Introduction to Statistics

Welcome to Introduction to Statistics!

Month One

February 2014

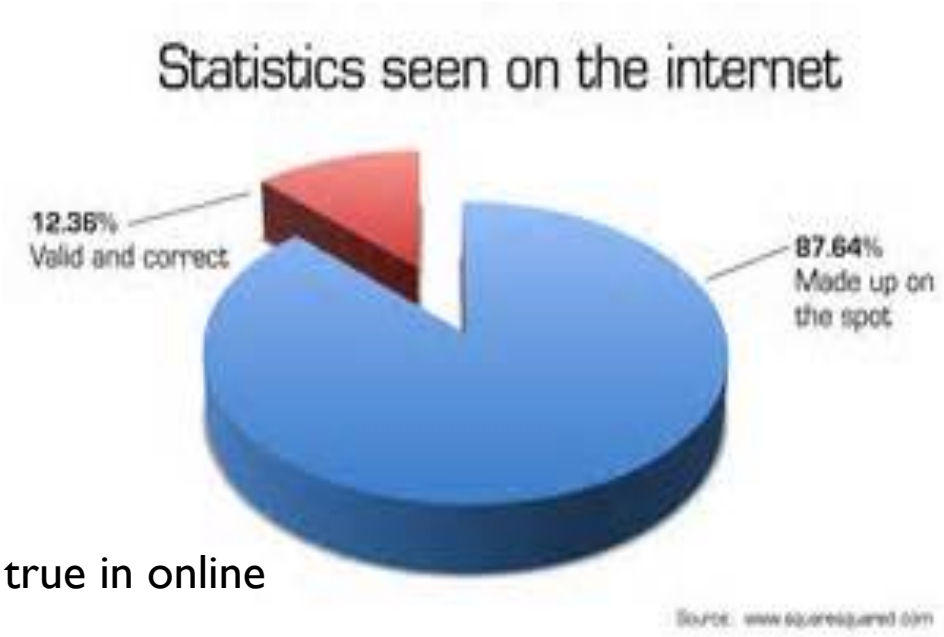
What are Statistics?

- Statistic: a numerical fact
  - A statistic is a quantifiable fact, such as a temperature, a total, an amount of money, etc.
  - The study of statistics is the manipulation of these facts to glean information

What aren't Statistics?

(Excuse the grammar!)

- Not guesses
- Not made up
- Not useless



Note: this is not always true in online forums!

Vocabulary

- Demographic: a vital or statistic about a population, such as age, gender, or racial identity
  - Demographics can be used to describe anything for which statistics are being gathered. For instance, a demographic for a series of temperatures could be the location at which the temperatures were recorded. However this term is usually applied to people.

Vocabulary

- Sample: a subset of the population
  - A sample is used when the population is too large for statistics to be useful or collected.
- Population: a group sharing one or more demographics
  - For instance, all US citizens
- Mean: average
- Mode: number that occurs the most in a range

Vocabulary

- Median: the middle number when a range of numbers are in order
- Probability: chances, odds
- Hypothesis: a proposition assumed as a premise in an argument; what you are trying to disprove
  - This is the educated guess for which we are gathering statistics.

Vocabulary

- Range: difference between largest number and smallest number in a set
- Bell (normal) distribution: theoretical frequency distribution represented by a normal curve
- Outlier: an observation, result, well outside the expected range of values
- Variance (mean square deviation): the square of the standard deviation

Vocabulary

- **Standard Deviation**: a measure of dispersion in a frequency distribution, equal to the square root of the mean of the squares of deviations from the arithmetic mean of the distribution
  - Represented by the Greek letter sigma: σ
  - This isn't as difficult as it sounds!
- **Z-score**: a measure that quantifies the distance a data point is from the mean in a range

Uses of Statistics

Statistics are used in every field, from the census, to lottery, to investments, to science, to computing, to television. In fact, you would be hard-pressed to find a field in which statistics are not used.

Understanding how statistics are determined will help you gauge the reliability of information presented to you.

Basic Formulae

The range in statistics is the difference between the largest and smallest numbers in a data set. For example, ages in a class may be as follows:

16, 15, 16, 14, 17, 16, 15, 17,13

To use them we first put them in numerical order:

13, 14, 15, 15, 16, 16, 16, 17, 17

Then we subtract: 17-13=4

13, 14, 15, 15, 16, 16, 16, 17, 17

Now that we have the numbers in order we can find the median and mode. The **median** is the number in the middle, in this case, 16.

The **mode** is the number that appears the most, which is 16 also.

To find the **mean**, or average, we add the numbers together and divide by the number of datum in the range:

13+14+15+15+16+16+16+17+17=139

139/9=15.4

We use these three figures to provide information about the dataset.

Population versus Sample

The population is the whole group and the sample is a portion that we look at for statistics. A population could be "all the sloths in Costa Rica," while the sample may be the sloths observed during a one-week period.

Statisticians use a sample that represents the entire group. This is because it is usually not realistic or possible to study an entire group.

Where we see the use of sample groups most often are exit polls during presidential elections, television ratings, and sales figures.

The Nielsen company provides television rating information by giving a sample group of Americans boxes that record what they watch on television. Programmers use this information to determine what they will show on TV and how much to charge for advertising during different shows. The more popular a show with the sample group, the more popular the show is assumed to be with the whole population, and the more expensive the advertising.

Samples are often used when studying animals, particularly endangered wild animals. It would be nearly impossible to find all of the animals in the wild, and dangerous to humans and the animals to encounter them. When you see statistics about endangered animals, they are usually based on sample studies.

Probability

We talk about **probability** as "chances," such as "there's a one in 1 million chance of winning the lottery." What does that really mean?

When we talk about chances we are referring to actual, calculable odds. For instance, when you flip a coin there are two possible outcomes: heads or tails.

We can determine how many possible outcomes there are with permutations. **Permutations** are the combinations of outcomes. For example, if I flip two coins I can get a heads or tails of the first coin and the heads or tails of the second coin. This gives us eight possible combinations:

First Coin	Second Coin
Heads	Heads
Tails	Tails
Heads	Tails
Tails	Heads

When each of the items, such as coins, are independent of each other, then we can determine the number of permutations by raising the possibilities to the number of items we have. In this case

If I had 5 coins, I would have $2^5 = 64$ possible outcomes.

But what if the combinations are not independent? What if I am making sandwiches and I can't repeat ingredients?

Let's say the possible ingredients are tomato, cheese, a burger patty, pickles, and mustard. I can have two on each sandwich. We compute it like this: 4+3+2+1=10 possibilities

This shows that the ingredients can't be repeated because after I use one the number of choices are reduced by one.

First Ing.	Second Ing.	Second Ing.	Second Ing.	Second Ing.
Tomato	Cheese	Burger patty	Pickles	Mustard
Cheese	Burger patty	Pickles	Mustard	
Burger patty	Pickles	Mustard		
Pickles	Mustard			
mustard				

Ten possible sandwich combinations, assuming order doesn't matter.

Companies can use information like this to determine possible outcomes. Combined with other information on historical data, they would be able to compute the possibility of different outcomes occurring. But if we don't know what possibilities there are we can't determine how likely they are to occur.

Therefore, I would use the chance of getting a result out of the possible results to determine the probability.

Example: I am flipping three coins. I want to know the probability of getting three heads at the same time.

The probability of getting a heads on a coin flip is 1:2 or 50% or .5. I have a 50% chance on each coin.

$.5^3$ =.5x.5x.5=.125 or 12.5%

(Multiply a decimal by 100 to make it a percent)

The coins are independent – a heads up on one doesn't impact the others.

Example: I need a computer to generate three random numbers between 1 and 10. I need three different numbers, so I program the computer to not repeat the digits.

Because I have three numbers that can't repeat, I add: 9+8+7= 24 permutations. The chances of getting any particular combination is 1:24.

Assignments

Each month of this course will have a presentation and assignments.

The assignments may include worksheets or research for you to do.

Don't forget, you can always look back in your notes or watch the presentation again if you need to!

Assignments Continued

- Week 1: Worksheet. Give the range, mean, median, and mode of series that are provided to you. Yes, you can use a calculator! Can you compute the z-scores? Are there any outliers? How do you know?
- Week 2: Find examples of statistics in your life. Give at least five examples and research where the statistics come from and who compiles them. Hint: we talked about one in the presentation! Are these organizations overseen by any agency? How do you know if you can trust them?

Assignments Continued

- Week 3: Describe a population versus a sample. Define each of these terms and give real life examples. What are some of your demographics? Who uses this information?

- Week 4: Probability experiment. Use Dice, coins, colored titles … and determine the probability of specific results. *Must be more than 1 coin, 1 die, etc…) Record the results.

How many permutations are possible? How many times do you have to repeat the experiment to show pure probability (the results you achieve=the computed probability)?