9
10
12
8
4
50

22S:30/105, Statistical Methods and Computing
Spring 2014, Instructor: Cowles
Final Exam

Name: Solutions Course no. (30 or 105) _____

1. Briefly explain the statistical error in each of the following statements.

   (a) The researchers found a strong positive correlation between employment status (working, unemployed, retired) and happiness.

   2 Correlation measures the strength of the linear relationship between 2 quantitative variables. Employment status is qualitative.

   (b) The researchers found a strong positive correlation between annual income and score on a happiness test. This proves that money makes people happy.

   2 "Association does not imply causation." If income is correlated with happiness, that does not prove that the income causes happiness.

   (c) The researchers reported a p-value of 0.032 for their hypothesis test. This means that there is 0.032 probability that the null hypothesis is true.

   2 No. A p-value is the probability of getting data with at least as much evidence against Ho as the current data, if Ho is true.

2. This problem uses data from collected in West Africa for the years 1971 - 1999. In each year two variables were recorded:

   - fish – the fish supply in kilograms per person
   - biomchg – the change in total weight of animals in nature reserves

   The expectation was that, in years in which the fish supply was smaller, declines in animals would be greater because the local populace would turn to such animal meat when other sources of protein were reduced.

   We will use linear regression to see whether the data confirm this expectation.

   The attached SAS output is needed to answer some of the following questions.

   (a) In the scatterplot, which variable – fish or biomchg – is treated as the response variable?

   1 biomchg

   (b) Does the scatterplot indicate a somewhat linear relationship? ([✓ yes] no) If so, is the relationship positive or negative?

   positive

   2

9)

(c) SAS proc corr gives the sample correlation coefficient between fish and biomchg as 0.67. Would this correlation get larger or smaller if the outlier were removed from the dataset. Explain why in one or two sentences.

2 The correlation would get larger if the outlier would be removed because correlation measures how closely points are clustered around the regression line.

(d) Does the residual plot give evidence of any violations of the assumptions of linear regression? Explain briefly.

There is one outlier. However, there is no curved pattern and no systematic increase or decrease in spread around the zero line

3

(e) I claim that the outlier will not be very influential on the regression analysis. Why is that the case?

2 The outlier is not at the extreme edge of the x-axis (not an outlier in predictor variable). There are many other points near it in x-axis. The line could not be adjusted closer to the outlier without becoming farther from these other points.

(f) Give the least squares regression line for the regression of biomchg on fish.

3 $\widehat{\text{biomchg}} = -21.09 + 0.63\,\text{fish}$

(g) Write the null hypothesis that means that there is no linear relationship between biomchg and fish. Use conventional statistical symbols.

2 $H_0: \beta = 0$

(h) Write the alternative hypothesis that expresses the expectation stated in the problem description.

2 $H_A: \beta > 0$ 1/2 if $\beta \neq 0$

(i) Write the test statistic value and p-value (numbers from SAS output) for carrying out the test of the hypotheses that you stated.

3 $t = 4.72$

p-value < 0.0001 ← divide by 2 for 1-sided test

2 if don't divide

17

(j) What is your conclusion? Does the data contain evidence in favor of the expectation stated? Explain briefly. Yes

3 We can reject the null hypothesis of no linear association and conclude that lower fish availability is associated with lower animal biomass.

3. Investigators are interested in whether men or women have better perception of color. They design a test of color perception, in which higher numeric scores are better. They know that age affects vision in general and may well affect color perception. To balance out the effects of age, they recruit the participants in their study in the following way. They first recruit 10 men into their study. Then for each male subject, they recruit one female subject of the same age. Thus, they end up with a sample of 10 men and a sample of 10 women. Which statistical procedure is most likely to be best for their study? (Circle one.)

(a) ANOVA

(b) chi square test

1 [✓ (c) paired t-test]

(d) two-independent-sample t-test

4. It is expected that rainfall in California will increase over the coming decades, but it is not known whether the increase will occur only during the winter wet season or whether there also will be more rain in the spring and summer, which historically have been dry.

Researcher Kenwyn Suttle of the University of California at Berkeley did an agricultural study to investigate the effect of three possible conditions on the plant biomass in open grassland:

- Control (no added water)
- Winter (add water equal to 20% of annual rainfall during January to March)
- Spring (add water equal to 20% of annual rainfall during April to June)

They randomly assigned 6 plots of open grassland to each of the three conditions. This problem concerns their results obtained during the 2001 growing season. SAS output is attached.

(a) The researchers wish to use ANOVA to determine whether the mean biomass produced per acre is different under the three conditions. Write the null and alternative hypotheses that they will test. Use standard statistical symbols, and briefly define the symbols that you use.

3 μ_c = population mean biomass in fields under control condition

$H_0: \mu_c = \mu_W = \mu_S$

$H_A: \mu_c \neq \mu_W$ and/or $\mu_c \neq \mu_S$ and/or $\mu_W \neq \mu_S$

(b) What assumptions must be met in order for the results of ANOVA to be trust-

5 4 if leave out SRS

12

worthy? State each one, and then comment whether the assumption appears to met with these data. Refer to specific SAS output when appropriate.

Sample is simple random sample. Cannot be verified by looking at data.

Each population distribution is normal. There is an outlier in the box plot of the control sample and winter sample is skewed so this may not be true.

Equal standard deviations in all populations. Smallest sample standard deviation (11.2) < $\frac{1}{2}$ largest one (49.6) No!

(c) Suppose that the researchers proceeded to do an ANOVA analysis and will do the overall hypothesis test at significance level alpha = 0.05.

i. Should they reject the null hypothesis that you stated above? Justify your answer referring to specific SAS output.

2 Yes. F statistic is 43.79, yielding a p-value < .0001, which is smaller than alpha.

ii. What does the result you gave in the previous question mean regarding biomass and extra water in different seasons?

2 There are real differences in mean biomass between at least 2 of the water conditions.

(d) Which means (if any) are significantly different from each other? Refer to specific SAS output.

$\mu_S \neq \mu_W$

2 $\mu_S \neq \mu_C$

In the "Bon grouping," μ_S has the letter A while the other 2 have the letter B.

(e) Are you convinced by these results? Why or why not?

2 No. With the violations of assumptions stated in part (b), these results cannot be trusted.

8

5. Sociologists want to determine the average number of hours per day that teenagers spend using social media such as Facebook. Suppose that they believe that the number of hours that individual teenagers spend daily on social media follows a normal distribution with known standard deviation 1.5 hours.

(a) How many teenagers will they need in their sample in order to obtain a 90% confidence interval of width no greater than 1? (Numeric answer. Show your work.)

3

width = 1 so margin of error is 0.5

90% confidence, so $z^* = 1.645$

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = \left(\frac{1.645\,(1.5)}{0.5}\right)^2 =$$

(b) What quantity will the sociologists be 90% confident lies in their interval? (Circle one)

1

[✓ i. μ]

ii. p

iii. $\hat{p}$

iv. s

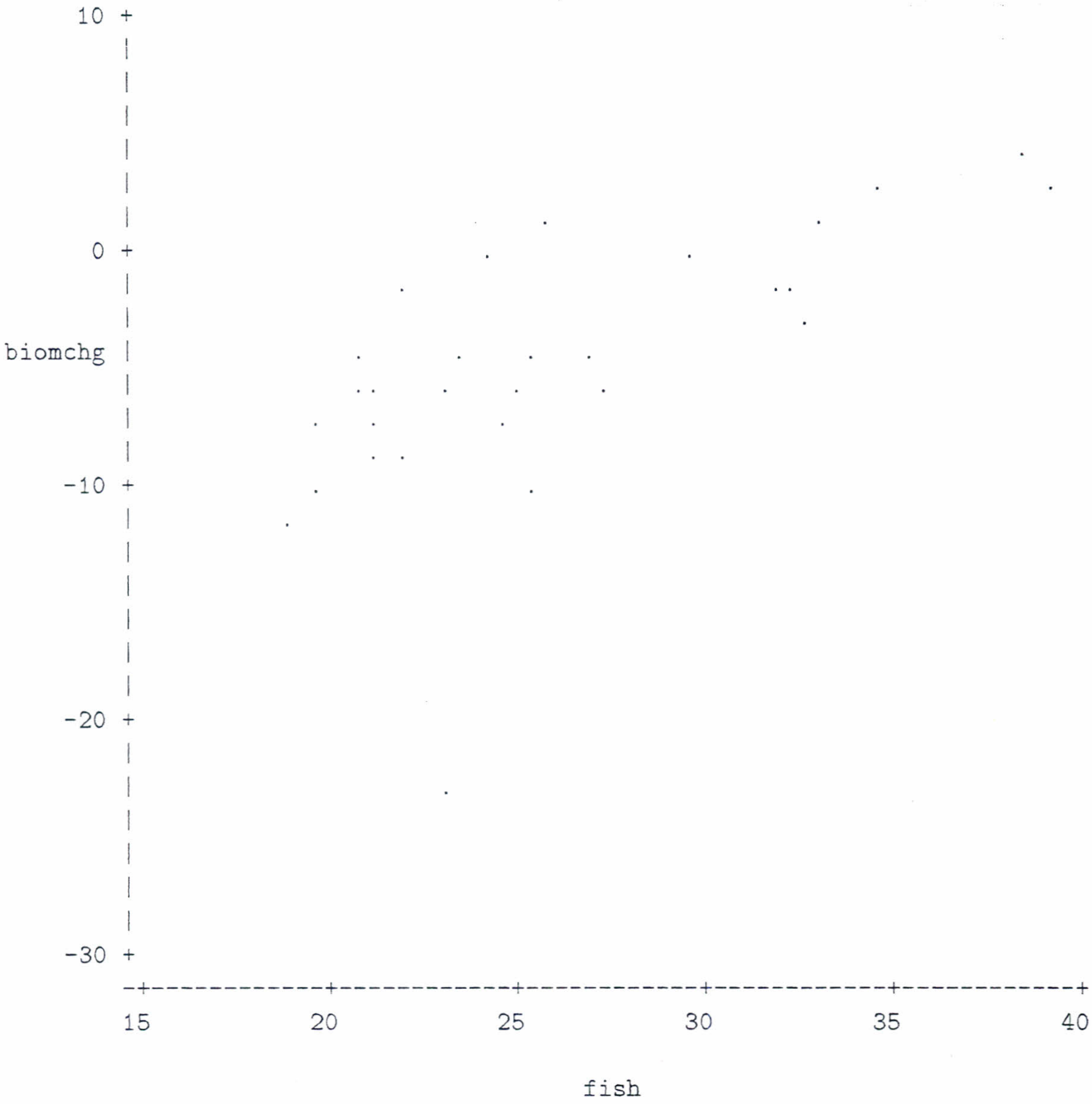
v. σ

vi. $\bar{x}$

4

SAS output for problem 2

Plot of biomchg*fish. Symbol used is '.'.


biomchg
fish