Sample Data

Math 530/630

Alison Presmanes Hill

2017-09-26

Today

Today

- What is "sample data"?

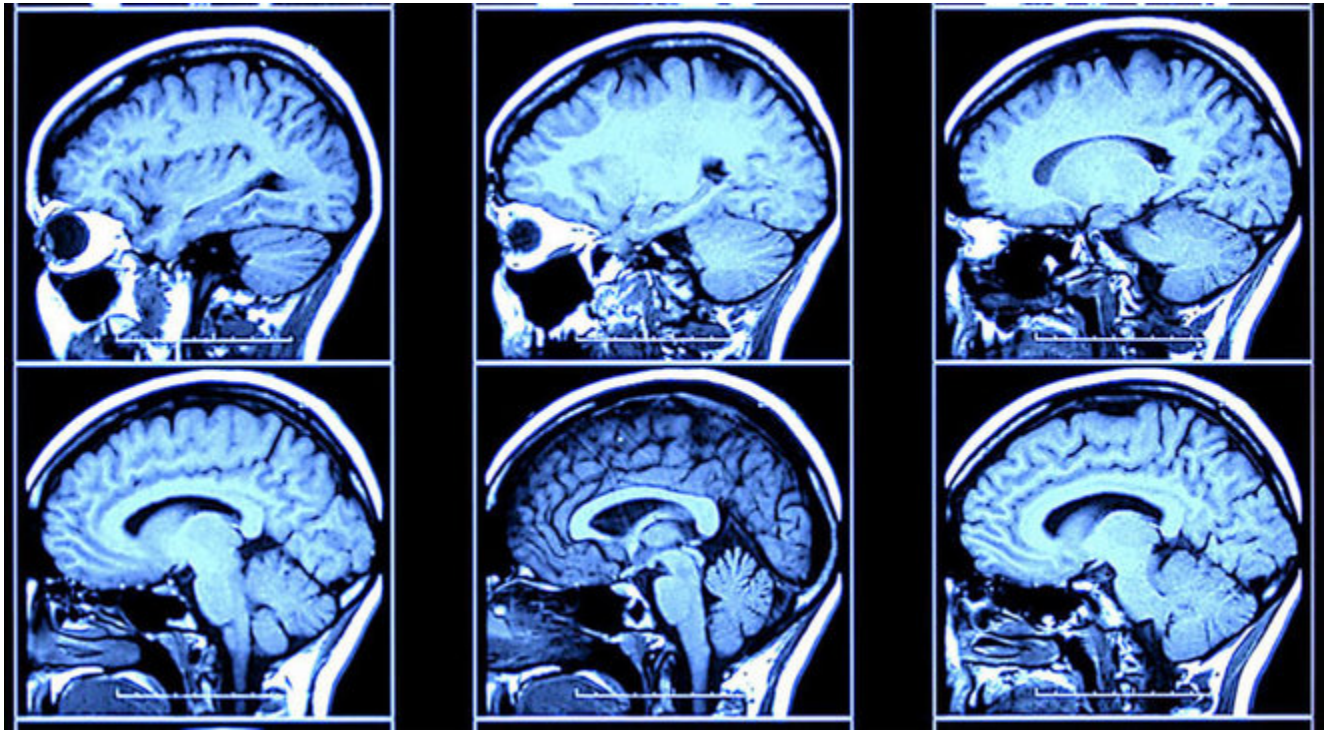
Today

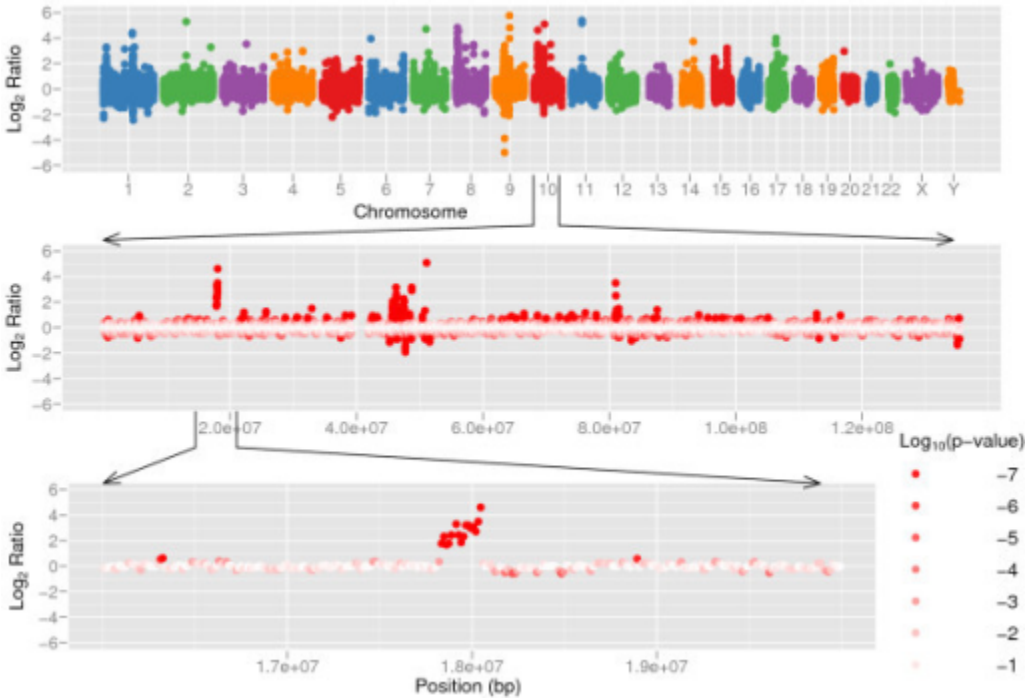
- What is "sample data"?
- What does it look like?

Today

- What is "sample data"?
- What does it look like?
- What properties of data are meaningful, vulnerable, and robust (i.e., invariant) under specific conditions?

What does data look like?




Log2 Ratio
6
4
2
0
−2
−4
−6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
Log2 Ratio
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Log10(p−value)
−7
−6
−5
−4
−3
−2
−1
Log2 Ratio
1.7e+07
1.8e+07
1.9e+07
Position (bp)

```
    "speaker": "CLINTON",
    "lines": [
      "If I could just follow up on that.",
      "There is no disagreement between us on universal coverage for health care, the disagreement is where do we start from and where do we end up.",
      "The Republicans want to repeal the Affordable Care Act, I want to improve it.",
      "I want to build on it, get the costs down, get prescription drug costs down.",
      "Senator Sanders wants us to start all over again.",
      "This was a major achievement of President Obama, of our country.",
      "It is helping people right now.",
      "I am not going to wait and have us plunge back into a contentious national debate that has very little chance of succeeding.",
      "Let's make the Affordable Care Act work for everybody. "
    ]
  },
  {
    "speaker": "SANDERS",
    "lines": [
      "Let me..."
    ]
  },
```

MLU	Sex	CA	Name
1.654	M	1.36345	Ross
1.475	M	1.36345	Ross
1.403	M	1.40178	Ross
1.463	M	1.40178	Ross
1.712	M	1.47964	Ross
1.561	M	1.52464	Ross
1.067	F	1.41667	Laura
1.176	F	1.41667	Laura
1.202	F	1.41667	Laura
1.2	F	1.41667	Laura
1.192	F	1.41667	Laura
1.06	F	1.41667	Laura
1.218	F	1.41667	Laura
1.127	F	1.41667	Laura

What does your data look like?

Take 2 minutes

At some point in the research process, "data" becomes a list of numbers

At some point in the research process, "data" becomes a list of numbers

This list is often very long!

At some point in the research process, "data" becomes a list of numbers

This list is often very long!

So how do you talk about your data with others?

Two ways to talk to others about your data

Two ways to talk to others about your data

Words

Two ways to talk to others about your data

Words

Plots

Talking about data

Categorical vs quantitative

Talking about data

Categorical vs quantitative

We need a common language

Talking about data

Categorical vs quantitative

We need a common language

For quantitative data:

1. Central tendency
2. Spread
3. Shape

A simple sample

Data as a list of numbers

```
x <- c(1, 4, 5, 8, 15)
x
```

```
## [1]  1  4  5  8 15
```

A simple sample

Data as a list of numbers

```
x <- c(1, 4, 5, 8, 15)
x
```

```
## [1]  1  4  5  8 15
```

List-wise operations

List-wise operations

What happens to a list of numbers when we apply the same transformation to every number in the list?

List-wise operations

What happens to a list of numbers when we apply the same transformation to every number in the list?

List-wise operations can be summarized in a simple equation

List-wise operations

What happens to a list of numbers when we apply the same transformation to every number in the list?

List-wise operations can be summarized in a simple equation

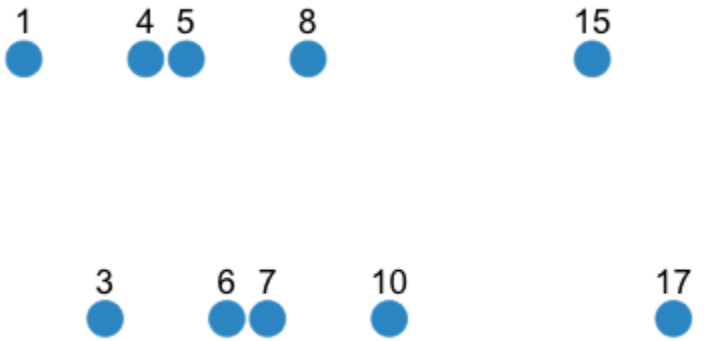
$$y = x + 2$$

"add 2 to every number in list x and put this in a new list called y"

List-wise addition

$$y = x + 2$$

What changed? What didn't?



List-wise subtraction

$$y = x - 2$$

What changes? What stays the same?

Take 2 minutes

List-wise subtraction

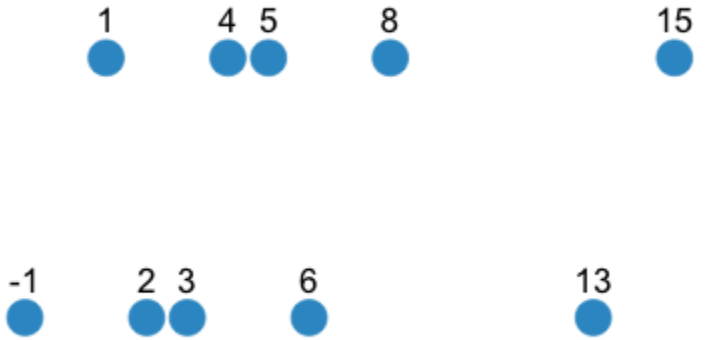
$$y = x - 2$$

What changes? What stays the same?

List-wise subtraction

$$y = x - 2$$

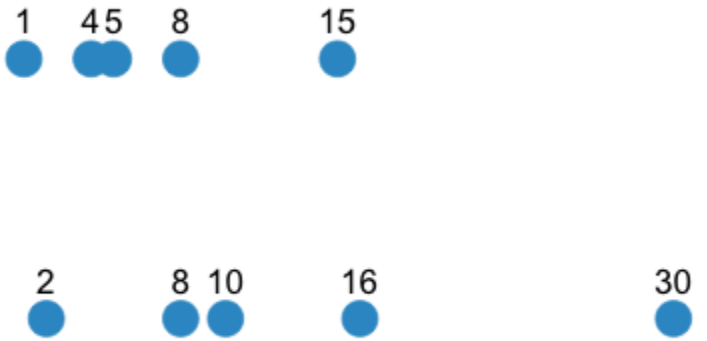
What changes? What stays the same?



List-wise multiplication

$$y = x \times 2$$

What changed? What didn't?



List-wise division

$$y = x \div 2$$

What changes? What stays the same?

Take 2 minutes

List-wise division

$$y = x \div 2$$

What changes? What stays the same?

List-wise division

$$y = x \div 2$$

What changes? What stays the same?

Invariance properties

The question of what does not change during statistical operations refers to invariance properties

Invariance properties

The question of what does not change during statistical operations refers to invariance properties

What stays constant under certain transformations is a key concept is modern statistics, as well as other branches of science

Invariance properties

The question of what does not change during statistical operations refers to invariance properties

What stays constant under certain transformations is a key concept is modern statistics, as well as other branches of science

Einstein actually wanted to call his theory of relativity the “theory of invariants”- what remains invariant in the space-time continuum
(http://www.economist.com/node/3518580)

What did we just show?

List-wise addition or subtraction

- Moves the numbers as a group, as though they were mounted on a rigid stick, and slid to the left or right.
- Does not change any of the distances between numbers.

What did we just show?

List-wise addition or subtraction

- Moves the numbers as a group, as though they were mounted on a rigid stick, and slid to the left or right.
- Does not change any of the distances between numbers.

List-wise multiplication or division

- Can move the numbers as a group
- But also causes them to "fan in" or "fan out."

Limitations of lists

Limitations of lists

Number of rows/columns

Limitations of lists

Number of rows/columns

We started by showing how all the information in a list of n numbers can be re-expressed in terms of n new numbers

$$list_x \longrightarrow list_y$$

Limitations of lists

Number of rows/columns

We started by showing how all the information in a list of n numbers can be re-expressed in terms of n new numbers

$$list_x \longrightarrow list_y$$

These new numbers contain all the information in the original list, and the original list can be perfectly reconstructed from the new list

Descriptive statistics

Measures of location or central tendency

- In general, in what region is the list located on the number line?
- What number is typical of the entire list?
- What number is in the center of the list?

Descriptive statistics

Measures of location or central tendency

- In general, in what region is the list located on the number line?
- What number is typical of the entire list?
- What number is in the center of the list?

Measures of spread or variability

- How far is the list spread out over the number line?

Descriptive statistics

Measures of location or central tendency

- In general, in what region is the list located on the number line?
- What number is typical of the entire list?
- What number is in the center of the list?

Measures of spread or variability

- How far is the list spread out over the number line?

Measures of shape

- Pattern of relative interval sizes, moving from left to right

The center

Mode

Median

Mean

The sample mode

The most frequently occurring value in a list

```
# compute the statistical mode of a vector of numbers
stat.mode <- function(x) {
    freqs <- tapply(x, x, length)
    as.numeric(names(freqs)[which.max(freqs)][1])
}
stat.mode(x)
```

```
## [1] 1
```

You don't need to understand the above R function

The sample mean

The arithmetic mean of the data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_1, x_2, x_3, \ldots, x_n$ represent the n observed values.

The sample mean

The arithmetic mean of the data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_1, x_2, x_3, \ldots, x_n$ represent the n observed values.

Number with the smallest sum of *squared* distances to the list of numbers

The sample mean

The arithmetic mean of the data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_1, x_2, x_3, \ldots, x_n$ represent the n observed values.

Number with the smallest sum of *squared* distances to the list of numbers

Hence, the mean is a *least squares estimator*

The sample mean

The arithmetic mean of the data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_1, x_2, x_3, \ldots, x_n$ represent the n observed values.

Number with the smallest sum of *squared* distances to the list of numbers

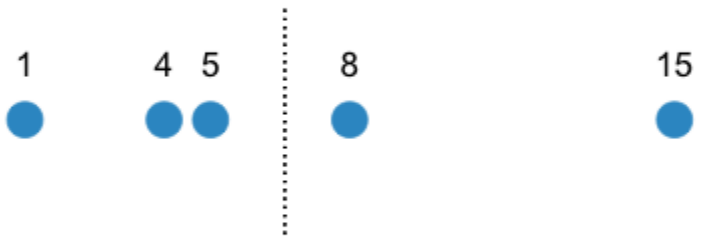
Hence, the mean is a *least squares estimator*

Unless ALL scores are identical, most if not all, scores will be different from the mean.

The sample mean

```
x <- c(1, 4, 5, 8, 15)
mean(x)
```

```
## [1] 6.6
```



What happens to the **center** if we do list-wise addition?

$$y = x + 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 8.6
```

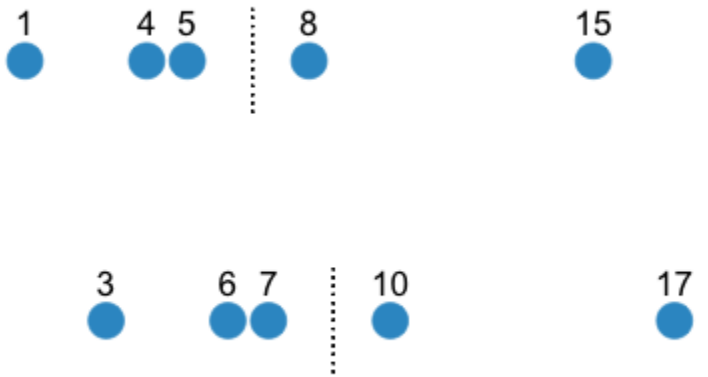
$$y = x + 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 8.6
```



What happens to the **center** if we do list-wise subtraction?

$$y = x - 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 4.6
```

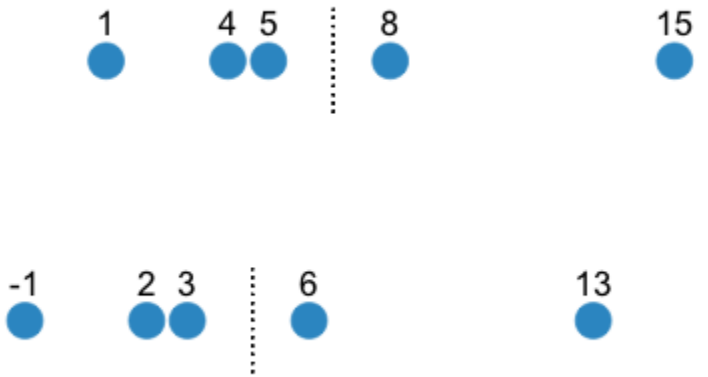
$$y = x - 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 4.6
```



What about the median?

Take 2 minutes

The sample median

- Order the numbers from highest to lowest
- If number of numbers is odd, choose the middle
- If number of numbers is even, choose the average of the 2 middle values

What about the median?

```
y <- x + 2
median(x)
```

```
## [1] 5
```

```
median(y)
```

```
## [1] 7
```

```
y <- x - 2
median(x)
```

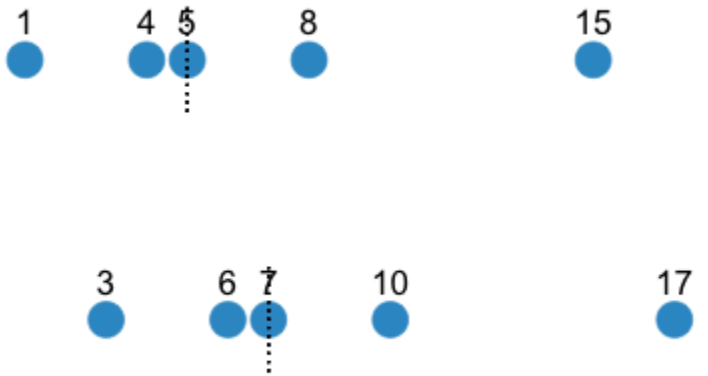
```
## [1] 5
```

```
median(y)
```

```
## [1] 3
```

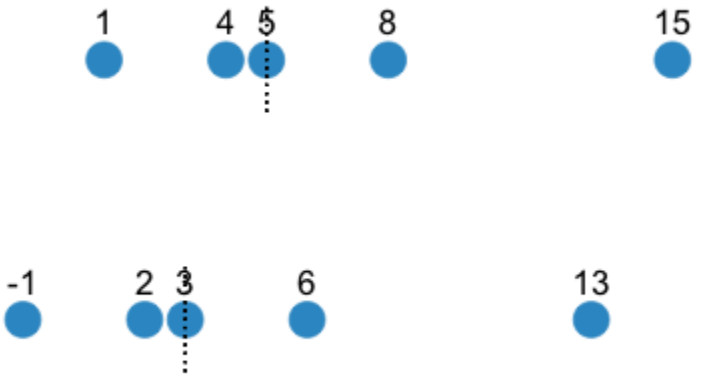
The median after addition

$$y = x + 2$$



The median after subtraction

$$y = x - 2$$



List-wise addition & subtraction come straight through to the center

$$c_y = c_x \pm b$$

What happens to the center if we do list-wise multiplication?

$$y = x \times 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 13.2
```

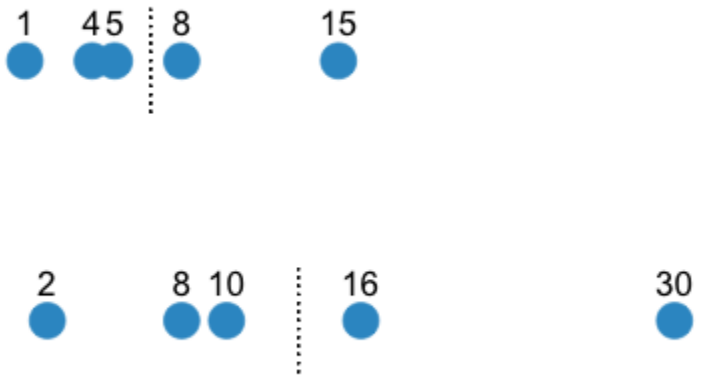
$$y = x \times 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 13.2
```



What happens to the center if we do list-wise division?

$$y = x \div 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 3.3
```

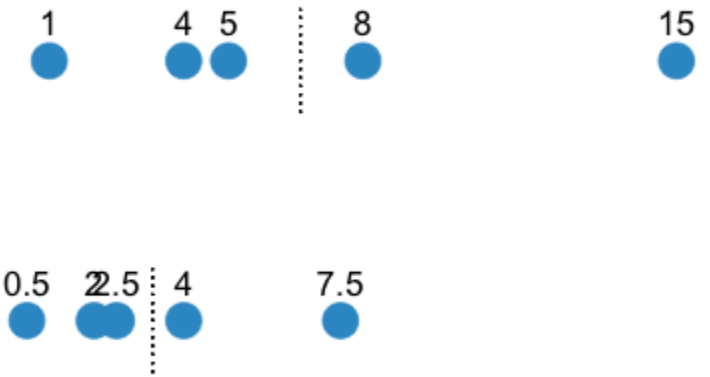
$$y = x \div 2$$

```
mean(x)
```

```
## [1] 6.6
```

```
mean(y)
```

```
## [1] 3.3
```



List-wise multiplication/division comes straight through to the **center**.

$$c_y = a \times c_x$$

Linear transformation theory

Operation	Center	Spread	Shape
$+$	$+$	?	?
$-$	$-$	?	?
$\times$	$\times$	?	?
$\div$	$\div$	?	?

Linear transformation theory

Operation	Center	Spread	Shape
$+$	$+$	?	?
$-$	$-$	?	?
$\times$	$\times$	?	?
$\div$	$\div$	?	?

So the central tendency of any list of numbers is affected by *all* possible linear transformations.

What happens to the **spread**?

Take 2 minutes

$$y = x + 2$$

$$z = (x \times 2) + 2$$

The spread

Standard deviation/variance

Median absolute deviation

Interquartile range

Min-to-max range

Standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
x <- c(1, 4, 5, 8, 15)
sd(x)
```

```
## [1] 5.319774
```

```
y <- x + 2
sd(y)
```

```
## [1] 5.319774
```

```
z <- x*2 + 2
sd(z)
```

```
## [1] 10.63955
```

List-wise addition/subtraction have no effect on **spread**...

List-wise addition/subtraction have no effect on **spread**...

But, list-wise multiplication/division come straight through to the **spread**.

List-wise addition/subtraction have no effect on **spread**...

But, list-wise multiplication/division come straight through to the **spread**.

$$s_y = a \times s_x$$

Linear transformation theory

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	?
$-$	$-$	nothing changes	?
$\times$	$\times$	$\times$	?
$\div$	$\div$	$\div$	?

Linear transformation theory

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	?
$-$	$-$	nothing changes	?
$\times$	$\times$	$\times$	?
$\div$	$\div$	$\div$	?

So you can only affect the spread of a list of numbers by using multiplication/division.

Shape

- Modality [think: mounds]

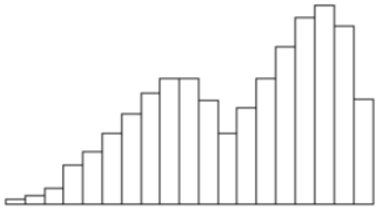
Shape

- Modality [think: mounds]
- Skewness [think: symmetry]

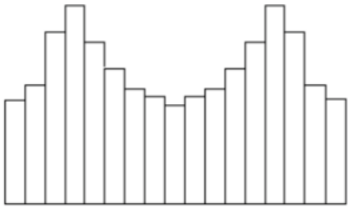
Shape

- Modality [think: mounds]
- Skewness [think: symmetry]
- Kurtosis [think: peakedness]

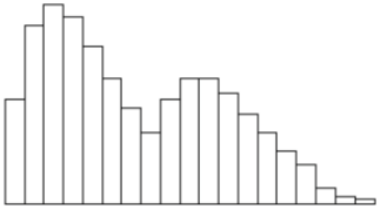
Modality



Two Modes



Bimodal



Bimodal

start here

```
x <- rnorm(100,50,12) #draws 100 samples from normal dist., M=50, SD=12
y <- 2*x + 5 #y is a linear transformation of x
z.x <- (x - mean(x))/sd(x) #makes z-scores using x
z.y <- (y - mean(y))/sd(y) #makes z-scores using y
head(z.x) #show the first six numbers in z.x
```

```
## [1] -1.30832402 -0.46948913  0.05628251 -1.41155574 -0.86059966  1.04955801
```

```
#[1]  0.3053039  0.4164851 -0.3180144 -0.6267830  0.3235671 -1.8508084
head(z.y) #show the first six numbers in z.y
```

```
## [1] -1.30832402 -0.46948913  0.05628251 -1.41155574 -0.86059966  1.04955801
```

```
#[1]  0.3053039  0.4164851 -0.3180144 -0.6267830  0.3235671 -1.8508084
```

What does the above show? What will be the location, spread, and shape of the new numbers in z.x? In z.y?

Skewness

Skewness statistics provide information about departures from symmetry

Skewness

Skewness statistics provide information about departures from symmetry

Draws on all 3 measures of location

Skewness

Skewness statistics provide information about departures from symmetry

Draws on all 3 measures of location

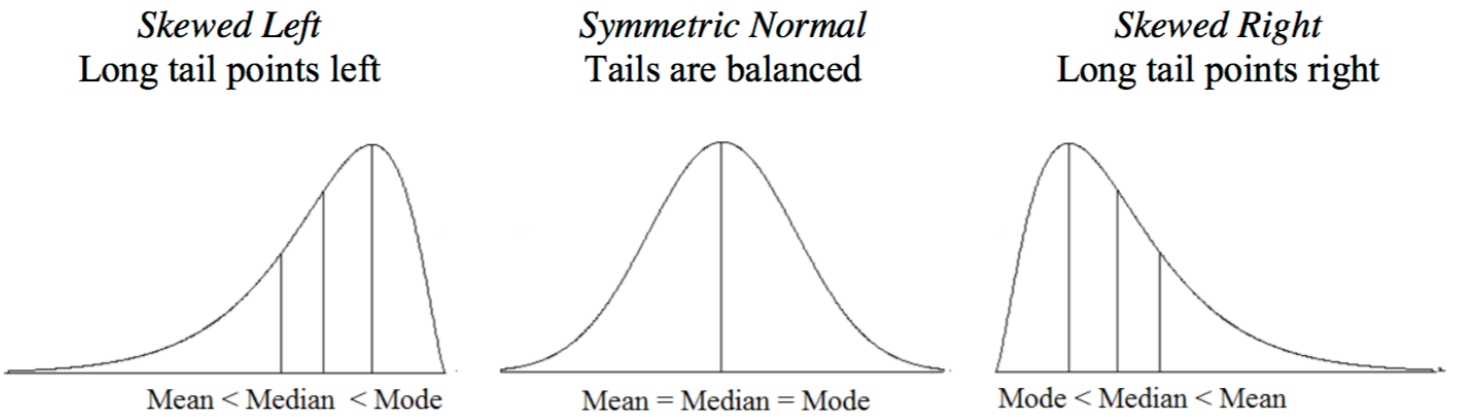
The standard definition of skewness in a population is the average cubed z-score

Skewness

Skewness statistics provide information about departures from symmetry

Draws on all 3 measures of location

The standard definition of skewness in a population is the average cubed z-score



Calculating skewness

Here is one way to calculate skewness, using the formula for the adjusted Fisher-Pearson standardized moment coefficient, G_1:

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{x}^3$$

```
g1_calculate <- function(x = data){
  s3.x <- sum((x-mean(x))^3/sd(x)^3) # right side of eq
  n <- length(x)
  g1 <- (n/((n-1)*(n-2)))*s3.x
  return(g1)
}
```

Calculating skewness

```
# install.packages("moments")
library(moments)
skewness(x)
```

```
## [1] 0.2649852
```

Kurtosis

The basic idea of kurtosis is that it is the average 4th power of the z-scores Kurtosis of a distribution is typically expressed relative to that of a normal distribution.

Kurtosis

The basic idea of kurtosis is that it is the average 4th power of the z-scores Kurtosis of a distribution is typically expressed relative to that of a normal distribution.

The normal distribution has a kurtosis of 3. So, the most common measure of kurtosis is the average 4th power of the z-scores minus 3.

Kurtosis

The basic idea of kurtosis is that it is the average 4th power of the z-scores Kurtosis of a distribution is typically expressed relative to that of a normal distribution.

The normal distribution has a kurtosis of 3. So, the most common measure of kurtosis is the average 4th power of the z-scores minus 3.

This can be computed in a number of ways. The simplest version is biased for a normal distribution, but is reported by some programs.

Kurtosis

The basic idea of kurtosis is that it is the average 4th power of the z-scores Kurtosis of a distribution is typically expressed relative to that of a normal distribution.

The normal distribution has a kurtosis of 3. So, the most common measure of kurtosis is the average 4th power of the z-scores minus 3.

This can be computed in a number of ways. The simplest version is biased for a normal distribution, but is reported by some programs.

If the data have high kurtosis due to long tails, then the sample mean may be an unreliable estimator of the population mean.

Kurtosis

The basic idea of kurtosis is that it is the average 4th power of the z-scores Kurtosis of a distribution is typically expressed relative to that of a normal distribution.

The normal distribution has a kurtosis of 3. So, the most common measure of kurtosis is the average 4th power of the z-scores minus 3.

This can be computed in a number of ways. The simplest version is biased for a normal distribution, but is reported by some programs.

If the data have high kurtosis due to long tails, then the sample mean may be an unreliable estimator of the population mean.

Standard formulas for the variance of the sample correlation and sample variance can be seriously in error if the data come from a population with high kurtosis.

Calculating kurtosis

```
s4.x <- sum((x-mean(x))^4) #using raw
s2.x <- sum((x-mean(x))^2) #using raw
n <- length(x)
n*s4.x/s2.x^2
```

```
## [1] 2.376262
```

```
# [1] 2.438948
s4.z <- sum((z.x-mean(z.x))^4) #using z
s2.z <- sum((z.x-mean(z.x))^2) #using z
n*s4.z/s2.z^2
```

```
## [1] 2.376262
```

```
# [1] 2.438948
```

All list-wise operations have no effect on shape.

Linear transformation theory

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Review: list-wise addition & subtraction

$$y = x \pm b$$

Review: list-wise addition & subtraction

$$y = x \pm b$$

Center

$$c_y = c_x \pm b$$

Review: list-wise addition & subtraction

$$y = x \pm b$$

Center

$$c_y = c_x \pm b$$

Spread

$$s_y = s_x$$

Review: list-wise addition & multiplication

$$y = (a \times x) \pm b$$

Review: list-wise addition & multiplication

$$y = (a \times x) \pm b$$

Center

$$c_y = (a \times c_x) \pm b$$

Review: list-wise addition & multiplication

$$y = (a \times x) \pm b$$

Center

$$c_y = (a \times c_x) \pm b$$

Spread

$$s_y = (a \times s_x)$$

Review: list-wise subtraction & division

$$y = \frac{x - b}{a}$$

Review: list-wise subtraction & division

$$y = \frac{x - b}{a}$$

Center

$$c_y = \frac{c_x - b}{a}$$

Review: list-wise subtraction & division

$$y = \frac{x - b}{a}$$

Center

$$c_y = \frac{c_x - b}{a}$$

Spread

$$s_y = \frac{s_x}{a}$$

Linear Transformations

Finding a and b, given $\bar{y}$ and s_y

Linear Transformations

Finding a and b, given $\bar{y}$ and s_y

How are the mean and sd affected by linear transformation?

$$\bar{y} = a\bar{x} + b$$

$$s_y = as_x$$

Linear Transformations

Finding a and b, given $\bar{y}$ and s_y

How are the mean and sd affected by linear transformation?

$$\bar{y} = a\bar{x} + b$$

$$s_y = as_x$$

There are two unknowns in these equations: a and b

Linear Transformations

Finding a and b, given $\bar{y}$ and s_y

How are the mean and sd affected by linear transformation?

$$\bar{y} = a\bar{x} + b$$

$$s_y = as_x$$

There are two unknowns in these equations: a and b

Take 2 minutes to solve for a and b

Try solving for a in the second equation first.

Linear Transformations

Linear Transformations

$$a = \frac{s_y}{s_x}$$

$$b = \bar{y} - a\bar{x}$$

What if we want z-scores?

So, given $\bar{y} = 0$ and $s_y = 1$, what are a and b?

Take 2 minutes

What if we want z-scores?

So, given $\bar{y} = 0$ and $s_y = 1$, what are a and b?

Take 2 minutes

$$a = \frac{1}{s_x}$$

and

$$b = -a\bar{x}$$

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Change center only?

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Change center only?

$$y = x \pm b$$

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Change center only?

$$y = x \pm b$$

Change center & spread?

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Change center only?

$$y = x \pm b$$

Change center & spread?

$$y = (a \times x) \pm b$$

Goal: linear rescaling

Operation	Center	Spread	Shape
$+$	$+$	nothing changes	nothing changes
$-$	$-$	nothing changes	nothing changes
$\times$	$\times$	$\times$	nothing changes
$\div$	$\div$	$\div$	nothing changes

Change center only?

$$y = x \pm b$$

Change center & spread?

$$y = (a \times x) \pm b$$

Change shape?

- We explored quantitative data using different types of graphic plots

- We explored quantitative data using different types of graphic plots
- We developed some intuition about properties of our data that are meaningful, vulnerable, and robust (i.e., invariant) under specific conditions

- We explored quantitative data using different types of graphic plots
- We developed some intuition about properties of our data that are meaningful, vulnerable, and robust (i.e., invariant) under specific conditions
- We discussed and calculated meaningful descriptive statistics to measure location, spread, and shape

- We explored quantitative data using different types of graphic plots
- We developed some intuition about properties of our data that are meaningful, vulnerable, and robust (i.e., invariant) under specific conditions
- We discussed and calculated meaningful descriptive statistics to measure location, spread, and shape
- We demonstrated how list-wise operations can impact measures of location, spread, and shape

- We explored quantitative data using different types of graphic plots
- We developed some intuition about properties of our data that are meaningful, vulnerable, and robust (i.e., invariant) under specific conditions
- We discussed and calculated meaningful descriptive statistics to measure location, spread, and shape
- We demonstrated how list-wise operations can impact measures of location, spread, and shape
- We have proven that for any list of numbers with non-zero spread, the z-score transformation produces numbers with the same shape (same skewness, same kurtosis) as the original numbers but with location of 0 and spread of 1.