

**DIRECTIONS:**

- This exam contains three parts:
  - Part 1. 15 Multiple Choice [45 points]
  - Part 2. 10 Term Identification [20 points]
  - Part 3. 3 Short Answer/Computation [35 points]
- On Part 1, circle the correct response for each question. Make sure that your answer is clearly marked. You will not receive partial credit for any work done in Part 1. On Part 3, show all of your work to receive full (or partial) credit.
- This is a closed-book examination. However, you may use one  $8.5 \times 11$  sheet of notes if you wish. You may also use a calculator.
- The standard normal distribution is attached at the end of this examination.
- Any discussion or otherwise inappropriate communication between examinees, as well as the appearance of any unnecessary material, will be dealt with severely.
- This exam is worth a total of 100 points. **Print** your name **at the top of this page in the upper right hand corner**. *Good Luck!!*

**PART 1: MULTIPLE CHOICE.** Circle the correct response for each question. Make sure that your answer is clearly marked.

1. A survey records many variables of interest to the researchers conducting the survey. Which of the following variables is **quantitative**?

- (a) occupation of household head
- (b) total household income, before taxes
- (c) county of residence
- (d) party affiliation

2. What is the 95th percentile of the standard normal distribution?

- (a)  $-1.96$
- (b)  $-1.65$
- (c)  $1.65$
- (d)  $1.96$

3. A stem and leaf display is shown below. The stem is the tens place and the leaf is the units place.

0	3	6		
1	0	2	2	8
2	1	2	5	6
3	1			

What is the **median** of this distribution?

- (a) 12
- (b) 15
- (c) 17
- (d) 18

4. Which of the following statements is **false**?

- (a) Completely randomized designs may be inappropriate if experimental units are not similar.
- (b) Matched-pairs experiments are experiments where each experimental unit provides two responses.
- (c) Randomization is not needed in block designs.
- (d) Blocking is used to increase the scope and precision of the experimental results.

5. What are the three basic principles of experimental design?

- (a) bias, non-response, and undercoverage
- (b) power, variability, and consistency
- (c) extrapolation, lack of realism, and stratification
- (d) control, replication, and randomization

6. True or False. If two variables have a correlation  $r$  close to 1, then there is a cause and effect relationship between the variables.
- (a) True
  - (b) False
7. What is **extrapolation**?
- (a) a technique used to predict normally distributed data
  - (b) a graphical display for longitudinal measurements
  - (c) a prediction that is outside the range of the available data
  - (d) a bias associated with faulty survey forms
8. We want to portray the distribution of a **qualitative** variable. Which graphical display would be most appropriate?
- (a) time plot
  - (b) pie chart
  - (c) stem and leaf plot
  - (d) boxplot
9. A large data set consisting of stock market rate of returns (measured in percentages) has mean 5.3 percent and standard deviation 3.6 percent. A histogram of the percentages is approximately symmetric. The Empirical Rule says that approximately 99.7 percent (or nearly all) of the rate of returns should fall between which two values?
- (a)  $-1.9$  and  $12.5$  percent
  - (b)  $-5.5$  and  $16.1$  percent
  - (c)  $3.6$  and  $7.0$  percent
  - (d)  $1.7$  and  $8.9$  percent
10. Each sentence below includes the words (or forms of the words of) “estimate,” “statistic,” “population,” “sample,” and “parameter.” Which sentence is technically correct?
- (a) In a recent experiment from the *population*, *parameters* were used *statistically* to *estimate sample* information.
  - (b) A *statistic* is used to quantify some measure concerning a *population*, and *sample parameters* can be used as *estimates*.
  - (c) It will often be the case that *population parameters* are unknown; however, we can use *sample statistics* to *estimate* them.
  - (d) For most *populations*, *parameters* and *statistics* should always be equal, assuming a random *sampling* model was used to *estimate* the variables of interest.

11. In the language of experiments, what is a **block**?
- (a) a data frame that consists of all experimental units
  - (b) a group of individuals that are similar in some way
  - (c) a treatment that is given to the control group
  - (d) an analysis for replicated studies
12. I have  $n = 20$  observations of a quantitative variable  $X$  yielding data  $x_1, x_2, \dots, x_{20}$ . These data represent the number of accidents (per month) at a busy intersection in Jackson, Mississippi.

6 4 3 4 6 6 4 6 0 7 6 6 4 6 3 7 6 6 5 5

I have computed that  $\sum x_i = 100$ . What is  $\bar{x}$ ?

- (a) 4
  - (b) 5
  - (c) 6
  - (d) 10
13. On a television show I saw recently on C-SPAN, 64 participants provided their opinions about the current India-Pakistan situation (they did so by calling in to the show). Which type of sampling model best describes this sample obtained?
- (a) voluntary response sample
  - (b) stratified sample
  - (c) random sample
  - (d) systematic sample
14. In an agricultural experiment, a total of  $n = 20$  plots of land are studied. For each plot, I measured  $x$ , the amount of fertilizer applied and  $y$ , the yield. From the data, I computed  $r = -0.02$ . What does this suggest?
- (a) The amount of fertilizer applied and the yield must not be related.
  - (b) The slope of the least-squares regression line of  $y$  on  $x$  must be larger than zero.
  - (c) The average of the residuals obtained from the least-squares regression of  $y$  on  $x$  must be larger than zero.
  - (d) None of the above statements is correct.
15. I have 4 factors and each factor has 2 levels. In an experiment with a factorial treatment structure, how many observations are recorded if there are three replicates?
- (a) 6
  - (b) 8
  - (c) 12
  - (d) 24

**PART 2: TERM IDENTIFICATION.** Simply indicate the letter that corresponds to your answer.

- |                        |                            |
|------------------------|----------------------------|
| A. bias                | L. population              |
| B. influential point   | M. cluster sampling        |
| C. lurking variable    | N. stratified sampling     |
| D. confounded          | O. residual                |
| E. observational study | P. frame                   |
| F. symmetric           | Q. IQR                     |
| G. skewed              | R. density curve           |
| H. boxplot             | S. stemplot                |
| I. range               | T. explanatory variable    |
| J. variance            | U. sample                  |
| K. response variable   | V. least-squares principle |

1. The entire group of individuals (e.g. people) that we want information about is called the \_\_\_\_\_.
2. The \_\_\_\_\_ statistic is defined to be the maximum value minus the minimum value.
3. The \_\_\_\_\_ is a graphical display that uses the 5-Number Summary.
4. In a(n) \_\_\_\_\_ scheme, the sampling frame is divided up into subpopulations. Then, a random sample of these subpopulations is chosen. Finally, each individual in the chosen subpopulations is then surveyed.
5. In a regression setting, the \_\_\_\_\_ is what we denote as  $y$ .
6. A unimodal data distribution that is not symmetric is said to be \_\_\_\_\_.
7. In regression, the idea of fitting models via the \_\_\_\_\_ is done by choosing a regression equation that minimizes the sum of squared residuals.
8. Two explanatory variables are said to be \_\_\_\_\_ when their effects on a response variable cannot be distinguished from each other.
9. In regression, an observation is called a(n) \_\_\_\_\_ if removing it would dramatically alter the equation of the regression line.
10. A(n) \_\_\_\_\_ is a non-negative smooth curve approximation to a histogram of data. The entire area under this curve is equal to one.

**PART 3: SHORT ANSWER/COMPUTATION.** Show all of your work, and explain your reasoning.

1. [15] *Net interest margin* is the difference between the rate banks pay on deposits and the rate they charge for loans. Suppose that the net interest margins for all US banks are **normally distributed** with mean 4.15 percent and standard deviation 0.5 percent. Let  $X$  denote the net interest margin.

**In each part, you must draw the corresponding picture. If you do not, then you will receive at most 1/2 credit.**

(a) Find the proportion of US banks that have a net interest margin of 4.92 percent or larger.

(b) Wachovia Bank wants its net interest margin to be **less than** the net interest margins of 70 percent of all US banks. Where should Wachovia Bank's net interest margin be set?

2. [15] A study in Finland in the early 1900s looked at the relationship between the length (measured in cm) and weight (measured in grams) of a certain kind of perch. The goal of the study was to model the length ( $y$ ) as a function of weight ( $x$ ). A sample of  $n = 56$  fish was used in the study. Each fish was caught, measured, and then thrown back. The scatterplot for the data is shown below in Figure 1. Minitab was used to compute the equation of the best-fit straight line, obtained using least squares. This output is also given.

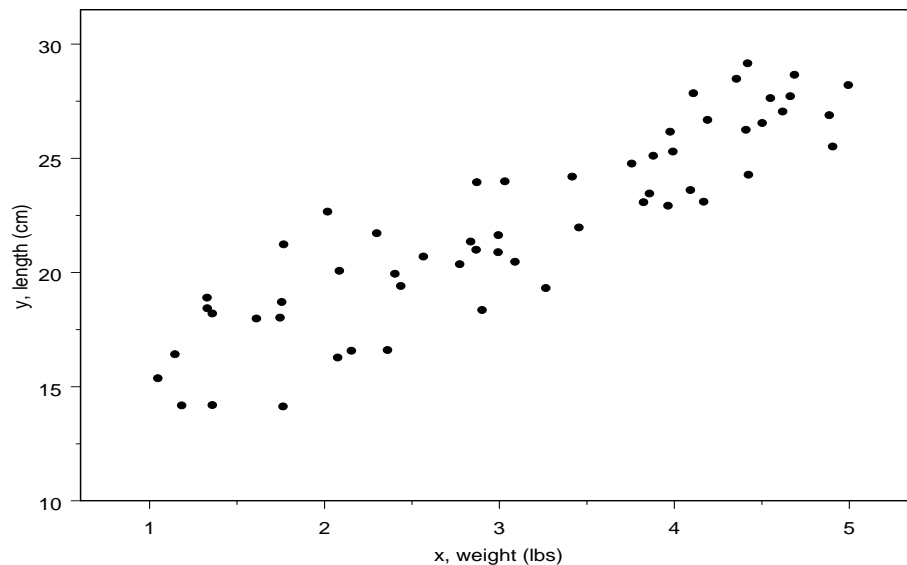


Figure 1: *Scatterplot for Finnish perch.*

The regression equation is  $\text{length} = 12.4 + 3.13 \text{ weight}$

Predictor	Coef	SE Coef	T	P
Constant	12.3728	0.6960	17.78	0.000
weight	3.1296	0.2100	14.90	0.000

S = 1.84939    R-Sq = 80.4%    R-Sq(adj) = 80.1%

(a) Is this an experiment or an observational study? Explain.

(b) Using the regression equation provided by Minitab, predict the length for a fish that weighs 3.4 lbs.

(c) From the printout, we see that the square of the correlation ( $r^2$ ) is equal to 80.4 percent. Interpret this statistic *in terms of the problem* (that is, talk about fish).

(d) Suppose that I had constructed the residual plot for this least-squares fit. What would you expect the residual plot to look like? Why?

3. [5] What is the difference between **exploratory data analysis** and **formal statistical inference**?