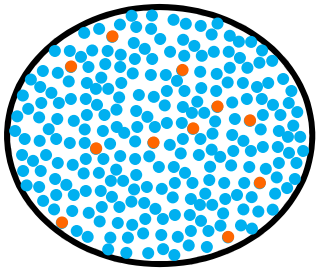
Descriptive statistics

PSLS chapters 1, 2 & 5

Part II: issues and examples (flipped lesson)

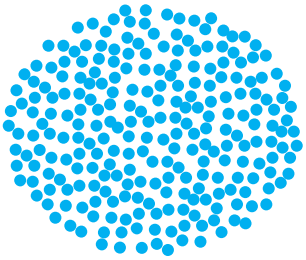
How do we learn from data?

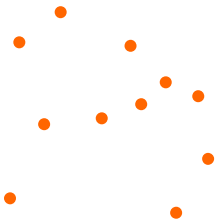
Target population



Population data
(data for all individuals)

Sample data
(data for some individuals)





Population data
(data for all individuals)
→ parameter

Expensive
Time consuming
Maybe impossible
Exact knowledge

Sample data
(data for some individuals)
→ statistic

Cheaper
Faster
Typically doable
Uncertainty

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

- the individuals ("units") studied
- whether the study 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 31.9% of US births in 2016 were delivered via cesarean (C-section) and that the mean age of mothers at first birth was 26.6 years.

These numbers were computed based on the births certificates for all 3,945,875 births registered in the United States in 2016.

- individuals ("units") studied:
- population or sample:
- variable(s) reported:

What are all the values cited?

www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_01.pdf

NVSS January 31, 2018

The polling organization Gallup interviewed a random sample of 1,023 American adults in July 2016 and found that 19% had smoked cigarettes in the past week and that smokers smoked on average 12 cigarettes per day.

- individuals ("units") studied:
- population or sample:
- variable(s) reported:

What are all the values cited?

GALLUP

www.gallup.com/poll/194216/cigarette-smokers-lighting-less-often.aspx

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).

Other than the two groups on the horizontal axis, what is the variable displayed here?
This variable is: **C)** categorical **D)** quantitative

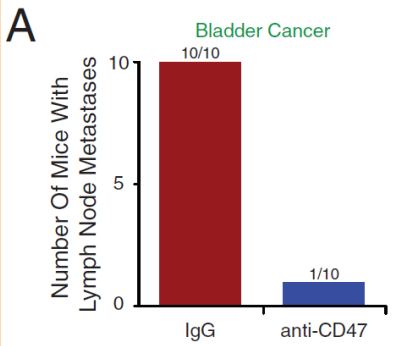


Fig. 5. Anti-CD47 mAbs prevent tumor metastasis. (*A*) The number of mice exhibiting lymph node metastases in each cohort. (*B*) The number of secondary lymph nodes detected in each mouse. The total number of secondary lymph nodes is indicated.

doi: 10.1073/pnas.1121623109

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).

Other than the two groups on the horizontal axis, what is the variable displayed here?
This variable is: **C)** categorical **D)** quantitative

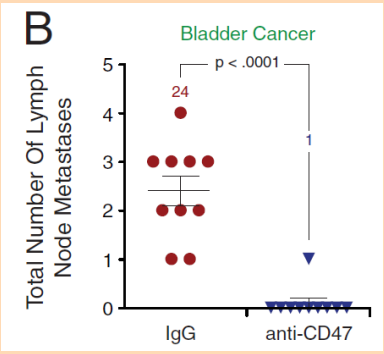


Fig. 5. Anti-CD47 mAbs prevent tumor metastasis. (*A*) The number of mice exhibiting lymph node metastases in each cohort. (*B*) The number of secondary lymph nodes detected in each mouse. The total number of secondary lymph nodes is indicated.

doi: 10.1073/pnas.1121623109

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?

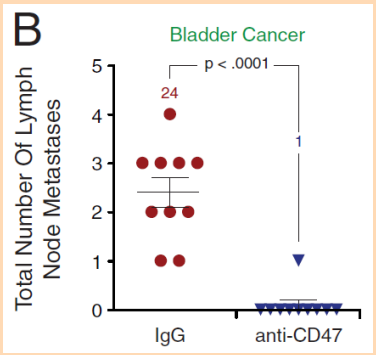
	Population (parameters)	Sample (statistics)
Count	$\text{count} = X$	$\text{count} = x$
Proportion	$p = \frac{X}{n}$	$\hat{p} = \frac{x}{n}$
Mean	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum x}{n}$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
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

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).

Which treatment group had the most variable results?

Which treatment group had the largest standard deviation?



Results - Descriptive Statistics

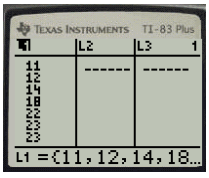
Export

	n	Sample Mean	Standard Deviation	Min	Q1	Median	Q3	Max
IgG-control	10	2.400	0.9661	1	2	2.500	3	4
anti-CD47	10	0.1000	0.3162	0	0	0	0	1

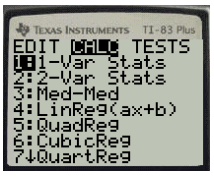
doi: 10.1073/pnas.1121623109

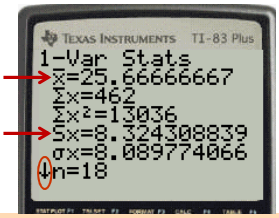
Summary statistics for quantitative data – TI calculator

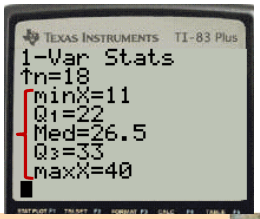
- Mean and standard deviation of data set *see flipped video*
- Five number summary (min, Q1, median, Q3, max)



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



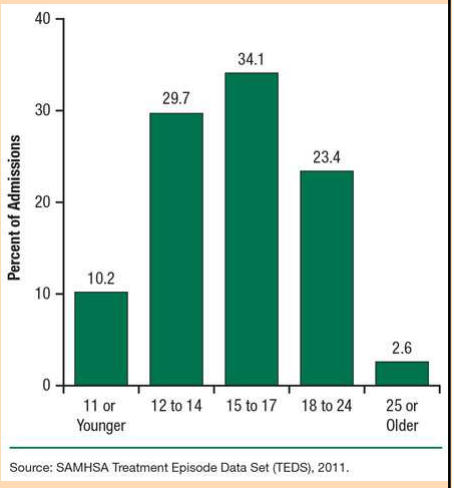




Skin healing rated (mcm/h): 11 12 14 18 22 22 23 23 26 27 28 29 30 33 34 35 35 40

The Treatment Episode Data Set (TEDS) is a national data system of annual admissions to substance abuse treatment facilities.

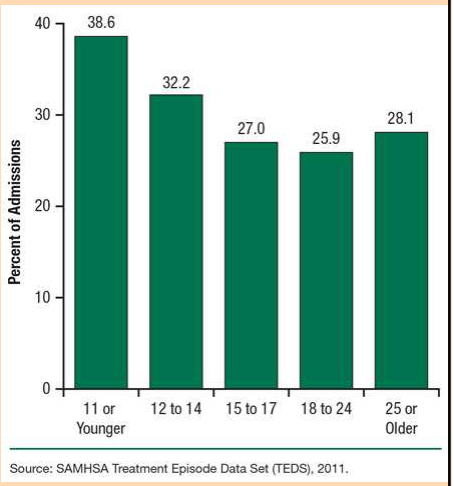
"Age of Substance Use Initiation among Treatment Admissions Aged 18 to 30: 2011"



What percent of young adults (ages 18-30) in treatment facilities initiated substance abuse before the age of 18?

A) 34.1%
B) Somewhere between 10.2% and 34.1%
C) 74%

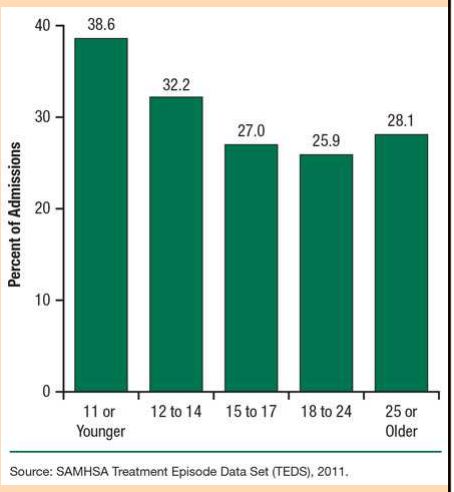
"Admissions Reporting Co-Occurring Mental Disorders, by Age at Substance Use Initiation among Treatment Admissions Aged 18 to 30: 2011"



What percent of young adults in treatment facilities who initiated substance abuse before the age of 18 report co-occurring mental disorders?

A) 27%
B) Somewhere between 27% and 38.6%
C) 38.6%
D) 97.8%

"Admissions Reporting Co-Occurring Mental Disorders, by Age at Substance Use Initiation among Treatment Admissions Aged 18 to 30: 2011"

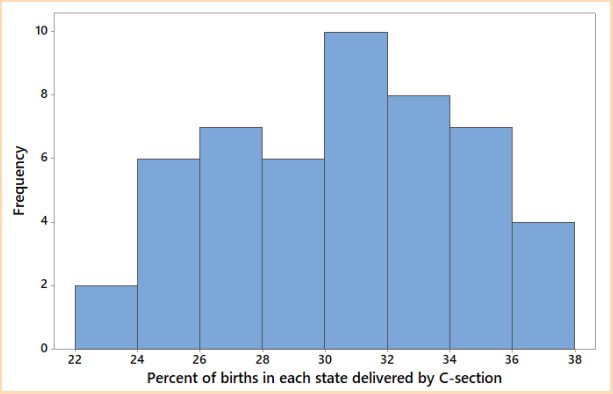


Always ask yourself:

How many variables are described in this graph?

		Age at Substance Use Initiation				
		≤ 11	12-14	15-17	18-25	≥ 25
Co-Occurring Mental Disorder	Yes	38.6%	32.2%	27.0%	25.9%	28.1%
	No	61.4%	67.8%	73.0%	74.1%	71.9%

Percent of all births delivered by C-section in 2015, for each of the 50 US states



www.cdc.gov/nchs/pressroom/sosmap/cesarean_births/cesareans.htm

-Who or what are the individuals described by this graph?

-This graph is a:

A) histogram

B) boxplot

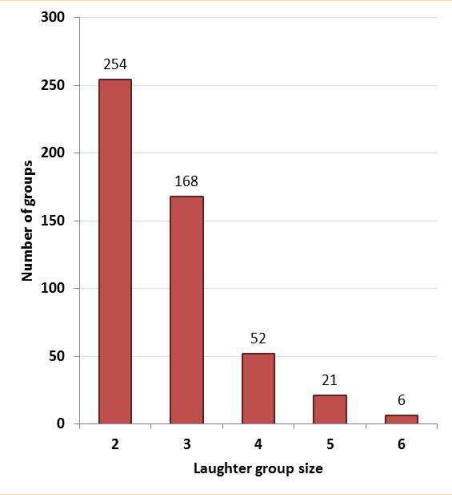
C) bar graph that could be turned into a pie chart

D) bar graph that could not be turned into a pie chart

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) 4 D) 52 E) 100.2



Is the average laughter group size

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

The five-number summary: min, Q1, med, Q3, max

Between min. and Q1	25% of the data	← Q_1: 25th percentile
Between Q1 and median	25% of the data	
Between median and Q3	25% of the data	← Q_3: 75th percentile
Between Q3 and max.	25% of the data	

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 = $Q_3 - Q_1$

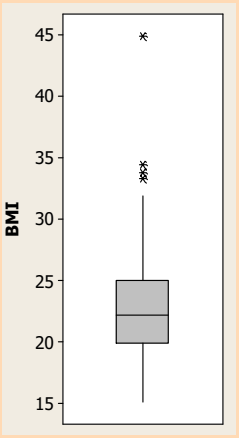
An observation is a **"suspected" outlier** if it is

$> Q_3 + (1.5)(IQR)$

or

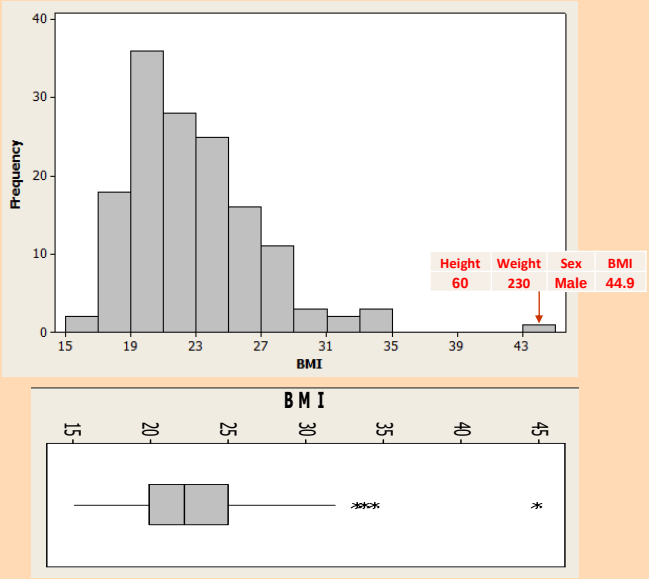
$< Q_1 - (1.5)(IQR)$

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.*



Class survey results: weight and height used to compute BMI

Class survey results: weight (lbs) and height (in) used to compute BMI.



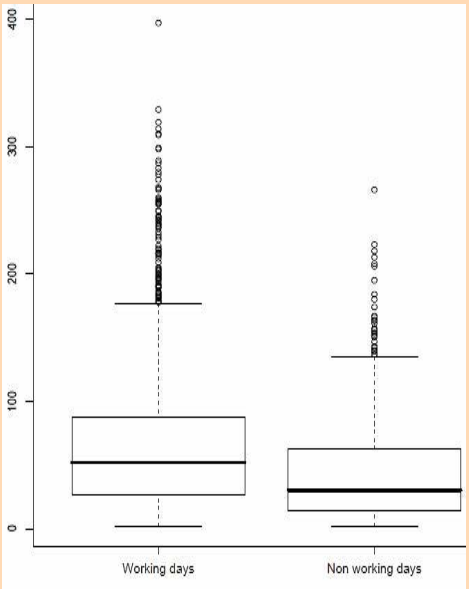
Nitrogen oxides (NOx) are greenhouse gases with a strong impact on global warming. A monitoring station in an urban area in Spain records NOx levels every hour of every day.

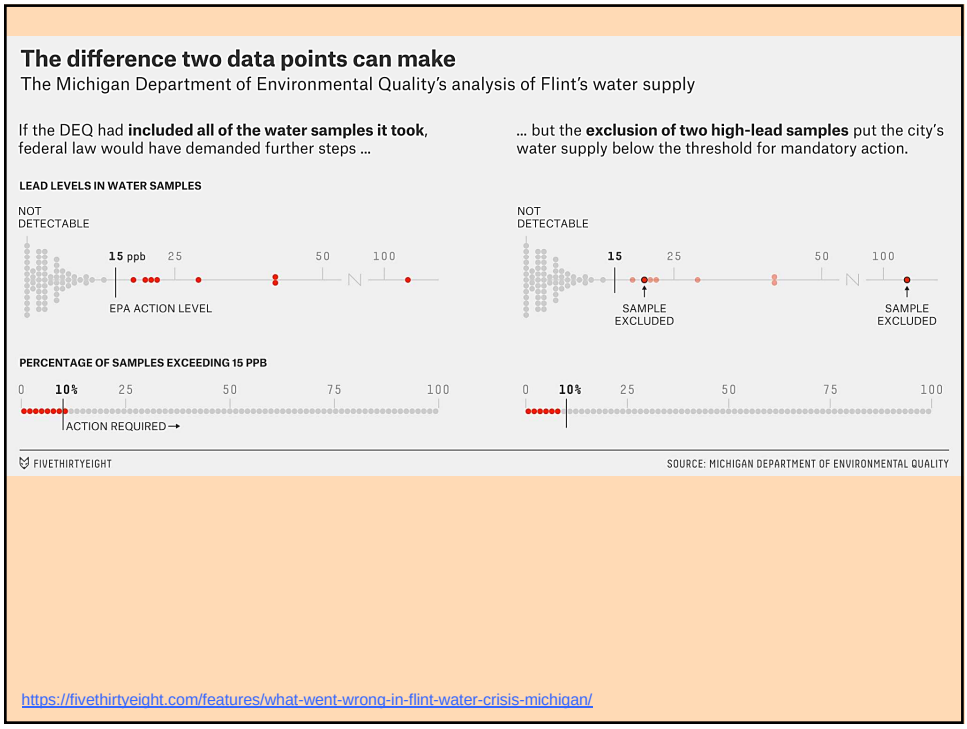
This graph compares recorded NOx levels (in µg/m^3) during working days and non-working days (weekends and holidays) over a six-month period.

What can we conclude?

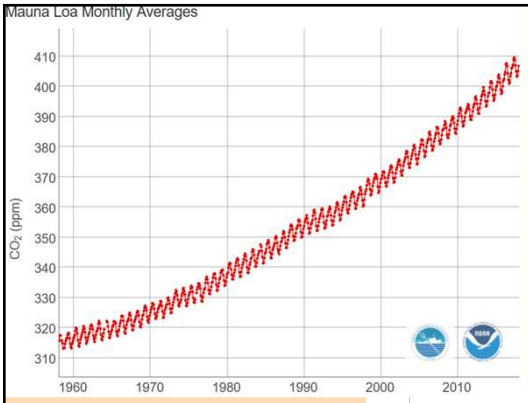
The distribution of NOx levels on <u>working days</u> has:

A) no outliers
B) 1 outlier
C) 2 outliers
D) many outliers

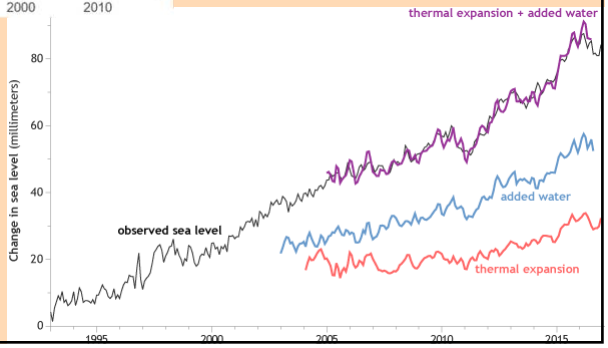




https://fivethirtyeight.com/features/what-went-wrong-in-flint-water-crisis-michigan/



Monthly mean atmospheric CO_2 at Mauna Loa Observatory, Hawaii (continuous records since 1958)
www.esrl.noaa.gov/gmd/ccgg/trends/graph.html



NOAA Climate.gov graphics, from *State of the Climate in 2016* report.
www.climate.gov/news-features/understanding-climate/state-climate-highlights/2016

1993: start of satellite records

FiveThirtyEight

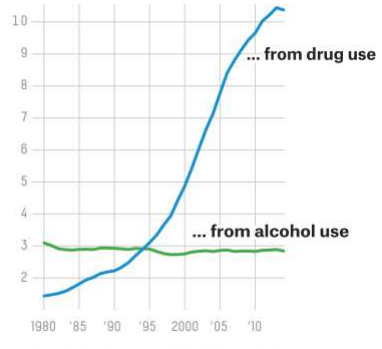
UPDATED 4:00 PM EDT | APR 2, 2016

35 Years Of American Death

Mortality rates for leading causes of death in every U.S. county from 1980 to 2014.

By Ella Koeze

Estimated **deaths per 100,000** people in the U.S. ...



Mortality rates are age-adjusted to account for higher mortality in older populations

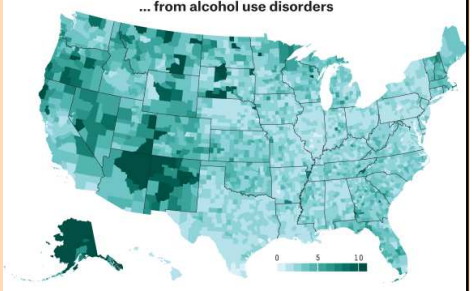
SOURCE: INSTITUTE FOR HEALTH METRICS AND EVALUATION

https://fivethirtyeight.com/features/more-americans-are-dying-from-suicide-drug-use-and-diarrhea/

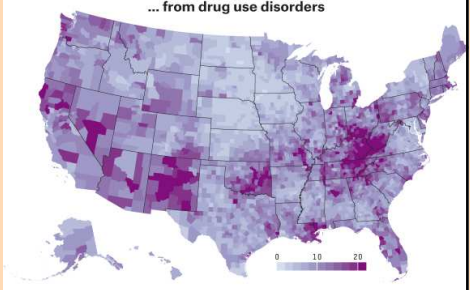
https://projects.fivethirtyeight.com/mortality-rates-united-states/drug/#2014

Estimated **deaths per 100,000 people** in 2014, by county ...

... from alcohol use disorders



... from drug use disorders



Mortality rates are age-adjusted to account for higher mortality in older populations

FiveThirtyEight

SOURCE: INSTITUTE FOR HEALTH METRICS AND EVALUATION