Introduction to Statistics

Introduction

Statistics are used everywhere in agriculture. For statistics to be useful you need to know how they were collected and summarized. Often a statistic by itself is meaningless or subject to misinterpretation. When you see a table of numbers, an average or a percentage stop and consider what the numbers really mean or you could be misled. When correctly calculated and interpreted, statistics allow us to make comparisons and to make valid decisions—and progress.

Statistics are important to scientists for three reasons:

1. Allow data to be quantitatively described and summarized
2. Allow generalized conclusions to be drawn based on relatively small sets of data
3. Differences and relationships between sets of data can be objectively analyzed

Key Terms

Mean
Median
Mode
Range
Standard Error
Standard Deviation

Learning Objectives

- ✓ To apply simple statistical measurements including the mean, median, mode, and standard deviation
- ✓ To make predictions or decisions based on data
- ✓ To select a representative sample and tell how near it approximates the whole
- ✓ To collect organize and display data
- ✓ To learn the basic concept of probability

Materials

Ruler (with centimeter and millimeter divisions)
Scales
Calculator
Computer and Excel program
Popping corn (yellow and white)
Hotplate
Stopwatch (or watch that can indicate seconds)
Aluminum foil
1000 ml beaker (1-quart Mason jar)

Graph paper,
Cooking oil
Popcorn seasonings (optional)
Container for microwave popping (optional)
Assortment whole peanuts

Background Information

Mean, Median and Mode

Living things are, by their nature, variable; a single individual, population, community etc. will not be exactly the same as any other. In order to describe any group of living things, statistics, descriptive measures derived from sample data, must be computed. One of the most common descriptive statistics is the mean, or average. If X represents a datum (for example the birth weight of calves in pounds) the mean of a sample of several calves is where(X bar) is the symbol of the mean,(sum of X's) is the total of all the calves' birth weights in the sample and (sample size) is the number of plants sampled.

For example, birth weights of five calves were 90, 85, 70, 90, and 100 pounds.

Thus the average birth weight of the calves in this sample was 87 pounds. If these 5 calves were randomly selected from a larger group of calves, we may assume that the average for the larger population is also approximated by 87 pounds.

The median and the mode give us more information about the sample. The median is the number that is exactly half way between the top and the bottom sample value. In the example, 85 is the median. A number in the sample may or may not be equal to the median.

The mode of a sample is the most frequently occurring number. In the example, 90 is the mode. A sample may or may not have a mode.

As a general rule, the mean should not be considered more accurate than 1 significant figure beyond the accuracy of the original data. In the above example each calf was weighed to the nearest pound. Therefore, the mean may be rounded to the nearest 1/10 (0.1).

	A	B	C
1		1	10
2		6	11
3		11	11
4		16	11
5		21	12
6			
7	**Sum**	**55**	**55**
8	**Mean**	**11**	**11**
9	**Median**	**11**	**11**
10	**Mode**		**11**
11	**S.D.**	**7.9**	**0.7**
12	**S.E.**	**4.1**	**4.1**

Standard Deviation and Standard Error

Calculating a mean only gives a partial description of the data - the average value. It is usually necessary to also describe how much variability there is around the mean. The following two sets of data have the same mean: 1, 6, 11, 16, 21, and 10, 11, 11, 11, 12. You will probably agree that the mean is not enough to meaningfully describe both sets (see Figure 1). Some measure of variability is also necessary. Two measures of variability are commonly used, standard deviation (SD) and standard error (SE).

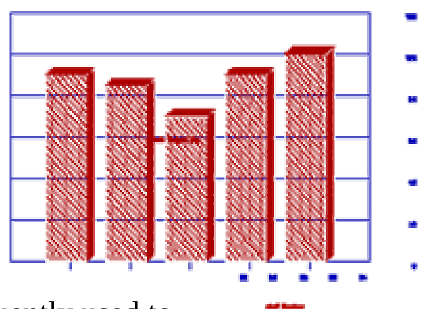
The magnitude of this variation is inversely proportional to the sample size. That is, the larger the sample size, the more precise the estimate of the population mean. Large samples generally give better results than small samples! Electronic spreadsheets such as Excel will calculate the SD and SE. In Excel the syntax is STDEV(number1,number2, etc, or range such as B1: B10) for calculating the SD of a sample.

Comparing Two Means

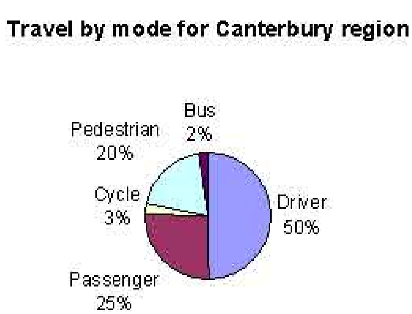
Frequently two means of two samples need to be compared to draw conclusions about similarities or differences. For instance, are the results of a particular experimental treatment significantly different from the control? In some cases the difference may be very large and obvious, but in other cases the means and variances may be quite similar and an objective method is required to determine the degree of difference or similarity.

Student's t-test is commonly used to compare two means where the null hypothesis (HO):1 = 2, is that the means are the same. Electronic spreadsheets such as Excel will calculate t-tests.

Graphs and Charts



Another way to summarize data is to create graphs or charts from the raw data or from the summarized data. For example, the birth weights of calves could be put into a bar graph (right).



Depending on the data, bar graphs, pie graphs (left) or line graphs are frequently used to describe data. Electronic spreadsheets such as Excel are able to produce all types of graphs. Almost everything we ready has some type of data summarized or described to help us reach conclusions and make decisions. We see political polls or opinion polls. We see graphs to help us recognize trends in commodity prices or interest.

Procedure

In this lab, you will be working with data that you collect and data that has been collected. To complete this lab, work in teams of two. Be certain that both of your names are on the lab report and any data sheets that you turn in.

How Popcorn Pops

A Bit of History

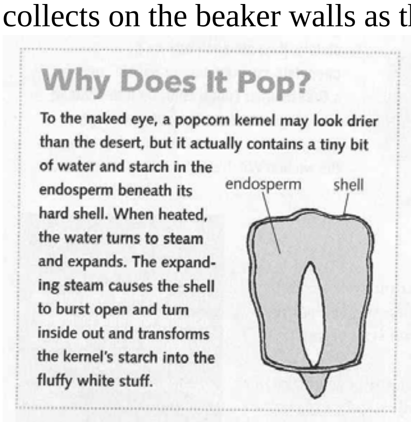
Corn or maize is native to the American continent. Paleobotanical and archeological findings indicate that wild corn not much different in botanical properties with modern corn existed in Southern Mexico about 5000 years ago. Corn was a staple grain for Native Americans for many centuries prior to arrival of European explorers.

Popcorn has been known and used by humans for thousands of years. It probably originated in Mexico as evidenced by multiple thousands- year-old pollen found 200 feet below Mexico City, which is indistinguishable from modern corn pollen. At about 5600 years old, small ears of popcorn found in Bat Cave New Mexico in 1948 and 1950, represent the oldest popcorn ears discovered. Grains of popcorn that still pop have been found in tombs on the east coast of Peru. These grains are estimated to be about 1000 years old. A maize god is represented on a funeral urn found in Mexico dating from about 300 A.D. Ancient popcorn poppers, shallow vessels with a hole on the top and a single handle date from the same time.

By the time European settlers arrived in the Americas, over 700 types of popcorn were in use by virtually all Native American Tribes. Natives tried to sell popcorn to the crew of Christopher Columbus. Cortez was first introduced to popcorn when he invaded Mexico in 1519. The Aztecs used popcorn for food, ornamentation, and ceremonial headdresses. Early French explorers in the 1600's noted that the Iroquois in the Great Lakes region popped popcorn and even made a popcorn soup. English colonists were given a deerskin bag filled with popcorn at the first Thanksgiving Feast at Plymouth Massachusetts. Popcorn snacks were brought to meetings with English colonists as goodwill tokens. Since that time popcorn has become a familiar favorite among children and adults alike.

Characterizing the Pop

The exact details of the popcorn pop are not completely understood. The most likely sequence of events that unfold as the oil surrounding the kernel heats is that the starch-protein matrix inside the kernel absorbs water, expands as it cooks, and with the aid of the tremendous pressure developed as water turns to steam, then explodes the kernel's shell. When popcorn is popped in the classroom in a beaker, condensed water readily collects on the beaker walls as the popcorn pops. The ideal water content of a viable popcorn kernel ranges from 11 to 14 percent.



In this investigation you will be counting the kernels of yellow corn popped in fifteen second intervals and then making a graph of the number of pops versus time. You will also be determining the mean, median, and mode for the number of pops.

The hotplate is turned to a medium heat and the 1000 ml beaker (or 1-quart Mason jar) is placed on it. A small amount of cooking oil is added and 2 test kernels.

An aluminum foil lid is fashioned for the beaker. Monitor the system until at least one of the kernels pops. At this time about 200 to 250 kernels are dumped in the beaker. (A film can canister holds just the right number of kernels, but using 100 counted kernels is an easy way to check calculations.) The foil lid is replaced, the system shaken, and timing begun. Count the number of kernels that pop in fifteen-second time intervals until most kernels pop. Do not try to pop every kernel as this might result in a burnt batch. (Record your data in a table similar to Table 1.) Remember to count the unpopped kernels and record this number. Also, repeat the popping of the yellow corn at least three times and then use the mean values for graphing.

What conclusions can you reach the characteristics of popping corn? When does most of the corn pop? What variables did you control? What variables needed more control? How does your sample compare with others? Include this information in your report.

Repeat the popping experiment, but compare the popping of yellow corn to that of white corn. Run three test of white corn popping and graph the mean. Graph the popping results of both types of corn on the same graph. How do they compare? Calculate the mean, median and mode. Include your raw data, your graphs and the means, medians, and modes in your laboratory report.

Lab Notes:
You might want to make some notes about the accuracy of your observation your methods. Did you hear or see all of the pops?

(**Note**: As a variation you could collect the same data and compare the popping of yellow and white corn in a microwave.)

Summarizing Previously Collected Data

1. Study the data in Table 3 and determine/calculate the high, low, mean, mode, and median for the data.
2. Decide how you will present the data in a chart or graph.
3. Include the summary of Table 3 (mean, mode, and median) in your lab report.
4. Include the chart or graph you develop in your lab report.
5. What conclusions can you reach from the data in Table 3 and from your graph or chart? Include your conclusions in your lab report.

More Collected Data

Working with class members complete Tables 4 and 5.
Decide how to graph the data from Table 4 and put this graph in your lab report.

Table 1. Sample Raw Data—Popcorn Pops

Time	Number of Pops							
	Test 1	Test 2	Test 3	Mean	Test 1	Test 2	Test 3	Mean
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
Mean								
Median								
Mode								
SD								

Table 2. Comparisons with Other Team Results

Team	Mean	Mode	Median	Standard Deviation
Your team				

Table 3. Egg Production by State

State	Millions of Eggs
AL	2,499
AR	3,215
CA	6,663

State	Millions of Eggs
CO	857
CT	917
FL	2,499
GA	4,867
HI	172
IL	841
IN	5,652
IA	5,527
ME	1,434
MD	882
MI	1,327
MN	2,957
MS	1,547
MO	1,719
NE	2,469
NY	931
NC	2,794
OH	6,976
OK	901
OR	783
PA	5,900
SC	1,228
SD	542
TX	4,186
VA	806
WA	1,379
WI	998
DE	112
ID	236
KS	323
KY	710
LA	460
MA	156
MT	88
NH	46
NJ	463
NM	302
ND	49
RI	24
TN	255
UT	483
VT	56
WV	245

STDEV uses the following formula:

$$\sqrt{\frac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}}$$

$$\sqrt{\frac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}}$$

Be glad For Electronic Spreadsheets Like Excel!

State	Millions of Eggs
WY	3.5
Statistics	
High	
Low	
Mean	
Mode	
Median	
Std. Dev	

Biological Variation

Measure the length of the right index finger from the first knuckle to the tip in millimeters and record the information in the following table. This is an example of allometry.

Table 4. Allometry

First Name	Sex	Length of Index Finger (mm)	Approx. Height (1, 2, or 3)*
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
Mean			
Median			
Mode			
Standard Deviation	---		

Lab Notes:
Be certain that you measure finger length the same each time!

* Code 1 for short; 2 for medium; and 3 for tall

Using your data, make a graph that plots the length of the index finger (mm) against the person's height coded as 1, 2, or 3.

Using your data from Table 4, complete Table 5.

Table 5 Comparison of Index Finger Length of Males and Females

Gender	Mean	Median	Mode	S.D.
Males				
Females				

Laboratory Report

After completing this laboratory, write up your laboratory using the following guidelines.
Your Name
Title of Lab:
Date:
Introduction
<State the purpose of the experiment and give some brief background information. Reading the laboratory handouts will provide information for this section.>
Materials and Methods
<Summarize the materials and general procedures used.>
Results and Discussion
<Include your observations, or include the actual data in table, chart or graph obtained in the laboratory with your interpretation of the data and the conclusions reached. Any possible sources of error should be discussed. Also, any questions asked in the lab manual or by the instructor should be answered here.>
Summary and Application
<Briefly summarize what you learned and how it is applied to food systems.>

Glossary

allometry. The relation between the size of an organism and the size of any of its parts. For example, an allometric relation exists between brain size and body size, such that (in this case) animals with bigger bodies tend to have bigger brains. Allometric relations can be studied during the growth of a single organism, between different organisms within a species, or between organisms in different species.

mean, median, mode. Three common ways to measure the center of a set of numerical data. The **mean** is the arithmetic average of the data. The **median** is the middle value of the sorted data set with an odd number of items or the average of the middle two (2) values when the data contains an even number of items. The **mode** is the most common data value, if it exists. Of the three, the mean and median are more useful and frequently used. In any particular application, whether the mean or median is more appropriate depends on the data set and the intended use.

probability. The branch of mathematics dealing with chance. The experimental model is one (1) illustration of probability. Imagine an experiment with outcomes. An event is a collection of outcomes. The probability of an event is the proportion of the experiments that result in an outcome in the event. The probability of an event is always a number between zero (0) and one (1). Events with probabilities near one (1) are very likely to occur, while those with probabilities near zero (0) are very unlikely. For example, to estimate the probability that a randomly selected, adult American female is between sixty (60) inches and sixty-six (66) inches tall, select an adult American female at random and measure her height. If one thousand (1,000) women are selected and measured, the probability would be the proportion of the experiments that selected a woman between those heights.

standard deviation. A measure of the spread of a set of numerical data. If a data set has a relatively large standard deviation, then the data is very spread out. If the standard deviation is small, the data is highly clustered.

statistics. The branch of mathematics dealing with collecting, analyzing, and reasoning from data. The process may involve collecting all of the possible data (a census), or it may involve collecting a subset or sample of the data. The analysis may involve organizing, condensing, calculating summary measurements (statistics), or constructing graphical displays. These descriptive tools help draw conclusions about the real world from which the data originated. When appropriate, probability models provide the framework for attaching a measure of confidence to the conclusions.

Appendix

Websites
http://www.shef.ac.uk/uni/projects/wrp/mjcwhy.html
http://www.robertniles.com/data/
http://www.fedstats.gov/
http://www.ers.usda.gov/statefacts/
http://usda.mannlib.cornell.edu/usda/usda.html
http://stats.bls.gov/
http://www.mlb.ilstu.edu/learn/stat/understanding.htm
http://www.minitab.com/resources/whitepapers/

On Your Own

Take a sample of peanuts. (Don't eat them!) Use the ruler and scales to collect data about your sample. Then use statistics to describe your sample. Include this with your lab report.

Academic Standards

Data Analysis, Probability, and Statistics.

Rationale: With society's expanding use of data for prediction and decision-making, it is important that students develop an understanding of the concepts and processes used in analyzing data.

	Standard - The student will:		**Content Knowledge and Skills**:
01.	Understand data analysis.	a.	Read and interpret tables, charts, and graphs (scatter plots, line graphs, three-dimensional graphs, and pie charts).
02.	Collect, organize, and display data.	a.	Collect and organize data, and display the data in tables, charts, and graphs (scatter diagrams, frequency tables, bar graphs, or pie charts).
03.	Apply simple statistical measurements.	a.	Understand basic statistical concepts including mean (average), median, mode, range, and standard deviation.
04.	Understand basic concepts of probability.	a.	Understand experimental and theoretical probability.
		b.	Distinguish between independent and dependent events.
		c.	Know that probability ranges from 0% to 100%. Understand randomness and chance.
05.	Make predictions or decisions based on data.	a.	Use appropriate technology to employ simulation techniques, curve fitting, correlation, and graphical models to make predictions or decisions based on data.
		b.	Design, conduct, and interpret results of statistical experiments.
		c.	Analyze the effect of biased data on statistical predictions.

Developed by:
Rick Parker, Ph.D.
Agricultural Consumer and Environmental Science (ACES) Department
College of Southern Idaho
rparker@csi.edu
208-732-6402