university of groningen

name: .....

stud. nr.: S.....

Exam

Statistics II, PSBE2–07

12 January 2015, 18.30 – 20.30 hrs

- This exam has two parts. Part 1 consists of **5 open questions** and Part 2 consists of **20 multiple-choice questions**.
- Part 1 contributes 25% to the final grade and Part 2 contributes 75% to the final grade.
- For each open question, write down the best answer to your knowledge on the separate white sheet. Remember: Incomplete or absent explanations **may lead to points deduction**. Clearly indicate to which of the questions your answers relate.
- For each multiple-choice question, write down the best answer to your knowledge on the separate pink answering sheet. Only one out of four answers is correct for each question.
- Write your name and student number **on both** answer sheets.
- At the end, **hand both answer sheets, as well as the questions set, over to the proctor**. Your grade will only be released if your questions set is returned, otherwise your score is invalid.
- The exam is closed book. No items are allowed on your desk other than the papers provided, your student card, pens/pencils, and a calculator.
- It is not allowed to use a graphical calculator. It is also **not allowed to use a mobile phone**, also not as a calculator.
- At the end of the exam there is a table with critical values and a formula sheet. Formulas from the formula sheet may or may not be used to answering questions.
- Fraud (such as looking into other's work, allowing others to look into your work, any communication) is prohibited and will be reported to the Examination Committee.

Good luck!!

Part 1 – Open questions (5 items in total)

A study has been performed to find out which of three teaching methods yielded the best results. To this end, 69 students were allocated at random among the three methods. At the end of the teaching period a multiple choice exam has been taken. The mean and standard deviation of the number of questions correct is as follows:

Method	Mean	SD	n
Method A	29.74	2.51	25
Method B	25.27	2.42	23
Method C	26.92	4.26	21
Total	27.39	3.61	69

Use this information in the next three questions.

A Compute s_p. Write down the computation details.

B Compute the upper bound of the 95% confidence interval for μ_2 (i.e., concerning teaching Method B) based on s_2. Write down the computation details.

C Next, two code variables, d_1 and d_2 are created with

$$d_1 = \begin{cases} 1 & \text{Method B} \\ 0 & \text{Methods A and C} \end{cases} \quad \text{and} \quad d_2 = \begin{cases} 1 & \text{Method C} \\ 0 & \text{Methods A and B} \end{cases}$$

and the model $\mu_y = \beta_0 + \beta_1 d_1 + \beta_2 d_2$ is formulated.

Compute a sample estimate for β_2. Explain your answer.

.....

A logistic regression model is used in order to better understand whether scores on two personality traits (variables X_1 and X_2) can be used to predict the display of aggressive behavior in stressing situations (variable Agress; Agress = 0: No aggressive behavior; Agress = 1: Aggressive behavior). Some results are shown below. Use this information in the next two questions.

Variables in the Equation

	B	S.E.	Sig.
Constant	−4.406	2.436	0.071
X1	0.290	0.200	0.147
X2	0.596	0.292	0.042

Dependent Variable: Y

D What is the predicted probability of displaying aggressive behavior for a subject with scores 4.58 and 6.83 on predictors X_1 and X_2, respectively? Write down the computation details.

E Provide a correct interpretation for the regression coefficient of X_1 in terms of log-odds.

End of Part 1. Go to next page for Part 2.

Part 2 – Multiple-choice questions (20 items in total)

1 The confidence interval for a population mean becomes narrower if ...

a. ... α increases.

b. ... β increases.

c. ... n decreases.

d. ... s increases.

.....

2 A sample of size 16 was randomly drawn from a normal population with known standard deviation $\sigma = 5$. The sample mean is equal to 20.5. The researcher wants to run the following z-test ($\alpha = 5\%$): $H_0 : \mu = 20$ versus $H_a : \mu < 20$.
What is the rejection region for this test?

a. $\overline{x} < 17.55$.

b. $\overline{x} < 17.94$.

c. $\overline{x} < 18.05$.

d. $\overline{x} < 18.44$.

.....

3 What is a Type I error?

a. It is the probability of correctly not rejecting the null hypothesis.

b. It is the probability of correctly rejecting the null hypothesis.

c. It is the probability of incorrectly not rejecting the null hypothesis.

d. It is the probability of incorrectly rejecting the null hypothesis.

.....

The test scores of three school groups are compared. Some results from the data analysis are shown in the following table (R_i = rank sum in group i).

Use this information in the next two questions.

Group	$\overline{y}_i$	SD	n_i	R_i
1	33.31	3.36	9	151
2	30.72	5.36	11	125
3	33.61	6.69	9	159
Total	32.42	5.3	29	

4 Is the information above sufficient to set up the ANOVA table?

a. Yes.

b. No, one cannot compute all degrees of freedom.

c. No, one cannot compute the Mean Squared Error (MSE).

d. No, one cannot compute the Total Sum of Squares.

5 A Kruskal-Wallis test is carried out using the same data. Determine the value of the test statistic H.

a. $H < 2.5$.

b. $2.5 \leq H < 5$.

c. $5 \leq H < 10$.

d. $H \geq 10$.

.....

Consider the following extract describing statistical conclusions derived from a two-way ANOVA analysis ($\alpha = 5\%$). The experiment factors are drug dosage and mice group; the dependent variable is the level of a specific toxin in the mice's blood.

> *"Drug dosage had a significant effect on the blood toxin levels ($F(4, 40) = 2.99$, $p = 0.030$). The differences across the mice groups were not significant ($F(2, 40) = 3.19$, $p = 0.052$). However, it was observed that increasing the levels of drug dosage had different effects across several groups of mice ($F(8, 40) = 3.09$, $p = 0.008$)."*

Use this information in the next three questions.

6 Which option is correct?

a. There are 2 groups of mice.

b. There are 4 drug dosage levels.

c. There are 4 groups of mice.

d. There are 5 drug dosage levels.

7 What is the total sample size?

a. 40.

b. 47.

c. 49.

d. 55.

8 The size of the drug dosage effect on the toxin level was also reported: $\omega^2 = .10$. According to the usual guidelines for ω^2, what can be concluded about the size of this effect?

a. The effect is small.

b. The effect is medium.

c. The effect is large.

d. Nothing can be concluded without knowing the sample sizes in each group.

.....

Scores measuring racial prejudice towards foreigners in a European western country were regressed on an indicator measuring group identity. Some results are shown below. Use this information in the next two questions.

Descriptive Statistics

	Mean	Std. Dev.	N
RacPr	9.657	2.294	12
GrId		7.210	12

Coefficients

	B	S.E.	t	Sig.
(Constant)	5.147	7.483	0.688	0.507
RacPr	2.042	0.756	2.702	0.022

Dependent Variable: GrId
Alpha: 5%

9 What is the value of the correlation r between GrId and RacPr?

a. $r = 0.21$.

b. $r = 0.39$.

c. $r = 0.63$.

d. $r = 0.65$.

10 What is the 95% confidence interval of β_{RacPr}?

a. $(0.36, 3.73)$.

b. $(0.40, 3.69)$.

c. $(0.56, 3.52)$.

d. $(0.67, 3.41)$.

.....

For a bivariate sample of size $n = 67$, the correlation is $r = 0.75$. Use this for the next two questions.

11 Consider these two claims about the 95% confidence interval for the population correlation coefficient ρ:
Claim A: "The interval is symmetric around $r = 0.75$."
Claim B: "One needs the Fisher Z-transformation to obtain this interval".
What can be concluded?

a. Claim A is incorrect. Claim B is incorrect.

b. Claim A is incorrect. Claim B is correct.

c. Claim A is correct. Claim B is incorrect.

d. Claim A is correct. Claim B is correct.

12 Compute r_z.

a. $r_z = 0.42$.

b. $r_z = 1.25$.

c. $r_z = 0.97$.

d. $r_z = 0.69$.

.....

A multiple regression analysis has been carried out after a sample of size 110 has been collected. The dependent variable is Y. $X1$ and $X2$ are continuous predictors, and D is a code variable with possible values 0 (group A) and 1 (group B). The model $Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 D_i + \varepsilon_i$ has been fitted and the following tables have been obtained. Use this information in the next five questions.

Coefficients

	B	SE	t	p-value	VIF
(Constant)	−113.736	65.391	−1.739	0.085	
$X1$	0.808	0.440	1.836	0.069	1.771
$X2$	9.019	0.943	9.563	0.000	1.885
D	2.212	2.666	0.830	0.408	1.156

Model Summary

Model	R	R Square	Std. Error of the Estimate
1	0.828	0.686	12.947

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	671.073	785.692	726.770	18.854	110
Residual	−35.380	28.920	0.000	12.768	110
Std. Residual	−4.943	2.254	−0.019	1.075	110
Cook's Distance	0.000	13.874	0.133	1.322	110

13 The second person in the data set has $Y_2 = 722.723$ and $\widehat{Y}_2 = 724.799$. What is the contribution of this person to the Error SS?

a. 3.89.

b. 4.31.

c. 16.38.

d. 163.02.

14 What claim about R^2 is true?

a. $R^2_{\text{Stein}} \leq R^2_{\text{Wherry}} \leq R^2$.

b. $R^2_{\text{Stein}} \leq R^2 \leq R^2_{\text{Wherry}}$.

c. $R^2_{\text{Wherry}} \leq R^2_{\text{Stein}} \leq R^2$.

d. $R^2_{\text{Wherry}} \leq R^2 \leq R^2_{\text{Stein}}$.

15 How should the value $R = 0.828$ in the second table be interpreted?

a. It is the correlation between Y and $\widehat{Y}$.

b. It is the correlation between Y and the residuals.

c. It is the correlation between $X1$ and $X2$.

d. It is the correlation between $\widehat{Y}$ and the residuals.

16 Consider now the regression model $X1 = \gamma_0 + \gamma_1 X2 + \gamma_2 D$. Compute R^2 for this model.

a. 0.23.

b. 0.44.

c. 0.47.

d. 0.63.

17 What interpretation for $b_3 = 2.212$ is correct?

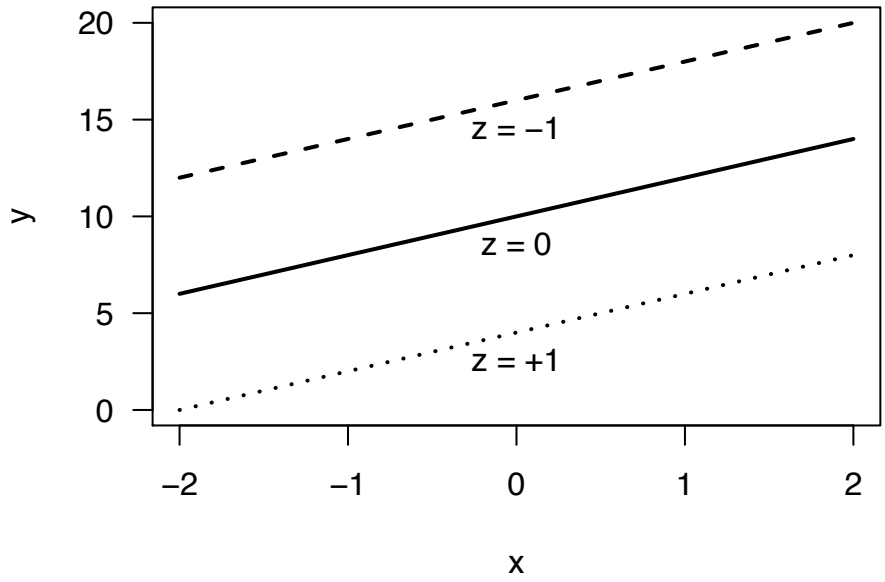
a. The population mean of Group B is 2.212 points higher than that of Group A.

b. The sample mean of Group B lies 2.212 points above average.

c. While keeping $X1$ and $X2$ constant, the predicted values for Group B are 2.212 points higher than those for Group A.

d. While keeping $X1$ and $X2$ constant, the predicted values of Group B are 2.212 times higher than those of Group A.

.....



18 Consider the plot above corresponding to a regression analysis $Y = B_0 + B_1x + B_2z + B_3xz$, where x and z are centered. What conclusion can be drawn?

a. The effect of x on Y is unrelated to z.

b. The plot displays an example of a compromised interaction between x and z.

c. The effect of x on Y gets stronger as z increases (for the given values of the moderator).

d. The effect of x on Y gets weaker as z increases (for the given values of the moderator).

.....

19 A regression has been performed on two continuous centered predictors. It is known that $n = 144$, $sd(x) = 2.73$, and $sd(z) = 2.01$. The estimated regression line is

$$\widehat{Y}_i = 286.27 + 17.88x_i + 28.18z_i + 7.53x_iz_i$$

Compute the simple regression equation of Y on x for z one standard deviation above mean.

a. $\widehat{Y}_i = 322.2 + 43.3x_i$

b. $\widehat{Y}_i = 314.5 + 25.4x_i$

c. $\widehat{Y}_i = 286.3 + 17.9x_i$

d. $\widehat{Y}_i = 342.9 + 33.0x_i$

.....

20 In the context of a one-way ANOVA, what is the best method to check the assumption of homoscedasticity?

a. A histogram of the residuals.

b. A boxplot per group of the residuals.

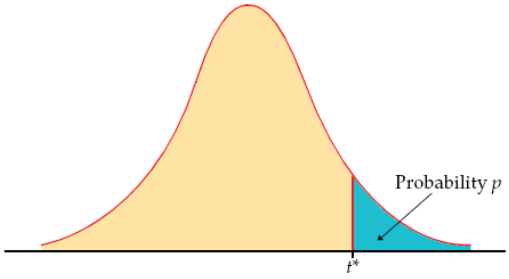
c. A QQ-plot of the residuals.

d. Testing whether the skewness and kurtosis of the residuals deviate from zero.

End of exam

t distribution: critical values

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.



	Upper-tail probability p								
df	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001
1	1.376	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3
2	1.061	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33
3	0.978	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21
4	0.941	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173
5	0.920	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893
6	0.906	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208
7	0.896	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785
8	0.889	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501
9	0.883	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297
10	0.879	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144
11	0.876	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025
12	0.873	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930
13	0.870	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852
14	0.868	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787
15	0.866	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733
16	0.865	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686
17	0.863	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646
18	0.862	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611
19	0.861	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579
20	0.860	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552
21	0.859	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527
22	0.858	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505
23	0.858	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485
24	0.857	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467
25	0.856	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450
26	0.856	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435
27	0.855	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421
28	0.855	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408
29	0.854	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396
30	0.854	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385
40	0.851	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307
50	0.849	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261
60	0.848	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232
80	0.846	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195
100	0.845	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174
1000	0.842	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098
z^*	0.841	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091
	60%	80%	90%	95%	96%	98%	99%	99.5%	99.8%
	Confidence level C								

Formulas

Pooled variance for i groups

$$s_p^2 = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i (n_i - 1)}$$

Confidence interval for μ

$$\overline{y} \pm t^* \frac{s}{\sqrt{n}}.$$

t-test for H: $\mu_1 = \mu_2$
Test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Test for H: $\rho = 0$
Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Contrasts

Sample estimation:

$$c = \sum_i a_i \bar{x}_i$$

Standard error:

$$\mathrm{SE}_c = s_p \sqrt{\sum_i \frac{a_i^2}{n_i}}.$$

Fisher Z-transformation

Transformation:

$$r_z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right).$$

Inverse transformation:

$$r = \frac{e^{2r_z} - 1}{e^{2r_z} + 1}.$$

Variance Inflation Factor

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

Adjusted R^2's

$$R^2_{\text{Wherry}} = 1 - \frac{n-1}{n-p-1} \left(1 - R^2\right).$$

$$R^2_{\text{Stein}} = 1 - \frac{(n-1)(n-2)(n+1)}{(n-p-1)(n-p-2)n} \left(1 - R^2\right).$$

Kruskal-Wallis test

Test Statistic:

$$H = \frac{12}{N(N+1)} \sum_i \frac{R_i^2}{n_i} - 3(N+1).$$

Effect sizes

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}, \quad \omega^2 = \frac{SS_{\text{effect}} - df_{\text{effect}} \times MSE}{MSE + SS_{\text{total}}}$$