2
14
10
12
15
7
60

STAT:2010/4200, Statistical Methods and Computing
Spring 2019, Instructor: Cowles
Final Exam

Name: Solutions new

1. A study assessed the effectiveness of a new drug designed to reduce repetitive behaviors in children affected with autism. A total of 8 children with autism enrolled in the study and the amount of time that each child engaged in repetitive behavior during a three-hour observation period was measured both before treatment and then again after the child took the new medication for a period of 1 week.

   The research question is whether the medication reduces the number of repetitive behaviors.

   Since this is a paired-sample situation, I calculated the differences (after - before) for the children and used those as the study variable. Here is some SAS output for this variable.

```
                Tests for Location: Mu0=0

Test           -Statistic-       -----p Value------

Student's t    t  -2.05996       Pr > |t|     0.0784
Sign           M        -2       Pr >= |M|    0.2891
Signed Rank    S       -14       Pr >= |S|    0.0625


Stem Leaf                        #            Boxplot
   1 0                           1               |
   0 5                           1               |
  -0                                          +-----+
  -1 500                         3            *--+--*
  -2 50                          2            +-----+
  -3
  -4
  -5
  -6 0                           1               0
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

2 (a) What is the lowest value in the dataset? (Numeric answer)

$-6.0 \times 10 = -60$

(b) What is the best statistical test for using these data to address the research question? (Circle one.)

2/3 ½ for paired t

i. paired t-test
ii. sign test
iii. Wilcoxon signed rank test — because we have small sample with outlier
iv. none of the above

(c) What is the p-value of the test you chose? (Numeric answer)

4 p-value for 2-sided test is .0625. The data goes in the direction of supporting the alternative hypothesis (mean difference is negative). So p-value for 1-sided test is .0625/2 = .0312.

2 (d) At the .05 significance level should you reject the null hypothesis that the medication makes no difference? (yes/no)

(e) The appropriate conclusion is: (Circle one)

i. The data proves that the medication makes no difference in the number of repetitive behaviors.
ii. The data proves that the medication reduces the number of repetitive behaviors.
2 iii. The data provides evidence that the medication reduces the number of repetitive behaviors.
iv. The data does not provide evidence that the medication reduces the number of repetitive behaviors.
v. None of the above

2. A researcher is interested in the effects of sleep deprivation and caffeine consumption on reaction times in a driving test. She will recruit 48 people into her study. Each study participant will be randomly assigned to one of three sleep conditions (3 hours, 6 hours, or 9 hours) the night before being tested in a driving simulator as well as to one of two caffeine levels (no coffee or two cups of coffee) before the test. The variable recorded for each participant is his or her mean reaction time in tenths of a second to various stimuli presented during the simulated driving test.

(a) This study is: (circle one)

i. an observational study
2 ii. an experiment
iii. insufficient information is provided to determine which

(b) The factors are: (circle one)

2 i. caffeine consumption and sleep deprivation
ii. the 48 participants
iii. mean reaction times
iv. coffee or no coffee and 3 hours, 6 hours, or 9 hours of sleep

14

(c) The response variable is: (circle one)

2

i. caffeine consumption and sleep deprivation

ii. the 48 participants

iii. mean reaction times

iv. coffee or no coffee and 3 hours, 6 hours, or 9 hours of sleep

(d) How many treatments are possible? (numeric answer)

2

6: 3 levels of sleep x 2 levels of caffeine

3. For each of the following scenarios, write which of the following statistical procedures would be most likely to be appropriate.

- ANOVA
- chi square test
- correlation
- Kruskal-Wallis test
- regression
- sign test
- t-test
- Wilcoxon rank sum test
- Wilcoxon signed rank test

(a) Agronomists wish to assess the strength of the linear relationship between pounds of herbicides applied to fields and numbers of honeybee deaths in nearby hives.

2 correlation 1 for regression

(b) Educators wish to determine whether the proportion of students receiving A grades is the same in undergrad courses in Physics, Math, and Computer Science.

2 chi square test

(c) Economists wish to assess whether the distribution of income is the same in workers in Actuarial Science, Statistics, and Data Science.

2 Kruskal Wallis test (better than ANOVA because distributions of income are skewed) 1/2 for ANOVA

10

4. Between birth and age 3 months, babies' heart rates follow a normal distribution with mean 143 beats per minute and standard deviation 19 beats per minute.

   (a) What proportion of babies in this age range have heart rates faster than 150 beats per minute? (Numeric answer; show your work.)

   $Z = \frac{150-143}{19} = .368$

   5

   $Pr(Z > 0.368) = 1 - 0.644 = 0.356$
   from Table A

   (b) Consider drawing simple random samples of size 36 from the population of babies between birth and age 3 months. In what proportion of such samples will the sample mean $\bar{x}$ be greater than 150 beats per minute? (Numeric answer; show your work.)

   $Z = \frac{150-143}{(19/6)} = 2.21$

   5

   $Pr(Z > 2.21) = 1 - 0.9864 = 0.0136$

5. The Central Limit Theorem says (circle one):

   (a) when the sample size n is large, the shape of the sampling distribution of the sample mean $\bar{x}$ is approximately normal regardless of the shape of the population distribution of individual values.

   2

   (b) when the sample size n is large,the shape of the population distribution of individual values is approximately normal.

   (c) if your sample size n is at least 30 you will live happily ever after.

   (d) none of the above.

12

6. In the 1970s, smoking rates in the U.S. were much higher than they are now. An insurance company wanted to study cigarette consumption rates in all 50 states and the District of Columbia. Among the data that they collected on each state and D.C. were the following three variables:

```
income -- per capita income in the state in dollars
price  -- average price of a pack of cigarettes in the state in cents
sales  -- per capita number of packs of cigarettes sold in the state
```

In answering the following questions, refer to the attached SAS output for two regression models using these data.

(a) In both models, sales is the (circle one):

i. predictor variable

ii. lurking variable

iii. response variable

iv. none of the above

2/2

(b) Which variable, income or price, explains more of the variability in sales? Justify your answer by citing specific SAS output.

Income explains slightly more. Its $R^2 = .1063$ whereas for price, $R^2 = .0904$

3

(c) Based on model 2, what is the expected value of sales in a state in which the price of a pack of cigarettes is 35 cents? (Numeric answer; show your work – and yes, cigarettes really were that cheap in the 1970s.)

3 $\hat{y} = 210 - 2.33521(35) = 128.72$

(d) Based on the following line from model 1, explain to a nonstatistician what the number 0.01758 means about sales and income.

```
income       1       0.01758       0.00728       2.41       0.0195
```

For every 1 dollar increase in per capita income we expect on average a 0.01758 unit increase in per capita number of packs of cigarettes sold.

5

(e) Based on model 1, what quantity are we 95% confident is in the interval (0.00295, 0.03222)? Use the conventional symbol.

2/2 The population slope β

15

(f) From model 1, what is the p-value for testing the null hypothesis that the population slope of `sales` on `income` is 0? (Numeric answer)

2

0.0195

(g) Circle **all** of the following that are assumptions of simple linear regression.

i. equality of variance of the errors

ii. normality of the predictor variable

5

iii. linear relationship between the response variable and the predictor variable

iv. independence

v. none of the above

7

Model 1

The REG Procedure
Model: MODEL1
Dependent Variable: sales

Number of Observations Read	51
Number of Observations Used	51

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5467.56714	5467.56714	5.83	0.0195
Error	49	45958	937.91584		
Corrected Total	50	51425			

Root MSE	30.62541	R-Square	0.1063
Dependent Mean	121.54118	Adj R-Sq	0.0881
Coeff Var	25.19756		

The REG Procedure
Model: MODEL1
Dependent Variable: sales

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.36245	27.74308	2.00	0.0516
income	1	0.01758	0.00728	2.41	0.0195

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-0.38936	111.11426
income	1	0.00295	0.03222

Model 2

The REG Procedure
Model: MODEL2
Dependent Variable: sales

Number of Observations Read	51
Number of Observations Used	51

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4647.52138	4647.52138	4.87	0.0321
Error	49	46778	954.65147		
Corrected Total	50	51425			

Root MSE	30.89743	R-Square	0.0904
Dependent Mean	121.54118	Adj R-Sq	0.0718
Coeff Var	25.42137		

The REG Procedure
Model: MODEL2
Dependent Variable: sales

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	210.45305	40.52848	5.19	<.0001
price	1	-2.33521	1.05837	-2.21	0.0321

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	129.00802	291.89808
price	1	-4.46208	-0.20833