# STAT 101 Final Examination

*Instructions: Answer all questions in the spaces provided. You are allowed to use writing equipment, erasers, your text, notes, and a non-programmable calculator. All other items including backpacks and other containers larger than a wallet, other electronic devices, overcoats, etc., are not permitted, and must be placed at the front of the room for the duration of the test. You will be required to show your student ID card and to sign an attendance sheet during the course of the examination.*

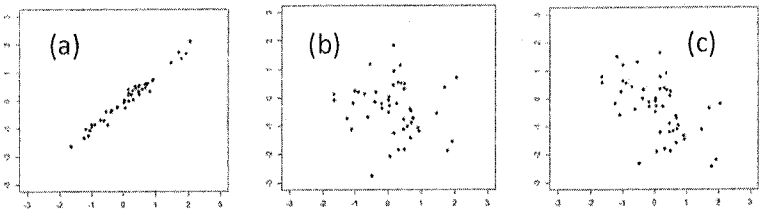
**Section A: Multiple Choice. Circle the correct answer(s) in the following questions.**

1. Which one of the following is a FALSE statement about density curves?

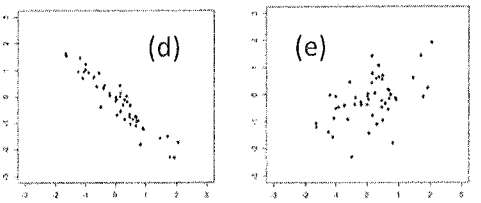
2

   a. The median divides the area under the curve in half.
   b. The mean is the balancing point of the density curve.
   c. The mean and the median are always the same number.

2. Match the following scatterplots to the correlations given below.
   (i) -0.20, (ii) 0.51, (iii) -0.95, (iv) 0.99, (v) -0.54




5




3. For the scatterplot with $r$ = -.99, what portion of the variation in the response variable can be accounted for through its association with the explanatory variable?

2

   (a) −0.99 (b) +0.99 (c) +0.98 (d) +0.995 (e) -0.98

Your Initials: _____ Last 4 digits of your student ID#: _____

4. Three students work independently on a homework problem. The probability that the first student solves the problem is 0.95. The probability that the second student solves the problem is 0.85. The probability that the third student solves the problem is 0.80. What is the probability that the first student solves the problem and the other two students do not?

   a. 0.95 + 0.15 + 0.20
   b. (0.95) (0.15) (0.20)
   c. 0.95 – 0.15 – 0.20
   d. 0.95 – 0.85 – 0.80
   e. 0.95

2

5. Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative? Indicate your choice with a 'c' for categorical and a 'q' for quantitative.

   a. Gender (female or male) ___
   b. Age (years) ___
   c. Race (Asian, black, white, or other) ___
   d. Smoker (yes or no) ___
   e. Systolic blood pressure (millimeters of mercury) ___
   f. Level of calcium in the blood (micrograms per milliliter) ___

3

6. Edwin Hubble collected data on the relationship between the distance a galaxy is from the earth and the velocity with which it is receding. He assumed that $\mu_y = \alpha + \beta x$, where $x$ is the distance from earth (megaparsecs) and $\mu_y$ is the mean velocity (km/sec) for all galaxies at that distance. What does $\alpha$ represent?

   a. The average velocity for a galaxy that is extremely close to earth.
   b. The change in mean velocity for a one-megaparsec increase in distance for Hubble's sample of galaxies.
   c. The slope of the least squares regression line for Hubble's sample.
   d. The mean velocity for all galaxies in the universe.
   e. The change in mean velocity for a one-megaparsec increase in distance for all galaxies in the universe.

2

7. For a simple random sample of college students, eye color and height were determined. Should regression inference be used to draw conclusions about the relationship between eye color and height for this data set?

   a. no, because the responses are not independent
   b. no, because the slope and intercept are unknown
   c. no, because one of the variables is categorical
   d. yes, if eye colors are normally distributed with the same standard deviation at each height
   e. yes, if mean eye color has a straight line relationship to height

2

Your Initials: _____ Last 4 digits of your student ID#: _____

8. Which clinic should you prefer?

| | Died | Survived | Total |
|---|---|---|---|
| Clinic A | 210 (35%) | 390 (65%) | 600 (100%) |
| Clinic B | 80 (20%) | 320 (80%) | 400 (100%) |
| Total | 290 (29%) | 710 (71%) | 1000 (100%) |

| Cancer | Clinic | Outcome | | |
|---|---|---|---|---|
| | | Died | Survived | Total |
| Remission | Clinic A | 5 (11%) | 40 (89%) | 45 (100%) |
| | Clinic B | 65 (18%) | 300 (82%) | 365 (100%) |
| | Total | 70 (17%) | 340 (83%) | 410 (100%) |
| | | Died | Survived | Total |
| Active | Clinic A | 205 (37%) | 350 (63%) | 555 (100%) |
| | Clinic B | 15 (43%) | 20 (57%) | 35 (100%) |
| | Total | 220 (37%) | 370 (63%) | 590 (100%) |

2

a. A
b. B

9. The data in the tables above are an illustration of which of the following?

2

a. The gambler's fallacy
b. The regression fallacy
c. The law of averages
d. Simpson's paradox
e. The central limit theorem

10. You are considering using the normal approximation to calculate a probability based on the binomial distribution. Which of the following would be important to take into account? (More than one may be appropriate.)

4

a. The law of large numbers
b. The "plus four rule"
c. The expected number of successes
d. The expected number of failures
e. The ½-integer correction factor
f. The possibility of replacing the standard normal distribution with the *t*-distribution

11. As the number of degrees of freedom increases indefinitely, the *t*-distribution approaches which of the following distributions?

2

a. The standard normal distribution
b. The binomial distribution
c. The *F*-distribution
d. The normal distribution with mean 0 and standard deviation equal to the number of degrees of freedom
e. None of the above

Your Initials: _____ Last 4 digits of your student ID#: _____

**Section B: Provide answers in the spaces provided.**

12. Comparing investments. Should you put your money into a fund that buys stocks or a fund that invests in real estate? The answer changes from time to time, and unfortunately we can't look into the future. Looking back into the past, the boxplots in the figure compare the daily returns (in percent) on a "total stock market" fund and a real estate fund over a year ending in November 2007.

    a. Read the graph: about what were the highest and lowest daily returns on the stock fund?

    2

    Highest: _____ Lowest: _____

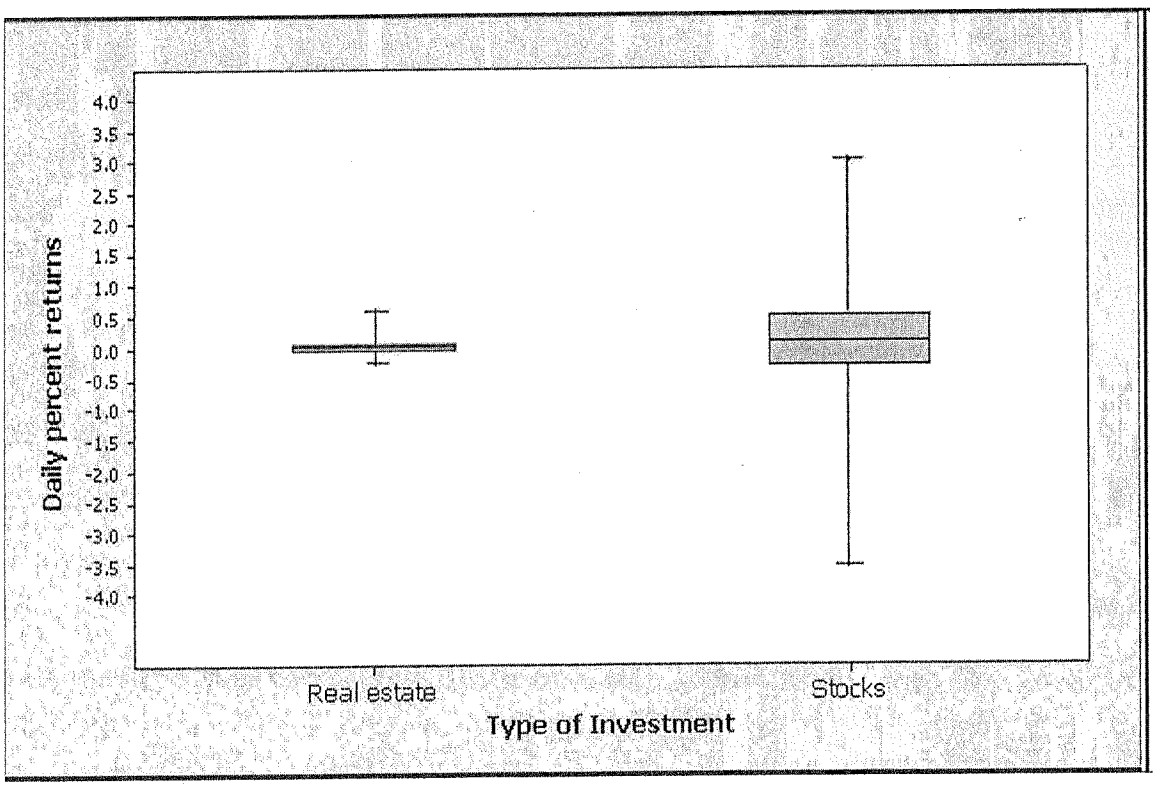
    b. Read the graph: the median return was about the same on both investments. About what was the median return?

    1

    Median: _____

    c. What is the most important difference between the two distributions?

    2



13. Three great hitters. Three landmarks of baseball achievement are Ty Cobb's batting average of .420 in 1911, Ted Williams's .406 in 1941, and George Brett's .390 in 1980. These batting averages cannot be compared directly because the distribution of major league batting averages has changed over the years. The distributions are quite symmetric and (except for outliers such as Cobb, Williams, and Brett) reasonably Normal. While the mean batting average has been held roughly constant by rule changes and the balance between hitting and pitching, the standard deviation has dropped over time. Here are the facts:

| Decade | Mean | Std.Dev. |
|---|---|---|
| 1910s | .266 | .0371 |
| 1940s | .267 | .0326 |
| 1970s | .261 | .0317 |

Compute the standardized batting averages for Cobb, Williams, and Brett to compare how far each stood above his peers. Draw an appropriate conclusion in one sentence.

6

14. Risks of playing soccer. A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and people of the same age who did not play soccer. Here is a two-way table that classifies these subjects by whether or not they had arthritis of the hip or knee by their mid-fifties:

| | Elite | Non-Elite | Did Not Play |
|---|---|---|---|
| Arthritis | 10 | 9 | 24 |
| No Arthritis | 61 | 206 | 548 |

a. How many people do these data describe?

1

b. How many of these people have arthritis of the hip or knee?

1

c. Provide a table of percents of people in each of the three soccer-playing categories who had arthritis, and comment on any apparent pattern.

5

d. What null hypothesis would a researcher typically test first on these data before commenting on the potential significance of the comparisons in part (c) above? What type of test statistic would you compute to perform the test of significance, and how many degrees of freedom (if any) does it have?

5

15. A random walk on Wall Street? The "random walk" theory of stock prices holds that price movements in disjoint time periods are independent of each other. Suppose that we record only whether the price is up or down each year, and that the probability that our portfolio rises in price in any one year is 0.65. (This probability is approximately correct for a portfolio containing equal dollar amounts of all common stocks listed on the New York Stock Exchange.)
    a. What is the probability that our portfolio goes up for three consecutive years?

3

    b. What is the probability that the portfolio's value moves in the same direction (either up or down) for three consecutive years?

3

16. Which data design? Is each of these designs (1) single sample, (2) matched pairs, or (3) two independent samples? Explain your choices.
    a. An education researcher wants to learn whether it is more effective to put questions before or after introducing a new concept in an elementary school mathematics text. He prepares two text segments that teach the concept, one with motivating questions before and the other with review questions after. He uses each text segment to teach a separate group of children. The researcher compares the scores of the groups on a test over the material.

3

    b. Another researcher approaches the same issue differently. She prepares text segments on two unrelated topics. Each segment comes in two versions, one with questions before and the other with questions after. The subjects are a single group of children. Each child studies both topics, one (chosen at random) with questions before and the other with questions after. The researcher compares test scores for each child on the two topics to see which topic he or she learned better.

3

Your Initials: _____ Last 4 digits of your student ID#: _____

17. An experiment has just been conducted, designed to test 3 treatments and a control. There were 36 observations generated in total, 9 within each of the four groups.
    a. Complete the following analysis of variance table for the results of this experiment.

7

| Source | Degrees of Freedom | Sum of Squares | Mean Square | *F* |
|---|---|---|---|---|
| Groups | | 145.33 | | |
| Error | | | | |
| Total | | 202.04 | | |

    b. If you were told that the experiment had involved counting numbers of insects caught in traps, and that a substantial number of traps caught no insects, would this cause you to question the validity of the analysis of variance? Explain in at most 2 sentences.

3

18. Sparrowhawk colonies. One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

| Percent return | 74 | 66 | 81 | 52 | 73 | 62 | 52 | 45 | 62 | 46 | 60 | 46 | 38 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New adults | 5 | 6 | 8 | 11 | 12 | 15 | 16 | 17 | 18 | 18 | 19 | 20 | 20 |

a. Biologists conjecture that the number of new adults joining colonies goes down as higher percents of adults return from the previous year. Refer to the JMP output on the following page.

i. Do the data show this effect? If so, how strong is it estimated to be? Explain how to interpret this estimate in the context of this specific example. Use at most one sentence.

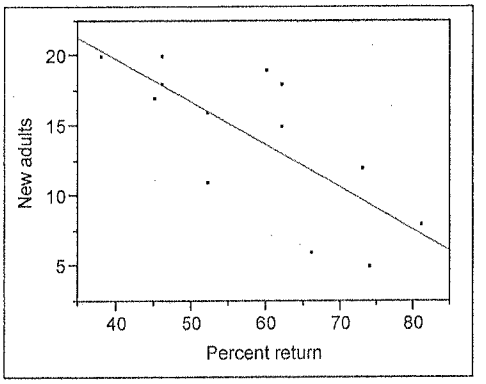
4

ii. Is the relationship statistically significant? Answer this question both (i) by referring directly to a *p*-value in the JMP output and (ii) with another test of significance described in the course.

6

Your Initials: _____ Last 4 digits of your student ID#: _____

## Bivariate Fit of New adults By Percent return




### Linear Fit

New adults = 31.934259 - 0.3040229*Percent return

### Summary of Fit

| | |
|---|---|
| RSquare | 0.560203 |
| RSquare Adj | 0.520222 |
| Root Mean Square Error | 3.666891 |
| Mean of Response | 14.23077 |
| Observations (or Sum Wgts) | 13 |

### Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 31.934259 | 4.837616 | 6.60 | <.0001* |
| Percent return | -0.304023 | 0.08122 | -3.74 | 0.0032* |

b. Use the additional output below to predict with 95% confidence the number of new birds in a colony to which 60% of the past year's adults return.

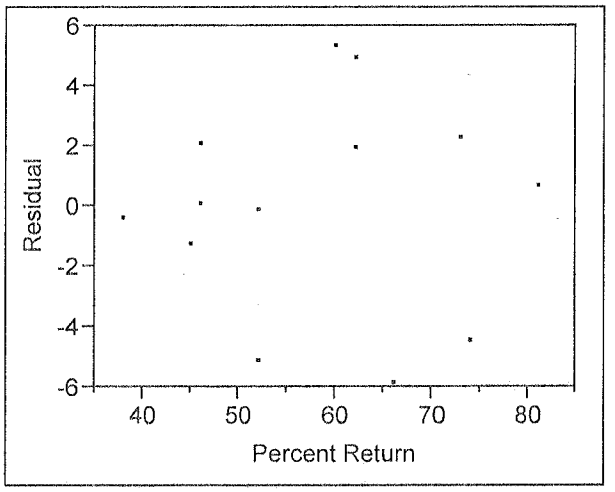
**Distributions**

| **Percent return** | | **New Adults** | |
|---|---|---|---|
| Mean | 58.230769 | Mean | 14.230769 |
| Std Dev | 13.032996 | Std Dev | 5.2939249 |
| Std Err Mean | 3.6147026 | Std Err Mean | 1.4682706 |
| Upper 95% Mean | 66.10653 | Upper 95% Mean | 17.429856 |
| Lower 95% Mean | 50.355009 | Lower 95% Mean | 11.031682 |
| N | 13 | N | 13 |

5

Your Initials: _____ Last 4 digits of your student ID#: _____

c. Does the following residual plot for this relationship show any clear signs of any problems with the conditions underlying the above inferences?




3

d. There is a related inference to that in part (b) above for which one of the above-mentioned conditions is far less important than it is for the inference in part (b). Which condition is this? Also, describe the related inference in the context of this particular example.

3

Total = 94

Your Initials: _____ Last 4 digits of your student ID#: _____

Blank sheet for rough work.

Your Initials: _____ Last 4 digits of your student ID#: _____

Blank sheet for rough work.

Your Initials: _____ Last 4 digits of your student ID#: _____

Blank sheet for rough work.