5 Project Connection

Analyzing Your Data

Statistical tools can help you analyze and interpret the data you collect. You need to think carefully about which statistical tool to use and when, because other people will be scrutinizing your data. A summary of relevant tools is given below.

Measures of Central Tendency

Selecting which measure of central tendency (mean, median, or mode) to use depends on the distribution of your data. As the researcher, you must decide which measure most accurately describes the tendencies of the population. Consider the following criteria when you are deciding which measure of central tendency best describes your set of data.

- Outliers affect the mean the most. If the data includes outliers, use the median to avoid misrepresenting the data. If you want to use the mean, remove the outliers before calculating the mean.
- If the distribution of the data is not symmetrical, but instead strongly skewed, the median may best represent the set of data.
- If the distribution of the data is roughly symmetrical, the mean and median will be close, so either may be appropriate to use.
- If the data is not numeric (for example, colour), or if the frequency of the data is more important than the values, use the mode.

Measures of Dispersion

Both the range and the standard deviation give you information about the distribution of the data in a set.

The range of a set of data changes considerably because of outliers. The disadvantage of using range is that it does not show where most of the data in a set lies—it only shows the spread between the highest and lowest values. The range is an informative tool that can be used to supplement other measures, such as standard deviation, but it is rarely used as the only measure of dispersion.

Standard deviation is the measure of dispersion that is most commonly used in statistical analysis when the mean is used to calculate central tendency. It measures the spread relative to the mean for most of the data in the set.

Outliers can affect standard deviation significantly. Standard deviation is a very useful measure of spread for symmetrical distributions with no outliers.

Standard deviation helps with comparing the spread of two sets of data that have approximately the same mean. The set of data with the smaller standard deviation has a narrower spread of measurements around the mean, and therefore usually has comparatively fewer high or low values.

Normal Distribution and *Z*-Scores

When working with several sets of data that approximate normal distributions, you can use *z*-scores to compare the data values. A *z*-score table enables you to find the area under a normal distribution curve with a mean of zero and a standard deviation of one. To determine the *z*-score for any data value in a set that is normally distributed, you can use the formula

$$z = \frac{x - \bar{x}}{\sigma}$$

where x is any observed data value, $\bar{x}$ is the mean of the set, and σ is the standard deviation of the set.

Margin of Error and Confidence Level

When analyzing the results of a survey, you may need to interpret and explain the significance of some additional statistics. Most surveys and polls draw their conclusions from a sample of a larger group. The margin of error and the confidence level indicate how well the sample represents the larger group. For example, a survey may have a margin of error of plus or minus 3% at a 95% level of confidence. This means that if the survey were conducted 100 times, the data would be within three percent points above or below the reported results in 95 of the 100 surveys.

The size of the sample that is used for a poll affects the margin of error. If you are collecting data, consider the size of the sample you need for a desired margin of error.

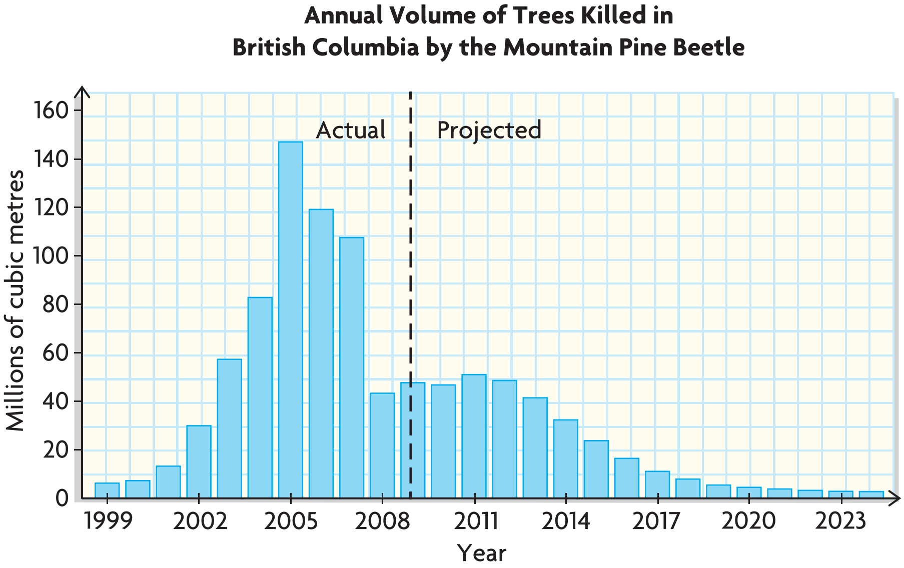
Project Example — Choosing tools to analyze your data

Walther chose as his topic pests and invasive species that cause the greatest damage to natural resources in the Western provinces and territories, and the possible preventive measures that can be taken to control these threats. In his analysis, Walther describes how he determined which statistical tools he might use.

Walther's Analysis

I identified the mountain pine beetle, native to the forests of British Columbia and Alberta but spreading, as one of the most significant pests in these provinces. The European green crab, native to western Europe and northern Africa, is a significant invasive species that threatens the native shellfish that live along the coast of British Columbia. There are many others, including species of plants.

I was able to find data that support my findings. For example, I found the number of hectares that have been damaged in B.C. by the mountain pine beetle for the period 1975 to 2008 at the National Forestry Database, a federal government website. This data is used to make projections about future damage. A graph that I found in an online article, and redrew, shows the actual and projected damage caused by the mountain pine beetle infestation.



Data: B.C. Forest Service

I could use a measure of central tendency to represent the average number of hectares damaged or wood damaged over this period. I am not interested in frequency, so the mode is not appropriate. I think the mean would be the best measure to use in this situation. I could also look at the spread in the number of hectares damaged in B.C. over this period using range and standard deviation. The data is not normally distributed, so I do not need to use *z*-scores. This is secondary data and was not collected by a sampling process, so no margin of error or confidence level needs to be reported.

Your Turn

A. Which statistical tools are appropriate for your data? Explain why.

B. Use the tools you selected, and calculate the statistics.

C. Use these statistics to analyze your data.