DRAFT SOLUTIONS, PROBLEMS 1, 5, 9, 13, 17, and 21

Please let me know if you find any errors or typos. My solutions are more detailed than your answers had to be.

1. a. *Find the conditional probability that zero watches are sold given zero pieces of jewelry sold.* (5p.)

We can call number pieces of jewelry sold X and number of watches sold Y. When we condition on zero pieces of jewelry sold, we limit the sample space to only the events where this happened:

		Jewelry			
		0	1	2	**Sum**
Watches	0	0,2	0,1	0,05	0,35
	1	0,1	0,2	0,1	0,4
	2	0,05	0,1	0,1	0,25
	Sum	0,35	0,4	0,25	1

Given that zero pieces were sold, what is the probability that zero watches were sold? It is the probability that both happen divided by the probability that zero watches were sold. Formally:

$$\mathrm{P}(Y = 0 \mid X = 0) = \frac{\mathrm{P}(\mathrm{Y} = 0 \cap X = 0)}{\mathrm{P}(X = 0)} = \frac{0{,}2}{0{,}35} \approx 0{,}571$$

b. *Find the covariance between the number of watches sold and the number of pieces of jewelry sold.* (5p.)

This is perhaps one of the harder problems in the whole exam. We are going to use the formula for covariance

$$\mathrm{Cov}(\mathrm{X},\mathrm{Y}) = \sum_{x=0}^{2}\sum_{y=0}^{2} x\,y\,\mathrm{P}(x,y) - \mu_X\mu_Y$$

$$\sum_{x=0}^{2}\sum_{y=0}^{2} x\,y\,\mathrm{P}(x,y) =$$
$$0\cdot 0\cdot \mathrm{P}(0{,}0) + \ 0\cdot 1\cdot \mathrm{P}(0{,}1) + \ 0\cdot 2\cdot \mathrm{P}(0{,}2) +$$
$$1\cdot 0\cdot \mathrm{P}(1{,}0) + \ 1\cdot 1\cdot \mathrm{P}(1{,}1) + \ 1\cdot 2\cdot \mathrm{P}(1{,}2) +$$
$$2\cdot 0\cdot \mathrm{P}(2{,}0) + \ 2\cdot 1\cdot \mathrm{P}(2{,}1) + \ 2\cdot 2\cdot \mathrm{P}(2{,}2)$$

The easiest way to calculate the double sum is to use a table

	0	1	2
0	0	0	0
1	0	0,2	0,2
2	0	0,2	0,4

And we see that the sum is 1.

We need μ_X and μ_Y but they are the same.

$$\mu_X = \mathrm{E}[\mathrm{X}] = \sum_{x=0}^{2} x\,\mathrm{P}(x) = \ 0\cdot \mathrm{P}(0) + \ 1\cdot \mathrm{P}(1) + \ 2\cdot \mathrm{P}(2)$$
$$= 0 + 0{,}4 + 2\cdot 0{,}25 \ = \ 0{,}4 + 0{,}5 = 0{,}9$$

So $\mu_X = \mu_Y = 0{,}9$ and

$$\mathrm{Cov}(X,Y) = \ 1 - \ 0{,}9\cdot 0{,}9 = 0{,}19$$

Getting this far will give you most of the points.
Now we need the variances; they are also the same for X and Y

Var(Y) = Var(X) = $\sum_{i=0}^{2} x^2\,\mathrm{P}(x) - {\mu_X}^2 = 0^2\cdot 0{,}35 + \ 1^2\cdot 0{,}4 + 2^2\cdot 0{,}25 - 0{,}9^2 = 0{,}59$

So

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{0{,}19}{\sqrt{0{,}59\cdot 0{,}59}} = \frac{0{,}19}{0{,}59} \approx 0{,}32$$

c. *Find the probability that profits will be more than 300, but less than 600, according to the owner's model.* (5p.)

In the problem, we are told that X = profit and that $X \sim N(400, 100^2)$

We want to find

$$\mathrm{P}(300 < X < 600)$$

Standardize

$$\mathrm{P}\left(\frac{300-400}{100} < Z < \frac{600-400}{100}\right) = \mathrm{P}(-1 < Z < 2)$$

Draw and use the symmetries of the normal distribution

$$\mathrm{F}(2) - \big(1 - \mathrm{F}(1)\big) = \mathrm{F}(2) + \mathrm{F}(1) - 1$$

Use table 1:

$$\mathrm{F}(2) = 0{,}97725\ ;\ \mathrm{F}(1) = 0{,}84134$$

So

$$\mathrm{F}(2) + \mathrm{F}(1) - 1 = 0{,}97725 + 0{,}84134 - 1 \approx 0{,}82$$

5. a. *Suppose that 90% of all Swedes are right-handed and that we draw a simple random sample of 18 Swedes. Find the probability that at most 16 of the Swedes in our sample are right-handed.* (5p.)

We can regard getting a right-handed person as "success" and the probability is the same for each element in our sample. Hence if X is the number of righthanded people, X is binomially distributed

$$X \sim Bin(18; 0{,}9)$$

Make a table. We seek $\mathrm{P}(X \leq 16)$ which is the orange event in the table:

X	0	1	2	3	…	11	12	13	14	15	16	17	18
Y	18	17	16	15	…	7	6	5	4	3	2	1	0

Since the probability of success is over 50%, we cannot use the table directly. We have to use the transformation $Y = 18 - X$
We see that at most 16 right-handed is the same as at least 2 left-handed (or not-right-handed). So

$$\mathrm{P}(X \leq 16) = \mathrm{P}(Y \geq 2) = 1 - \mathrm{P}(Y \leq 1)$$

Since $Y \sim Bin(18; 0{,}1)$ we can get this from the table (page 18):

$$1 - \mathrm{P}(Y \leq 1) = 1 - 0{,}45028 \approx 0{,}55$$

b. *A television production company has produced a new reality TV-show. They show the first "pilot episode" to a random sample of 16 potential viewers, to find out if the show will be a success or not. Suppose that a randomly selected viewer has a 20% probability of liking the episode. What is the probability that at most 4 viewers in the sample like the show?* (5p.)

X	0	1	2	3	4	5	…	12	13	14	15	16

Here we seek the yellow part, fewer than 4 successes. We can get this straight from the binomial table (page 17)

$$\mathrm{P}(X \leq 4) = 0{,}83577 \approx 0{,}84$$

c. *You decide to flip a coin 50 times for some reason. Each time, the coin will come up heads with probability 0.5 and tails with probability 0.5. Find the probability that your coin will come up heads exactly 25 times.* (5p.)

The number of heads is binomially distributed.

First alterative: use the formula on page 4:

$$\mathrm{P}(X = x) = \binom{n}{x} P^x (1-P)^{n-x}$$

$$\mathrm{P}(X = 25) = \binom{50}{25} 0{,}5^x (1-0{,}5)^{50-x} = \frac{50!}{25!\,25!} 0{,}5^{25}(1-0{,}5)^{50-25} = \frac{50!}{25!\,25!} 0{,}5^{25} 0{,}5^{25}$$

Put this in your calculator to get the probability 0,112 or 11,2%

Second alternative. Use the normal approximation of the binomial distribution. Then if X is the number of heads:

$$\mathrm{E}[X] = nP = 50 \cdot 0{,}5 = 25;\ \mathrm{Var}(X) = nP(1-P) = 50 \cdot 0{,}5 \cdot 0{,}5 = 12{,}5$$

Find P(X ≤ 25).

$$\mathrm{P}(X \leq 25) = \mathrm{P}\left(Z \leq \frac{25-25}{\sqrt{12{,}5}}\right) = \mathrm{P}(Z \leq 0) = 0{,}5$$

Then find Find P(X ≤ 24).

$$\mathrm{P}(X \leq 24) = \mathrm{P}\left(Z \leq \frac{24-25}{\sqrt{12{,}5}}\right) = \mathrm{P}(Z \leq -.28) = 1 - \mathrm{P}(Z \leq .28) = 0{,}61026$$

Then $\mathrm{P}(X = 25) = \mathrm{P}(X \leq 25) - \mathrm{P}(X \leq 24) = 0{,}61026 - 0{,}5 = 0{,}110 =$ 11,0%

This time the approximation was pretty close to the true probability, but this is not always the case. Search “continuity correction” for a better method.

9. a. *Assuming that both samples were representative, find a 90% confidence interval for the change in proportion of students who have used cannabis in the last 12 months.* (5p.)

Apply the formula on page 5

$$\hat{p}_x - \hat{p}_y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

$\hat{p}_x = 0.11$

$\hat{p}_y = 0.13$

α = 0,10

$\alpha/2$ = 0,05

$z_{\alpha/2} = z_{0,05} = 1{,}6449$

$n_x = n_y = 200$

Putting it all together we get $(-0{,}073;\ 0{,}033)$ or (-7,3%; 3,3%)

So in this case, the small sample sizes mean that we do not know much about the actual change.

b. *In a larger survey the proportion of students who answered "Yes" to the same question was 12% and the 95% margin of error was less than 1%. Find the minimum sample size that could have been used.* (5p.)

The formula used to find this confidence interval is

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

Where this is the margin of error:

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

So set the margin of error to be smaller than 1% and then we solve for n

$$0{,}01 > 1{,}96\sqrt{\frac{0{,}12(1-0{,}12)}{n}}$$

$$\frac{0{,}01}{1{,}96} > \sqrt{\frac{0{,}12(1-0{,}12)}{n}}$$

$$\left(\frac{0{,}01}{1{,}96}\right)^2 > \frac{0{,}12(1-0{,}12)}{n}$$

$$n > \frac{0{,}12(1-0{,}12)}{\left(\frac{0{,}01}{1{,}96}\right)^2}$$

$$n > 4056.7$$

So they must have used a sample size of at least 4057 to get a margin of error that of less than 1%.

13. a. *State the hypotheses and the decision rule* (5 p.)

We can call the treatment group x and the control group y

$$H_0: \mu_x - \mu_y = 0$$

$$H_1: \mu_x - \mu_y > 0$$

Decision rule:

- small sample sizes
- unknown variance ("standard deviation is calculated from the sample")
- normal distribution

So the test variable will be *t*-distributed (see page 6):

σ_X^2, σ_Y^2 unknown and assumed equal:

$$t_{n_x+n_y-2} = \frac{\bar{X}-\bar{Y}-D_0}{S_p\sqrt{1/n_x+1/n_y}}$$

$$\text{where } s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$

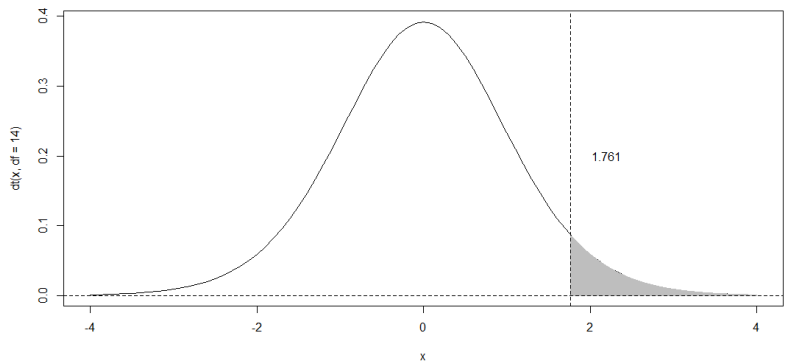
(as it says in the problem, we can also assume that the population variances are equal)

The degrees of freedom is equal to $n_x + n_y - 2 = 8 + 8 - 2 = 14$

α = 0,05

$t_{0,05;14} = 1.761$

Rule: reject if $t_{obs} > 1.761$



It is a good idea to draw a picture, but you do not have to.

b. *Calculate the test variable and state the outcome of the test* (5p.)

So as mentioned above, we will use

σ_X^2, σ_Y^2 unknown and assumed equal:

$$t_{n_x+n_y-2} = \frac{\bar{X}-\bar{Y}-D_0}{S_p\sqrt{1/n_x+1/n_y}}$$

$$\text{where } s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$

$\bar{x} = 76$
$\bar{y} = 74$
$D_0 = 0$
$s_x^2 = 2{,}8^2$
$s_x^2 = 2{,}6^2$
$n_x = n_y = 8$

So we get

$s_p^2 = \frac{(8-1)2{,}8^2 8+(8-1)2{,}6^2}{8+8-2} = \frac{2{,}8^2+2{,}6^2}{2} = 7{,}3$

And

$t_{obs} = \frac{76-74}{\sqrt{7{,}3}\sqrt{\frac{1}{8}+\frac{1}{8}}} = 1{,}48$

Conclusion: since $t_{obs} = 1{,}48$, which is not greater than 1.761, we fail to reject the null. We have not found statistically significant evidence that access to a hamster wheel improves the VO2-max in mice.

This problem was inspired by the paper:

Davidson SR, Burnett M, Hoffman-Goetz L. Training effects in mice after long-term voluntary exercise. *Med Sci Sports Exerc*. 2006;38(2):250-255.

This study spanned more weeks and the sample sizes were larger. They found significant evidence that access to a hamster wheel increases VO2-max in mice.

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

The appropriate test is a Chi-square test, a so-called independence test. The idea is that if website version is independent of how users rate the experience, then we should see the same kind of distribution of ratings for each for each of the three versions.

Hypotheses:

H_0: user experience and version are independent

H_1: user experience and version are dependent

Test variable

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{R_i C_j}{n}$$

The degrees of freedom is equal to (number of rows – 1)*(number of columns – 1) = 8 and α=0,05.

Critical value

$X^2_{8;0,05} = 15.507$

Chi-square test are always one-sided in this course.

Decision rule: Reject the null if $X^2_{obs} > 15{,}507$.

b. The test variable is best calculated by making three tables:

Remember that the *i*'s are the rows and the *j*'s are the columns. Start by taking the observed frequencies and find the row sums and column sums:

O_{ij}-table:

	light	**dark**	**blue**	**C_j**
1	12	9	9	30
2	20	12	28	60
3	30	34	26	90
4	11	18	10	39
5	7	7	7	21
R_j	80	80	80	n=240

Then make a table for $E_{ij} = \frac{R_i C_j}{n}$ where R_i is the row sum and C_j is the column sum. I choose the values in your problem carefully, so that you would get all whole numbers in this table.

E_{ij}-table:

	C1	C2	C3
R1	10	10	10
R2	20	20	20
R3	30	30	30
R4	13	13	13
R5	7	7	7

Last, we use these first two tables to find the table of values from the double sum.

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

	C1	C2	C3	
R1	0,4	0,1	0,1	
R2	0	3,2	3,2	
R3	0	0,533333	0,533333	
R4	0,307692	1,923077	0,692308	
R5	0	0	0	
				10,98974

In the corner, I have calculated the sum of the last table, so $X^2_{obs} \approx 10{,}99$

Conclusion: Since X^2_{obs} is not larger than $15{,}507$, we fail to reject the null. We have not found statistically significant evidence at the 5% level that user experience is dependent of website version.

c. *Before the study started, the developer worried that the study would result in a Type-2 error. Someone suggested that this would be less likely to happen with a 1% significance level instead of 5%. Does this make sense? Explain.* (5.)

A Type-2 error is the error of failing to reject the the null hypothesis when the null is true. If we lower the level of significance to 1%, we need stronger evidence to reject the null, so this suggestion makes no sense (the critical value would be 20,090 at the 1% level).

d. *Make a table of the conditional relative frequencies of user ratings. Condition on website version.* (5p)

	light	dark	blue
1	12	9	9
2	20	12	28
3	30	34	26
4	11	18	10
5	7	7	7
Sum	80	80	80

When we condition on version, we consider each version as the whole. We then ask, what proportion of within this group, e.g. the light version, belongs to rating 1, 2, 3, and so on.

	light	dark	blue
1	12/80	9/80	9/80
2	20/80	12/80	28/80
3	30/80	34/80	26/80
4	11/80	18/80	10/80
5	7/80	7/80	7/80
Sum	80/80	80/80	80/80

Then convert to percentage (or fraction):

	light	dark	blue
1	15,0%	11,3%	11,3%
2	25,0%	15,0%	35,0%
3	37,5%	42,5%	32,5%
4	13,8%	22,5%	12,5%
5	8,8%	8,8%	8,8%
R_j	100,0%	100,0%	100,0%

We see that 15% of people who saw the “light theme” version gave the website a rating of 1 out of 5. Whereas only 11,3% of users who saw the “dark theme” version gave the worst rating.

21. a. *Use the estimated MODEL 1 to find a 95% prediction interval for the fuel consumption of a truck given a cargo weight of 18. Interpret the result* (5p.)

We will use

Prediction interval for the prediction of y given $X = x$:

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

We get b_0, b_1, n, and s_e^2 from the output

$$b_0 = 7{,}921$$
$$b_1 = 1{,}825$$
$$n = 34$$
$$s_e^2 = MSE = 5{,}24$$

We are also given

$$\bar{x}_1 = 15{,}5 \; ; \; s_x^2 = 7{,}8$$
$$x = 18$$

And finally

$$n - 2 = 34 - 2 = 32$$
$$\alpha = 0{,}05$$
$$\frac{\alpha}{2} = 0{,}025$$
$$t_{n-2;\alpha/2} = t_{32;0{,}025} \approx 2{,}042$$

We do not have 32 degrees of freedom in the table, so you can use either 30 or 35 (I used 30).

We get

$$(7{,}921 + 1{,}825 \cdot 18) \pm 2{,}042 \sqrt{5{,}24 \left(1 + \frac{1}{34} + \frac{(18 - 15{,}5)^2}{(34-1)7{,}8}\right)}$$

This gives: (35,97 ; 45,57)

Interpretation: There is a 90% probability that a randomly chosen truck will consume between 35,98 and 45,56 liters gasoline /100 km, given that the cargo weight is 18,000 kg.

b. *Find the coefficient of determination of MODEL 1. Interpret the result.* (5p.)

We will use

Coefficient of determination: $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

From the output:

ANOVA

	df	*SS*	*MS*
Regression	SSR 1	965,1439479	965,1439479
Residual	32	2276,359973	71,13624915
Total	SST 33	3241,503921	

$$R^2 = \frac{965{,}1}{3242} \approx 29{,}8\%$$

Interpretation: cargo weight explains 29,8% of the variation in fuel consumption.

c. *Test at the 5% level of significance whether* $\beta_2 > 0$, *given that cargo weight is included in the model. Clearly state hypotheses, test variable, critical value and decision rule, calculations, and conclusion.* (10p.)

Hypotheses:

$$H_0\colon\ \beta_2 = 0 \mid \beta_1 \neq 0$$
$$H_1\colon\ \beta_2 > 0 \mid \beta_1 \neq 0$$

Test variable

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

Critical value

$$n - K - 1 = 34 - 2 - 1 = 31$$
$$\alpha = 0{,}05$$
$$t_{31;0{,}05} \approx 1{,}697$$

Decision rule

Reject the null if $t_{obs} > 1{,}697$

Calculations (values from the output)

$$t_{obs} = \frac{17{,}917 - 0}{0{,}8918} \approx 20{,}09$$

Conclusion: since $t_{obs} = 20{,}09 > 1{,}697$, we reject the null. We find that $\beta_2 > 0$, i.e. that hilly roads increase fuel consumption when cargo weight is held constant.

d. *Explain why one might include this new term. What is the interpretation of β_3 ?* (5p.)

An interaction term like this gives us a different slope for cargo weight depending on whether the road is hilly or not. The coefficient β_3 tells us how much more fuel consumption increases with cargo weight when the road is hilly compared to when it is not hilly.

When the road is not hilly, $\beta_2 = 0$:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 0 + \beta_3 \cdot x_1 \cdot 0 = \beta_0 + \beta_1 x_1$$

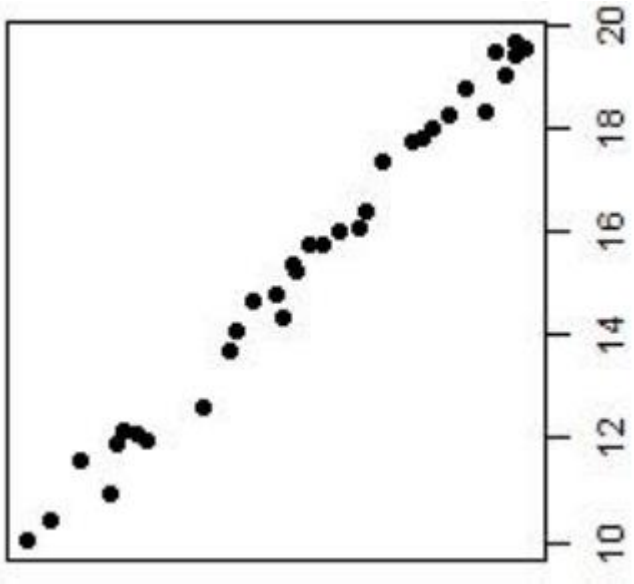
When the road is hilly, $\beta_2 = 1$:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 1 + \beta_3 \cdot x_1 \cdot 1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$$

We see that the intercept changes by β_2 and the slope by β_3, when the road is hilly. If we believe that each extra 1000 kg cargo weight may have more of an effect on fuel consumption when the road is hilly, we might want to try to include this interaction term.

e. *Finally, the analyst considers to include a fourth variable x_4. She decides against including x_4 in her model after studying the scatter plots in the figure "scatter plots" below. Explain why.* (5p.)

The scatter plots reveal this relationship between x1 and x4:



This suggests that x_4 and x_1 are strongly correlated. Including both them causes problems of **multicollinearity**, so this variable should be left out. The variable $x_1$is cargo weight, so maybe x_4 is total weight? Either way, we should not include both.