

# Overview

Stephanie Ranft S2459825

June 6, 2020

## Statistics 3 PSBE2-12

### Exercises

#### Statistics 2 Exam 16th April 2020

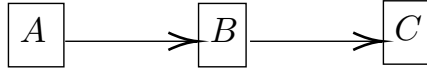
- What statement about the adjusted  $R^2$  is **not** correct ( $n$  = sample size;  $p$  = number of predictors)?
  - When  $n/p$  is very large,  $R_{\text{adj}}^2$  and  $R^2$  are very similar.
  - When  $p$  increases then  $R_{\text{adj}}^2$  will also increase.
  - $R_{\text{adj}}^2$  estimate the proportion of variance accounted for in the population.
  - $R_{\text{adj}}^2$  is always smaller than  $R^2$ .
- In regression analysis, a categorical predictor with three categories is included through dummy variables. No other predictor variables included. Let  $\beta_i$  denote the regression coefficient associated to the  $i$ -th dummy variable. What null hypothesis is tested by the omnibus  $F$  test?
  - $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
  - $H_0 : \beta_1 = \beta_2 = \beta_3$
  - $H_0 : \beta_1 = \beta_2 = 0$
  - $H_0 : \beta_1 = \beta_2$
- A study as 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 teach 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	31.69	2.29
Method B	27.72	2.76
Method C	24.74	4.82
Total	28.05	4.46

Compute  $s_p$ .

- $s_p = 4.94$
  - $s_p = 3.47$
  - $s_p = 4.46$
  - $s_p = 3.29$
- 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?
    - $df_{\text{between}} < df_{\text{error}}$
    - $SS_{\text{between}} = 0$
    - $MSE = 0$
    - $df_{\text{total}} = 59$

5. What is not a questionable research practice?
- (a) Doing more observations as the result is not yet significant.
  - (b) In reporting the results, focusing entirely on the significant results.
  - (c) Silently removing values as outliers because they do not fit the model.
  - (d) Combining two variables into one because of multicollinearity.
6. Consider the following relation between the variable  $A$ ,  $B$ , and  $C$ :



How can this relation be best described?

- (a) Variable  $B$  moderates the relation between  $A$  and  $C$ .
  - (b) Variable  $B$  mediates the relation between  $A$  and  $C$ .
  - (c) Variable  $A$  mediates the relation between  $B$  and  $C$ .
  - (d) Variable  $A$  moderates the relation between  $B$  and  $C$ .
7. Which of the alternatives below corresponds to the concept of ‘power’?
- (a)  $\mathbb{P}(\text{not rejecting } H_0 | H_0 \text{ is true})$
  - (b)  $\mathbb{P}(\text{rejecting } H_0 | H_0 \text{ is true})$
  - (c)  $\mathbb{P}(\text{not rejecting } H_0 | H_0 \text{ is false})$
  - (d)  $\mathbb{P}(\text{rejecting } H_0 | H_0 \text{ is false})$
8. A multiple regression with one dependent variable,  $Y$ , and two predictors,  $X_1$  and  $X_2$ , has been carried out on a sample size  $n = 31$ . This yields  $R^2 = 0.71$ ,  $r_{Y,X_1} = -0.11$ , and  $r_{Y,X_2} = 0.47$ . Compute the partial correlation coefficient for  $X_1$ .
- (a) 0.84
  - (b) 0.70
  - (c) 0.79
  - (d) 0.68
9. What is an advantage of centering predictors when performing a regression with interaction between continuous variables?
- (a) It improves the interpretability of  $B_3$
  - (b) It improves the interpretability of  $B_1$  and  $B_2$
  - (c) The interaction will get a lower  $p$ -value and will be significant quicker
  - (d) All the other alternatives are correct
10. 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:

$$\mathbb{E}(Y) = 12.4 + 2.6a + 3.4b - 1.3ab.$$

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

- (a)  $17.5 + 0.65a$
- (b)  $15.8 + 1.3a$
- (c)  $9 + 3.9a$
- (d)  $7.3 + 4.55a$

11. A sample of size  $n = 56$  yields the following bivariate correlations:

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

Which of the following four relations is true?

- (a)  $sr_1 < r_{1,2} < pr_1$
  - (b)  $sr_1 < pr_1 < r_{1,2}$
  - (c)  $r_{1,2} < sr_1 < pr_1$
  - (d)  $pr_1 < r_{1,2} < sr_1$
12. 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 incorrect, claim B is incorrect
  - (c) Claim A is incorrect, claim B is correct
  - (d) Claim A is correct, claim B is incorrect
13. 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)  $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$
  - (b)  $H_0 : R^2 = 0; H_A : R^2 \neq 0$
  - (c)  $H_0 : \rho = 0; H_A : \rho \neq 0$
  - (d)  $H_0 : r = 0; H_A : r \neq 0$
14. Which claim is true?
- (a) Which alternative is correct depends on the situation.
  - (b) The median of the likelihood always lies between the median of the prior distribution and the median of the posterior distribution.
  - (c) The median of the prior distribution always lies between the median of the posterior distribution and the median of the likelihood.
  - (d) The median of the posterior distribution always lies between the median of the prior distribution and the median of the likelihood.

15. 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	51.356		25.678	2.488	0.095
Within Groups	433.500		10.321		
Total	484.856	44			

How many groups were included in the analysis?

- (a) 5
  - (b) 4
  - (c) 2
  - (d) 3
16. 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 incorrect; claim B is incorrect.
  - (b) Claim A is correct; claim B is correct.
  - (c) Claim A is incorrect; claim B is correct.
  - (d) Claim A is correct; claim B is incorrect.
17. A researcher collects data on 100 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 60+ are negative towards the environment'.
- What has happened here?
- (a) The researcher is HARKing.
  - (b) The result suggested by the title needs to be validated in future research.
  - (c) All the other alternatives are correct.
  - (d) The study is likely to have low power.
18. Consider the multiple regression model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ . Testing the null hypothesis  $H_0 : R^2 = 0$  is equivalent to testing which hypothesis?
- (a)  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ .
  - (b)  $H_0$  : All regression coefficients are different from 0.
  - (c)  $H_0$  : At least one regression coefficient is different from 0.
  - (d)  $H_0 : \beta_1 = \beta_2 = \beta_3$ .
19. 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, given that there is a traffic jam?
- (a) 42%
  - (b) 50%
  - (c) 60%
  - (d) 25%

20. A regression has been performed on two continuous centred predictors  $x$  and  $z$ . It is known that  $n = 81$ ,  $s_x = 2.27$  and  $s_z = 1.85$ . The estimated regression line is

$$\hat{y}_i = 25.55 + 6.81x_i + 3.15z_i - 1.83x_iz_i.$$

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

- (a) 35.7
  - (b) 10.2
  - (c) 3.4
  - (d) 1.0
21. The values of the dependent variable in a study are denoted by  $Y$ . A prior distribution is set up such that  $\mathbb{P}(Y > 10) = 0.25$ . 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) The posterior probability is zero.
  - (b) The posterior probability is larger than 0.25.
  - (c) The posterior probability is between 0 and 0.25.
  - (d) The posterior probability is 0.25.
22. 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

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

- (a) 8.78
  - (b) 11.57
  - (c) 9.27
  - (d) 9.55
23. 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)  $t = 10.12$
  - (b)  $t = 7.13$
  - (c)  $t = 6.99$
  - (d)  $t = 9.32$
24. 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 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

25. The population multiple regression model with  $p$  predictors
- (a) is built on the principle  $\text{DATA} = \text{FIT} + \text{RESIDUAL}$ .
  - (b) is a straight line through the value of  $x_1, \dots, x_p$  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.
26. Which of the following correctly describes an assumption in linear regression?
- (a) There is a perfect linear relation between the scores of the predictor and the dependent variables in the population.
  - (b) The set of all scores from the dependent variable is normally distributed.
  - (c) All observations are independent of each other.
  - (d) There is a perfect linear relation between the scores of the predictor and the dependent variables in the sample.
27. In a sample, two variables  $A$  and  $B$  are positively correlated. What is **not** a possible explanation for this?
- (a) Coincidence.
  - (b)  $A$  and  $B$  are common causes of  $C$ .
  - (c)  $A$  causes  $B$  through  $C$ .
  - (d)  $B$  causes  $A$ .
28. A one-way ANOVA experiment has been carried out. There were  $I = 4$  groups with  $n_i = 20$  measurements per group. 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) 75
  - (c) 76
  - (d) 77
29. 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  | ...   | ... |      |     |
- Compute the  $F$ -value for Factor B.
- (a)  $F < 0.5$
  - (b)  $0.5 < F < 1.0$
  - (c)  $F > 2$
  - (d)  $1 < F < 2$
30. The file drawer problem has to do with ...
- (a) Mediator analysis
  - (b) Influential points
  - (c) Hidden moderators
  - (d) Publication bias

31. 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 metres. For compatibility with a similar American study, this variable is converted into yards (1 yard = 0.9144 metres). How many of the values  $b_0$ ,  $b_1$  and  $r_{xy}$  change because of this rescaling?

- (a) 0
- (b) 2
- (c) 1
- (d) 3

32. 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)	10	9.91	3.54
Experiment (E)	18	11.95	1.85

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)  $\hat{y} = 9.91 + 2.04d$
- (b)  $\hat{y} = 11.95 - 9.91d$
- (c)  $\hat{y} = 11.95 + 2.04d$
- (d)  $\hat{y} = 9.91 + 11.95d$

33. For a bivariate sample of size  $n = 67$  the correlation is  $r = 0.52$ . What is the distribution of the test statistic for  $H_0 : \rho = 0.5$  versus alternative  $H_A : \rho > 0.5$  if the null hypothesis is true?

- (a) Standard normal distribution.
- (b)  $t$ -distribution with 65 degrees of freedom.
- (c)  $t$ -distribution with 64 degrees of freedom.
- (d)  $t$ -distribution with 66 degrees of freedom.

34. In a two-way ANOVA setting Factor A has three levels and Factor B has 6 levels. What is  $df_{A \times B}$ , the degrees of freedom for the interaction term?

- (a) 17
- (b) 28
- (c) 18
- (d) 10

35. What is a Type I error?

- (a) It is the probability of incorrectly not rejecting the null hypothesis.
- (b) It is the probability of correctly rejecting the null hypothesis.
- (c) It is the probability of incorrectly rejecting the null hypothesis.
- (d) It is the probability of correctly not rejecting the null hypothesis.

36. For a given sample, the Fisher  $Z$  correlation is computed as  $r_z = -0.44$ . Compute the regular correlation coefficient  $r$ .

- (a)  $r = -0.47$
- (b)  $r = -0.41$
- (c)  $r = -0.17$
- (d)  $r = -0.14$