university of groningen

name: .....

stud. nr.: S.....

Resit (SAMPLE)

Statistics II, PSBE2–07

16–04–2020, 2 hours

- This resit consists of **36 multiple-choice 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 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!!

1 Consider the following two claims about simple linear regression:
A: 'Prediction intervals for y at x-values close to $\bar{x}$ are smaller than those at x-values far from $\bar{x}$.'
B: 'The homogeneity assumption states that the variance of x is fixed.'

a. Claim A is correct, claim B is correct.

b. Claim A is correct, claim B is incorrect. √

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

d. Claim A is incorrect, claim B is incorrect.

.....

2 For a given sample, the Fisher Z correlation is computed as $r_z = -0.27$. Compute the regular correlation coefficient r.

a. $r = -0.28$.

b. $r = -0.26$. √

c. $r = -0.14$.

d. $r = -0.07$.

.....

3 The population regression line in the multiple regression model with p predictors:

a. Is built on the principle DATA = FIT + RESIDUAL.

b. Is a straight line through the values of $x_1, \ldots, x_p$ and y.

c. Returns the value of the average of the dependent variable, for given values of the predictors. √

d. Provides p predictions for the mean dependent variable.

.....

4 A multiple regression with one dependent variable, Y, and two predictors, X_1 and X_2, has been carried out on a sample of size $n = 40$. This yields $R^2 = 0.59$, $r_{Y,X_1} = 0.13$ and $r_{Y,X_2} = 0.58$. Compute the partial correlation coefficient for X_1.

a. 0.76.

b. 0.50.

c. 0.61. √

d. 0.08.

.....

5 Which of the following options correctly describes an assumption in simple linear regression?

a. All observations are independent of each other. √

b. The set of all scores from the dependent variable is normally distributed.

c. There is a perfect linear relation between the scores of the predictor and the dependent variables in the population.

d. There is a perfect linear relation between the scores of the predictor and the dependent variables in the sample.

.....

6 A researcher collects data on 500 men and women aged 18 to 65 on their attitudes towards the environment. No predefined research hypotheses were stated. After studying the data, the researcher decides to write a manuscript entitled 'Grumpy old men: Why men aged 50+ are negative towards the environment'.

What has happened here?

a. The researcher is HARKing.

b. The study is likely to have low power.

c. The result suggested by the title needs to be validated in future research.

d. All of the other alternatives are correct. √

.....

A study has been performed to find out which of three teaching methods yielded the best results. To this end, 20 students were randomly allocated to each method. 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
Method A	33.83	2.87
Method B	24.57	2.97
Method C	30.43	4.60
Total	29.61	5.21

7 Compute s_p.

a. $s_p = 3.48$.

b. $s_p = 5.22$.

c. $s_p = 3.57$. √

d. $s_p = 5.21$.

.....

A one-way ANOVA model was used to compare a number of groups with each other. The corresponding ANOVA table is as follows:

	Sum Sq	df	Mean Sq	F	Sig.
Between Groups	682.240		341.120	36.966	0.000
Within Groups	525.986		9.228		
Total	1208.227	59			

8 How many groups were included in the analysis?

a. 2.

b. 3. √

c. 4.

d. 5.

.....

Scores of a response variable are collected for two independent groups (Control, Experiment). Some summary statistics are displayed below.

Source	n	Mean	SD
Control (C)	12	9.45	1.92
Experiment (E)	17	13.69	1.59

9 Consider code variable d such that $d_i = 0$ for subjects in the Control group and $d_i = 1$ for subjects in the Experiment group. What is the estimated regression model $\mu_y = \beta_0 + \beta_1 d$?

a. $\widehat{y} = 9.45 + 4.24d$. √

b. $\widehat{y} = 9.45 + 13.69d$.

c. $\widehat{y} = 13.69 - 9.45d$.

d. $\widehat{y} = 13.69 + 4.24d$.

.....

A study has been performed on the effectiveness of three types of mindfulness training. In total 60 participants participated in the experiment. The results are summarised in the following table:

Type	$\bar{y}$	sd	n
A	10.00	3.20	15
B	14.00	3.60	30
C	8.00	2.80	15

10 Compute the upper bound of the 95% confidence interval for group C, based on that group's standard deviation.

a. 8.775.

b. 9.273.

c. 9.551. √

d. 11.565.

.....

A one way ANOVA experiment has been carried out. There were $I = 4$ groups, with $n_i = 20$ measurements per group.

11 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. 76. √

c. 77.

d. 79.

.....

12 For a sample of size $n = 52$ the correlation coefficient between two variables has been computed as $r = 0.71$. Compute the test statistic for H_0: $\rho = 0$ versus the two-sided alternative.

a. $t = 6.99$.

b. $t = 7.13$. √

c. $t = 9.32$.

d. $t = 10.12$.

.....

13 Which of the alternatives below corresponds to the concept of 'power'?

a. P(not rejecting $H_0|H_0$ is false).

b. P(not rejecting $H_0|H_0$ is true).

c. P(rejecting $H_0|H_0$ is false). √

d. P(rejecting $H_0|H_0$ is true).

.....

14 In the context of simple linear regression, tests for three out of the following four null hypotheses always yield the same p-value. Which one does not?

a. H_0: $R^2 = 0$ vs. $R^2 > 0$.

b. H_0: $\beta_1 = 0$ vs. $\beta_1 \neq 0$.

c. H_0: $r = 0$ vs. $r \neq 0$. √

d. H_0: $\rho = 0$ vs. $\rho \neq 0$.

.....

15 A simple linear regression is carried out in order to predict y from x. This provided estimates $b_0 = 20.00$, $b_1 = 84.00$ and $r_{xy} = 0.36$. Originally, x was distance measured in meters. For compatibility with a similar American study, this variable is converted into yards (1 yard = 0.9144 meter). How many of the values b_0, b_1 and r_{xy} change because of this rescaling?

a. 0

b. 1 √

c. 2

d. 3

.....

16 Consider the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Testing the null hypothesis $\mathcal{H}_0 : R^2 = 0$ is equivalent to testing which hypothesis?

a. $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3$.

b. $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$. √

c. $\mathcal{H}_0$: At least one regression coefficient is different from 0.

d. $\mathcal{H}_0$: All regression coefficients are different from 0.

.....

17 What is *not* a questionable research practice?

a. Removing values as outliers because they do not fit the model.
b. Combining two variables into one because of multicollinearity. √
c. Doing more observations as the result is not yet significant.
d. In reporting the results, focussing entirely on the significant results.

.....

18 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. √

.....

19 In a two-way ANOVA setting, Factor A has 4 factor levels and Factor B has 4 factor levels. What is $df_{A\times B}$, the degrees of freedom for the interaction term?

a. 9. √
b. 15.
c. 16.
d. 25.

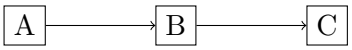
.....

20 The file drawer problem has to do with...

a. Publication bias. √
b. Hidden moderators.
c. Mediator analysis.
d. Influential points.

.....

21 Consider the following relation between the variables A, B, and C. How can this relation be best described?



a. Variable A moderates the relation between B and C.

b. Variable A mediates the relation between B and C.

c. Variable B moderates the relation between A and C.

d. Variable B mediates the relation between A and C. √

.....

22 What is an advantage of centering predictors when performing a regression with interaction between continuous variables?

a. The interaction will get a lower p-value and will be significant quicker.

b. It improves the interpretability of B_1 and B_2.√

c. It improves the interpretability of B_3.

d. All alternatives are correct.

.....

For a bivariate sample of size $n = 84$, the correlation is $r = 0.66$.

23 What is the distribution of the test statistic for H_0: $\rho = .5$ versus alternative H_A: $\rho > .5$ if the null hypothesis is true?

a. t-distribution with 81 degrees of freedom.

b. t-distribution with 82 degrees of freedom.

c. t-distribution with 83 degrees of freedom.

d. Standard normal distribution. √

.....

A two-factor fixed-effects ANOVA has been carried out, with two levels for both Factor A and Factor B. Based on four measurements per cell, the following ANOVA table was obtained.

Source	SS	df	MS	F
A	24.06	...	...	...
B	19.06	...	...	...
AB	8.56	...	...	...
Within	...	...	5.73	
Total	...	...		

24 Compute the F-value for Factor B.

a. $F < 0.5$.
b. $0.5 < F < 1.0$.
c. $1 < F < 2$.
d. $F > 2$. √

.....

25 In a sample, two variables, A and B, are positively correlated. What is **not** a possible explanation for this?

a. A and B both are common causes of C. √
b. A causes B through C.
c. Coincidence.
d. B causes A.

.....

26 A sample of size $n = 56$ yields the following bivariate correlations. Which of the following four relations is true?

	y	x_1	x_2
y	1.00	0.42	0.62
x_1		1.00	0.28
x_2			1.00

a. $pr_1 < r_{1,2} < sr_1$.
b. $r_{1,2} < sr_1 < pr_1$.
c. $sr_1 < r_{1,2} < pr_1$. √
d. $sr_1 < pr_1 < r_{12}$.

.....

27 A regression has been performed on two continuous centered predictors x and z. It is known that $n = 116$, $sd(x) = 1.73$, and $sd(z) = 2.01$. The estimated regression line is

$$\widehat{Y}_i = 28.63 + 7.15x_i + 2.82z_i - 1.12x_iz_i$$

Compute the simple slope for the regression of Y on x when z is one standard deviation below mean.

a. 1.5
b. 4.9
c. 9.4 √
d. 38.0

.....

28 In order to find the relation between exam grade (on the scale from 1 to 10) and time spent preparing for the exam (measured in hours), the following regression model is set up: $\text{grade}_i = \beta_0 + \beta_1 \text{time}_i + \varepsilon_i$. Consider the following two claims: A: 'Since both grade and time must be positive, the intercept must be positive as well.'
B: 'If preparation time is measured in days rather than hours, the slope will become a factor 24 larger.'

a. Claim A is correct, claim B is correct.
b. Claim A is correct, claim B is incorrect.
c. Claim A is incorrect, claim B is correct. √
d. Claim A is incorrect, claim B is incorrect.

.....

29 Consider the following claims about the Kruskal-Wallis test. Which alternative is correct?
Claim A: 'The Kruskal-Wallis test is a non-parametric alternative to two-way ANOVA'.
Claim B: 'Under H_0, the test statistic follows an F-distribution with $I - 1$ and $N - I$ degrees of freedom.'

a. Claim A is correct. Claim B is correct.
b. Claim A is correct. Claim B is incorrect.
c. Claim A is incorrect. Claim B is correct.
d. Claim A is incorrect. Claim B is incorrect. √

.....

30 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. $MSE = 0$ √
c. $df_{total} = 59$
d. $df_{between} < df_{error}$.

.....

31 In a regression analysis, a categorical predictor with three categories is included through dummy variables. No other predictor variables are included. Let β_i denote the regression coefficient associated to the i-th dummy variable. What null hypothesis is tested by the omnibus F test?

a. $\mathcal{H}_0 : \beta_1 = \beta_2$.
b. $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3$.
c. $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$.√
d. $\mathcal{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

32 What statement about the adjusted R^2 is *not* correct?

- **a.** R^2_{adj} is always smaller than R^2.
- **b.** When n/p is very large, R^2_{adj} and R^2 are very similar.
- **c.** R^2_{adj} estimates the proportion of variance accounted for in the population.
- **d.** When p increases then R^2_{adj} will also increase. √

.....

33 In a certain town, the probability that it is raining at any given moment is 30%. The probability that there is a traffic jam at the town's main junction is 25%. The probability of a traffic jam during rain is 50%. What is the probability that it is raining when there is a traffic jam?

- **a.** 60%. √
- **b.** 25%.
- **c.** 42%.
- **d.** 50%.

.....

34 Which claim is true?

- **a.** The median of the prior distribution always lies between the median of the posterior distribution and the median of the likelihood.
- **b.** The median of the posterior distribution always lies between the median of the prior distribution and the median of the likelihood. √
- **c.** The median of the likelihood always lies between the median of the prior distribution and the median of the posterior distribution.
- **d.** Which alternative is correct depends on the situation.

.....

35 The values of the dependent variable in a study are denoted by Y. A prior distribution is set up such that $P(Y > 10) = 0.25$. Subsequently, a sample of size $n = 25$ is drawn, with all values being below 10. What can you say about the posterior probability of $Y > 10$?

- **a.** That posterior probability is zero.
- **b.** That posterior probability is 0.25.
- **c.** That posterior probability is between 0 and 0.25.
- **d.** That posterior probability is larger than 0.25.

.....

36 A study has collected data on 50 participants. Predictor variable A has a mean value of 12.0 and $SD = 2.0$. Predictor variable B has a mean of 7.5 and $SD = 1.5$. Dependent variable Y is regressed onto the standardised predictors, denoted a and b, yielding the regression equation:

$$E(Y) = 12.4 + 2.6a + 3.4b - 1.3ab$$

Compute the simple regression equation for Y on a for b one SD below the mean.

a. $9 + 3.9a$ √

b. $7.3 + 4.55a$

c. $15.8 + 1.3a$

d. $17.5 + 0.65a$

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

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:

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

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

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

(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's

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

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

Effect sizes

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