

## Stat 2000 International – Sample Final Exam

The Final Exam consists of 50 questions: 25 multiple-choice questions (with exactly 1 correct answer), 10 text-based questions where you have to provide a verbal explanation or calculate one or multiple numerical values, and 15 questions that include the use of WebStats, data analysis, and interpretation of results.

The Final Exam is worth a total of **450 points**. The number of points for each question is indicated in parentheses at the beginning of each question. You have **3 hours** to complete the Final Exam. Try to correctly answer as many questions as possible during this time period. **You are allowed to answer questions in any order.** Start with a question that seems the easiest for you. If you cannot answer a question within a short time, move to another question, and come back to the previously unanswered questions toward the end of the exam.

Obviously, you are allowed to correct your answers. However, only your last submitted answer will be graded. If you change a previously correct answer and your last submitted answer is incorrect, you will obtain 0 points for your last submitted answer.

**Note: This Sample Final Exam includes only 10 multiple-choice questions, 5 text-based questions, and 15 WebStat questions. It does not include multiple-choice or text-based questions from Units that were covered in the first two Midterms. However, the actual Final Exam WILL include that portion of the material, covered in the remaining 20 questions.**

The actual exam will be fully given within CyberStats. This means you will have to mark your choices of multiple-choice questions and fill in answers to text-based questions within CyberStats. In the actual exam, interactivities will be directly linked to the questions. Make sure to memorize your CyberStats password for the exam.

1. (8 Points) A coin is tossed 10,000 times to see if it is fair (i.e., 'heads' and 'tails' are equally likely). In particular, the investigator thought that a head came up more than it should. Let  $p$  be the probability of a head. If the coin is fair, then  $p = 1/2$ .

What is the null hypothesis?

- a.  $H_0: p < 1/2$
- b.  $H_0: p = 1/2$
- c.  $H_0: p > 1/2$
- d.  $H_0: p = 1/2$
- e. no answer or skip this item

2. (8 Points) Researchers are interested in testing whether there are an excessive number of rat hairs in jars of peanut butter produced at a particular factory. They examine a random sample of 144 jars, and find an average of 6.3 rat hairs in each jar. The sample standard deviation is 2. They would like to do a one-sided z-test of whether the population average is equal to five (the maximum permitted by law) versus the alternative that it is greater than five.

What is the z-statistic?

- a. 7.8
- b. -7.8
- c. 0.65
- d. 0.35
- e. no answer or skip this item

3. (8 Points) A salt-free diet is examined to determine the effect upon blood pressure. The change in diastolic blood pressure for 64 subjects was examined after the subjects had followed a salt-free diet for one month. The researcher measured the differences in blood pressure before going on the salt free diet and after the subjects were on the salt-free diet for one month:

Difference = Blood pressure after - Blood pressure before.

If the salt-free diet changes the diastolic blood pressure, the population mean of the differences is different from 0. The sample mean of the differences was -1.2 mm and the sample standard deviation of the differences was 2.6.

What is the null hypothesis?

- a. The sample mean of the differences is -1.2
- b. The sample mean of the 64 subjects is 0
- c. The population mean of the population is 0
- d. The population mean of the population is -1.2
- e. no answer or skip this item

4. (8 Points) Which statement best describes a t-interval?
- a. A t-interval is a significance testing procedure used to evaluate hypothesis statements about means
  - b. A t-interval is a tool used to answer questions about the population mean when the population standard deviation is known
  - c. A t-interval is a confidence interval estimate of the population mean
  - d. A t-interval is a useful model for describing the behavior of sample means when the population standard deviation is known
  - e. no answer or skip this item

5. (8 Points) The output from a hypothesis test was:

| Test of $\mu = 12$ vs $\mu \text{ not } = 12$ |       |       |         |      |       |
|---|-------|-------|---------|------|-------|
| N   | Mean  | StDev | SE Mean | T    | P     |
| 23  | 14.61 | 4.11  | 0.857   | 3.04 | 0.006 |

Which sentence interprets the p-value of this test?

- a. If the population mean is really 14.61, the probability is 0.006 that the mean of 23 values could be different from 12.
  - b. If the population mean is really 14.61, the probability is 0.006 that the mean of 23 values is 12.
  - c. If the population mean is really 12, the probability is 0.006 that the mean of 23 values is 12.
  - d. If the population mean is really 12, the probability is 0.006 that the mean of 23 values could be different from 12.
  - e. no answer or skip this item
6. (8 Points) For which of the samples in this table would it be all right to use the p-value based on the z-test?

| Sample # | # who party | # who don't party |
|----------|-------------|-------------------|
| 1        | 4           | 25                |
| 2        | 235         | 53                |

- a. Sample #1 is large enough to use the z-test.
- b. Sample #2 is large enough to use the z-test.
- c. Both sample #1 and sample #2 are large enough to use the z-test.
- d. Neither sample #1 and sample #2 are large enough to use the z-test.
- e. no answer or skip this item

7. (8 Points) Suppose you were given a 95% confidence interval for the difference in two population means.

What could you conclude about the two population means if the confidence interval contained both negative and positive numbers?

- a. The first mean is larger than the second mean
  - b. The first mean is smaller than the second mean
  - c. The means differ, but it is not clear which is larger
  - d. The means do not differ
  - e. no answer or skip this item
8. (8 Points) A chi-square test of the relationship between alcohol consumption and depression led to rejection of the null hypothesis, indicating that there is a relationship between these two variables.

One conclusion that can be made is:

- a. Depression leads to alcohol consumption
  - b. Alcohol consumption leads to depression
  - c. The more alcohol one consumes, the more likely they are to be depressed
  - d. There are likely to be confounding variables related to both alcohol consumption and depression
  - e. no answer or skip this item
9. (8 Points) Imagine a population of grasshoppers that starts with eight grasshoppers. Every day, the number of grasshoppers doubles. (Thus after one day, there are 16 grasshoppers, and after two days, there are 32 grasshoppers.) Let Y be the number of grasshoppers and X be the number of days that have passed.

Is there a linear model  $Y = a + b \times X$  that describes this relationship? If so, what are a and b?

- a. No, there is no linear relationship
  - b. Yes.  $a = 10$  and  $b = 2$
  - c. Yes.  $a = 10$  and  $b = 10$
  - d. Yes.  $a = 0$  and  $b = 20$
  - e. no answer or skip this item
10. (8 Points) For 11 students, the least squares line for  $X = \text{ShoeSize}$  versus  $Y = \text{Height}$  is

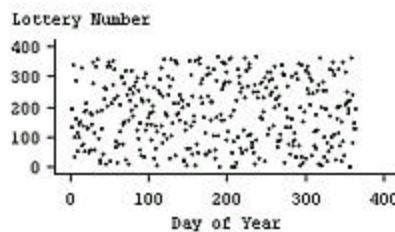
$$\text{Height} = 53.17 + 1.70 \times (\text{ShoeSize}).$$

The correlation coefficient is 0.92, the sum of squared deviations of the ShoeSizes is 15.68, sum of squared deviations of the Heights is 53.34, and the sample standard deviation of the residuals is 0.95.

The least squares estimate of the slope is:

- a. 53.17
- b. 1.70
- c. 0.92
- d. 0.95
- e. no answer or skip this item

11. (10 Points) **Example 1** from **Unit C-3: The Concepts of Hypothesis Tests** looked at the 1969 draft lottery and concluded that it was not totally random. The correlation coefficient between the day of the year and the lottery number was not 0. In 1970, the lottery was repeated, but a better randomization procedure was allegedly used. Here is the plot for 1970:



If the lottery is totally random, then the population correlation coefficient is 0, so we wish to test

$H_0$ : Population correlation coefficient = 0

- a. Suppose any deviation from the null hypothesis is bad. What is the alternative hypothesis?

- b. Suppose one rejects the null hypothesis if the p-value is no more than 0.05. If possible, give the probability of a Type I error. If possible, give the probability of a Type II error.

12. (10 Points) What would the temperature, in degrees Fahrenheit, be for you on an ideal summer day? A meteorologist posed this question to a random sample of 17 people. He analyzed their data using MINITAB. Use the meteorologist's output to the following questions:

| Test of mu = 80.00 vs mu < 80.00 |    |       |       |         |       |      |
|----------------------------------|----|-------|-------|---------|-------|------|
| Variable                         | N  | Mean  | StDev | SE Mean | T     | P    |
| Temp                             | 17 | 78.59 | 5.70  | 1.38    | -1.02 | 0.16 |

- a. What are the meteorologist's null and alternative hypotheses?

- b. How likely is it that a sample of 17 people would have an average ideal temperature of 78.6 degrees or less if, in fact, the population average ideal temperature were 80 degrees?

- c. Should the meteorologist reject or not reject his null hypothesis? Explain.

- d. Write one sentence that summarizes the meteorologist's conclusion.

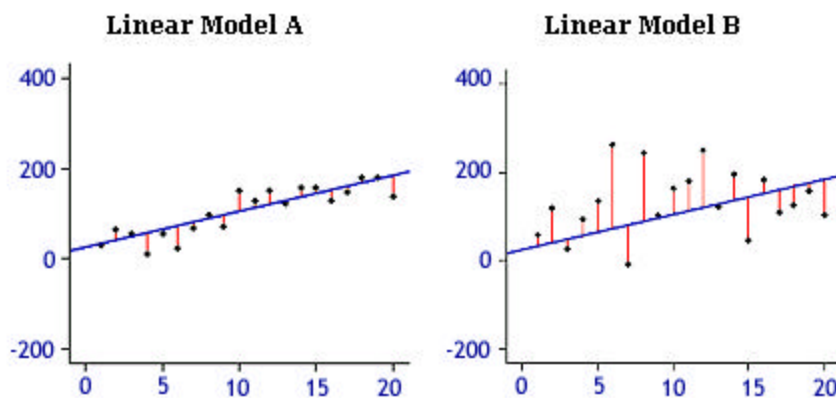
13. (10 Points) A taste test is conducted to determine if people can distinguish the tastes of two different soft drinks. Each participant tastes beverages from three different cups, randomly presented.

- In two cups, the beverages are the same.
- In a third cup, the beverage is different from the beverage in the other two cups.

Each participant attempts to identify the cup that contains the unique beverage. There are 90 participants, of whom 36 are able to correctly identify the unique beverage.

The researchers want to see if people generally are able to identify the unique beverage. Write an alternative hypothesis about the proportion of the population who can identify the unique beverage.

14. (10 Points) Here are two linear models with error. They both have  $Y = 25 + 8X + E$ , but the standard deviations of the residuals "E" are different.



Which linear model has the higher standard deviation of residuals E?

15. (10 Points) For the following examples, specify the appropriate null and alternative hypotheses:
- a. Fifty cars coming off of the paint assembly line on each day of the week (Monday, Tuesday, Wednesday, Thursday and Friday) were classified according to whether they had more than 5 obvious defects in the paint job (Yes, No).

- b. A random sample of 200 children beginning kindergarten were classified according to whether they could spell their names and whether they had attended preschool.

36. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set "Milk\_pH.dat". This is one out of 8 questions that will work with this data set.

Calculate (and report) the mean, median, and standard deviation for the variable milk-pH.

37. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set "Milk\_pH.dat". This is one out of 8 questions that will work with this data set.

Construct a boxplot of the variable milk-pH, using fences to identify outliers. Are there any outliers - and does the boxplot suggest that the data is fairly symmetric or whether there is skewness?

38. Use WebStat. Load from "Data -> Sample Data" the data set "Milk\_pH.dat". This is one out of 8 questions that will work with this data set.

Conduct a test for the mean milk-pH (assuming the data originates from a random sample), where  $H_0$ : mean = 6.7 versus  $H_1$ : mean not equal to 6.7. Clearly state whether you have to use a z-statistic or a t-statistic (in this case, indicate the degrees of freedom). Report the p-value and provide a verbal conclusion.

39. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set "Milk\_pH.dat". This is one out of 8 questions that will work with this data set.

Calculate (and indicate) the correlation between temperature and milk-pH. What does this value indicate?

40. Use WebStat. Load from "Data -> Sample Data" the data set 'Milk\_pH.dat'. This is one out of 8 questions that will work with this data set.

Take temperature as explanatory (independent) variable and milk-pH as response (dependent) variable. State the regression line for predicting milk-pH based on temperature.

41. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Milk\_pH.dat'. This is one out of 8 questions that will work with this data set.

Use your regression line from the question above to predict milk-pH for temperatures of 30 degrees and 100 degrees. Which of these 2 predicted values is more reliable?

42. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Milk\_pH.dat'. This is one out of 8 questions that will work with this data set.

Can we really assume that the slope is different from zero? State how you could conduct a test for  $H_0$ : slope = 0 versus  $H_1$ : slope not equal to 0 and report your conclusions.

43. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Milk\_pH.dat'. This is one out of 8 questions that will work with this data set.

If someone tells us that for a temperature of 40 degrees, a milk-pH value of 6.60 was observed, would this be surprising?

44. Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Construct a histogram of BMI (Body Mass Index), using a bin width of 5 and starting bins at 20. What do you notice?

45. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Construct a scatterplot of Height (x) versus BMI (y). Keep in mind that these are 2 groups of athletes (sport = 1 and sport = 2). What do you notice? Is there a distinction possible among the 2 sports?

46. Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Based on your observations in the previous question, for which sport does a simple linear regression model that predicts BMI from Height make more sense? Explain your answer.

47. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Construct histograms of Height for the two sports separately (group by sport and choose a separate graph for each group - start bins at 60 with a bin width of 5). Explain how these 2 histograms look.

48. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Construct (and report) regression lines for both sports separately, predicting BMI from Height.

49. (10 Points) Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Based on your regression calculations above, for which sport(s) do we have a significant slope, i.e., for which sport can we use Height to predict BMI. How does this relate to our initial graphical interpretation?

50. (10 Points) Use WebStat. Load from "Data -> Sample Data" the data set 'Male\_athletes\_heights-weights.dat'. This is one out of 7 questions that will work with this data set.

Based on your regression calculations above, what would be the predicted BMI value for someone from sport 2 with a Height of 75?