Gathering Data

Previously, you

- Displayed data sets with dot plots, histograms and boxplots in Algebra I
- Summarized data sets with numerical descriptions like mean, median and standard deviation in Algebra I

In this unit you will

- Learn about methods for collecting data
- Recognize bias and how important randomization is during sampling and experimentation
- Learn the difference between random samples and random assignment of treatments
- Recognize the difference between observational studies, surveys and experiments
- Classify data as categorical or quantitative
- Interpret data in dot plots, histograms and boxplots by discussing shape, center and spread
- Analyze how outliers and skew can affect measures of center and spread

You can use the skills in this unit to

- Select the type of sampling for a given study
- Set up an experiment including randomization of treatments and measurement of the outcome
- Select an appropriate data display for real-life data
- Compare two or more data sets using center and spread

Vocabulary

- **Bias –** A property of a statistical sample that makes sample results unrepresentative of the entire population..
- **Convenience sample** – One of the main types of non-random sampling methods. It is a sample made up of people who are easy to reach or readily available.
- **Experiment** – A scientific procedure undertaken to make test a hypothesis or known fact where treatments are deliberately imposed on the subjects.
- **Observational study** – A type of study in which individuals are observed or certain outcomes are measured. No attempt is made to affect the outcome.
- **Parameter** – A measureable characteristic of a population.
- **Qualitative vs**. **Quantitative –** A qualitative variable places values into categories and can be described with bar graph, pictographs and pie charts. Quantitative or numerical values are best displayed using histograms, dot plots, boxplots or stemplots.
- **Simple Random Sample (SRS) –** A randomly selected sample from a population giving all the individuals in the sample an equal chance to be chosen.
- **Skewness –** Data that is not symmetric and has more data in one of the tails of the distribution. Skewed to the left has a longer tail to the left and skewed to the right has a longer tail to the right.
- **Standard Deviation –** A measure of the spread of a data set that finds the average amount the data varies from the mean.
- **Stratified Random Sample –** A type of sampling procedure that requires the population to be divided into smaller, homogeneous groups and then taking a random sample from each group (strata).

- **Statistic –** A measurable characteristic of a sample.
- **Symmetric –** A distribution where one side is a mirror image or reflection of the other.
- **Systematic Random Sample –** A type of random sampling that selects every nth item from the population.
- **Treatment –** A specific experimental condition applied to the experimental units or subjects.
- **Voluntary Response Sample** – A non-random sampling procedure that involves only those who want to participate.

Essential Questions

- How can a researcher select a method of collecting data with as little bias as possible?
- How can patterns in sets of data be discovered?
- What shapes can distributions have and how are statistics affected by shape and outliers?

Overall Big Ideas

The researcher selects a method of gathering data for a random sample after the type of study has been determined.

Data displays are useful in making sense of data.

Data distributions can be described by their shapes. The shape and presence of extreme values may affect center and spread.

Skills

To identify and perform an appropriate method of gathering data (experiment, simulations, observational studies including sample surveys).

To understand the importance of randomization, and the difference between random sampling and random assignment.

To organize data.

To know the difference between a parameter and a statistic.

Related Standards

S.IC.B.3
Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S.ID.A.1
Represent data with plots on the real number line (dot plots, histograms, and box plots).

S.ID.A.2
Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S.ID.A.3
Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (Outliers).

S.ID.A.4
Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

S.IC.A.1
Understand statistics as a process for making inferences about population parameters based on a random sample from that population. *(Modeling Standard)

Notes, Examples, and Exam Questions

Sec. 9.1, 9.2, 9.6 To identify and perform an appropriate method of gathering data and to understand the importance of randomization, and the difference between random sampling and random assignment. To know the difference between a parameter and a statistic.

Study vs. Experiment:

1. Observational Study – Observes individuals and measures variables of interest but does not attempt to influence responses. **Sample Surveys** are one kind of observational study. There are many different types or techniques of observation and conducting a survey (face-to-face, mail survey, telephone survey) is just one technique. Retrospective observational studies take a look back at events that have already taken place whereas prospective studies watches for future outcomes, such as the development of a disease, during the study period.

2. An Experiment, on the other hand, deliberately imposes some **treatment** on individuals in order to observe their responses.
   ****When our goal is to understand cause-and-effect, experiments are the only source of fully convincing data.**

Important vocabulary when discussing observational studies:

1) Population - the entire group of individuals that we want information about.
2) Parameter – a measureable characteristic of the population.
3) Sample – a part of the population that we will actually examine in order to gather information.
4) Statistic – a measureable characteristic of the sample.
5) Census – attempts to contact every individual in the entire population.
6) Sample Design – refers to the method used to choose the sample from the population.
7) Sampling Frame – a complete list of all the members of the population that we wish to study.
8) Bias – systematic error – systematically favoring certain outcomes or some parts of the population over others.
9) Simple Random Sample – (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected. **Also**, it gives every possible sampling **group** an equal chance to be chosen.
10) Table of random digits – a long string of the digits 0,1,2,3,4,5,6,7,8,9 with these properties:
    a) Each entry is equally likely to be one of the 10 digits.
    b) Entries are independent. Knowledge of one part of the table gives no information about any other part.
8) Choosing an SRS: 1) LABEL. Assign a numerical label to every individual. Must be the **same number of digits** for each person. 2) TABLE . Use the random number table to select people at random. (01 – 99), (001 – 999), (0001 – 9999). You can also start at 0. Ten people – (0 – 9), 100 people, (00-99)….

*****The use of chance to select the sample is the essential principle of statistical sampling. This is why we randomize. Randomizing protects us from the influences of ALL the features of our population, even ones that we may not have thought about. It does that by making sure that ON AVERAGE the sample looks like the rest of the population.**

Bad Sampling techniques – these are both very biased:

1) **Convenience sample** – using results that are readily available or choosing the individuals easiest to reach. This is a type of nonrandom sampling.

2) **Voluntary response sample** – consists of people who **choose themselves** to be part of the sample by responding to a general appeal. They are biased, because people with strong opinions, usually negative, are more likely to respond.

Cautions about sampling: Sampling data should be done very carefully and systematically. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased. It is better for the person conducting the survey to select the sample respondents.

Using a random number table:

Ex 1: An accounting firm serves 90 business clients. They want to interview a sample of 8 clients in detail to find ways to improve client satisfaction. To avoid bias, they choose an SRS of size 8. We can do this using our random number table. What is the population and the sample? How will we label the 90 clients? How do we choose the numbers?

Population: All 90 business clients

Sample: The 8 clients we choose to survey

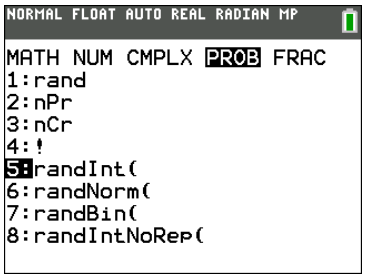
Randomization: Take a list of the business clients (this is our sampling frame) and number them 01 – 90. I will then read two digits at a time from the random number table. I will ignore repeats and any numbers from 91-99. The first eight numbers that represent the business clients will become my sample.

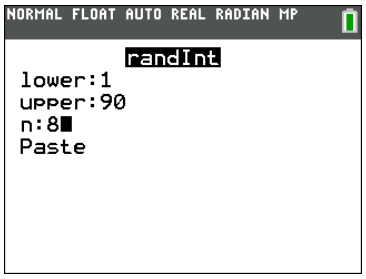
Table 1— Random Numbers

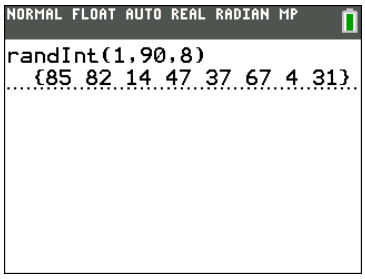
92630	78240	19267	95457	53497	23894	37708	79862	76471	66418
79445	78735	71549	44843	26104	67318	00701	34986	66751	99723
59654	71966	27386	50004	05358	94031	29281	18544	52429	06080
31524	49587	76612	39789	13537	48086	59483	60680	84675	53014
06348	76938	90379	51392	55887	71015	09209	79157	24440	30244
28703	51709	94456	48396	73780	06436	86641	69239	57662	80181
68108	89266	94730	95761	75023	48464	65544	96583	18911	16391
99938	90704	93621	6634	393	0534	517	91616	32	

I will arbitrarily start on the second row. The companies in my sample are labelled: 79, 44, 57, 87, 35, 71, 54, and 48.

To use a random number generator on the TI-84 calculator, use the following commands: MATH, PRB, randInt(. For the arguments, choose the first number, last number and how many random numbers as the third argument.







Ex 2: For each of the following situations, identify the sampling method used. Then explain how the sampling method could lead to bias.

a) A farmer brings a juice company several crates of oranges each week. A company inspector looks at 10 oranges from the top of each crate before deciding whether to buy all the oranges.

This is a convenience sample. This could lead the inspector to overestimate the quality of the oranges if the farmer puts the best oranges on top.

b) The ABC program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. Viewers were invited to call one telephone number to respond "Yes" and another for "No". There was a charge for calling either number. More than 186,000 callers responded and 67% said "No".

This is a voluntary response sample. Those who are happy that the UN has its headquarters in the United States already have what they want and so are less likely to respond. The proportion who answered "No" in the sample is likely to be higher than the true proportion in the United States who would answer "No".

Errors in Sampling: There are two types: sampling errors and nonsampling errors.

Sampling errors: There are errors caused by the act of taking a sample.

1) **Random sampling error** is the deviation between the parameter and the statistic. It is unavoidable because the sample will never full match the entire population.

2) **Bad sampling methods** – these can be avoided. Sampling begins with a list of individuals from which we will draw our sample. This list is called the **sampling frame**. If this list is incomplete, our sample can suffer from **undercoverage.**

3) **Undercoverage bias** occurs when some groups in the population are left out of the process of choosing the sample. (For example, if our sampling list is a list of all addresses in the community, we are leaving out the homeless. If our list is a list of all listed telephone numbers, we are leaving out people without phones and with unlisted numbers.)

Nonsampling errors: There are errors that are made after the sampling has been done.

1) **Processing errors** – mistakes in mechanical tasks (arithmetic, computer error).

2) **Response bias** – this occurs when the interviewer or the interviewee's behavior affects the response. This could be poor wording of questions, the interviewee may lie or guess. Question wording is very important and can influence responses.

3) **Nonresponse bias** – this is the failure to obtain data from an individual selected for a sample. This occurs when you pick someone, but they refuse to cooperate or refuse to be in your sample. (example – hanging up on a phone survey or not returning a mail survey).

Sampling Designs: We need to use probability samples. These are samples that are chosen by chance. Some type of randomness is involved. Here are some types of probability samples:

1. **Simple random sample** – Each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample. Every possible sample of a given size also has the same chance of selection.

2. **Stratified random sample** – divide the sampling frame into distinct groups of individuals, called **strata.** These should be **homogeneous** (like) groups of some variable of interest. Then, take a separate SRS in each stratum and combine these to make the complete sample. This differs from an SRS. For example, 100 students are grouped by gender and there are 40 males and 60 females. A random sample of 10% of each group is taken to get a representation, so 4 males and 6 females are chosen. If it was an SRS, it would be possible to get 8 females and 2 males, but with the stratified sample, this cannot occur and every possible sample of a given size does NOT have the same chance of selection.

3. **Systematic random sample** – select a starting point and take every nth piece of data from a listing of the population. Choose the first item in the list at random. Then, choose the next items in the same intervals. For example, I will pick the every 10^{th} person who walks in the door and ask them my question.

4. **Cluster sample** – a design in which entire groups, or clusters are chosen at random. We split the population into similar clusters to simplify the design. Each cluster should be **heterogeneous**. ALL the members from these selected sections are in the cluster sample.

5. **Multistage** – most large-scale sample surveys use this technique that combine two or more sampling method, like combining a stratified sample and a cluster sample.

Ex 3: Name the type of sampling method used in each of the studies below.

a) An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.

Stratified Random Sample

b) A large nuclear power company needs information about how the staff feel about the facilities they offer. The company workers are spread out over 20 sites. The company manager decides to ask all workers from a random selection of 5 sites.

Cluster Sample

c) Jenny decides to interview every 100^{th} baseball fan that comes into the stadium.

Systematic Sample

d) From a class of 25 students the teacher selects the last 5 to enter the room to interview.

Convenience Sample

e) In a class of 12 boys and 12 girls a teacher selects 5 students by numbering the boys 01 to 12 and the girls 13 to 24 and uses a random number table to choose 5 numbers between 011 and 24.

Simple Random Sample

f) A teacher is interested in finding out the opinions of Algebra II students at her school about homework. The teacher gets a list of Algebra II students from the registrar and interviews each of them.

Census

Ex 4: Name the bias the survey suffers from and state how this bias might affect the study results.

a) A study in Summerlin looked at seat belt use by drivers. Drivers were observed at randomly chosen convenience stores. After they left their cars, they were invited to answer questions that included questions about seat belt use. In all, 75% said they always used seat belts, yet only 61.5% were wearing seat belts when they pulled into the store parking lots.

This is an example of response bias. People will claim to wear their seat belts because they know they should, even if they don't. Because they want to show they are doing "the right thing", our sample statistic will be biased and overestimated.

b) A survey of drivers began by randomly sampling all listed residential telephone numbers in the United States. Of 45,956 calls to these numbers, 5029 were completed. The goal of the survey was to estimate how far people drive, on average, per day.

This is an example of nonresponse bias. There was an 89.1% nonresponse rate which is very high. Because the people who have long commutes are less likely to be at home and be included in the sample, this will likely produce an estimate that is too small.

c) Suppose you want to know the average amount of money spent by the fans attending opening day for the Los Angeles Dodgers baseball season. You get permission from the team's management to conduct a survey at the stadium, but they will not allow you to bother the fans in the club seating or box seats (the most expensive seating). 500 seats were randomly selected from the rest of the stadium.

This is an example of undercoverage. Part of the population was left out of our sampling frame. Because the sample is only from the lower-priced ticket holders, this will likely produce an estimate that is too small, as fans in the club seats and box seats probably spend more money at the game than the fans in cheaper seats.

Ex 5: Identify the parameter and the statistic in each case.

a) On Tuesday, the bottles of tomato ketchup filled in a plant were supposed to contain an average of 14 ounces of ketchup. Quality control inspectors sampled 50 bottles at random from the day's production. These bottles contained an average of 13.8 ounces of ketchup.

Parameter: 14 ounces $(\mu = 14)$ **Statistic: 13.8 ounces** $(\bar{x} = 13.8)$

b) On a New York to Denver flight, 8% of the 125 passengers were selected for random security screening prior to boarding. According to the TSA, 10% of airline passengers are chosen for random screening.

Parameter: 10% $(p = 0.10)$ **Statistic: 8%** $\left(\hat{p} = 0.08\right)$

Important vocabulary when discussing experiments:

An experiment deliberately **imposes some treatment** on individuals in order to observe their responses. The goal of experiments is to understand cause and effect.

1. Experimental units – objects or individuals on which the experiment is done.
2. Subjects – when the experimental units are individuals.
3. Treatment – a specific experimental condition applied to the units.
4. Factors – the explanatory variables in an experiment.
5. Level – the specific values that the experimenter chooses for a factor.
6. Explanatory variable – this variable attempts to explain the observed outcomes. This is often called the independent variable.
7. Response variable – this is the variable measured in the outcome of the study. This is often called the dependent variable.

Advantages of experiments:

1. They can give good evidence of causation.
2. We are able to study the effects of the specific treatments we are interested in.
3. We can control the environment to hold constant factors we are not interested in.
4. We can study the combined effects of several factors simultaneously.

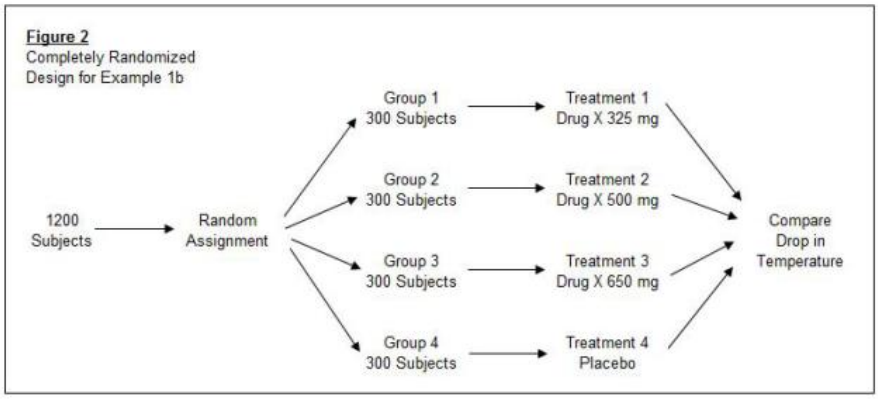
Confounding: A **lurking variable** is a variable that has an important effect on the relationship among the variables in a study, but it is not one of the explanatory variables studied. Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

Placebo effect: This effect is the response to a dummy treatment. A **placebo** is a dummy treatment is a treatment that has no physical effect or no active ingredients (sugar pills).

Randomized Comparative Experiment:

A. Control Group – This is a group that we can use to compare our results to. They might be receiving the dummy treatment or the "old" treatment. Comparing the treatment and control groups allows us to control the effects of lurking variables.
B. Randomization – The use of chance to divide experimental units into groups. It is the second major principle of statistical design of experiments.
C. Completely randomized design (CRD) – In this design, all the experimental units are allocated at random among all the treatments.

CRD diagram:



THE THREE PRINCIPLES OF EXPERIMENTAL DESIGN:

1) **CONTROL** – the effects of lurking variables on the response variable by comparing two or more treatments. We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups.

2) **RANDOMIZE** – use impersonal chance to assign experimental units to treatments. Randomization allows us to equalize the effects of unknown or uncontrollable sources of variation.

3) **REPLICATE** – repeat each treatment on many units to reduce chance variation in the results. Also, repeat the experiment if possible. **Replication** of an entire experiment with the controlled sources of variation at different levels is also essential.

Ex 6: In order to investigate a rumor that there is a greater than expected number of girls among the children of chemists, *Science* magazine conducted an informal survey of eight chemistry departments. A secretary in the chemistry department at Indiana University, Bloomington thought there might be something to this rumor and made sure that every one of the 34 faculty members in her department who have children responded to the *Science* survey. Altogether, these Indiana chemists have 53 (56%) girls and 41 (44%) boys.

a) Was this an observational study or an experiment? Justify your answer.

This is an observational study because there was no treatment imposed on the subjects.

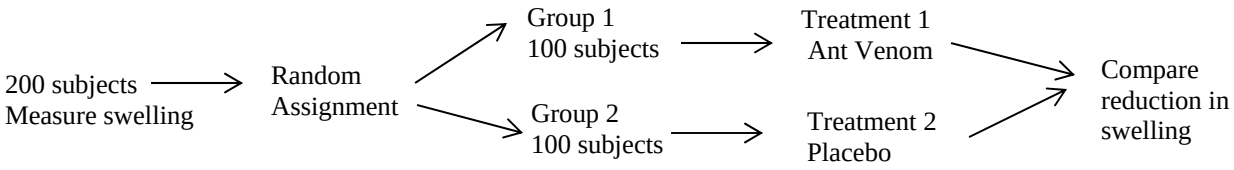
b) Does this study give convincing evidence that chemists produce more girl babies than boy babies? Why or why not?

No, this study does not provide evidence of cause-and-effect. We would need a controlled experiment to show evidence.

Ex 7: A researcher at Ohio State University believes that a certain component of ant venom can be used to lessen the amount of swelling in the knuckles of people suffering from arthritis. The ant venom treatment has been made into a capsule form that can be swallowed.

a) Explain carefully how you would design an experiment to investigate whether this new treatment, when taken orally each day for one week, causes a lower degree of swelling in arthritis sufferers. (You may suppose that 200 people suffering from arthritis have already volunteered to be experimental subjects.)

First, measure the amount of swelling in each volunteer's hands. Then, randomly assign the 200 volunteers into two groups by placing all of their names in a hat, shaking the hat and picking out 100 names. These 100 people will be Group 1 and will receive the ant venom. The remaining names in the hat will be Group 2 and they will receive a placebo – a capsule looking just like the ant venom treatment. At the end of the week, again measure the swelling in all 200 particpants' knuckles and compare the reduction in swelling between Group 1 and Group 2.



b) Identify the explanatory variable and the response variable in your experiment.

The explanatory variable is the ant venom and the response variable is the amount of swelling in the knuckles of the subjects at the end of one week.

Cautions About Experimentation:

1) If possible, an experiment should be **double-blind.** In a double-blind experiment, neither the subjects nor the people who have contact with them know which treatment a subject received. In a **single-blind** experiment, the subjects do not know which treatment they received, but the people in contact do know. This is not as good as a double-blind.

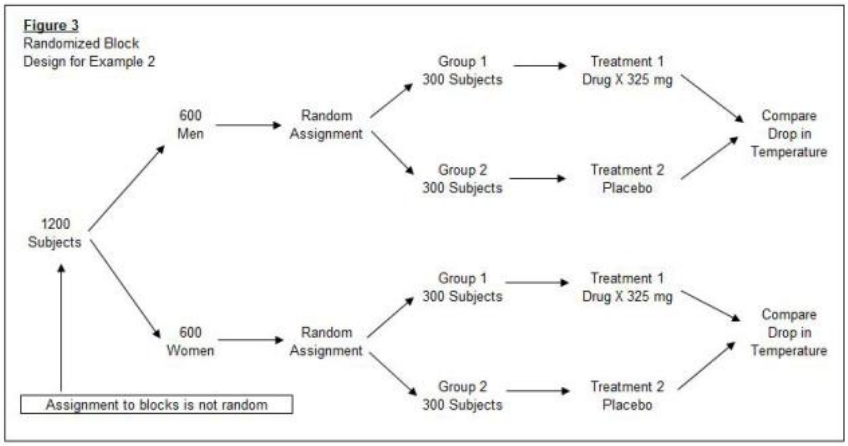
2) The most serious potential weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study.

Matched Pairs Design:

Matched pairs design compares just two treatments. We choose blocks of two units that are as closely matched as possible (often these are twins) and we RANDOMLY assign one of the treatments. Alternatively, each block in a matched pairs design may consist of just one subject, who gets BOTH treatments one after the other. Each subject serves as his or her own control. **The order of the treatments can influence a subject's response, so the order for each subject must be randomized!** A good example of the matched pairs test is a taste test. Each subject tries both items and, therefore, gets both treatments.

Block Designs:

A **block** is a group of experimental units or subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments. In a **block design**, the random assignment of units to treatments is carried out separately within each block. That means, we RANDOMIZE AFTER we block.



Difference between block and stratified:

Blocks are another form of CONTROL. Blocks allow us to draw separate conclusions about each block. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the experimental units. We block by a trait in common – like gender. Blocking is like stratifying. When subjects are grouped by traits in an OBSERVATIONAL STUDY it is called STRATIFYING. In an experiment, it is called BLOCKING Same technique – just a different name.

Ex 8: **A report in the April 26, 2001 *New England Journal of Medicine* studied a new treatment for children with a severe anxiety disorder. The study was a randomized double-blind comparative experiment. Data from the study showed that 76% of the children treated with the new drug had a reduced anxiety level. Of the children who were given a placebo, 29% had a reduced anxiety level. Almost none of the patients in the study exhibited an increase in anxiety levels.**

a) Explain the meaning of the word "placebo" in the above description. Then, discuss why it was important to administer a placebo as part of the design of the experiment.

There was a group that was given a pill that looked like the treatment drug, but was a "dummy" pill in that it did not have any active ingredients. It is important in the experiment because some participants might exhibit or feel the effect just because they are given a drug (this is known as the placebo effect). We now have a group to compare our results to. It is hard to study the 76% effect unless we have the 29% to compare it to.

b) Explain what is meant by "double-blind" in the above description. Then, discuss why it was important to make this experiment double-blind.

Double blind means that the children were unaware of which treatment they received and also the experimenters who interact with the children do not know which group they are in. This is important because our results could be biased if the subjects knew they were taking the treatment and not the placebo. Also, the experimenters could inadvertently or subconsciously enter bias into the results if they know whether they are in the treatment or placebo group.

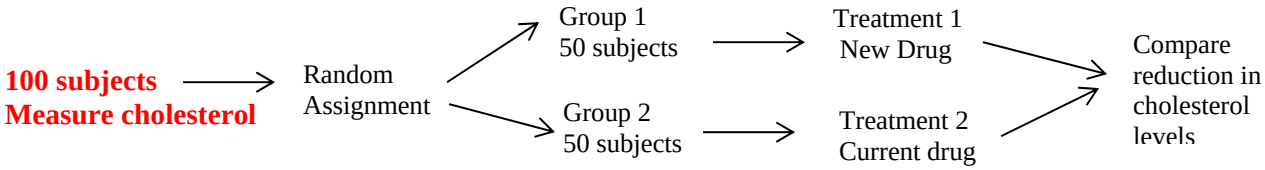
Ex 9: An experiment compares the taste of a new spaghetti sauce with the taste of a successful sauce. Each of a number of tasters tastes both sauces (in random order) and says which tastes better. This is called a
(a) simple random sample.
(b) stratified random sample.
(c) completely randomized design.
(d) matched pairs design.

Ans: D

Ex 10: High cholesterol levels in people can be reduced by exercise or by drug treatment. A pharmaceutical company has developed a new cholesterol-reducing drug. Researchers would like to compare its effects to the effects of the cholesterol-reducing drug that is currently available on the market. One hundred volunteers who have a history of high cholesterol and who are currently not on medication will be recruited to participate in a study.

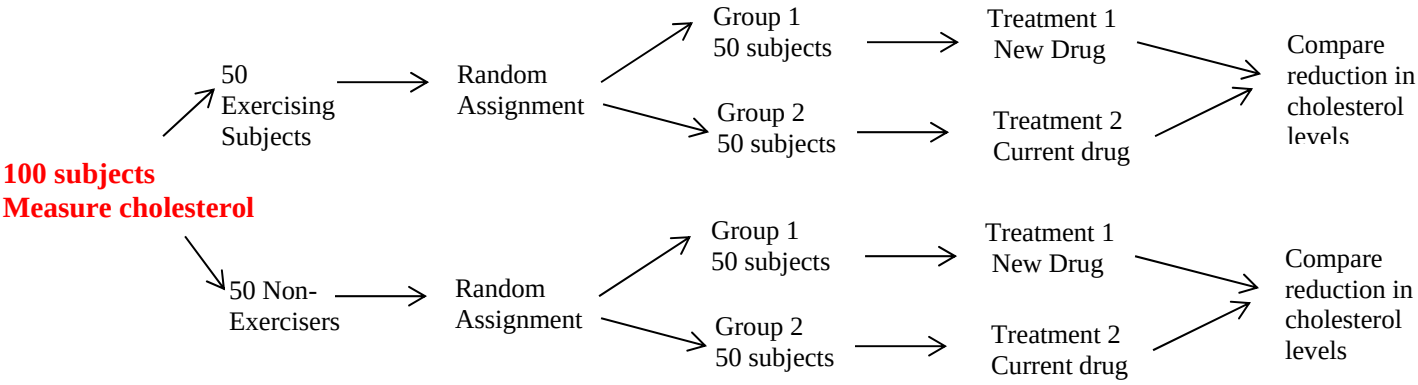
a) Explain how you would carry out a completely randomized experiment for the study.

Measure all of the volunteer's cholesterol levels. Then number the volunteers from 00 to 99. Use a random number table and read two digits at a time, ignoring repeats. Pick 50 two-digit numbers to be in Group 1 that gets the new drug. The remaining 50 numbers that were not selected will be Group 2 and get the drug currently on the market.



b) Describe an experimental design that would improve the design in (a) by incorporating blocking.

Because exercise might reduce cholesterol levels, it would be a good idea to block the participants by exercise or lack of exercise. This will help account for the variation in reduction that is due to exercise.



(c) Can the experimental design in (b) be carried out in a double-blind manner? Explain.

Yes, it can be carried out in a double-blind manner. The subjects do not need to know which drug they are receiving and the experimenters that are in contact with the subjects also do not need to know which treatment they received.

Ex 11: As dogs age, diminished joint and hip health may lead to joint pain and thus reduce a dog's activity level. Such a reduction in activity can lead to other health concerns such as weight gain and lethargy due to lack of exercise. A study is to be conducted to see which of two dietary supplements, glucosamine or chondroitin, is more effective in promoting joint and hip health and reducing the onset of canine osteoarthritis. Researchers will randomly select a total of 300 dogs from ten different large veterinary practices around the country. All of the dogs are more than 6 years old, and their owners have given consent to participate in the study. Changes in joint and hip health will be evaluated after 6 months of treatment.

a) What would be an advantage to adding a control group in the design of this study?

A control group gives the researchers a comparison group to be used to evaluate the effectiveness of the treatments. The control group allows the impact of the normal aging process and other variables on joint and hip health to be measured with appropriate response variables. The effects of glucosamine and chondroitin can be assessed by comparing the responses for these two treatment groups with those for the control group.

b) Assuming a control group is added to the other two groups in the study, explain how you would assign the 300 dogs to these three groups for a completely randomized design.

Each dog will be assigned a random number from 001 to 300. Then, using a random number generator on the calculator, pick 100 unique numbers in that range and the dogs with those numbers become Group 1 which receives glucosamine. The next 100 unique numbers picked will be the dogs that form Group 2 and they will be in the control group. The 100 dogs not chosen become Group 3 and receive chondroitin.

c) Rather than using a completely randomized design, one group of researchers proposes blocking on clinics, and another group of researchers proposes blocking on breed of dog. How would you decide which <u>one</u> of these two variables to use as a blocking variable?

The goal of blocking is to create groups of homogeneous experimental units. The key question is which variable has the strongest association with joint and hip health. It is reasonable to assume that most clinics will see all kinds and breeds of dogs, so there is no reason to suspect that joint and hip health will be strongly associated with a clinic. On the other hand, different breeds of dogs tend to come in different sizes. The size of a dog is associated with joint and hip health, so it would be better to block on breed.

Ex 12: Two essential features of all statistically designed experiments are
(a) compare several treatments; use the double-blind method.
(b) compare several treatments; use chance to assign subjects to treatments.
(c) always have a placebo group; use the double-blind method.
(d) use a block design; use chance to assign subjects to treatments.
(e) always use a large number of subjects; use the double-blind method.

Ans: B

Ex 13: The reason that block designs are sometimes used in experimentation is to
(a) prevent the placebo effect.
(b) allow double blinding.
(c) eliminate confounding with another factor.
(d) eliminate sampling variability.
(e) stratify the sample.

Ans: D

Ex 14: The most important advantage of experiments over observational studies is that
(a) experiments are usually easier to carry out.
(b) experiments can give better evidence of causation.
(c) confounding cannot happen in experiments.
(d) an observational study cannot have a response variable.
(e) observational studies cannot use random samples.

Ans: B

SAMPLE EXAM QUESTIONS

1. **In order to assess the effects of exercise on reducing cholesterol, a researcher sampled 50 people from a local gym who exercised regularly and 50 people from the surrounding community who did not exercise regularly. They each reported to a clinic to have their cholesterol measured. The subjects were unaware of the purpose of the study, and the technician measuring the cholesterol was not aware of whether subjects exercised regularly or not. This is**

   A. An observational study

   B. An experiment, but not a double-blind experiment

   C. A double blind experiment

   D. A matched-pairs experiment

   Ans: A

2. **Can pleasant aromas help a student learn better? Two researchers believed that the presence of a floral scent could improve a person's learning ability in certain situations. They had 22 people work through a pencil-and-paper maze; three times while wearing a floral-scented mask and three times wearing an unscented mask. The three trials for each mask closely followed on another. Testers measured the length of time it took subjects to complete each of the six trials. They reported that, on average, subjects wearing the floral-scented mask completed the maze more quickly than those wearing the unscented mask, although the difference was not statistically significant. This study is**

   A. A convenience sample

   B. An observational study, not an experiment

   C. An experiment, but not a double-blind experiment

   D. A double-blind experiment

   Ans: C

3. **In order to assess the opinion of students at the University of Michigan on campus snow removal, a reporter for the student newspaper interviews the first 12 students he meets who are willing to express their opinion. The method of sampling used is**

   A. Simple random sampling

   B. Convenience sampling

   C. Voluntary response

   D. A census

   Ans: B

4. In order to assess the opinion of students at the University of Michigan on campus snow removal, a reporter for the student newspaper interviews the first 12 students he meets who are willing to express their opinion. In this case, the sample is

A. All those students favoring prompt snow removal

B. All students at universities receiving substantial snow

C. The 12 students interviewed

D. All students at the University of Michigan

Ans: C

5. In order to select a sample of undergraduate students in the United States, I select a simple random sample of four states. From each of these states, I select a simple random sample of two colleges or universities. Finally, from each of these eight colleges or universities, I select a simple random sample of 20 undergraduates. My final sample consists of 160 undergraduates. This is an example of

A. Simple random sampling

B. Stratified random sampling

C. Multistage sampling

D. Convenience sampling

Ans: C

6. A marketing research firm wishes to determine if the adult men in Boise, Idaho, would be interested in a new upscale men's clothing store. From a list of all residential addresses in Boise, the firm selects a simple random sample of 100 and mails a brief questionnaire to each. The sample in this survey is

A. All adult men in Boise, Idaho

B. All residential addresses in Boise, Idaho

C. The members of the marketing firm that actually conducted the survey

D. The 100 addresses to which the survey was mailed

Ans: D

7. A study of the effects of running on personality involved 231 male runners who each ran about 20 miles a week. The runners were given the Cattell Sixteen Personality Factors Questionnaire, a 187-item multiple-choice test often used by psychologists. A news report (*New York Times*, Feb. 15, 1988) stated, "The researchers found statistically significant personality differences between the runners and the 30-year-old male population as a whole." A headline on the article said, "Research has shown that running can alter one's moods." Which of the following statements is true?

A. This study was not a designed experiment.

B. This study was an experiment, but not a double-blind experiment.

C. This study was a double-blind experiment, but not a randomized experiment.

D. This study was a randomized, double-blind experiment.

Ans: A

8. A call-in poll conducted by *USA Today* concluded that Americans love Donald Trump. *USA Today* later reported that 5640 of the 7800 calls for the poll came from the offices owned by one man, Cincinnati financier Carl Lindner, who is a friend of Donald Trump. The results of this poll are probably

A. Surprising, but reliable since it was conducted by a nationally recognized organization

B. Biased, but only slightly since the sample size was quite large

C. Biased, understating the popularity of Donald Trump

D. Biased, overstating the popularity of Donald Trump

Ans: D

9. In a recent study, a random sample of children in grades two through four showed a significant negative relationship between the amount of homework assigned and student attitudes. This is an example of

A. An experiment

B. An observational study

C. The establishing of a causal relationship through correlation

D. A block design, with grades as blocks

Ans: B

10. For a sample to be a simple random sample of size *n*,

A. *n* must be a large number

B. every item in the population must be selected

C. every collection of *n* individuals must have the same chance to be the sample actually chosen

D. the size of the population must be smaller than *n*

Ans: C

11. High blood pressure adds to the workload of the heart and arteries and may increase the risk of heart attacks. If not treated, this condition can also lead to heart failure, kidney failure, or stroke. We wish to test the effectiveness of Angiotensin-converting enzyme (ACE) inhibitors as a treatment for high blood pressure.

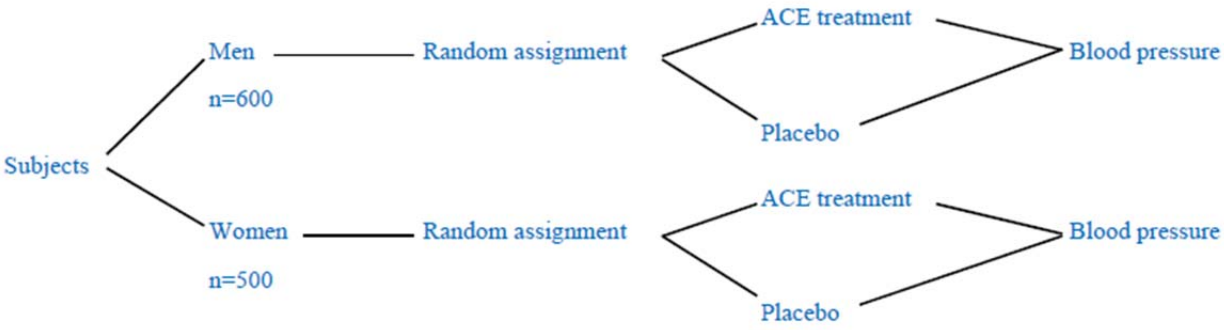
(1) It is well known that men and women may react differently to common cardiovascular drug treatments. What sort of experimental design would you choose for this study, and why?

A randomized block design by gender to reduce the variability due to gender arising from a completely randomized design.

(2) Explain why an experiment involving 600 men and 500 women is preferable to one involving 60 men and 50 women.

A larger number of subjects will decrease the impact of random variation on the results of the experiment.

(3) Assume that 600 men and 500 women suffering from high blood pressure are available for the study. Describe a design for this experiment. Be sure to include a description of how you assign individuals to the treatment groups.



Within the block of men, assign numbers 001-600. Within the block of women, assign numbers 001-500. Choose three digit numbers from a random digit table until you have selected 300 men, ignoring repeats and numbers 000, 601-999. Move to a different area of the table and repeat the process to select 250 women. These subjects will be in the ACE treatment group. All others will be in the control group.

12. Read the following article about the connection between vitamin E and heart bypass surgery.

Vitamin E may have special health benefits

Large doses of vitamin E apparently can reduce harmful side effects of bypass surgery in heart patients. A study involving 28 bypass patients found that the 14 randomly-assigned patients who took vitamin E for two weeks before their operations had significantly better heart function after the procedure than the 14 randomly-assigned patients who took placebos.

The vitamins apparently prevent damage to the heart muscle by destroying the toxic chemicals, called free radicals, that form when blood is cut off during surgery, said Dr. Terrance Yau of the University of Toronto.

(1) Explain why this is an experiment and not an observational study.

This is an experiment because treatments are being imposed on the subjects.

(2) Identify the explanatory and response variables.

Explanatory variable: vitamin E Response variable: heart function

(3) Identify the type of experimental design used in this study. Justify your answer.

This is a completely randomized design. All subjects are randomly assigned to a treatment group.

(4) In the second sentence above is the phrase, "…the 14 patients who took vitamin E for two weeks before their operations had significantly better heart function after the procedure…" What is the statistical meaning of the word "significantly" in the context of this study?

"Significantly" means the differences in heart function between the groups is unlikely to have occurred by chance.

(5) This was a controlled experiment. Describe how it was controlled and explain the purpose of doing so.

They are comparing a placebo group (control) to the vitamin E treatment group. This isolates the impact of the vitamin E on heart function.

Sec. 9.3 To organize data

The standards covered in this learning target were all covered in the Algebra I curriculum. This unit is meant to be review of Algebra I learning targets and should go quickly. It is important that students can describe distributions, discuss center and variability, and interpret graphical representations of data. The notes provided should help review these topics.

DISTRIBUTION: The distribution of a variable tells us what values the variable takes and how often it takes these values.

VARIABLES: **Categorical (Qualititative):** Places an object/individual into one of several groups or categories.

Types of graphs: Pictograph, bar chart, pie chart, dotplot

Quantitative (Numerical): Takes a numerical value for which arithmetic operations such as finding the mean, make sense.

1) Discrete – numerical data that are isolated on the number line (counting numbers).

2) Continuous – The set of possible values forms an entire interval on the number line (measurement).

Types of graphs: Bar chart, histogram, stemplot, dotplot, boxplot, line graph

MAKING GOOD GRAPHS: Make sure to title and label your graphs. Use legends or a key if needed. Make the data stand out and pay attention to what the eye sees. Avoid clutter. Avoid three-dimensional effects and have an **appropriate scale**.

Displaying Quantitative Data: We can use a frequency table, histogram (no spaces), dotplot, stemplot or boxplot. Dotplots and stemplots show all of the values and a lot of detail, but are bad for large sets of data. Histograms are better for large data sets. Boxplots only show the five-number summary.

Histograms: Most common graph for quantitative variables. We first divide the range of the data into classes of equal width and count the number in each class (make a frequency table), label and scale your axes and title your graph and then draw bars to represent the count in each class.

Tips: Five classes is a good minimum – take the square root of the number of observations and that will give you a good number of classes. Ex: If you have 50 pieces of data, use 7 classes. If the data is continuous, the bars should be touching.

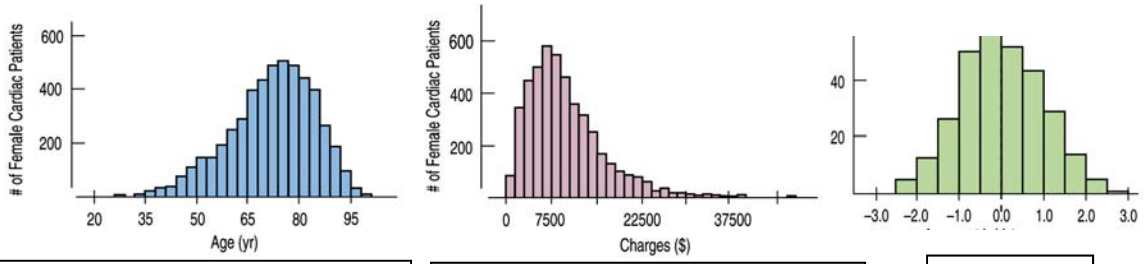
Stemplots: First, separate each observation into a stem, consisting of all but the final (rightmost) digit, and a leaf, the final digit. Stems may have as many digits as needed, but leafs contain a single digit. Then write the stems in a vertical column with the smallest or largest at the top. The leaf is then written next to its corresponding leaf in increasing order out from the stem. We can also make back-to-back or comparative stemplots if we have two data sets that we want to compare. Stems are in the middle and the leaves are written to the right and left of the stems. Make sure to label which side is which!

Boxplots: A graph that uses the five number summary. It can be drawn horizontally or vertically. Always provide a number line with it. A central box spans the quartiles. A line in the box marks the median. Lines extend from the box out to the smallest and largest observations that are not outliers. The modified boxplot shows outliers as dots or asterisks and the lines ("whiskers") then extend to the smallest and largest observations that are not outliers.

Describing a distribution: To describe the overall pattern, remember the acronym: **SOCS**
S – shape **O** – outliers **C** – center **S** – spread

Describing Shape:

1) Symmetric – left and right sides are mirror images.
2) Skewed to the right (positively skewed) – the right side extends much farther out than the left side.
3) Skewed to the left (negatively skewed) – the left side extends much farther out than the right side.



Skewed left (negatively skewed) | Skewed right (positively skewed) | symmetric

Outliers: An outlier in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Measuring Center

Median – is the midpoint of a distribution. The median is the middle number, so it is NOT affected by outliers. When a statistic is not affected by outliers, we call it a **resistant measure**.

The most common measure of center is **mean**, the arithmetic average. $\bar{x} = \frac{\sum x}{n}$

The mean is sensitive to extreme observations (outliers), so we say that it is NOT a **resistant measure** of center.
When a distribution is symmetric, mean and median are the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

Measuring Spread

The simplest measure of spread is **range**, which is the difference between the largest and smallest observation. It is GREATLY affected by outliers and NOT Resistant.

Quartiles:

The First Quartile, Q_1, is the 25th percentile. It is the median of the first half of the observations. The Third Quartile, Q_3, is the 75th percentile and is the median of the second half of the observations. The Second Quartile, Q_2, is also known as the median.

Interquartile Range (IQR) is the distance between the two quartiles.

$IQR = Q_3 - Q_1$. This gives the RANGE COVERED BY THE MIDDLE HALF OF THE DATA. The IQR IS resistant.

Formal Outliers:

Any observation that falls more than $1.5 \times IQR$ above the third quartile or below the first quartile. $[Q_3 + (1.5 \times IQR)]$ and $[Q_1 - (1.5 \times IQR)]$

5 Number Summary:

The five number summary: Minimum, Q_1, Median (Q_2), Q_3, Maximum

Standard Deviation:

Standard deviation is the average of the deviations from the mean. It is the average amount that the data varies from the mean.

Formula: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

The most common measure of spread is standard deviation. It's symbol is **s**. It is S_x on the TI-84 calculator under 1-VAR STATS.

S.D. Properties:

1. s = 0 only when there is no spread. Example: data set is 5, 5, 5 - no spread!

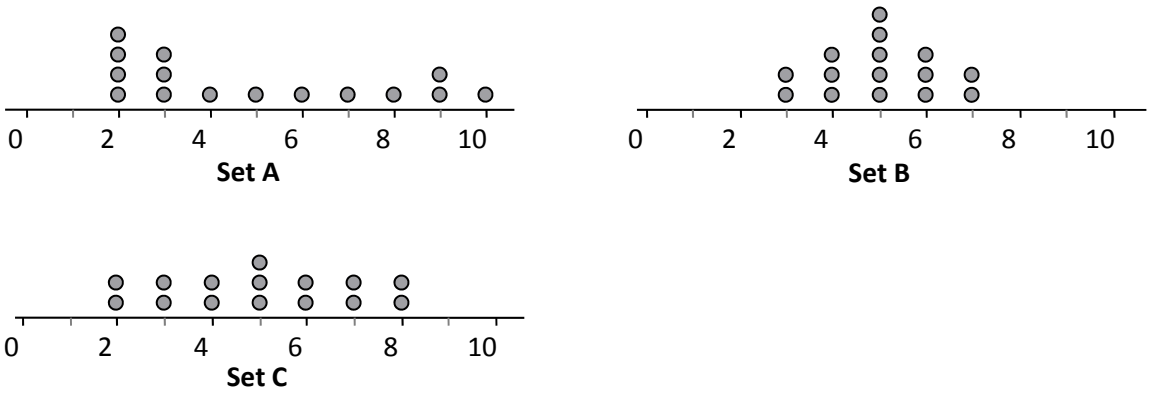
2. s, like $\bar{x}$, is NOT resistant. It is affected by outliers.

3. s is small if the observations are close to the mean, large if they are far from the mean. The more spread out the data, the larger the standard deviation.

4. The sum of the deviations from the mean are always equal to zero. This is why we square the deviations in the formula.

Which to report? The 5-number summary is better to report center and spread for skewed distributions or distributions with strong outliers. Use $\bar{x}$ and s to describe the data only for reasonably symmetric distributions without outliers. A graph is always the best way to describe a distribution. Always plot your data first and then report numerical summaries.

Ex 15: Examine the dotplots below from three sets of data.



The mean of each set is 5. The standard deviations of the sets are 1.3, 2.0, and 2.9. Match each data set with its standard deviation.

(A) Set A: 1.3 Set B: 2.0 Set C: 2.9

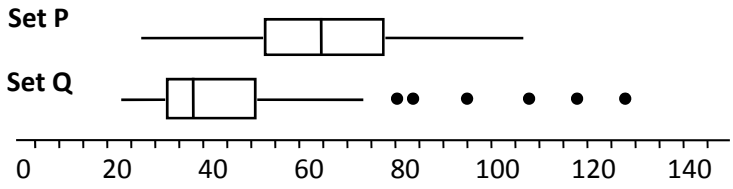
(B) Set A: 2.0 Set B: 1.3 Set C: 2.9

(C) Set A: 2.0 Set B: 2.9 Set C: 1.3

(D) Set A: 2.9 Set B: 1.3 Set C: 2.0

Ans: D

Ex 16: Use the boxplots of two data sets, P and Q, below.



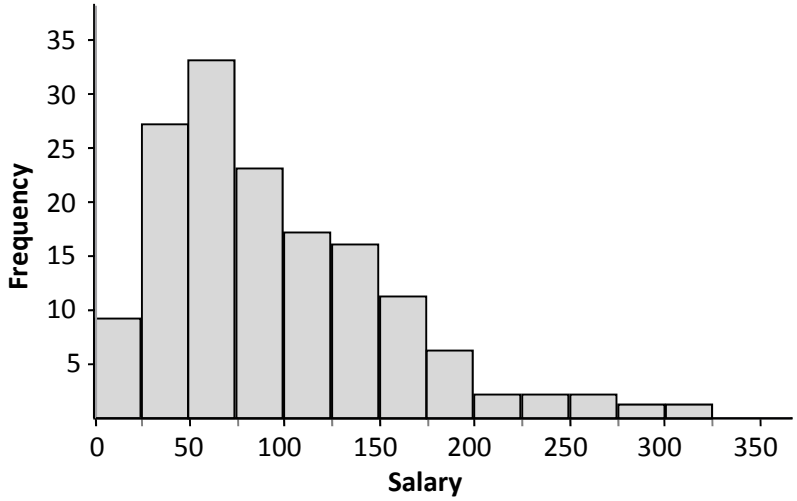
a) Which data set has the larger interquartile range? Set P

b) Describe the shape of both distributions.

Set P appears to be fairly symmetric, while Set Q is skewed to the right.

c) Which data set has values that are considered outliers? Only Set Q

Ex 17: This graph shows annual salaries (in thousands of dollars) for all workers in a certain city.



a) Describe the distribution.

The distribution of salaries is skewed to the right with no gaps and no apparent outliers. The center of the distribution is around $75,000 and the spread of the data has a range of $325,000.

b) If the median salary is $80,500, approximate the mean.

Since the data is skewed to the right, the larger values on the right will pull the mean toward its tail. That means the mean will be larger than the median at approximately $94,000.

SAMPLE EXAM QUESTIONS

1. **A company data base contains the following information about each employee: age, date hired, sex (male or female), ethnic group (Asian, black, Hispanic, etc.), job category (clerical, management, technical, etc.), yearly salary. Which of the following lists of variables are all categorical?**

   A. Age, sex, ethnic group

   B. Sex, ethnic group, job category

   C. Ethnic group, job category, yearly salary

   D. Age, date hired

   Ans: B

2. **Were the extinctions that occurred in the last ice age more frequent among species of animals with large body sizes? A researcher gathers data on the average body mass (in kilograms) of all species known to have existed at that time. These measurements are values of**

   A. A categorical variable

   B. A quantitative variable

   C. An invalid variable

   D. A margin of error

Ans: B

3. **The figure below is the density curve of a distribution. This distribution is**

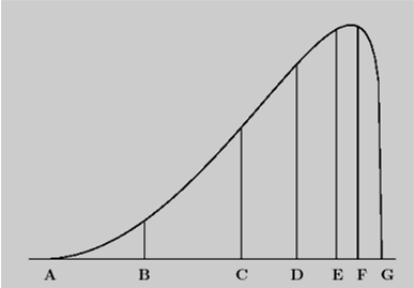
   A. Roughly symmetric

   B. Skewed to the left

   C. Skewed to the right

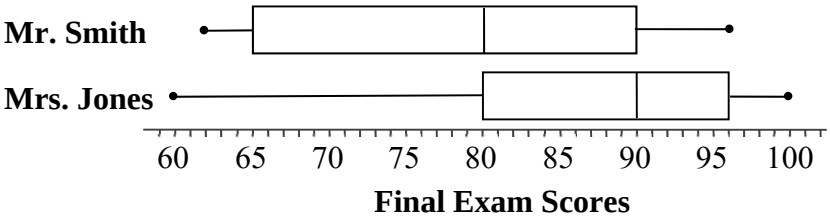
   D. Positively correlated

   E. Negatively correlated



Ans: B

4. **The distributions of two classes' final exam scores are shown below.**



Which statement about the box-and-whisker plots is true?

A. 50% of the scores for Mr. Smith's class are between 65 and 80.

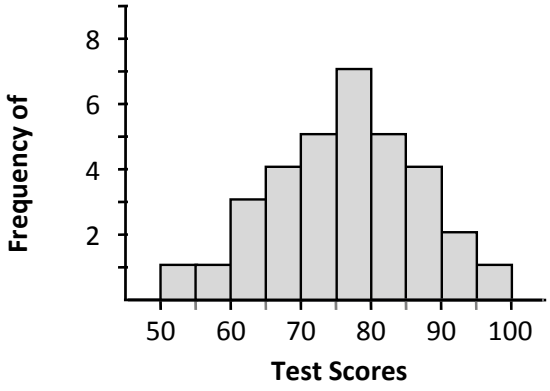
B. 50% of the scores for Mrs. Jones' class are between 80 and 100.

C. The median scores for the two classes are the same.

D. The interquartile range of scores for Mr. Smith's class is greater than the interquartile range of the scores for Mrs. Jones' class.

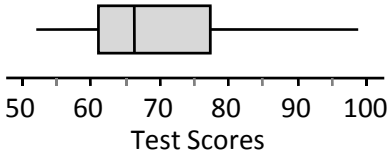
Ans: D

5. Mrs. Johnson created this histogram of her 3rd period students' test scores.

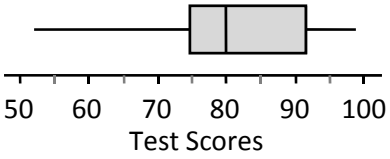


Which boxplot represents the same information as the histogram?

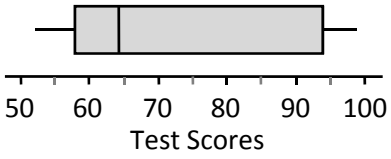
(A)



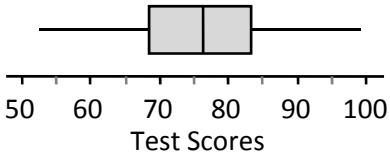
(B)



(C)



(D)



Ans: D

For questions 6 – 8, use the following scenario.

A survey was made of high-school-aged students owning cell phones with text messaging. The survey asked how many text messages each student sends and receives per day. Some results are shown in the table below.

		Number of text messages sent/received per day among teens who text	
Group	Number Surveyed	Mean	Median
Girls, 14–17 years old	270	187	100
Boys, 14–17 years old	282	176	50
Total	552		

6. A histogram of the girls' responses (not shown) has a strong right skew. Which statement would support that observation?

(A) The number of girls' surveyed is greater than the mean number of texts sent by girls.
(B) The mean number of texts sent by girls is greater than the median number of texts sent by girls.
(C) The mean number of texts sent by girls is greater than the mean number of texts sent by boys.
(D) The median number of texts sent by girls is greater than the median number of texts sent by boys.

Ans: B

7. Which expression shows the mean number of text messages for <u>all</u> girls and boys, 14–17 years old?

(A) $\frac{187+176}{2}$

(B) $\frac{187+176}{552}$

(C) $\frac{270\times187+282\times176}{552}$

(D) It cannot be computed from the information given.

Ans: C

8. Which group's data has the larger interquartile range?

(A) Boys
(B) Girls
(C) Neither, they are equal.
(D) It cannot be computed from the information given.

Ans: D