university of groningen

Second Partial Exam (SAMPLE)

Statistics II, PSBE2–07

09–99–2099, 1 hour

- The exam for this course exists of two partial exams. This is the second partial exam.
- The **real** partial exam will consist of **18 multiple-choice questions**. This sample exam may have more or less than 18 questions.
- Do not infer that the weight or type of content of each topic in the real exam will be the same as in this sample exam. Simply, use this exam to *get an idea* about the real exam.
- 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 the answer sheet.
- At the end, **hand the questions set and your answers over to the proctor**.
- 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!!

1 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).

Group	$\bar{y}_i$	SD	n_i	R_i
1	33.4	3.58	8	157.5
2	31.38	5.31	11	159
3	32.07	6.16	14	244.5
Total	32.16	5.26	33	

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

a. Yes. $\surd$

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.

Solution:
B is incorrect: $I-1$, $n-I$, and $n-1$ are easy from the table ($I=3$, $n=33$).
D is incorrect: MS-total $= 5.26^2$. SS-total = MS-total × df-total.
C is incorrect: from SS-Total and SS-Groups you get SS-error. From SS-error and df-error you get MSE. Alternatively, MSE $= s_p^2$.
Thus, A is correct.

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 three 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

2 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. $\surd$

c. 16.38.

d. 163.02.

Solution:
$(Y_2 - \widehat{Y}_2)^2 = (722.723 - 724.799)^2 = 4.31.$

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

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

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.

Solution:
By definition.

4 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. $\surd$

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

Solution:
Note that:

- X_1 and X_2 should be kept fixed (and not necessarily at zero, since no interaction with D exists).
- D coding is 0 (Group A) and 1 (Group B).

5 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. $\surd$

c. A QQ-plot of the residuals.

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

Solution:
Options A, C, and D aim at checking the normality assumption.

6 In the context of one-way ANOVA, which formula describes the residual sum of squares? (Here I denotes the number of groups and n_i denotes the sample size of group i.)

a. $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_j - \bar{y}_i)^2$.

b. $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$.

c. $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$.

d. $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$. $\surd$

7 A oneway ANOVA has been carried out on a data set containing three groups and 20 measurements per group. All three sample means are exactly equal. Which claim below is **not** true?

a. $SS_{between} = 0$

b. $MSE = 0$ $\surd$

c. $df_{total} = 59$

d. $df_{between} < df_{error}$.

Regression was used to run a one-way ANOVA with four groups (Group 1 through 4) and dependent variable Y. The experiment factor was coded as follows:

Group	D1	D2	D3
Group 1	1	0	0
Group 2	0	1	0
Group 3	0	0	1
Group 4	0	0	0

The regression model $\widehat{Y} = b_0 + b_1\mathrm{D1} + b_2\mathrm{D2} + b_3\mathrm{D3}$ was estimated. Some results are shown next.

$\boxed{\textbf{Use this setting to answer the next 4 questions.}}$

ANOVA

	Sum Sq	df	Mean Sq	F	Sig.
Regression	355.028	3	118.343	29.803	0.000
Residual	186.628	47	3.971		
Total	541.656	50			

Dependent Variable: Y
Predictors: (Constant), D1, D2, D3

Coefficients

	B	S.E.	Beta	t	Sig.	CI.Low	CI.Upp
(Constant)	19.602	0.553		35.468	0.000	18.490	20.714
D1	0.163	0.798	0.021	0.205	0.839	−1.442	1.768
D2	−4.763	0.755	−0.666	−6.308	0.000	−6.282	−3.244
D3	−5.694	0.816	−0.719	−6.975	0.000	−7.337	−4.052

Dependent Variable: Y
Alpha: 5%

8 What is the sample mean value in Group 1?

a. $\overline{Y}_1 = 13.91$.
b. $\overline{Y}_1 = 19.77$. $\surd$
c. $\overline{Y}_1 = 14.84$.
d. $\overline{Y}_1 = 19.60$.

Solution:
$M_1 = b_0 + 1 \times D1 + 0 \times D2 + 0 \times D3 = 19.602 + 0.163 = 19.77$.

9 What is the value of the t-test statistic used to test the contrast $\mathrm{H}_0 : \mu_2 - \mu_4 = 0$?

a. $t = -6.98$.
b. $t = 0.20$.
c. $t = -6.31$. $\surd$
d. $t = 35.47$.

Solution:
t test associated to D2.

10 Predictors D1, D2, and D3 are not centered. Let $\mathrm{D1}_c$, $\mathrm{D2}_c$, and $\mathrm{D3}_c$ denote the centered predictors. What can be concluded concerning the model $\widehat{Y} = d_0 + d_1\mathrm{D1}_c + d_2\mathrm{D2}_c + d_3\mathrm{D3}_c$?

a. $d_0 = 19.602$.
b. $d_1 = 0.163$. $\surd$
c. This model explains less variance of Y than the model with uncentered predictors.
d. This model explains more variance of Y than the model with uncentered predictors.

Solution:

Centering preserves the higher-order regression coefficients.

11 The researcher wants to perform all pairwise comparisons using the Bonferroni correction with an *overall error rate* of 10%. At what level of significance must each individual test be performed?

a. $\alpha = 0.001$.

b. $\alpha = 0.017$. $\surd$

c. $\alpha = 0.025$.

d. $\alpha = 0.100$.

Solution:
4 groups $\longrightarrow$ 6 pairs, hence $\alpha/6 = .10/6 = .017$.

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.

> *"It was observed that increasing the levels of drug dosage did not have different effects across several groups of mice ($F(10, 29) = 1.32$, $p = 0.266$). Drug dosage had a significant overall effect on the blood toxin levels ($F(5, 29) = 4.21$, $p = 0.005$). The differences across the mice groups were not significant ($F(2, 29) = 1.65$, $p = 0.210$)."*

Use this setting to answer the next 4 questions.

12 Which option is correct?

a. There are 2 groups of mice.

b. There are 5 drug dosage levels.

c. There are 4 groups of mice.

d. There are 6 drug dosage levels. $\surd$

Solution:
Number of drug dosage levels: 6; obtainable from the first df=$6-1$ of the second reported F-value;
Number of groups of mice: 3; obtainable from the first df=$3-1$ from the third reported F-value.

13 What is the total sample size?

- **a.** 29.
- **b.** 37.
- **c.** 40.
- **d.** 47. √

Solution:
Degrees of freedom:

Source	df
A (mice)	2
B (drug dos.)	5
A × B	10
Error	29
Total	46

$\Rightarrow n = 47.$

14 The size of the drug dosage effect on the toxin level was also reported: $\omega^2 = .01$. 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.

Solution:
By definition (small: $\omega^2 < .06$; medium: $.06 < \omega^2 < .14$; large: $.14 \leq \omega^2$).

15 Due to the non-significance of both the mice groups and the interaction term, it is decided to rewrite the manuscript such that it only reports the significant effect of different toxin levels. How can this approach be labelled?

- **a.** As a questionable research practice.
- **b.** As a contribution to the replicability crisis.
- **c.** Both A and B are correct. √
- **d.** Neither A nor B are correct.

A one way ANOVA experiment has been carried out. There were $I = 4$ groups, with $n_i = 24$ measurements per group. The mean squared error is 6.36, and the sample means and standard-deviations are as follows:

Group	Mean	SD
1	15.73	2.05
2	21.04	2.53
3	22.36	2.89
4	23.17	2.55

Use this setting to answer the next 2 questions.

16 Compute the t-value for the contrast for the null hypothesis H_0: $\mu_1 = \mu_4$, versus the two-sided alternative.

a. -10.22 $\surd$

b. -35.40

c. -7.23

d. -7.29

Solution:
$t = \frac{15.73-23.17}{s_p\sqrt{\frac{1}{24}+\frac{1}{24}}} = -10.22$, where $s_p = \sqrt{\frac{(24-1)\times(2.05^2+2.53^2+2.89^2+2.55^2)}{4\times(24-1)}} = 2.5228$.

17 Now consider the contrast that compares group 1 to the mean of the three other groups. How many degrees of freedom does the corresponding t-test have?

a. 3

b. 92 $\surd$

c. 93

d. 95

Solution:
$4 \times 24 - 4$.

t distribution: critical values[a]

	$\alpha_1 = .10$	.05	0.025	0.010	.005
ν	$\alpha_2 = .20$	.10	0.050	0.020	.010
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
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
75	1.293	1.665	1.992	2.377	2.643
100	1.290	1.660	1.984	2.364	2.626
1000	1.282	1.646	1.962	2.330	2.581
∞	1.282	1.645	1.960	2.326	2.576

[a] α_1 holds the one-sided upper-tail value of the distribution with ν degrees of freedom; α_2 holds the corresponding two-sided value.

Formula sheet

At the exam, you will receive this formula sheet.

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:

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

(Semi-)partial correlations
Formula's valid when working with DV y and two predictors.

$$pr_1 = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} = \sqrt{\frac{R^2 - r_{y2}^2}{1-r_{y2}^2}}$$

$$sr_1 = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1-r_{12}^2}} = \sqrt{R^2 - r_{y2}^2}$$

Adjusted R^2

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

Effect sizes

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

Binomial model

$$p(X = x|N,\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}, \quad \binom{N}{x} = \frac{N!}{(N-x)!x!}$$