**Barron’s Let’s Review Regents – Algebra II**

# Chapter 11: Statistics

## 11.1 Types of Statistical Studies

**Key Ideas**

Three ways to collect data for statistical study are , *observational study*, and *experiment*. *Bias* in a statistical study is when something about how the data are collected may have caused inaccurate results. Causes of this bias should be identified and, if possible, eliminated before collecting the data.

**Statistical Survey**

A survey is a question or series of questions that participants, also called subjects, in the study are asked to answer. The survey can have a simple yes/no question like “Do you like pistachio ice cream?” or a question with a numerical answer like “How much time did you spend on homework last night?” A survey is the simplest way of collecting data, but there are many ways that it can lead to bias.

The wording of the survey question can cause bias. For example, if the question says “Do you not hate pistachio ice cream?” rather than “Do you like pistachio ice cream?” the negative phrasing of the question could change the results even if the questions are supposed to mean the same thing.

The participants of the survey should be randomly selected, otherwise the results can also be distorted. The results of a survey about pop music, for example, will not be accurate if most of the people asked to complete the survey are under 20 years old. If the participants in the survey do not accurately reflect the total population, the survey has *selection bias*.

If the survey is voluntary, there is a chance that people who respond to the survey are more likely to answer the question a certain way. A survey that is conducted by text messaging, for example, might be answered more by younger people, causing bias in the results.

**Observational Study**

An *observational study* is like a survey. Instead of subjects being asked to answer questions, the person conducting the study observes the behavior of the participants and records the results. To learn about whether or not people like pistachio ice cream, the observer could go to an ice cream shop and watch what different people order. An important aspect of an observational study is that the person conducting the study cannot do anything that could interfere with or control what the subjects do.

An observational study could have bias if the subjects are not randomly selected. For example, if the ice cream shop observations happen during school hours, people under 18 will not be adequately represented.

**Experimental Study**

In an experimental study, the person conducting the study randomly chooses some of the subjects and exposes them to some kind of treatment. An example is a study to see if taking vitamin C prevents colds. Fifty-people are randomly selected and then, from those fifty, twenty-five are randomly selected to take vitamin C pills daily while the other twenty-five are not given the pills.

It is important in an experimental study to choose the group of people getting the treatment randomly. If there is some kind of bias in who receives the treatment, the experiment might lead to an inaccurate conclusion.

### Check Your Understanding of Section 11.1

1. *Show how you arrived at your answers*.
2. You want to find out what type of movie is the favorite among the residents of New York City. If you want to do this by conducting a survey, what are some ways you can reduce bias.

Randomly select people. Two approaches come to mind. A random selection of people with phone numbers having an New York City area code could be called and asked what their favorite type of movie is.

You could go to a movie theatre and count the number of people in each movie, and determine the type of movie that might be their favorite.

1. You want to study how the amount of sleep a student gets the night before the Regents relates to his or her Regents score. How can this be studied as an observational study? How can this be studied as an experiment?

For an observational study, one could either perform a sleep study the night before a Regents test, or ask the student just prior to taking the test. The teacher scoring the students test would have to provide test results with the student’s permission, or the student would have to provide test results.

For an experimental study, we could randomly choose 50 students scheduled to take a Regents test, a party could be scheduled for half the class that keeps them up to 2 am in the morning, while the other half is sent home and asked to get a good night’s sleep.

1. There are 500 10th graders at Regentsville High School. You want to do an experiment where 20 students study for the Regents with just the red Barron’s book and 20 other students study the Regents with both the blue and red Barron’s books. How can the groups be chosen to reduce the chance of skewed data through bias?

It would be possible to get a list of 40 students willing to take part in the study, and randomly assign the students to either the red Barron’s book or both books group.

1. A survey is conducted to learn what subjects students enjoy most at school. One question reads, “What is your most boring subject?”. Could the wording of this question skew the results of the survey? Explain.

First, the question is worded negatively and change the results.

Secondly, and more importantly, it does not identify what subjects enjoy most, just what subjects students are bored with.

1. Is this describing a survey, an experiment, or an observational study? Explain.

A teacher teaches two Algebra II math classes, one during period 5 and another during period 6. In the period 5 class, the temperature in the room is set to 70 degrees for a two-week period. In the period 6 class, the temperature in the room is set to 85 degrees for the same two-week period. At the end of the two weeks, the students are given a test to see how well they learned the math. The results are collected an analyzed.

This is an experimental study, where the person conducting the study randomly chooses some of the subjects and exposes them to some kind of treatment.

## 11.2 Inferential Statistics

**Key Ideas**

If much of the data in a data set is unknown, it is still possible to estimate different information about it by using a smaller collection of the total set known as a *sample*. Statistics about the sample can provide valuable information about the complete data set.

A complete data set can contain thousands, if not millions, of data points. The complete set of data is known as the *population*. A much smaller subset of the population is known as a *sample*.

Two types of measures about the population that can be estimated with samples are the *population proportion* and the *population mean*. The population proportion is the percent of the total number of things in the population that has a certain characteristic. The population mean is the average of some characteristic of all the things in the population. The percent of some sample of things that has a certain characteristic is called the *sample proportion*. The *sample* *mean* is the average value of some characteristic of all things in that sample.

To estimate the population proportion or the population mean, you generally find a sample proportion or sample mean and then add and subtract the *standard error*.

For example, if the sample mean is 20 and the standard error is 7, the population mean will be between and . If the sample proportion is 0.40 and the standard error is0.15, the population mean estimate will be between   
 and .

Calculating the standard error depends on the type of question. Four different types are presented below.

**Using One Sample Set to Approximate the Population Proportion**

In an auditorium with 800 students, a certain percent of them will be boys and a certain percent will be girls. If 450 of those students are boys, then the percent of total students who are boys is .

This percent is the population proportion, represented by the symbol , of boys for the entire population of students.

If, for whatever reason, we are only able to sample a subset of 75 students, we can calculate the sample proportion for that sample. If 45 out of those 75 students were boys, the sample population for the percent of boys, represented by the symbol , would be .

For this type of problem, the standard error, SE, can be calculated with the following formula.

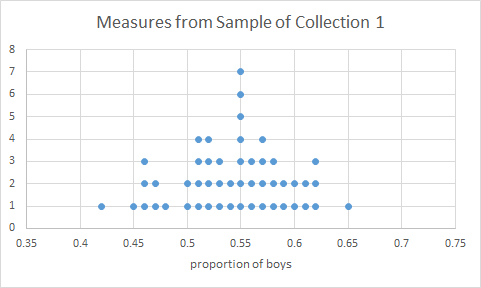
The sample proportion is different from the population proportion in this example. However, there is a formula that allows you to find a range of likely values for the population proportion based on this one sample of 75 students.

The population proportion is, almost surely, between and   
.

This is a pretty large range of values. However, the actual value of the population proportion, 56.25% is between 49% and 71% as the formula predicted.

**Using Multiple Sample Sets to Approximate the Population Proportion**

For a more accurate estimate of the population proportion, more than one sample set is needed. In the auditorium example, one sample set of 75 students with a sample mean of 0.60 was used. To get a better approximation, repeat this process 50 times. The first sample of 75 had 45 boys, which was 60%. The second sample of 75 had 35 boys, which as approximately 47%. A computer is used to do this process 50 times. When the percent of boys for all 50 sets of 75 students are plotted on a dot plot, it looks like this.



This plot is called the *sampling distribution of the sample proportion*.

Two statistics about this plot are also provided. The mean value for the numbers in this plot is approximately 0.54, and the standard deviation for the the numbers in this plot is 0.05.

These two numbers can now be used to approximate the population proportion of the complete data set.

For this type of problem, the standard error, SE can be calculated with the following formula.

The estimated population proportion is then .

For this example, the actual value of the population proportion is, almost surely, between and . The actual proportion is 56.25%, which is, indeed, between 44% and 64%.

On the Regents, the information will be provided so you can do the analysis. Creating the dot plot and finding the mean and the standard deviation of the numbers in the dot plot is something that requires a computer.

Making this plot required numbers. Even though this is a lot of numbers for calculating the percent of boys out of just 800 people, this method of using a computer to do this analysis is very powerful when the population is either extremely large or when, for various reasons, it is not possible to get access to all the data.

**Using One Sample Set to Approximate the Population Mean**

At a high school graduation party there are 600 people, including parents and students. There are more parents than students at the party, and a dot plot of the ages of the party goers looks like this.

The approximate age of the 600 people at the party an be approximated by first taking a random sample of 30 people at the party. Here is one random sampling of 30 people from the total population of 600 people.

51, 45, 42, 13, 40, 20, 45, 49, 41, 15, 18, 45, 41, 14, 35, 41, 17, 40, 42, 45, 51, 18, 31, 11, 19, 47, 20, 14, 17, 47

To use these 30 numbers to approximate the mean age of the 600 people, first find the mean and the standard deviation of these 30 values. For this sample. For this sample set, the mean value, represented by , is approximately 32.4. The standard deviation of the sample, represented by is approximately 13.8.

For this type of problem the standard error, SE, can be calculated with the following formula.

The problem type for  is calculating the standard error of the mean (SEM), which is a measure of how much the sample mean varies from the population mean. The formula is used when you have a sample standard deviation (s) and the sample size (n), and you want to estimate how much the sample means will vary if you were to take many samples.

### Check Your Understanding of Section 11.2

1. *Multiple-Choice*
2. An ecologist wants to know what percent of the 10,000 fish in a lake are cod. She conducts a study in which 50 fish are randomly caught. Of those 50 fish, 15 are cod. Based on this data, using a 95% confidence interval, the research can determine what percent of fish that are cod in the lake is which of the following?

**(1) Between 17% and 43%**

1. Out of 30,000 households in a town, 100 are selected at random and asked if they have a landline or not. 70 out of 100 households replied they do have a landline. Based on this data, using a 95% confidence interval, you can determine that the percent of households in the town that have a landline is which of the following?

**(4) Between 61% and 79%**

1. A sports physician conducts and observational study to learn the average amount of time that 3,000 swimmers in the town can hold their breath underwater. 60 swimmers are chosen at random and tested. The average amount of time those 60 swimmers can hold their breath is 70 seconds with a standard deviation of 6. Based on this data, using a 95% confidence interval, the researcher can determine that for the entire 3,000 swimmers, the average amount of time they can hold their breath underwater is which of the following?

– Sample size

– Sample mean

– Sample standard deviation

**(1) Between 71.5 seconds and 68.5 seconds**

1. A concert promoter wants to estimate the average age of the 20,000 people attending a Taylor Swift concert. At the concert, 80 people are selected at random. The average age of the 80 people is 24.3 with a standard deviation of 11.9. Based on this data, using a 95% confidence interval, the researcher can determine that the average age for all 20,000 people at the concert is which of the following?

– Sample size

– Sample mean

– Sample standard deviation

**(2) Between 21.64 and 26.96**

1. Use the same population as from question 1 except take 500 samples of size 50. The average for all these samplings is 0.24 with a standard deviation of 0.06.

This is a histogram of the sampling distribution of the sample proportion.

– Sample size

– Sample mean

– Sample standard deviation

**(1) Between 0.12 and 0.36**

1. Using the sample population as in question 2 but take 300 samplings of size 100. The average of all the means is 0.73, and the standard deviation is 0.05.

This is a histogram of the sampling distribution of the sample proportion.

Using this data, with a 95% confidence interval, we can determine the percentage of homes in the town that have a land line in which of the following?

– Sample size

– Sample mean

– Sample standard deviation

**(1) Between 0.63 and 0.83**

1. The sports physician uses the same population as in question 3 but uses 150 samplings of 60 people. The average of the means of all the samplings is 72.7, and the standard deviation is 0.92.

This is a histogram of the sampling distribution of the sample mean.

Based on this data, with a 95% confidence intervale, the researchers can determine that the actual average amount of time the entire population can hold its breath under water is which of the following?

– Sample size

– Sample mean

– Sample standard deviation

**(4) Between 70.86 and 74.54**

1. The concert promoter uses the same population as in question 4 but takes 150 samplings of 80 people. The average of the means of all the samplings is 25.5, and the standard deviation is 1.5.

This is the histogram of the sampling distribution of the sample mean.

Based on this data, with 95% confidence interval, the researchers can determine that the average age of the entire 20,000 person population is which of the following?

– Sample size

– Sample mean

– Sample standard deviation

Range 22.5 to 28.5

**(2) Between 22.5 and 28.5**

1. Amy Hogan rolls a six-sided die 12 times and lands on 6 on four of the rolls. She wants to examine if this might suggest the die is defective, since on average, the number of expected 6s in 12 rolls is 2. She uses a computer to simulate 500 times of rolling a fair die 12 times. A histogram of the results is shown below.

Based on this histogram, Amy concludes which of the following?

The probability of rolling 4 sixes out of 12 rolls is about:

The probability of rolling 2 sixes out of 12 rolls is about: .

**(3) No, the die is not necessarily defective because two 6s is not very common.**

1. A new plant food is developed that the inventors claim helps plants grow faster. They cite an experiment they did where out of 40 randomly chosen plants, 20 of them were randomly chosen to receive the new plant food while the other 20 plants received a different kind of plant food.

The 20 plants that received the new plant food grew an average of 2 inches more than the plants that received a different plant food. To test if this difference is statistically significant, the data for the 40 plants are entered into a computer that randomly separated them into two groups of 20. This is done 50 times. The results are show in the histogram below, where the number is the difference between the first group and the second group of the randomly chosen groupings.

Based on this data, can you conclude that the new plant food causes plants to grow faster?

For the value 2, the number of occurrences is .

**(2) No, the difference is not statistically significant. +2 is fairly common.**