output from MODEL 2 to calculate the test variable. Clearly state the conclusion and interpretation of the test result.*** *(5p)*

From the problem text and the output:
$\beta_2^* = 500$
$s_{b_2} = 270.817$
$b_2 = 1086.982$
$t_{obs} = \frac{1086.982 - 500}{270.817} = 2.167$

Conclusion: Since $2.167 < 2.366$, we fail to reject the null at the 1% level of significance.

Interpretation: We have not found sufficient evidence that the price effect of "best or second-best clarity" is more than plus $500, with CARAT and CUT being held equal.

Note: this model would likely be better with **interaction terms** between carat and clarity, and carat and cut. It is reasonable to assume that good clarity and good cut contributes more to the price of large diamond than it does to small diamonds. The current model does not capture this relationship.

d. Find the adjusted coefficient of determination for both models. Interpret the results and the difference in results. *(5p)*

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

We can get SSE and SST from the output.
MODEL 1

$$SSE = 112036353.3$$
$$SST = 1078766503$$
$$R^2_{adj1} = 1 - \frac{\frac{112036353.3}{100-4-1}}{\frac{1078766503}{100-1}} = 0.89177$$

MODEL 2
This is found in the same way

$$R^2_{adj2} = 0.89288$$

The coefficient of determination measures how much of the variation in the dependent variable, here this is price, is explained by the model. So, we see that both the models explain around 89% of the variation in price, which is good.
The adjusted coefficient of determination takes into account how many variables you have in your model. If you add unnecessary independent variables, the adjusted coefficient of determination can decrease. This is not what we see here, so by this metric, MODEL 2 is better than MODEL 1.

e. Note that the coefficient for COLOR is negative, even though a "1" means "best or second-best color." Use your statistics knowledge to explain how this is possible. *(5p)*

To explain this, you can just to compare the t-stat of of COLOR and compare it to the coefficient. The standard error: $s_{b_4} = 241.200$ is much larger than the coefficient which is -28.244. There is a lot of uncertainty here. We can also calculate a 95% confidence interval for the slope:

$$t_{95;\,0.025} = 1.985$$

$$-28.244 \pm 1.985 \cdot 241.200$$

Which gives us the interval

$$(-507, 451)$$

We can say with 95% confidence that the true relationship between COLOR and price is between these values. In other words: the estimated model does not tell us what the true sign of the coefficient is. Most likely, a larger sample would give us a positive coefficient for COLOR.