Statistics, Day 1

Objectives: To identify a statistical question

To recognize the roll of randomness in statistical studies

To distinguish between surveys, observational studies and experiments

Statistics is the study of data. It includes collecting data, organizing data, and drawing conclusions about data.

A statistical question is one for which you don't expect to get a single answer. Instead, you expect to get a variety of different answers, and you are interested in the distribution and tendency of those answers.

For example, "How tall are you?" is not a statistical question. But "How tall are the students in our school?" is a statistical question.

The data to answer a statistical question may be collected by a survey, observational study, or experiment. In a survey, the data is reported directly by the subjects. In an observational study, researchers observe and/or measure the data of interest. In an experiment, the researcher manipulates something in order to observe and measure a response.

Example: A student wants to know if the students at his high school prefer Coke or Pepsi. Describe how this data could be collected in a survey? An observational study?

An experiment?

What might be some of the advantages and disadvantages of each type of study?

The populationin a statistical study is the entire group that we want information about. A sample is the part of the population that we actually collect the information about. In order for a survey to be unbiased, the sample must be selected randomly.

For example, for the statistical question “How tall are the students in our school?" the population is all of the students in our school. It would take a long time to measure every student in the school, so we might choose to measure a smaller random sample of all the students.

Example: Which of the following methods could be used to produce a random sample of 30 students to survey?

1. Put all of the students names in a hat and randomly choose 30 names
2. Put all of the teachers names in a hat and randomly select a teacher and measure the students in her class
3. Label all of the students in the school with a number and use a calculator to generate 30 random numbers.
4. Pick the first 30 students on an alphabetical list of all students.
5. Go to the gym and choose 30 students who are practicing basketball.

A census is when everyone in the population is sampled. A census is frequently time-consuming and expensive. In the US, a census to count the population is required every 10 years.

Unlike a survey, the subjects of an experiment are usually not selected at random. Instead, for an experiment, the treatments must be assigned randomly.

Example: A biology teacher wants to know if his students learn as well watching a virtual lab as when doing an actual lab. He allows each student to decide which lab to participate in. Both the virtual lab and the actual lab take one class period, and the next day each student takes a test on what they have learned. The teacher finds that the students doing the actual lab do better on the test. Can he conclude that the virtual lab is not as effective as the real one? Explain.

Example: How could you improve the experiment above?

In general, surveys and observational studies can show association (or correlation) but only randomized experiments can prove causation.

Statistics, Day 2

Objectives: To find the mean and median of a data set

To make a histogram and box plot of a data set

Suppose we have a set of data, in this case the test scores of one student in Algebra 2. What might we conclude?

 55, 84, 99, 60, 73, 99, 76, 95, 91, 97

Sometimes it is convenient to have one number to describe a set of data. We call this number a measure of central tendency. The most commonly used measures of center are mean, median, and mode.

Mean – use when the data are spread out and you want an average value.

=

Median – arrange data numerically and take the middle value. If there are an even number of data points, average the two data in the middle.

 55, 60, 73, 76, 84, 91, 95, 97, 99, 99

It is best to use the median as a measure of center when the distribution has outliers or a long tail.

Mode – the number that appears most often. If no number is most often, the data has no mode. This measure of center is not widely used.

Which measure of center do most teachers use as grades? Does this seem fair?

Usually, a single number does not tell us all we want to know, and it is helpful to look at the data *graphically*.

Frequency Distributions and Histograms

A frequency distribution summarizes the number of data points that fall within a specific interval. Each interval has to have the same width. A histogram forms a picture of the distribution.

55, 60, 73, 76, 84, 91, 95, 97, 99, 99



|  |  |
| --- | --- |
| Interval | # of Data Pts |
| 50-60\* |  |
| 60-70\* |  |
| 70-80\* |  |
| 80-90\* |  |
| 90-100\* |  |

\*The upper endpoints are not

included in the interval

The way your histogram looks will depend greatly on the intervals you choose.

Box Plot (Box and Whisker Plot)

In a set of data, the quartiles divide the distribution into four equal parts. The first quartile, or Q1, is the median of the lower half of the data and the third quartile, or Q3, is the median of the upper half of the data. By definition, 25% of the data points are contained in the first quartile, 50% of the data is within the first two quartiles, and 75% of the data is within the first three quartiles.

To find Q1, Q2 (which is the same as the Median), and Q3, examine the data in numerical order:

55, 60, 73, 76, 84, 91, 95, 97, 99, 99

To make the box plot, place the quartile values over a number line, along with the minimum and maximum values. Draw a box from Q1 to Q3, a vertical line at the Median, and whiskers to the min and max values.



The minimum, Q1, Median, Q3 and maximum values are called the five number summary.

Statistics, Day 3

Objectives: To find measures of variation in a set of data; To identify outliers; To use the graphing calculator to find summary statistics and make Histograms and Box Plots.

We have already talked about measures of central tendency in a data set. What are they?

Today we will look at measures of variation (or spread). Let’s look at the test score data again:

 55, 60, 73, 76, 84, 91, 95, 97, 99, 99

The simplest measure of spread is the range, the difference between the maximum and minimum value in the data set. It is strongly influenced by outliers.

 Range = Max – Min =

The Interquartile Range (IQR) is another measure of variation that is not as influenced by outliers. The IQR is the difference between Q3 and Q1. It represents the middle 50% of the data.

 IQR = Q3 – Q1 =

A data point is considered an outlier if it lies more than 1.5\*IQR away from Q3 or Q1. Outliers are represented as individual data points on a box plot.

Example: Use the 1.5 IQR rule to determine if any of the test scores are outliers.

The IQR is used as a measure of spread when the median is used as a measure of center. When the mean is used as a measure of center, the standard deviation is used as the measure of spread. It measures an average distance of each data point in a sample from the mean. It is calculated by

Example: Find the standard deviation of the data set:

0, 5, 5, 10.

The graphing calculator can be used to find summary statistics, such as the mean or standard deviation. The calculator can also be used to make plots of the distribution.

55, 84, 99, 60, 73, 99, 76, 95, 91, 97

Example: Enter the test score data into your calculator by selecting: STAT: EDIT. If you already have data in List 1, you can clear it by placing your cursor on L1, and then selecting Clear, Enter.

To find the mean, standard deviation and five number summary:

 STAT CALC 1:1-Var Stats (Enter) L1 (Enter)

To make a histogram:

 2nd STAT PLOT 1 (Enter)

 Turn PLOT 1 On, Type: histogram, Xlist: L1

 Use WINDOW to set the intervals. Then GRAPH.

 TRACE gives the number of data points in each interval. (You could use Trace to help make the frequency table)

Example: Use several different intervals to change the look of your graph. Which one seems the most useful?

To make a box plot: Return to 2nd STATPLOT 1 and select the first box plot icon.

Example: Does the data have any outliers? How can you tell?Statistics, Day 4

Objectives: To use the calculator to do a simulation

To review summary statistics, histograms and box plots.

A simulation is the imitation of chance behavior. Simulations are often done using calculators or computers.

Let’s suppose that your kid brother’s favorite cereal advertises that each box of cereal contains one of six different matchbox cars. How many boxes of cereal will your mother have to buy so that your brother gets one of every kind? (Assume that there is an equal number of each type of car placed randomly in the boxes). First, make a guess.

We can use the random integer feature of the calculator to simulate which type of car (#1 – 6 ) is in a box of cereal. We will need to keep track of how many boxes we’ll purchase until we get all 6 types of cars. Will everyone in the class get the same answer?

Run your simulation until you get one of each type of car. Cross out the car number as you get each one. Don’t forget to keep track of how many boxes you’ve purchased.

 Car: 1 2 3 4 5 6

Tally:

Enter each student’s results into L1 in your calculator.

Make a histogram of the results. What do you notice about the shape of the distribution? Can you explain why the graph has that shape?

What are the mean and median of the distribution?

Which would be a better indicator of how many boxes your mother should expect to have to buy to get all six cars? Why?

What is the IQR of the data? Explain what this number tells you.

Does the data set have any outliers?

25% of the time your mother can expect to buy \_\_\_\_\_ boxes or more.

25% of the time your mother can expect to buy \_\_\_\_\_ boxes or fewer.Statistics, Day 5

Objective: To compare the shapes, center and spreads of two distributions; To determine if a distribution is symmetric or skewed

What does the standard deviation tell you about a data set?

If two data sets have the same mean, the data with the larger standard deviation is more spread out.

Example: A company that makes raisin bread advertises that there are at least 10 raisins in every slice of bread. Sample loaves are tested 20 times per day, and a slice of bread from each loaf is examined. If there are fewer than 8 raisins in the slice, the loaf cannot be sold. Examine the following data to determine which baker is doing a better job.

Baker 1 Baker 2

6 12 8 13 15 10 12 8 13 10

7 11 9 15 10 6 12 10 9 7

13 5 7 11 13 14 11 10 9 11

13 10 12 3 12 10 12 8 9 14

 

Distributions are frequently described by their shape.

Look at the histogram for Baker 2. The graph is relatively symmetric. For a symmetric distribution, the mean and the median are relatively the same.



Baker 2

= 10.25

Median = 10

s = 2.17

Now look at the histogram for Baker 1. This graph is called skewed left (or negatively skewed) because there is a long tail of data on the left. For this data, the mean is less than the median because it is being pulled down (to the left) by the lower numbers.

Baker 1

= 10.25

Median = 11

s = 3.34

The shapes of the distributions can also be compared using box plots. Your calculator can make up to three box plots stacked on top of each other.

Example: Describe the shape of the following set of data

 75, 104, 72, 71, 95, 69, 75, 73, 77, 76, 90

What are the values of the mean and the median?

Which would be a better measure of center for the distribution?

Are any of the data points outliers?

Statistics, Day 6

Objectives: To solve problems with normally distributed data

Many data sets have a symmetric distribution with one peak. If you have a lot of data and use very small intervals, the histogram becomes a continuous bell-shaped curve, called a normal distribution.

Normal distributions occur quite frequently. Standardized test scores and population heights can both be represented by normal distributions. In both cases, you must have a large number of data points before the distribution looks normal.

In ALL normal distributions, about

\*68% of the data fall within one standard deviation of the mean

\*95% of the data are within two standard deviations of the mean

\*99.7% of the data are within three standard deviations







-3 -2 -1 0 1 2 3

Example: The math scores on the SAT are approximately normally distributed with a mean of 500 and a standard deviation of 100. What percent of students score

 a) between 400 and 600?

 b) between 500 and 700?

 c) greater than 700?

 d) less than 600?

Example: The amount of coffee dispensed from a vending machine is normally distributed with a mean of 10.50 oz. and a standard deviation of 0.75 oz.

a) 68% of the amount of coffee dispensed falls within what range?

b) About what percent of the time will the machine overfill a 12 oz cup?

Example: The useful life of a tire is normally distributed with a mean of 30,000 miles and a standard deviation of 5000 miles. A company manufactures 10,000 tires a month.

a) Draw the normal curve and label the mean, and + 1, 2, 3 standard deviations from the mean.

a) How many tires will last between 25,000 and 35,000 miles?

b) How many tires will last more than 40,000 miles?

c) How many tires will last less than 20,000 miles?

Statistics, Day 7

Objectives: To use technology to solve problems with normally distributed data

To determine if a data set can be represented by a normal distribution.

Warm-up: Young women’s heights are approximately normally distributed with a mean of 64.5 and a standard deviation of 2.5.

Draw a sketch of the normal distribution and label the mean, and + 1, 2, 3 standard deviations from the mean.

What percentage of young women have heights between 62” and 67”?

What percentage is taller than 69.5”?

What is the probability that a randomly selected woman is less than 62”?

What if we wanted to know what percent of women is between 63” and 66” tall? Can you make a guess?

To determine the percentage of women between 63” and 66”, we need to use a calculator, computer, or tables. We will use the calculator’s normal cumulative distribution function. Go to 2nd VARS and under the DISTR menu select 2:normalcdf. After the parenthesis, enter the lower bound of the interval you want, the upper bound, the mean and the standard deviation. It should read normalcdf (63, 66, 64.5, 2.5). Draw the curve and shade the appropriate area.

Example: What percentage of women are taller than 68 inches? (Use 1E99 as the upper bound). Draw the curve and shade the appropriate area.

Example: What percentage of women are shorter than 60 inches? (Use -1EE99 as the lower bound).

Example: A teacher gave a test and recorded the following scores for a random sample of 16 of his students. Do the test scores seem to follow an approximately normal distribution? What are the mean and standard deviation?

85, 76, 83, 82, 75, 76, 73, 76, 72, 66, 80, 74, 70, 78, 70, 79

Based on this sample of 16 students, it would be reasonable to represent the population of all of the test scores for that teacher as an approximately normal distribution with a mean of about 76 and a standard deviation of about 5. Draw the normal curve below.

If a passing grade on the test is a 70%, about what percent of all the teacher’s students can be expected to pass the test?

Would you be surprised if a student scored 95% or more? Explain your reasoning.

Statistics, Day 8

Objective: To use simulation to make statistical inferences

Statistical inference is the process of drawing conclusions about population parameters from sample data.

Suppose you are running for student body president. On the morning of the election, your campaign manager asks a random sample of 20 students if they are planning to vote for you. Eight of the students asked say yes.

What proportion of the sample of students will vote for you?

If your manager had asked a different sample of 20 students, do you think the exact same number of students would vote for you?

Do you think it’s possible you will still get 50% of the vote in today’s election?

To answer such a question, we need to develop a probability model. The model is based on the hypothesis that 50% of all students will vote for you. The hypothesis tells us that the probability a student votes for you is the same as the probability you get a head when you flip a coin. We can use the random integer feature of the calculator to simulate how many times you might expect to get a head when you flip a coin 20 times.

Example: Use randInt (1,2) to determine whether a student votes for you. Repeat 20 times and count the number who will vote yes. Then make a dot plot of the results from everyone in the class. This simulates taking repeated samples of 20 students.

Here’s a dot plot showing the results of 100 simulations. Remember that the plot is based on the hypothesis that 50% of the voters will vote for you!



What is the most likely number of people in a sample of 20 who will vote for you?

Would you be surprised if 17 people vote for you?

How likely is it that 8 or fewer people will vote for you?

Based on the probability model, is it possible that you will get at least 50% of the vote?