Section 001 CDS 101 *Introduction to Computational and Data Sciences*

Class 20: Statistical distributions I

April 5, 2018

General

Annoucements

- Reading 10: *Representing distributions* will be posted on website this afternoon
- Don't forget to participate in the Question/Answer discussion for each Reading!
  - Answer post count reset for second half of course
  - Review the course syllabus for credit requirements
- Homework 3 will be posted within the next day (will be shorter so it can be completed in one week)

Statistical distributions

Variance

Variance is roughly the average squared deviation from the mean.

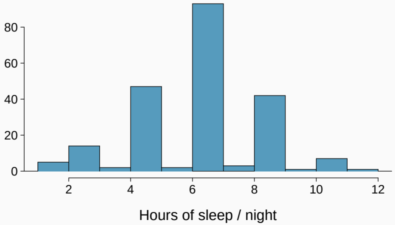
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.

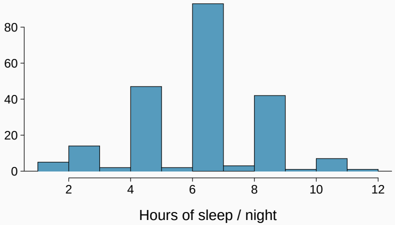


Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \cdots + (7 - 6.71)^2}{217 - 1} = 4.11\ hours^2$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- *To get rid of negatives so that observations equally distant from the mean are weighed equally.*
- *To weigh larger deviations more heavily.*

Standard deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.s

$$s = \sqrt{s^2}$$

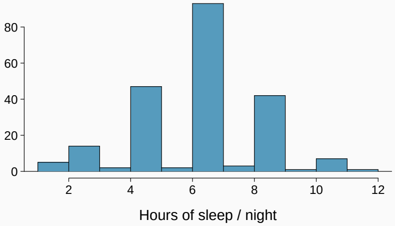
Standard deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.s

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03\ hours$$



Standard deviation

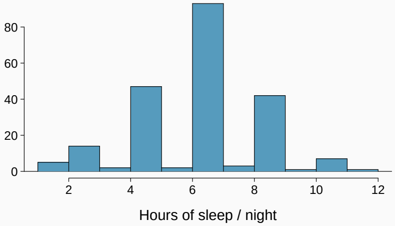
The *standard deviation* is the square root of the variance, and has the same units as the data.s

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \ hours$$

- We can see that all of the data are within 3 standard deviations of the mean.



Median

- The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} *percentile*.

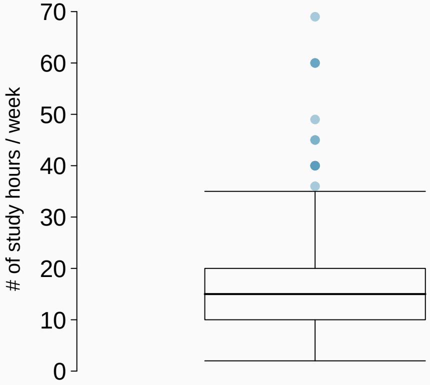
Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

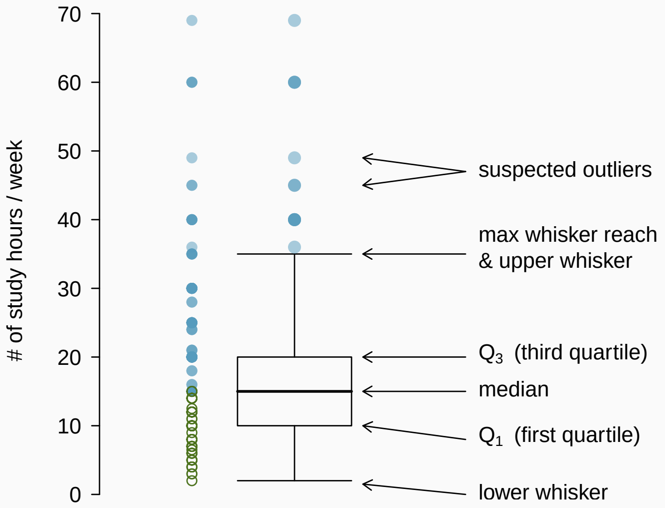
$$IQR = Q3 - Q1$$

Box plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a box plot



Whiskers and outliers

- *Whiskers*
  of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

Whiskers and outliers

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

Whiskers and outliers

- *Whiskers* of a box plot can extend up to $1.5\times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

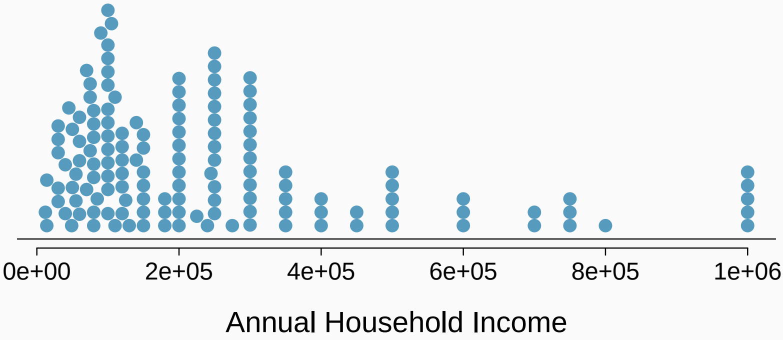
Outliers (cont.)

Why is it important to look for outliers?

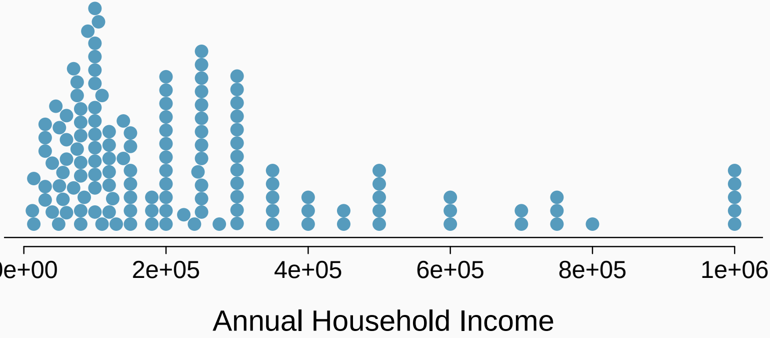
- *Identify extreme skew in the distribution.*
- *Identify data collection and entry errors.*
- *Provide insight into interesting features of the data.*

Extreme observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million? What if the smallest value was replaced with $10 million?



Robust statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to $10 million	190K	200K	309K	853K
move smallest to $10 million	200K	200K	316K	854K

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

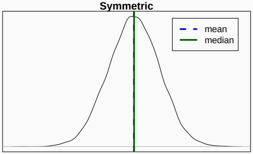
- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

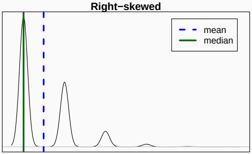
Median

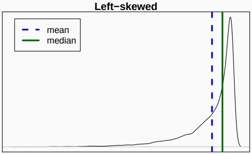
Mean vs. median

- If the distribution is symmetric, center is often defined as the mean: mean $\approx$ median



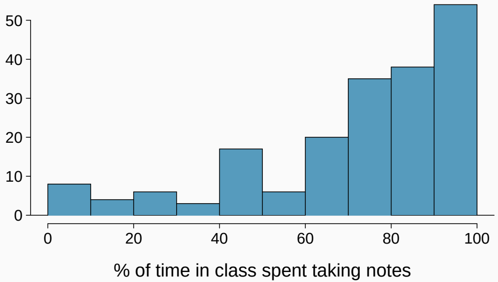
- If the distribution is skewed or has extreme outliers, center is often defined as the median
  - Right-skewed: mean $>$ median
  - Left-skewed: mean $<$ median





Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean$>$ median

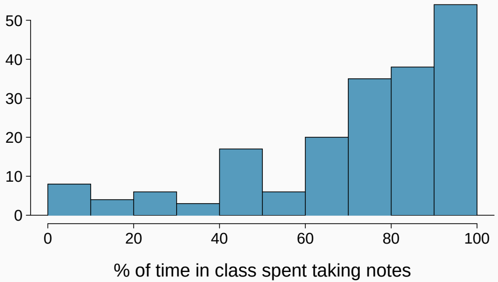
(b) mean $<$ median

(c) mean $\approx$ median

(d) impossible to tell

Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



median: 80%

mean: 76%

(a) mean$>$ median

(b) *mean $<$ median*

(c) mean $\approx$ median

(d) impossible to tell

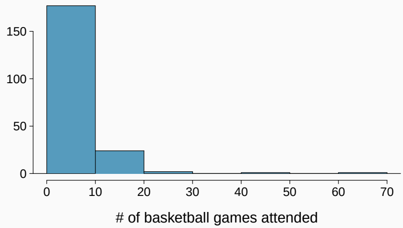
Extremely skewed data

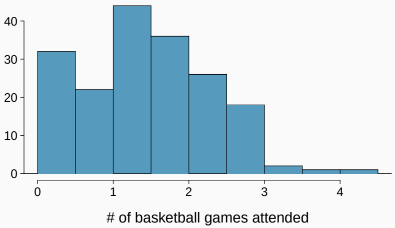
When data are extremely skewed, transforming them might make modeling easier. A common transformation is the *log transformation*.

Extremely skewed data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the *log transformation*.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.





Pros and cons of transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

Pros and cons of transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Pros and cons of transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

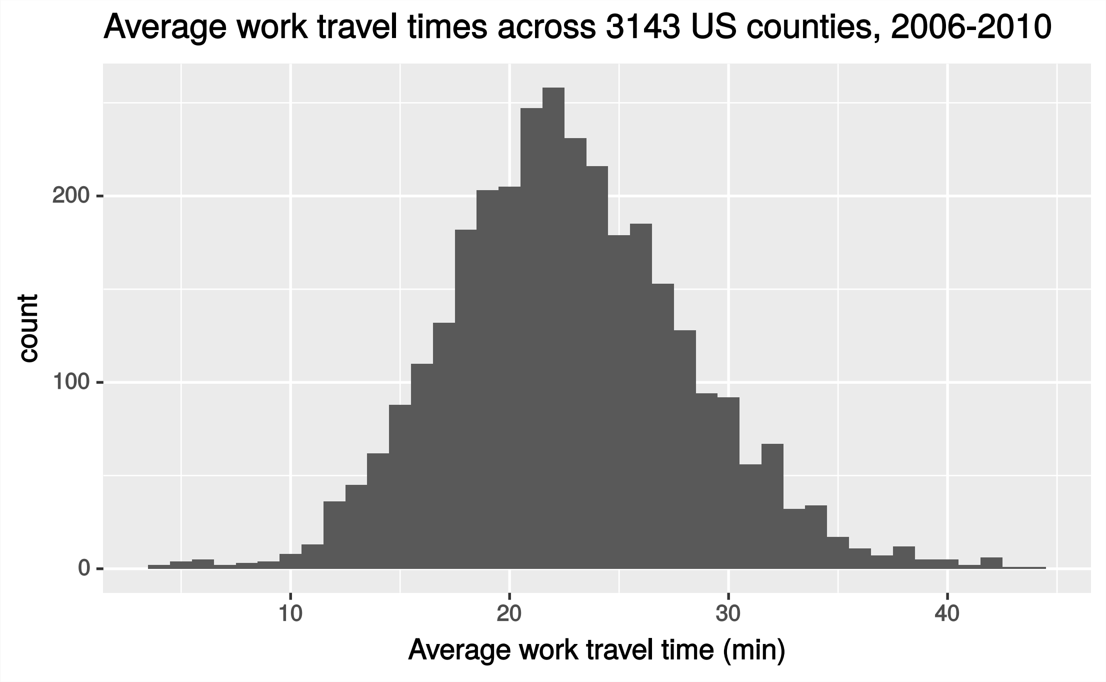
- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.

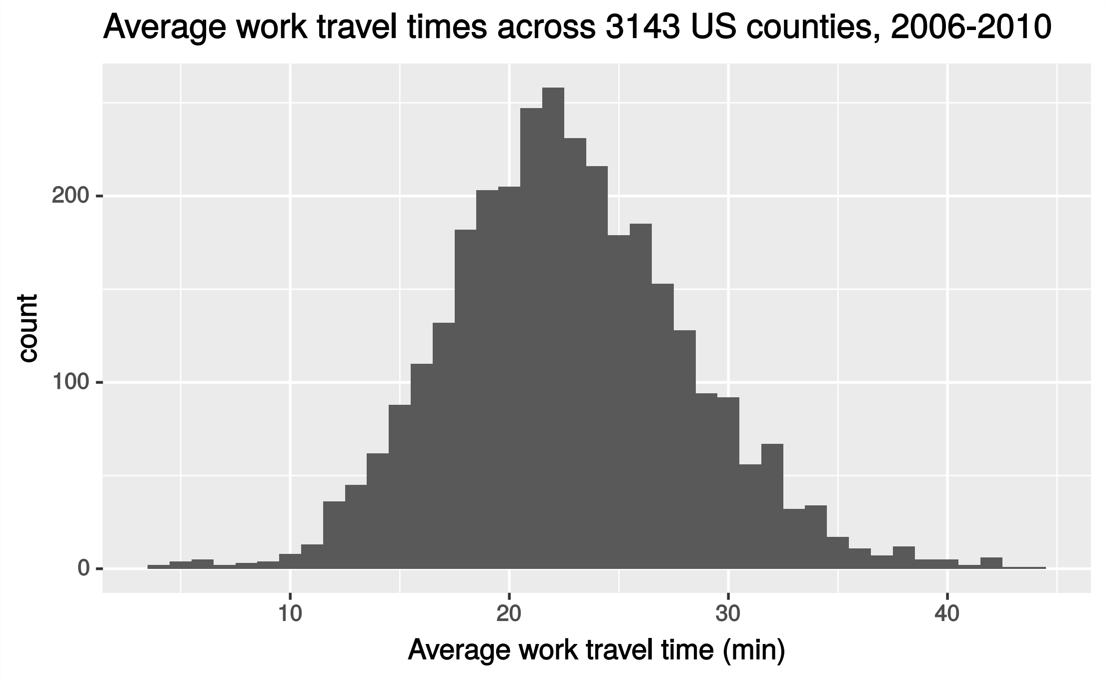
Example data distribution

The following distribution comes from data posted by the US Census Bureau:



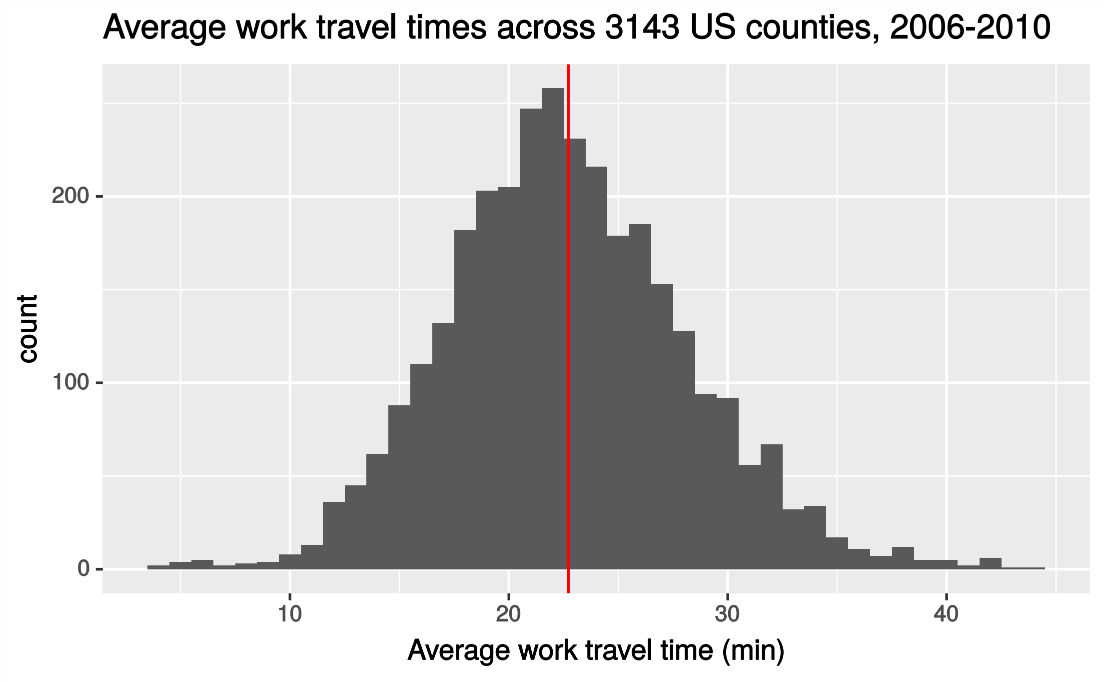
Example data distribution

The following distribution comes from data posted by the US Census Bureau:

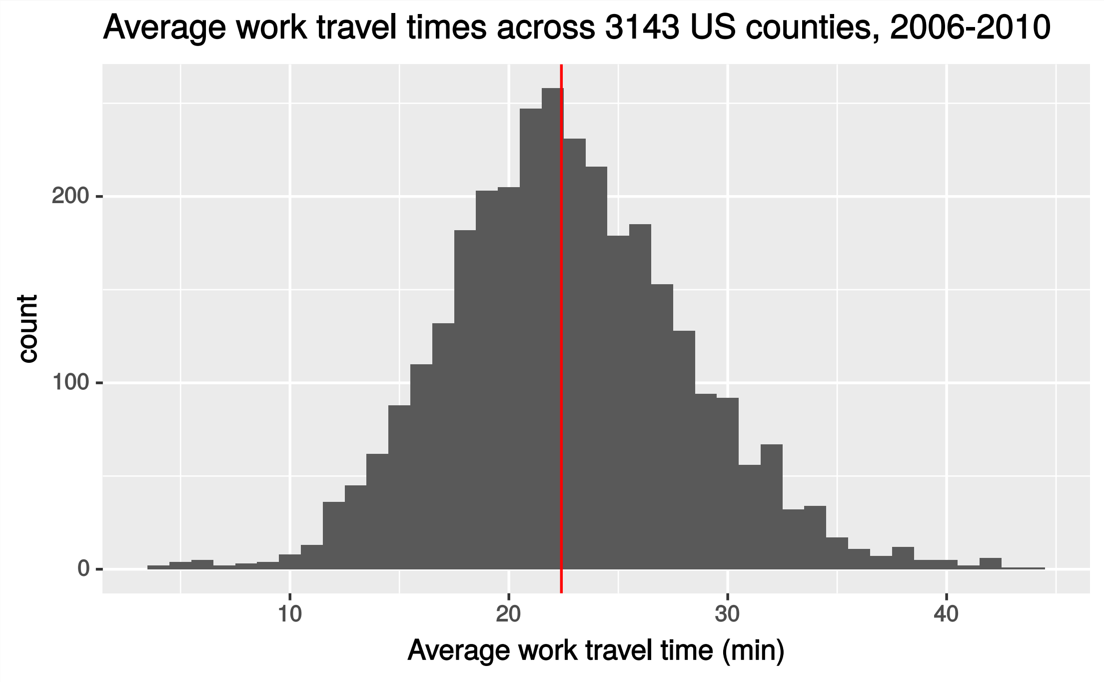


How can we quantify the shape of this distribution?

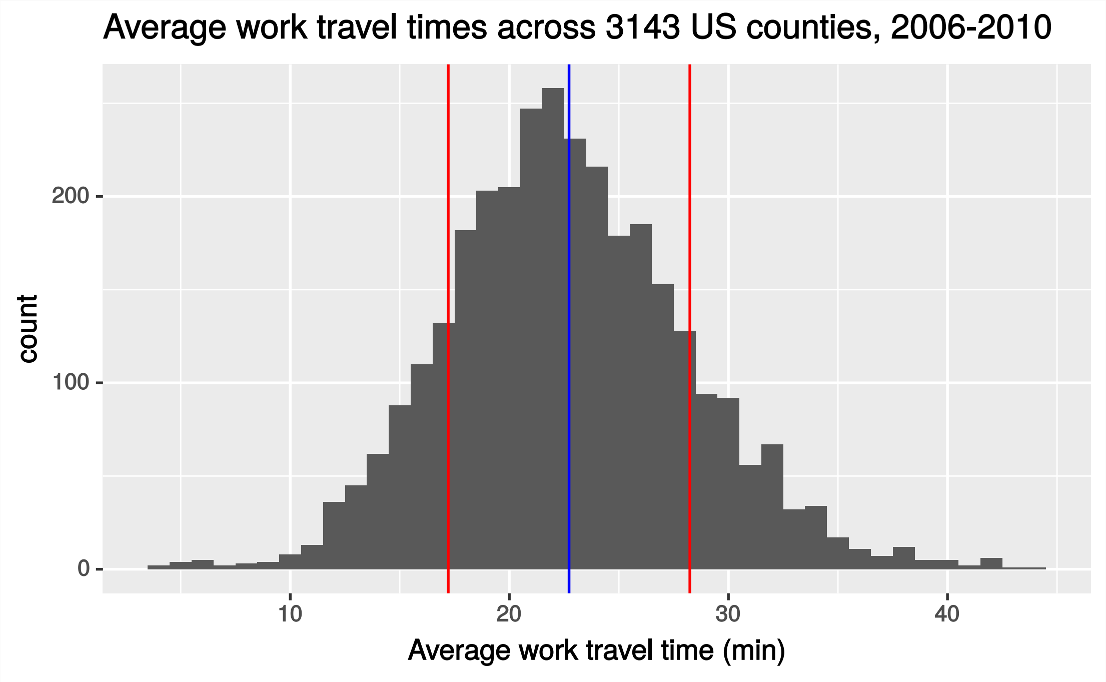
Distribution mean



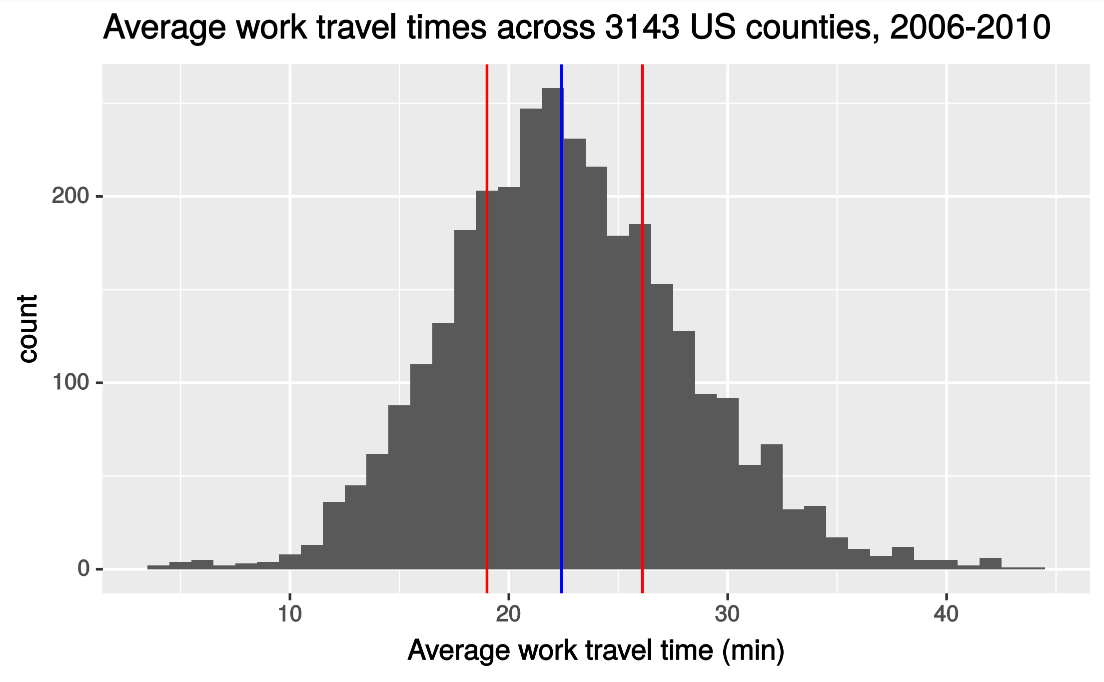
Distribution median



One standard deviation from mean



Inter-quartile range around median



Brown University's "Seeing Theory"

http://students.brown.edu/seeing-theory/index.html

Credits

The slides with blue headers originate from the following source:

- The Chapter 1 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.