Copenhagen Business School

Exam in Statistics

HA-IB 2nd year
BSc International Shipping and Trade 1st year

Friday 22 January 2016 9:00-13:00

Solutions

Problem 1

Put

$$T = \textit{smokes tobacco}$$
$$H = \textit{smokes hashish}$$

1.

$$P(H|T) = \frac{P(H \textbf{ and } T)}{P(T)}$$
$$\Rightarrow \qquad P(T) = P(H \textbf{ and } T)/P(H|T) = \frac{0.08}{0.71} = \frac{8}{71} = 0.1127$$

2.

$$P(T|H) = \frac{P(T \textbf{ and } H)}{P(H)} = \frac{P(T \textbf{ and } H)}{P(H|T)P(T) + P(H|\textbf{not } T)P(\textbf{not } T)}$$
$$= \frac{0.08}{0.71 \cdot \frac{8}{71} + 0.02 \cdot (1 - \frac{8}{71})} = \frac{8}{8 + 2 \cdot 63/71} = 0.8184$$

Problem 2

Probabilities:

$$P(100) = P(\text{win 100 DKK}) = P(\text{exactly one 6}) = \binom{2}{1} \cdot \frac{1}{6} \cdot \frac{5}{6} = 2 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{10}{36}.$$

or

$$
\begin{aligned}
P(0) = P(\text{win 0 DKK}) = P(\text{zero 6s}) &= \left(\frac{5}{6}\right)^2 = \frac{25}{36} \\
P(300) = P(\text{win 300 DKK}) = P(\text{two 6s}) &= \left(\frac{1}{6}\right)^2 = \frac{1}{36} \\
P(100) = P(\text{win 100 DKK}) = 1 - P(0) - P(300) &= 1 - \frac{25}{36} - \frac{1}{36} = \frac{10}{36}
\end{aligned}
$$

Thus the probability of winning 100 Danish kroner equals 10/36 ($\approx 27.77\%$)

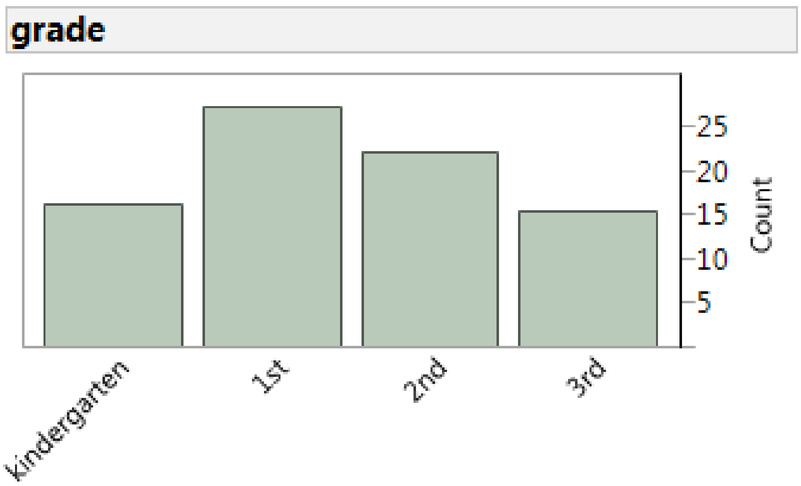
Expected winnings:

$$
\begin{aligned}
\mu &= \sum xP(x) = 0P(0) + 100P(100) + 300P(300) \\
&= 100 \cdot \frac{10}{36} + 3000 \cdot \frac{1}{36} = \frac{1000 + 300}{36} = \frac{1300}{36} = 36.11
\end{aligned}
$$

i.e. 36.11 Danish kroner.

Problem 3

Bar plot:



Confidence intervals:

Kindergarten	$]2.608; 3.527[$
1st grade	$]3.589; 4.518[$
2nd grade	$]4.676; 6.074[$
3rd grade	$]5.027; 6.397[$

Problem 4

$$\hat{p}_1 = \frac{103}{103+57} = 0.64375$$
$$\hat{p}_2 = \frac{56}{56+52} = 0.51852$$

Confidence interval for the difference is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\hat{p}_1(1-\hat{p}_1)/(103+57) + \hat{p}_2(1-\hat{p}_2)/(56+52)} = \ ]0.00529; 0.24518[$$

Thus the difference is between 0.005 and 0.245 with 95% confidence.

As 0 is not contained in the confidence interval the probability of having a psychosis is significant larger for cannabis users, who have started using cannabis when younger than 17 years old.

Problem 5

As each student tries to solve both problems, we have paired data. The "expanded data set" is[1]

*P(3)**P(4)*	Correct	Wrong	Total
Correct	**30**	*19*	**49**
Wrong	*4*	92	96
Total	**34**	111	**145**

McNemar's test statistic equals $(4-19)/\sqrt{4+19} = -15/\sqrt{23} = -3.127$. It is approximately normally distributed under the null hypothesis of no difference, which gives us a p-value equal to 0.0018 (0.00176 exact). Thus we reject the null hypothesis and conclude that *P(4)* is significantly more difficult to solve than *P(3)*.

The problems P(3) and P(4) dates back to 1560 when Cardano wrote down an incorrect solution to both problems. P(3) is the following problem: Roll 3 fair dice. What is the probability that you by adding some (or all) of the values you can obtain the value of 3? The answer to this is 116/216. P(4) is the probability of being able to get a sum of 4 when rolling 3 dice; the correct answer is 131/216.

Problem 6

The difference between the two means equals 1.156.

Assuming same standard deviation in the two groups, we get a standard error of 0.67765 and a test statistic of 1.7059. The test statistic is t-distributed with 161 degrees of freedom yielding a p-value of 0.08996.

Not assuming same standard deviations, we get a standard error of 0.69594 and a test statistic of 1.6611. The test statistic is t-distributed with 129.66 degrees of freedom yielding a p-value of 0.09912.

[1] bold face numbers are from the problem text, italics are the numbers needed for the test

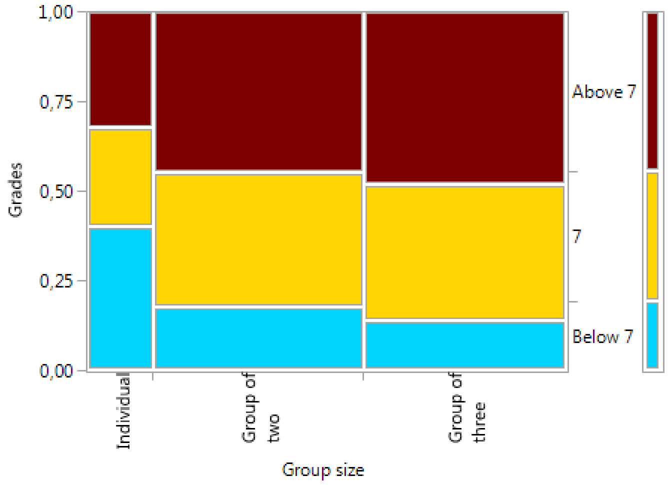
Using tables (100 degrees of freedom) we get a p-value between 5% and 10%; the corresponding quantiles equal 1.984 and 1.660.

We do not reject the null hypothesis of no difference and may conclude thatit seems that the average grade does not differ between female and male students.

Problem 7

Contingency Analysis of Grades By Group size

Mosaic Plot



Freq: Count

Contingency Table

Group size / Row %	Below 7	7	Above 7	Total
Individual	40,32	27,42	32,26	
Group of two	17,77	37,56	44,67	
Group of three	13,76	38,10	48,15	

Tests

N	DF	-LogLike	RSquare (U)
448	4	9,4743978	0,0202

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	18,949	0,0008*
Pearson	21,794	0,0002*

The Pearson test statistic equals 21.794. It is approximately χ^2-distributed under the null hypothesis of no dependence with 4 degrees of freedom. The p-value equals 0.0002 leading us to reject the null hypothesis. Hence we conclude that the exam grades depend on the number of students in the group.

Based on the mosaic plot or the proportions, it is evident that students having completed their hand-ins individually typically get lower grades (40% below 7) whereas students working in groups tend to get higher grades (only 18% and 13% get a grade below 7); there is little difference between groups of two and groups of three, though the latter get higher grades.

Problem 8

1. The oneway ANOVA F-test statistic equals 16.174. Under the null hypothesis of no difference, the test statistic is F-distributed with (3,76) degrees of freedom. The p-

value is smaller than 0.0001, so we reject the null hypothesis and conclude that the results of the math test depends on which grade the students are in.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
grade	3	77,01999	25,6733	16,1736	<,0001*
Error	76	120,63917	1,5874		
C. Total	79	197,65916			

Looking at the estimates, it is clear that the results increase as the students progress through elementary school. A Tukey-test shows that the results are significantly lower in kindergarten and 1st grade than they are in 2nd and 3rd grade, but we cannot with certainty conclude that the results differ between kindergarten and 1st grade, nor between 2nd and 3rd grade.

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
kindergarten	16	3,06738	0,31498	2,4400	3,6947
1st	27	4,05333	0,24247	3,5704	4,5363
2nd	22	5,37477	0,26861	4,8398	5,9098
3rd	15	5,71200	0,32531	5,0641	6,3599

Std Error uses a pooled estimate of error variance

Comparisons for all pairs using Tukey-Kramer HSD

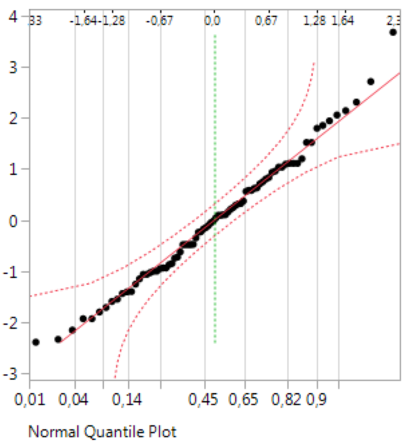
Connecting Letters Report

Level			Mean
3rd	A		5,7120000
2nd	A		5,3747727
1st		B	4,0533333
kindergarten		B	3,0673750

Levels not connected by same letter are significantly different.

The assumptions of the F-test are independent normally distributed observations with the same standard deviations regardless of group (here grade). A QQ-plot of the residuals indicate that we may regard the observations as normally distributed. The standard deviations in the four groups (0.862, 1.175, 1.576, 1.238) differ somewhat. A Levene's test does however not reject that they may be equal, and the "ratio of at most two"-rule of thumb is not violated.

result centered by grade



Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	2,0495	3	76	0,1140
Brown-Forsythe	2,3063	3	76	0,0834
Levene	2,3205	3	76	0,0819
Bartlett	1,9930	3	.	0,1126

It is (almost) impossible to tell from data whether the observations are independent. We should worry a lot if the results from the same student are included more than once (from different grades for instance).[2] *The fact that the students come from three different schools may be a cause for concern; some schools may be better, i.e. students from these schools may have better results, and this will make the observations (look) dependent (and this can be seen in the data). The following question handles this potential problem. Having students belonging to the same classes may also be a problem.*

[2] All the measurements are from different students; this is not explicitly stated in the problem formulation.

2. The test statistic for testing interaction equals 1.1471. Under the null hypothesis of no interaction, it is F-distributed with (6,68) degrees of freedom, giving us a p-value of 0.3450. Hence, we do not reject the null hypothesis and conclude that there is little evidence to support that the way the results depend on grade differ between schools.

Sequential (Type 1) Tests

Source	Nparm	DF	Seq SS	F Ratio	Prob > F
school	2	2	55,684300	22,8020	<,0001*
grade	3	3	50,540121	13,7970	<,0001*
grade*school	6	6	8,404053	1,1471	0,3450

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
grade	3	3	50,540121	13,6344	<,0001*
school	2	2	29,204432	11,8179	<,0001*

Using either sequential tests from the model with interaction (test statistic equals 13.797, F-distributed with (3,68) degrees of freedom) or marginal tests from the model without interaction (test statistic equals 13.634, F-distributed with (3, 74) degrees of freemdom), we reject the null hypothesis of no difference between grades; the p-value is below 0.0001. So the results differ between grades even after correcting for differences between schools.

The parameter estimates show that (even after correcting for school differences) the results increase as the students progress through elementary school. After correcting for school differences, the difference between the grades appear to be slightly smaller, though.

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	4,937525	0,359259	74,00	13,74	<,0001*	4,221686	5,653364
grade[kindergarten]	-2,282294	0,406658	74,00	-5,61	<,0001*	-3,092577	-1,472011
grade[1st]	-1,408585	0,365982	74,00	-3,85	0,0002*	-2,13782	-0,67935
grade[2nd]	-0,379608	0,372836	74,00	-1,02	0,3119	-1,122501	0,3632844
school[d]	0,0601485	0,378857	74,00	0,16	0,8743	-0,694741	0,815038
school[m]	1,2707421	0,286419	74,00	4,44	<,0001*	0,7000407	1,8414436