Math 217 - Winter 2017
Review packet for Exam 1

Module 1 Review

Four steps in a statistical investigation:

1. Ask a question that can be answered by collecting data
2. Decide what to measure and then collect data
3. Summarize and analyze
4. Draw a conclusion and communicate the results

Two types of statistical research questions:

1. Questions about characteristics of a **population**
   - To answer a question about a population we conduct an *observational study* with a *random sample*.
2. Questions about **cause-and-effect**
   - To answer a question about cause-and-effect we conduct an *experiment* with a *random sample* and *random assignment*.

Two types of statistical studies:

1. ***Observational study****:* Observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an *association* between two variables. An observational study alone does **not** provide convincing evidence of a cause-and-effect relationship.
2. ***Experiment:*** Intentionally *manipulates* one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a *cause-and-effect* relationship between two variables.

Two types of conclusions:

- Generalize from sample to population.
  - Valid when sample is randomly selected and *representative* of the population
  - Observational study can be used to draw this type of conclusion
- Cause-and-effect conclusion
  - Valid when a well-designed experiment is conducted
  - Groups assigned to experimental conditions should be similar –use random assignment.

Types of variables:

- ***Explanatory*** **variable:** the variable that is being modified or manipulated in the study.
- ***Treatment*** **variable*:*** the treatment being applied in an experiment - changed intentionally (manipulated variable).
- ***Response*** **variable:** the output variable which is used to measure the impact of changes to the explanatory variable.
- ***Confounding*** **variable:** other factors that may influence response.
- ***Quantitative*** **variable:** can be measured or counted – expressed as a number
- ***Categorical*** **variable:** cannot be measured or counted, is instead expressed as a quality or membership in a group

Population and Sample:

- The **population** is *all* the members of interest.
- A **sample** is a subset of the population that we study to collect data.

Sampling Techniques:

- **Convenience sampling:** uses an easily available group to form a sample. ***Not random.***
- **Voluntary Response Sampling:** participants are self-selected. ***Not random.***
- **Random samples**
  - **Simple Random Sampling:** every possible sample has an equal chance of being chosen for the sample. Good example: random number generator.
  - **Systematic Sampling:** individuals are arranged in order, then the sample is selected by choosing every k^{th} individual.
  - **Stratified Random Sampling:** the population is divided into subgroups, and then a simple random sample is taken from each subgroup. Members within each subgroup should have similar characteristics of interest.

Sample Size:

- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly
- *The size of the population does not affect the accuracy of a random sample as long as the sample is sufficiently large.*

Terms related to Experimental Design:

- **Significant Difference:** difference is large enough that it is unlikely to have been caused by chance.
- **Direct Control:** controlling other variables that might affect the response.
- **Random Assignment:** randomly assigning subjects to different treatment groups.
- **Control Group:** a group that does not get treatment. Used for baseline comparison.
- **Single Blinding Technique:** when participants (or researchers, but not both) do not know which experimental condition they are assigned to. Prevents prior beliefs influencing results.
- **Double Blinding Technique:** when both participants and people measuring the response are blinded.
- **Placebo:** a treatment with no active ingredient.
- **Placebo Group:** a group in an experiment receiving placebos.
- **Placebo Effect:** when a participant receiving a placebo reports that the placebo had an effect.

Module 2 Review

To analyze the distribution of a quantitative variable, we describe the **overall pattern of the data** (**shape**, **center**, **spread**) and any **deviations from the pattern** (**outliers**).

Three types of **graphs** to analyze the distribution of a quantitative variable:

1. **Dotplots**
   - Individual variable values are visible, particularly when the data set is small.
   - Descriptions of shape, center, and spread are not affected by how the dotplot is constructed.
   - We can accurately calculate the overall range (largest value – smallest value).
2. **Histograms**
   - Individual variable values are not visible.
   - Groups individuals into *bins* of equal-sized intervals. This is particularly useful when analyzing large data sets.
   - We can easily use percentages, also called relative frequencies, to describe the distribution.
3. **Boxplots**
   - Displays the five-number summary (min, Q1, med, Q3, and max).
   - Highlights any points that are considered outliers
   - Commonly used to compare two data sets.

Four ways to describe the **shape** of a distribution:

1. **Skewed left**
2. **Skewed right**
3. **Bell-Shaped** (symmetric)
4. **Uniform** (symmetric)
   - *Note: Not all distributions have a simple shape that fits into one of these categories.*

The **center** of a distribution is a typical value that represents the group. We studied 2 measurements for determining the center of a distribution:

1. **Mean:** Calculate the mean by adding the data values and dividing by the number of individual data points. The mean is also referred to as *"the balancing point"* of a distribution. (If we measure the distance between each data point and the mean, the distances are balanced on each side of the mean.) The notation for mean is $\bar{x}$.
2. **Median:** The physical center of the data when we make an ordered list. *It has the same number of values above it as below it.* It is either the *middle number* or the *average of the 2 middle numbers.*

General guidelines for choosing a measure of center:

- Always plot the data. We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measure of center best describes the data.
- Use the **mean** as a measure of center for distributions that are reasonably **symmetric** with a central peak. *When outliers are present, the mean is not a good choice.*
- Use the **median** as a measure of center when distribution is **not very symmetric** (*skewed*).

The **spread** of a distribution is a description of how the data varies. We studied 3 ways to measure spread:

1. **Range** = max – min
2. **Interquartile range (IQR)** = Q3 – Q1: measures the variability in the *middle half* of the data
   - **Five number summary** {min, Q1, Median, Q3, max}
3. **Standard deviation (SD)**: approximately the average distance of the data from the mean
   - **Empirical Rule for bell-shaped data**
     - 68 % of the data is within 1 standard deviation of the mean
     - 95 % of the data is within 2 standard deviations of the mean
     - 99.7 % of the data is within 3 standard deviation of the mean
   - **z-scores for data values**: how many standard deviations above or below the mean
     - $\boldsymbol{z - score = \frac{data\ value - mean}{standard\ deviation} = \frac{x - \overline{x}}{s}}$

Outliers: "Extreme values" that fall outside the fences

- **Lower fence**: $Q1 - 1.5(IQR)$
- **Upper fence**: $Q3 + 1.5(IQR)$

Alternate definition of **outliers**: "Extreme values" have a **z-score** greater than 3 or less than -3.

Notes:

- Mean and standard deviation go together,
- Median and IQR go together.
- The mean and SD are strongly affected by extreme values (outliers), while median and IQR are not.

Linear Functions Review

Basic mechanics and skills with linear functions:

- The formula for the equation of a linear function is $y = a + bx$
- Know how to fill out a table of values, given the equation of a linear function, and use it to sketch the graph of the corresponding line
- The meaning of a and b from the formula on the graph
    - $a =$ the **y-intercept**: the value of y where the graph crosses the y-axis
    - $b =$ the **slope**: a measure of the steepness of the line, *"rise over run"*
- Know how to find the *slope* of a line given two points
- Know how to find the *equation* of a line given two points
- Use the equation of the line to find the value of y when x is given, and the value of x when y is given

Applications of linear functions:

- *Interpret* the slope and the y-intercept in the context of the problem
- Find the equation of the line given *verbal description* of slope and y-intercept
- Perform unit analysis on the linear equation

Exam 1 Practice Problems

1. In each of the following scenarios, identify the population and the sample.

   a. 50 De Anza College students were sampled to find out the amount of sleep they get each night.
   b. 300 registered voters in a large city were asked whether they plan to vote in the next election.
   c. 75 randomly selected light bulbs made a factory were tested for defectiveness.

2. For each of the following, identify if it's an observational study or an experiment. If it's an observational study, identify the population and sample. If it's an experiment, identify the treatment group(s), control group (if any), placebo group (if any), the explanatory variable and the response variable.

   a. Researchers wish to see if studying with peers before a math final exam affects performance on the exam. They randomly assign 100 randomly selected students to 3 groups. The members of the first group spend 6 hours studying with peers for the final exam, the members of the second group spend 2 hours studying with peers for the final exam, while the members of the third group study alone. The final exam scores of each group are analyzed.

   b. 5 randomly selected teachers from 10 different randomly selected elementary schools in a city are interviewed. The researchers find that teachers in higher grades spend more time, on average, in preparing lessons than teachers in lower grades.

3. For each of the following, decide if it is a categorical variable or a numerical (quantitative) variable. If it's numerical, decide if it's discrete or continuous.

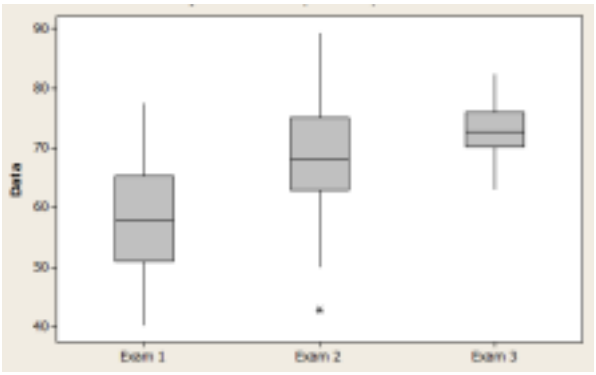
   a. The GPA of a student
   b. The major of a student
   c. The number of units a students is currently enrolled in
   d. Whether the student uses public transportation to get to campus
   e. The amount of money in a student's wallet
   f. How many quarters the student has been at De Anza

4. The three boxplots shown display midterm exam scores for students taking a statistics course.

   a. Which exam has the smallest IQR?

   b. If the outlier was removed from the data for Exam 2, which of the following statistics could be significantly affected? **Choose from:** IQR, Mean, Median, Range, Standard Deviation

   c. Describe all of the similarities and differences between the three exam scores. Include center, shape and spread in your descriptions.

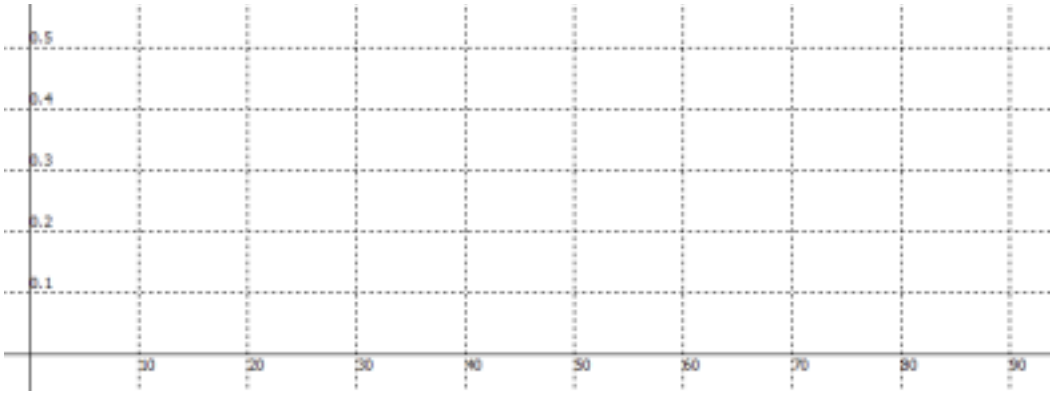


5. The following data represents the amount of time 50 students take to complete an exam:

Minutes	**Frequency**	**Relative Freq**
31-40	17	
41-50	13	
51-60	10	
61-70	8	
71-80	2	

a. Determine the relative frequencies for each bin and fill them in the column of the chart above.

b. Make a relative frequency histogram in the graph below.



c. **Without calculating**, what can you say about mean and median of this data?

6. Describe at least one problem you might have in obtaining a representative sample if you were to do a mail-in survey. In this type of survey, you would mail the surveys and each person would have to mail back their completed survey to you.

7. Suppose you are to find a sample of 100 De Anza students. Also suppose you have access to any data that the college president would have.

a. Describe a method of finding a simple random sample.
b. Describe a method of finding a systematic sample.
c. Describe a method of finding a stratified random sample.

8. **Sixty** randomly selected students were asked the number of phone calls they received yesterday. The results are as follows:

# phone calls	**Frequency**	**Rel. frequency**
1	6	
2	25	
3	12	
4		
6	9	

a. Fill in the blanks in the above table. Round to 4 decimal places.
b. Find the sample mean, $\overline{x}$.
c. Find the standard deviation.
d. Find the third quartile.
e. What percent of the students received at least 4 phone calls?
f. What percent of students received less than 3 phone calls?

g. Carefully construct a histogram for this data.
h. Carefully construct a boxplot for this data.

9. Given the points $(0,-3)$ and $(6,1)$
   a. Write the equation of the line y = a + bx through the two points
   b. Graph the line
   c. Find y when x = -9
   d. Find x when y = 5

10. A person earns a starting salary of \$44000 at a company. Each year, she receives a \$4500 raise. Let s be the salary after she has worked at the company for t years.

    a. State the units of t: _____ State the units of s: _____

    b. Find the slope. Include units.

    c. Find the formula for the linear function relating t and s in the form $s = a + bt$.

    d. Perform a unit analysis on the equation in part c.

Solutions

1.

a. Population: All De Anza College students.
Sample: The 50 students who the data was actually collected from.

b. Population: All registered voters in that city.
Sample: The 300 people who the data was actually collected from.

c. Population: All light bulbs made in that factory
Sample: The 75 light bulbs the data was actually collected from.

2.

a. Experiment. The treatment groups are the first and second group, the control group is the third group, and there is no placebo group. Explanatory variable: the number of hours spent studying with peers. Response variable: performance on the final exam

b. Observational study. Population: All teachers working in elementary schools in the city. Sample: the 50 teachers from whom the data was actually collected.

3.

a. Numerical
b. Categorical
c. Numerical
d. Categorical
e. Numerical
f. Numerical

4.

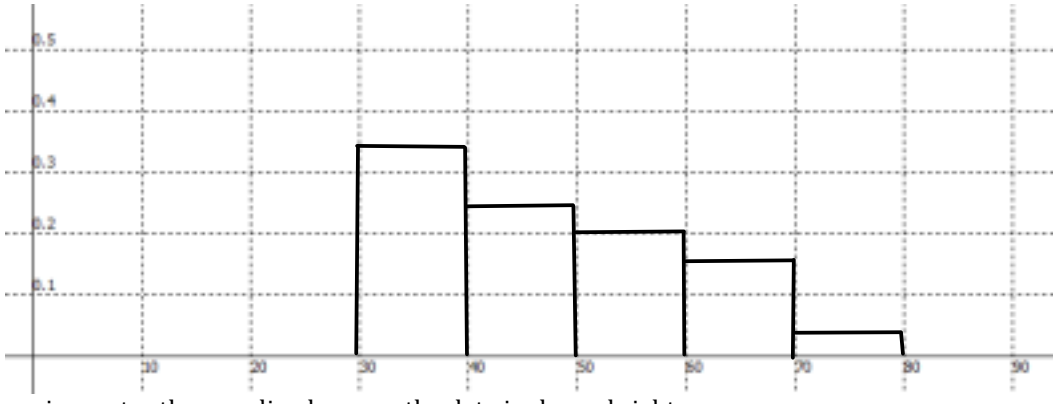
a. Exam 3
b. Mean, Range, and Standard deviation
c. All graphs are fairly symmetric (especially if we remove the outlier from Exam 2). Exams 1 and 2 have a greater spread than Exam 3. The median for Exam 1 is the lowest and that for Exam 3 is the highest. Most likely, Exam 1 was a fairly difficult exam since the entire boxplot is quite a bit lower than the other two. All students for Exam 3 were fairly evenly prepared (and therefore performed similarly) since the range of scores is so small, whereas there were many more differences in the student performances for Exam 2.

5.

a.

Minutes	**Frequency**	**Relative Freq**
31-40	17	17/50 = 0.34
41-50	13	13/50 = 0.25
51-60	10	10/50 = 0.2
61-70	8	8/50 = 0.16
71-80	2	2/50 = 0.04

b.



c. Mean is greater than median because the data is skewed right

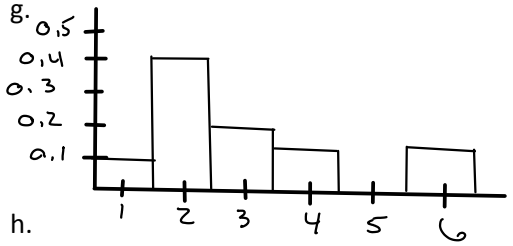
6. One problem is that you will have a non-response bias.

7.
   a. Make a list of all students and assign each a unique number from 1 through N (total number of students). Then use a random number generator until you have a list of 100 distinct students for your sample.
   b. In the list above, pick a random starting point and select every N/100-th student until you have 100 distinct students for your sample.
   c. Stratify the students by gender (male and female), and select proportionally from each group or stratify the students by majors and select proportionally (and of course randomly) from within each major until you have a sample of 100.

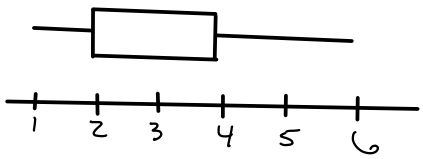
8.
   a.

# phone calls	Frequency	Rel. frequency
1	6	0.1
2	25	0.4167
3	12	0.2
4	**8**	0.1333
6	9	0.15

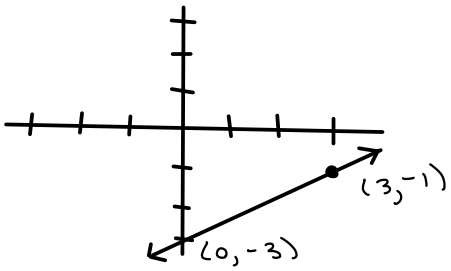
   b. Put # calls in L1 and frequencies in L2. Hit 1-var Stats L1, L2 and get $\bar{x} = 2.97$
   c. The sample standard deviation is $s = 1.53$
   d. $Q3 = 4$, which means that 75% of the data values are 4 or lower.
   e. Percent of students who received at least 4 phone calls: 0.1333 + 0.15 = 0.2833 = 28.33%
   f. Percent of students who received less than 3 phone calls: 0.1 + 0.4167 = 0.5167 = 51.67%
   g.



   h.



9.
   a. $y = -3 + \frac{2}{3}x$
   b.



   c. y=-9
   d. x = 12

10.
   a. t=years; s=dollars
   b. 4500 dollars per year
   c. $y = 44000 + 4500x$