Descriptive statistics

MOS5e chapter 2

Part II: issues and examples

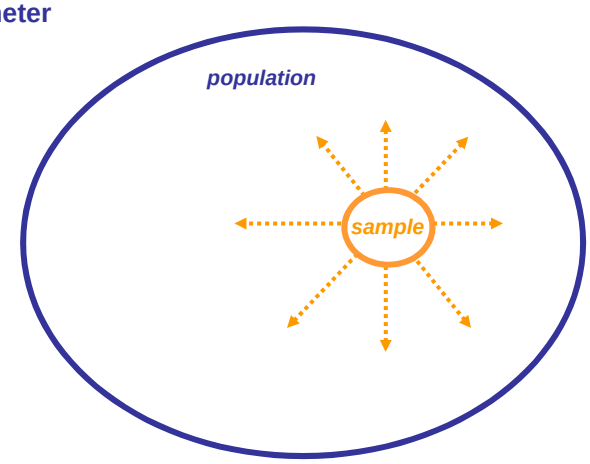
Two sources of information

Population
(data for all individuals)

→ parameter

Sample
(data for some individuals)

→ statistic



For every study, every news report, we need to identify

- the individuals ("units") studied
- whether the individuals are an entire *population* or just a *sample*
- the variable(s) studied
- whether each variable is quantitative or categorical
- the type and design of the study

Best way to determine whether each variable is quantitative or categorical: Imagine what a table of the raw data would look like

For each individual:

- if a meaningful number is recorded (→ **quantitative**)
- if a statement or attribute is recorded (→ **categorical**)

The National Center for Health Statistics reports that 8% of births in 2014 were low birthweight infants (weighing less than 2,500 grams) and that the mean age of mothers at first birth was 26.3 years.

"These number were computed based on the births certificates for all 3,988,076 births registered in the United States in 2014."

www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_12.pdf

NVSS December 23, 2015

PewResearchCenter

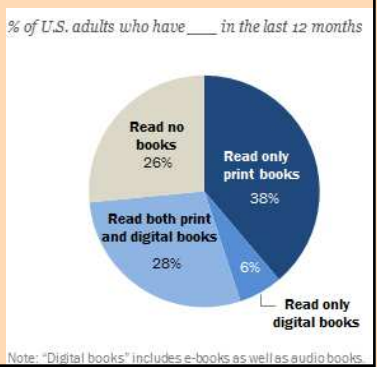
www.pewinternet.org/2016/09/01/book-reading-2016/

September 1, 2016
Book Reading 2016

"Some 73% of Americans report that they have read at least one book in the last year."

"Americans read an average (mean) of 12 books per year, while the typical (median) American has read 4 books in the last 12 months."

"The analysis in this report is based on a Pew Research Center survey conducted March 7-April 4, 2016, among a national sample of 1,520 adults, 18 years of age or older, living in all 50 U.S. states and the District of Columbia."



Researchers grafted human cancerous cells onto 20 healthy adult mice. Then 10 of the mice were injected with tumor-specific antibodies (anti-CD47) while the other 10 mice were not (IgG). Here are some published results.

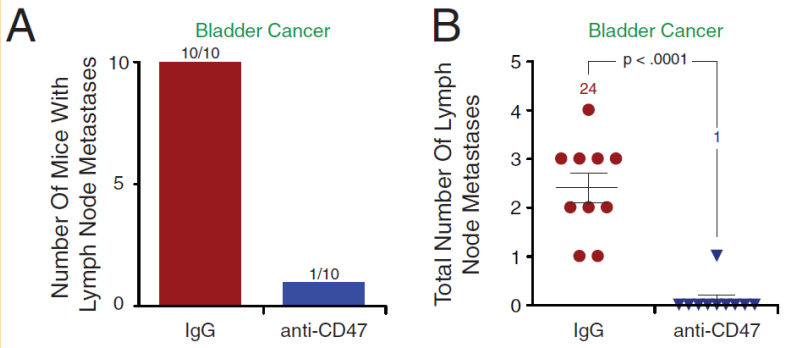


Fig. 5. Anti-CD47 mAbs prevent tumor metastasis. (*A*) The number of mice exhibiting lymph node metastases in each cohort. (*B*) The number of secondary lymph nodes detected in each mouse. The total number of secondary lymph nodes is indicated.

2012, doi: 10.1073/pnas.1121623109

The variable displayed is: **C)** categorical **D)** quantitative

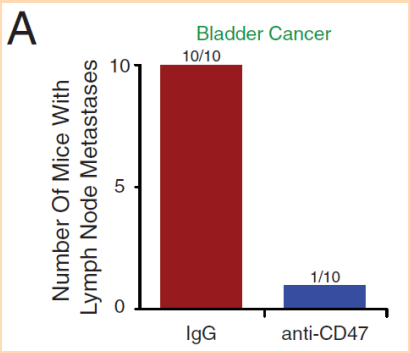


Fig. 5. Anti-CD47 mAbs prevent tumor metastasis. (*A*) The number of mice exhibiting lymph node metastases in each cohort. (*B*) The number of secondary lymph nodes detected in each mouse. The total number of secondary lymph nodes is indicated.

The variable displayed is: **C)** categorical **D)** quantitative

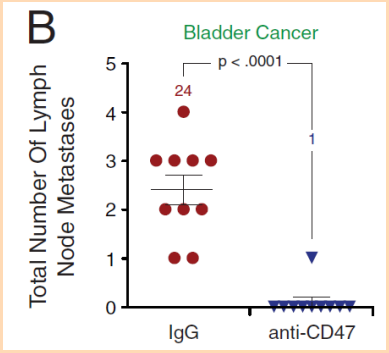


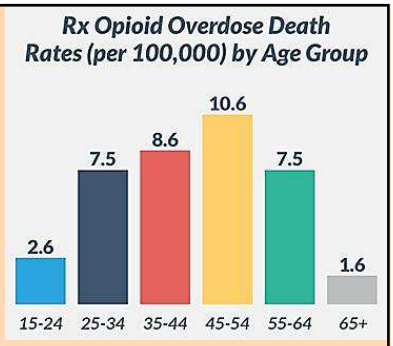
Fig. 5. Anti-CD47 mAbs prevent tumor metastasis. (*A*) The number of mice exhibiting lymph node metastases in each cohort. (*B*) The number of secondary lymph nodes detected in each mouse. The total number of secondary lymph nodes is indicated.

Researchers grafted human cancerous cells onto 20 healthy adult mice. Then 10 of the mice were injected with tumor-specific antibodies (anti-CD47) while the other 10 mice were not (IgG). Here is what a table of the raw data would look like.

Mouse	Treatment	Presence of metastatses	Number of metastases
1	IgG	yes	1
2	IgG	yes	1
3	IgG	yes	2
4	IgG	yes	2
5	IgG	yes	2
6	IgG	yes	3
7	IgG	yes	3
8	IgG	yes	3
9	IgG	yes	3
10	IgG	yes	4
11	anti-CD47	no	0
12	anti-CD47	no	0
13	anti-CD47	no	0
14	anti-CD47	no	0
15	anti-CD47	no	0
16	anti-CD47	no	0
17	anti-CD47	no	0
18	anti-CD47	no	0
19	anti-CD47	no	0
20	anti-CD47	yes	1

Appropriate summaries?

The CDC reports that, on average, 44 people die as a result of prescription opioid overdose every day in the United States.
Shown here is the rate of prescription opioid overdose deaths (in number of deaths per 100,000 people) in each of six age groups.

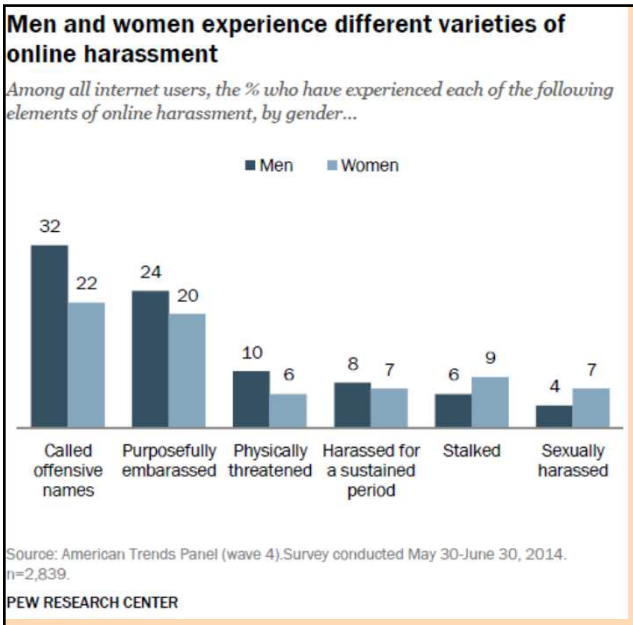


A) This is a histogram.
B) This is a boxplot.
C) This is a bar graph that could be displayed in 1 pie chart.
D) This is a bar graph that cannot be displayed in 1 pie chart.

The rate of prescription opioid overdose deaths among Americans between 35 and 64 years of age is

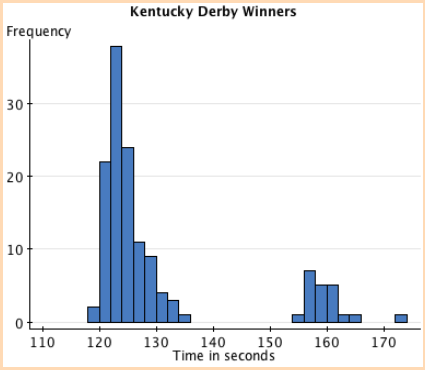
A) 8.9 per 100,000.
B) 10.6 per 100,000.
C) 26.7 per 100,000.
D) somewhere between 7.5 per 100,000 and 10.6 per 100,000.

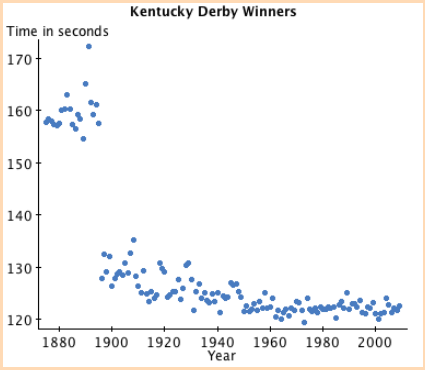
www.cdc.gov/drugoverdose/data/overdose.html



Can these data be displayed into 1 pie chart for the men and 1 pie chart for the women?

A) Yes **B)** No





Would it make sense to compute mean and standard deviation or five-number summary for these data?

According to kentuckyderby.com, in 1896 the race length was shortened from 1 ½ to 1 ¼ miles (the race was considered too long for 3-year-olds).

http://www.statcrunch.com/5.0/viewreport.php?reportid=10067

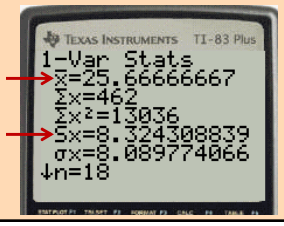
Summarizing quantitative data

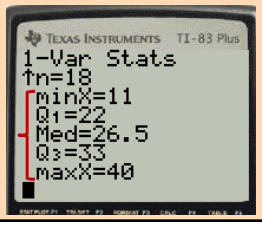
- Mean and standard deviation of data set
- Five number summary (min, Q1, median, Q3, max) of data set

Skin healing rate data from 18 newts, in micrometers/ hour:

11 12 14 18 22 22 23 23 26 27 28 29 30 33 34 35 35 40

STAT **CALC** **1-Var Stats** (select the list containing your data)



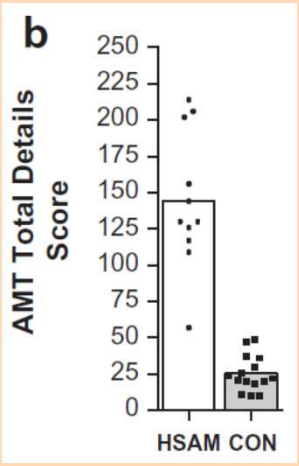


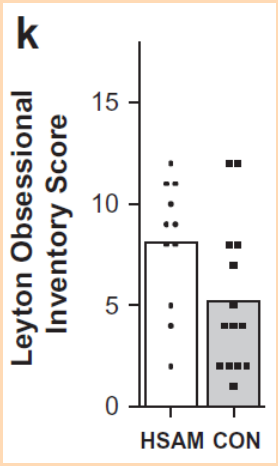
	Population (parameters)	Sample (statistics)
Mean	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum x}{n}$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

You need to know the symbols, but not the formulas

Comparing individuals with Highly Superior Autobiographical Memory (HSAM) and control individuals (CON): Individual scores and group mean scores on

- an autobiographical memory task (AMT)
- a test of common obsessional symptoms

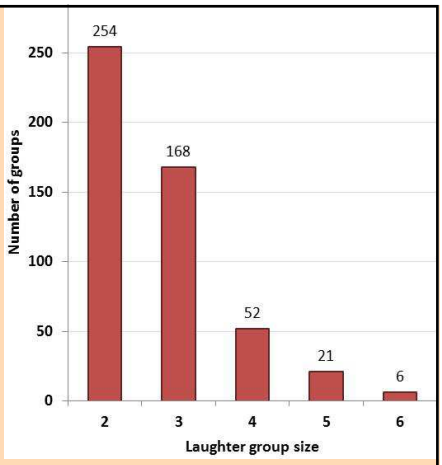




A study of freely forming groups in bars all over Europe recorded the group size (number of individuals in the group) of all 501 groups in the study that were naturally laughing.

Median laughter group size = ?

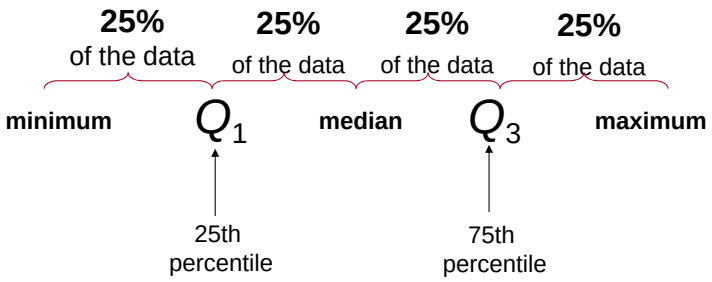
A) 2 B) 2.5 C) 3 D) 3.5 E) 4

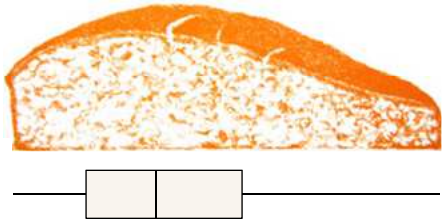


Would the average laughter group size be

A) smaller than the median?
B) about the same as the median?
C) larger than the median?

The five-number summary





The median does not have to be half-way between the min and max of a dataset!

Here is a statement from a study of 4,484 pregnant women enrolled in the Avon Longitudinal Study of Parents and Children:

"Blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th centiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively."

Draw a boxplot of mercury concentrations among the pregnant women in the study.

In the study, 25% of the women had blood mercury levels of ____ μg/l or greater.

A) 0.81 B) 1.35 C) 1.86 D) 2.52 E) 3.33

2013, DOI:10.1289/ehp.1206115

Spotting "suspected" outliers

Interquartile range IQR = $Q_3 - Q_1$

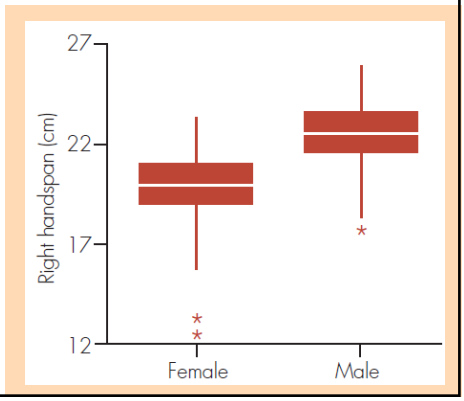
An observation is a **"suspected" outlier** if it is

$> Q_3 + (1.5)(IQR)$

or

$< Q_1 - (1.5)(IQR)$

Some stats software mark "suspected" outliers with an asterisk on a "**modified boxplot.**" *You should know how to interpret modified boxplots, not how to make them by hand.*



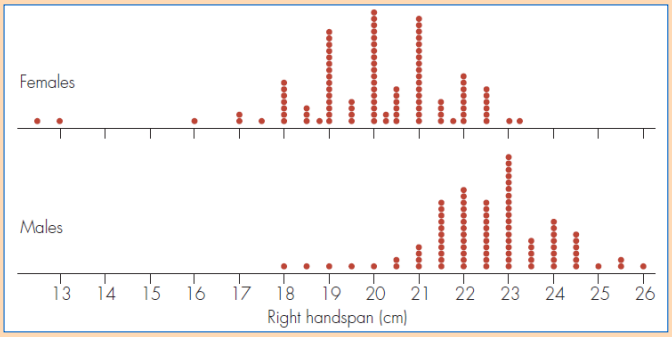
Dealing with outliers

What should you do if you find outliers in your data? It depends in part on what kind of outliers they are:

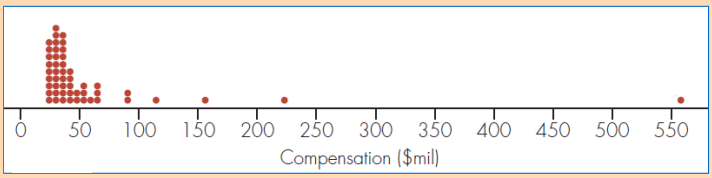
- Human error in recording information (typos)
- Problem with in experimentation (different conditions) or data collection (individuals that belong to a different group)
- Unexplainable but apparently legitimate wild observations
  - Are you interested in ALL individuals?
  - Are you interested only in typical individuals?

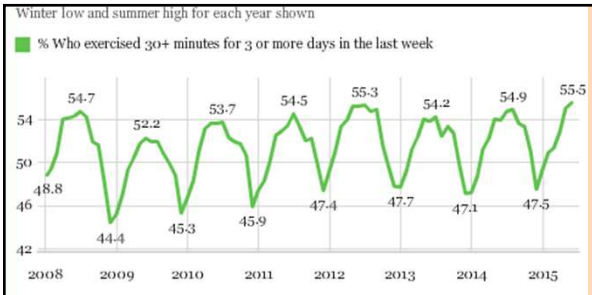
Don't discard outliers just to make your data look better, and don't act as if they did not exist.

Stretched right handspans (in centimeters) of female and male college students



50 highest-paid CEOs in *Fortune Magazine's* 2008 Top 500 companies

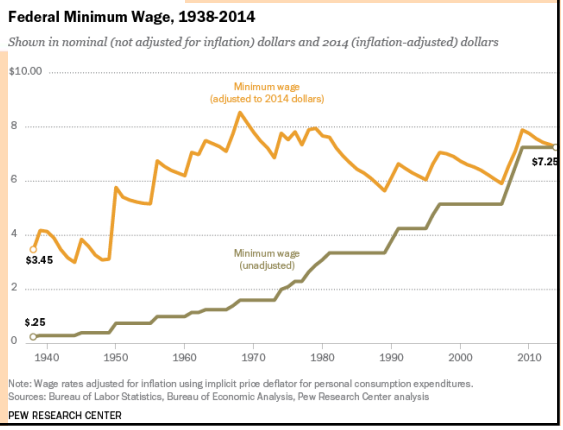




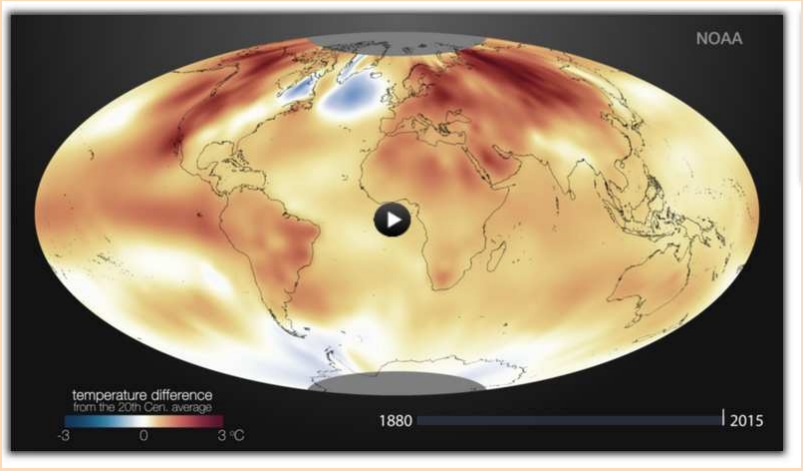
Results for June 2015 are based on telephone interviews with a random sample of 14,683 adults living in all 50 U.S. states.

www.gallup.com/poll/184403/far-2015-americans-exercising-frequently.aspx

Monthly averages
Gallup-Healthways Well-Being Index

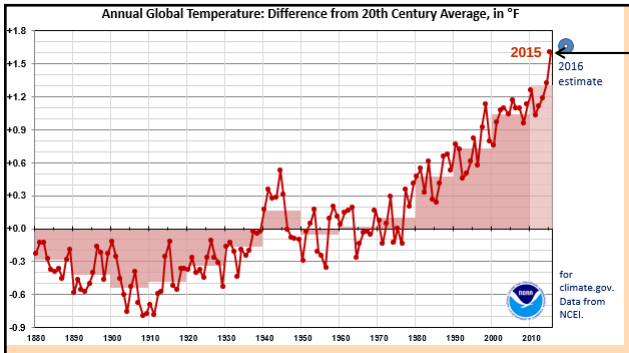


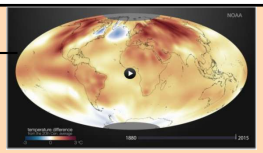
www.pewresearch.org/fact-tank/2015/07/23/5-facts-about-the-minimum-wage/



Animation at www.nnvl.noaa.gov/MediaDetail2.php?MediaID=1835&MediaTypeID=3&ResourceID=104973

Global annual temperature anomaly (land-ocean index) since 1880. (Anomalies are relative to a 1951-1980 reference.)





http://data.giss.nasa.gov/gistemp

"The hiatus period of 18 years 2 months, or 218 months, is the farthest back one can go in the RSS satellite temperature record and still show a sub-zero trend."
www.climatedepot.com