# Week 5: Exploring Data

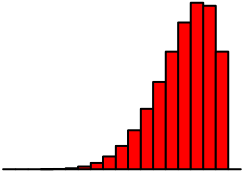
## Session 1

Spring 2020

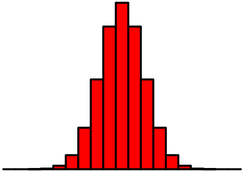
# iClicker quiz: Question 1

Which of the following appears Normally-distributed?

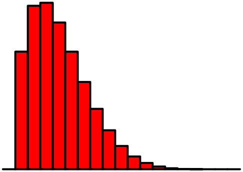
A



B



C



iclicker®

# iClicker quiz: Question 2

What is the **median** value of the following sequence of numbers?

14, 6, 8, 8, 21, 10, 9, 13, 5, 10, 5, 6, 14

A 11

B 9.5

C 8

D 9

E 10

# iClicker quiz: Question 3

What is the **3rd Quartile**, also known as the **75th percentile**, rounded to the nearest integer, of the following sequence of numbers?

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

A 6

B 7

C 9

D 11

E 12

# Announcements

**Some students were not included in salamander description groups.**

**Please verify your group membership and check the Moodle gradebook.**

Bear peer-feedback forms

# Announcements

This is a short week.

Chapter 4 has a lot of important information.

We'll continue chapter 4 materials into next week.

The pre-class exercises for next week will include reading questions from chapters 4 and 5.

# Questions from the salamander exercise:

- What is SVL?
- Why is there a `$` in `mander`?
- How is *central tendency* related to the *spread*?
- How do we define *quartiles*?
- Loading data files into R
- In-class R instruction

# Describing Data

What are two ways to summarize a collection of numbers?

- Central tendency
- Dispersion

# Central Tendencies

**Mean** is the arithmetic average.

- The sum of all the quantities divided by the count of quantities

**Median** is the middle value. There are the same number of observations that are less than and greater than the **median**.

- If there are an *even* number of observations, the median is the average of the 2 central values.

**Mode** is the quantity that appears most frequently.

- The mode is not always well-defined, for example if there are no repeated values, or if there are ties.

# Mean

Mean values are often written as variables with a *bar* symbol. The value of x-bar can be calculated from a collection of quantities **x**.

$$\bar{x} = \frac{\sum x_i}{n}$$

How do we read this equantion?

# Properties of the mean

Mean is a good measure of central tendency when data have a **symmetrical** distribution.

- For example, Normally-distributed data

It's easy to calculate the mean value of a vector in R using the funciton `mean()`.

# Median

Median is a good choice when data are skewed.

Why?

In R it's super easy with the `median()` function.

# Mode

Mode is the most commonly occurring quantity.

Mode can be more difficult to use.

What is the mode of the sequence of numbers?

3, 4, 8, 123, 6, 1239888

There isn't a simple function in R! We can write a custom function... but we probably won't need to calculate modes in this course.

# Dispersion

Dispersion, or spread, quantifies the degree to which the individual values in a set of quantities are different from one another.

There are several ways we could measure dispersion. We've talked about some dispersion measures in class:

- range
- standard deviation (and variance)

Calculating the variance and standard deviation is more complicated than calculating means.

# Variance and standard deviation

Standard deviation is the square root of **Variance.

Let's look at the formula for variance and see if we can figure out why the square root is useful:

$$var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

That looks a lot more complicated than the mean formula!

What are the units of *variance*?

# Variance and standard deviation

If we are measuring height in meters, what are the units of *variance*?

If we measure mass in grams, how can we interpret grams squared?

Variance is difficult to interpret directly. The square root, i.e. standard deviation, is usually more meaningful.

# Properties of variance and standard deviation

Variance cannot be negative.

A verbal definition of sample variance is: "The average squared deviation from the mean value."

The *squared deviation* is a key property. Can the square of a number ever be *negative*?

This definition for variance will make more intuitive sense when we look at Simple Linear Regression.

# Properties of variance and standard deviation

If data are Normally-distributed, the standard deviation has some very useful properties:

- approximately 68% of all observations fall within 1 standard deviation of the mean.
- approximately 95% of all observations fall within 2 standard deviations of the mean.

# Group Activity: Random numbers

Self-select groups of 3 or 4.

Follow the instructions on Moodle.

Submit a single report for the group.

# Quantiles and Quartiles

Sometimes called *percentiles*.

Quantilies are calculated by placing all of the observations in ascending order.

Let

$$x$$

be the value for which we want to calculate a quantile.

The quantile of %%x%% is the count of observations less than %%x%% divided by the total count of observations ($n$).

- Quantiles may be calculated with $n$ or $n$ - *1*.

# Quantiles and quartiles

What is another name for the 50% quantile?

What are quartiles?