

Statistics B

prof. Paolo Coletti – exercises prototypes A.Y. 2010/11

Note: all exercises must include the necessary calculations and explanations of the calculations. When a test or a calculation needs assumptions or prerequisites, they must be explicitly cited. Writing the conclusions of the test, what has been proven must be explicitly written.

Tests

1. Why do we use a stratified sample? Explain. (3 points)
2. The significance of a statistical test is 2% with $n = 12$. Estimate how many cases do we need to get a significance of 1%. (1 point)
3. You have to compare the expected values of three distributions. Which test would you use and under which assumptions? Specify how does it work. (3 points)
4. You have a sample divided into two groups of 20 and 100 cases with sample averages 3 and 4 and sample variances 4 and 9. Test whether the expected values are different with significance level 5%. Are there other tests you can perform with this information? (3 points)
5. You have a sample of 9 cases with two variables:

x	2	3	4	5	6	7	8	9	10
y	3	4	5	6	8	9	10	11	12

- a. what is the value of Spearman rank correlation coefficient? And its significance? (2 points)
 - b. according to your experience, what is an approximated value for Pearson's correlation coefficient? (1 point)
 - c. test, using all the methods you know, whether the distributions of variables X and Y have different positions (5 points)
 - d. test using only one method whether the distribution of variable X is shifted to the left with respect to distribution of variable Y (3 points)
 - e. test whether the median of X is larger than 5 (2 points)
 - f. test whether the expected value of $Y - X$ is different from 0 (3 points)
 - g. test whether the expected value of $Y - X$ is different from 2 (3 points)
 - h. test whether X and Y have the same variance (3 points)
 - i. test, using only one method, whether X is normally distributed (3 points)
 - j. explain how does Kolmogorov-Smirnov test works when testing whether X distribution comes from a uniform distribution from 2 to 10. You do not need to perform all the test, just explain through examples how does it work (3 points)
 - k. group x and y into two intervals of your choice and test whether variables X and Y are independent (3 points)
6. To use the t statistic to test for a difference between the expected values of two populations, what assumptions must be made? (1 point)

7. Two populations are described in each of the following situations. In which situations would it be appropriate to apply the small-sample t test to investigate the difference between the population means? (3 points)
- Population 1: Normal distribution with variance σ^2 . Population 2: Not normal with variance σ^2 .
 - Population 1: Normal distribution with variance σ^2 . Population 2: Normal distribution with variance different from σ^2 .
 - Population 1: Not normal with variance σ^2 . Population 2: Not normal with variance σ^2 .
 - Population 1: Normal distribution with variance σ^2 . Population 2: Normal distribution with variance σ^2 .
 - Population 1: Uniform distribution with variance σ^2 . Population 2: Uniform distribution with variance σ^2 .
8. Suppose you have performed Mann-Whitney test on two samples and you have rejected with $\alpha = 10\%$. What can you conclude? (2 points)
9. Independent random samples from 2 normal populations produced the results Sample 1: 1.2 3.1 1.7 2.8 3.0; Sample 2: 4.2 2.7 3.6 3.9. Test whether the positions of the two distributions are the same. (3 points)
10. Independent random samples from 2 normal populations produced the results Sample 1: 1.2 3.1 1.7 2.8 3.0; Sample 2: 4.2 2.7 3.6 3.9. Test with level 5% whether the expected values of the two distributions are the same. (4 points)
11. Consider the goals scored by 3 football players in different years of professional career (data are not paired):

Estiarte	18	17	18	15	16	18
Ferretti	21	19	18	22		
Budavari	21	20	20	19		

Test whether the expected value of the goals is the same for every player or whether it is different. Write clearly prerequisites, hypotheses, statistic, decision and conclusion.

12. Suppose you have performed Student's t test for two variables on paired data and you have accepted with $\alpha = 10\%$. What can you conclude? (2 points)
13. What is the significance? (2 points)
14. Student t test for paired data yield these results. What is the rejection region for one-tailed test $H_1: E > 0$? (4 points)
- ~~$n = 12, \alpha = 5\%$~~
 - ~~$n = 24, \alpha = 10\%$~~
 - ~~$n = 4, \alpha = 2.5\%$~~
 - ~~$n = 8, \alpha = 1\%$~~
15. A paired difference experiment produced the following data: ~~$n = 18, \bar{x}_{1+8} = 92, \bar{x}_{2+8} = 95.5,$~~
 ~~$\text{Var}(x_1 - x_2) = 21.$~~ You want to prove that ~~$E(X_1 - X_2) < 0$~~ using ~~$\alpha = 10\%$~~ . Perform Student's t test. (3 points)
16. Suppose you have performed Fisher test on variances on two groups and you have got a confidence of 96%. What can you conclude? (3 points)
17. For ~~$\alpha = 5\%$~~ and ~~$\alpha = 10\%$~~ identify the rejection region that should be used to test ~~$H_1: \text{Var population 1} > \text{Var population 2}$~~ . Assume ~~$n_1 = 30$~~ and ~~$n_2 = 20$~~ . (4 points)

18. ~~Suppose you have a sample 3000 data with sample skewness equal to -0.02 and sample kurtosis equal to +0.1. Test whether data come from a normal distribution (writing clearly prerequisites, hypotheses, statistic, decision and conclusion).~~
19. Suppose you have performed ANOVA on four groups (and prerequisites hold) and you have got a significance of 30%. What can you conclude? (2 points)
20. ~~Independent random samples of 12 and 27 cases were selected from each of two normally distributed populations. The means and variances for the two samples are $\bar{x}_{12} = 31.7$, sample variance 1 = 3.87, $\bar{x}_{27} = 37.4$, sample variance 2 = 8.75. Test with significance level 10% whether the variances of the populations are different. (3 points)~~
21. A multinomial experiment with $k = 3$ sells and $n = 310$ produced the following data: $n_1 = 62$, $n_2 = 79$, $n_3 = 169$. Test using $\alpha = 0.05$ for probability distribution $p_1 = p_2 = 0.25$, and $p_3 = 0.5$. (4 points). Hint: you must insert 310 data into SPSS
22. How many dummy variables do we need to model the seasonal component in a regression model in the case when the "seasons" are the months of a year? Explain. (2 points)
23. Compute Spearman's rank correlation coefficient for the following pairs of observations:
 x 1.7 3.5 6.2 1.7
 y 2.4 1.6 5.4 2.9 (3 points)
24. You are doing a Wilcoxon signed rank test on two variables on the same sample. Explain how can you deal with the problem of couples having the same value on the two variables, for two-tailed and for one-tailed tests. (2 points)
25. What is a non-parametric test? And a rank test? (2 points)
26. ~~Suppose your data from a multinomial experiment are:~~

	A	B	C
X	10	20	10
Y	40	50	20

~~Test whether the two variables are dependent (writing clearly prerequisites, hypotheses, statistic, decision and conclusion).~~ Hint: you must insert 150 data into SPSS

27. Suppose your sample data are:

X	4	5	5.5	4.5	3.5	10	8	9	16	17
---	---	---	-----	-----	-----	----	---	---	----	----

Test whether the median is smaller than 5.3 (writing clearly prerequisites, hypotheses, statistic, decision and conclusion). ~~Perform the test again this time using a large sample approximation.~~

28. You know the 2007 income of 30 subjects. The next year you manage to interview 20 of these subjects and ask their 2008 income. Explain all the method which you would use to test whether the income has increased (5 points).
29. What are the major complications in applying the sign test to a population having a discrete distribution? (2 points)
30. The following random sample was selected from a population with a continuous distribution: 4.31, 3.99, 3.43, 3.72, 2.78, 2.65, 3.13, 2.53, 2.92, 3.76. Test whether the median of the distribution is 3.80 with significance level of 5%. Suggest several ways to compute the p-value for small sample and a possible way for large sample approximation. Can we be almost sure that the median is smaller than 3.80? (6 points)

31. Suppose you test two fertilizers, A and B. Applying A, you have got the following crops (per hectare): 37, 40, 33, 29, 42, 33, 35, 28, while with B, the yields were: 65, 35, 47, 52. You do not have any idea on the crops' distributions. Test whether the fertilizers are equally efficient. Use $\alpha = 0.1$. (4 points)
32. ~~Data were collected from three populations A, B and C. The numbers of observations are $n_A = n_B = n_C = 15$ and the ranks sums are $R_A = 230$, $R_B = 440$, $R_C = 365$. Test whether the corresponding distributions differ in location. Use $\alpha = 0.05$. (3 points)~~
33. A multinomial experiment with 4 cells and $n=205$ produced the data shown in the table. Do these data provide sufficient evidence to conclude that the multinomial probabilities are not uniform? Test using $\alpha = 5\%$. What are the Type I and Type II errors in this case? (4 points) Hint: you must insert 205 data into SPSS

1	2	3	4
43	56	59	47

34. ~~Consider this 2 by 3 contingency table. Is row classification independent from column classification? (4 points)~~

9	34	53
16	30	25

35. Determine whether the distribution for population B is shifted to the right of the distribution for population A. Suppose you obtained the following independent random samples of observations on experimental units: Sample A: 37 40 33 29 42 33 35 28 34 Sample B: 65 35 47 52. (5 points)
36. ~~A paired difference experiment with $n = 30$ pairs yielded $t_{\bar{}} = 354$ (with B - A). Test whether the probability distribution for population A is located to the right of that for population B. (4 points)~~

Regression models

1. What can you tell about the estimated slope, $\hat{\beta}_1$, of a simple linear regression model if the Pearson moment product coefficient of correlation equals -0.5 ? Explain. (2 points)
2. What can you tell about the usefulness of a simple linear regression model, if the Pearson moment product coefficient of correlation equals -0.81 ? Explain. (1 point)
3. Explain why the residuals of a linear regression model are dependent. (2 points)
4. Suppose you fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + E$ to $n = 20$ data points and obtain $\sum_{i=1}^{20} (y_i - \hat{y}_i)^2 = 12.37$ and $\sum_{i=1}^{20} (y_i - \bar{y}_i)^2 = 23.75$. Calculate r^2 . Does this value of r^2 suggest that the model provides a good fit to the data? Test the usefulness of this linear regression model. Use $\alpha = 0.05$. (5 points)
5. Explain how does the least squares approach work for a simple linear regression model. (4 points)
6. Explain how does the least squares approach work for the nonlinear regression model $Y = \beta_0 + \beta_1 \ln x$ (you do not need to solve the linear system) (4 points)
7. Suppose your data are

x	0	1	2	3	4	5	6	7	8	9	10
y	1.1	2.3	5.0	9.9	17.1	26.0	37.3	50.4	64.8	81.9	100.9

Which regression model would you use to estimate Y using x ? Which coefficient of determination do you expect to have? (3 points)

8. Assume that you have performed the Durbin-Watson test and it revealed that the error are correlated. Would you still use the coefficient of determination to assess the quality of your linear regression? Explain. (2 points)
9. You have performed a Durbin - Watson test and its statistic is 0.08 . Would you use the F (or t in the one-dimensional case) distribution to assess the usefulness of the linear regression model in this case? Explain. (3 points)
10. Suppose you fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + E$ with $n = 30$ data points and obtain $sse = 0.46$ and $d = 0.87$. Do the values suggest that the model provides a good fit to the data? Explain. Is the model of any use in predicting Y ? (4 points)
11. Write a model relating Y to one qualitative independent variable with 4 categories. Illustrate the meaning of all the terms in your model. (3 points)
12. The next table presents the quarterly sales index for one brand of calculator at a campus bookstore.

Year	First Quarter	Second Quarter	Third Quarter	Fourth Quarter
1992	438	398	252	160
1993	464	429	376	216
1994	523	496	425	318
1995	593	576	456	398
1996	636	640	526	498

We defined the time variable as $t = 1$ for the first quarter of 1992, $t = 2$ for the second quarter of 1992, etc. The seasonal dummy variables are Q_1 , Q_2 and Q_3 . The resulting least squares equation is $\hat{y} = 119.85 + 16.51 t + 262.34 Q_1 + 222.83 Q_2 + 105.51 Q_3$ with $r^2 = 0.9692$. Interpret the least squares estimates, and evaluate the usefulness of the model. Find the forecasts for the 1997 third quarter sales. (5 points)

13. For each situation indicate the decision regarding the test with alternative hypothesis of positive first order autocorrelation. k is the number of independent variables. (4 points)

- a) ~~$k = 2, n = 20, \alpha = 0.05, DW \text{ statistic} = 1.1$~~
- b) ~~$k = 2, n = 20, \alpha = 0.01, DW \text{ statistic} = 1.1$~~
- c) ~~$k = 5, n = 65, \alpha = 0.05, DW \text{ statistic} = 0.95$~~
- d) ~~$k = 1, n = 31, \alpha = 0.01, DW \text{ statistic} = 1.35$~~

14. Consider the following sample of data with X as intensity of a physical exercise and Y as heartbeats

X	0	0	0	0	0	5	5	5	10	10
Y	4	5	5.5	4.5	3	10	8	9	14	17

Estimate the coefficients for a linear regression model and explain their meaning. Calculate coefficient of determination. Test whether the model is useful (writing clearly prerequisites, hypotheses, statistic, decision and conclusion).

- 15. Explain how does the least squares approach work for estimating β_0 and β_1 in the nonlinear regression model $y = \beta_0 + \beta_1 \text{Exp } x$ (you must write also the linear system but you do not have to solve it).
- 16. Suppose you are building a simple linear regression model $y = \beta_0 + \beta_1 x$ where your data y_j and x_j are sums of data calculated on a different amount of elements (you are using aggregated data with different n_j). Can you perform usefulness test? Explain.