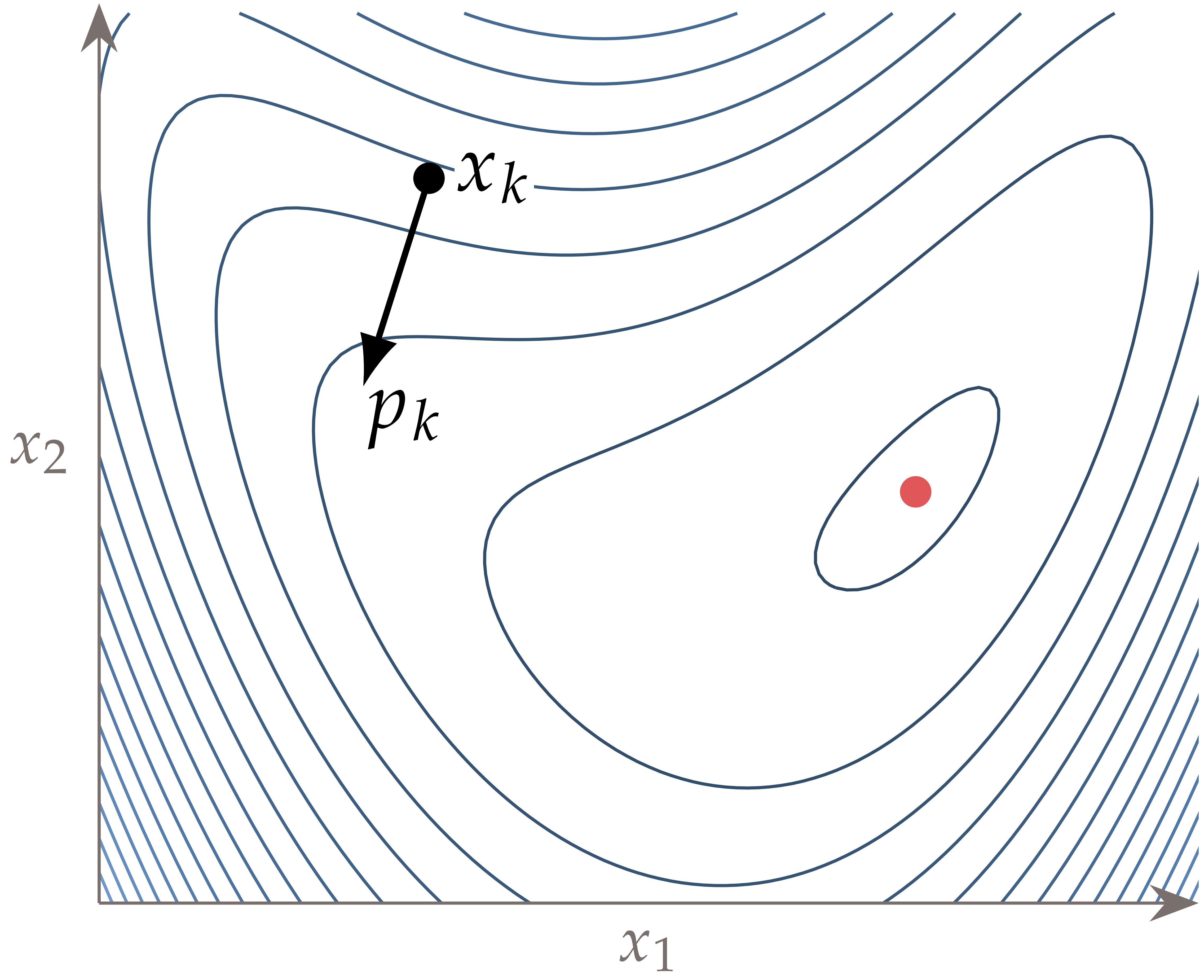
Intro to Statistics



ME EN 275
Andrew Ning
aning@byu.edu

Statistics Flaws

1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

from onlinestatbook.com

Statistics Flaws

1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.
2) The more churches in a city, the more crime there is. Thus, churches lead to crime.

from onlinestatbook.com

Statistics Flaws

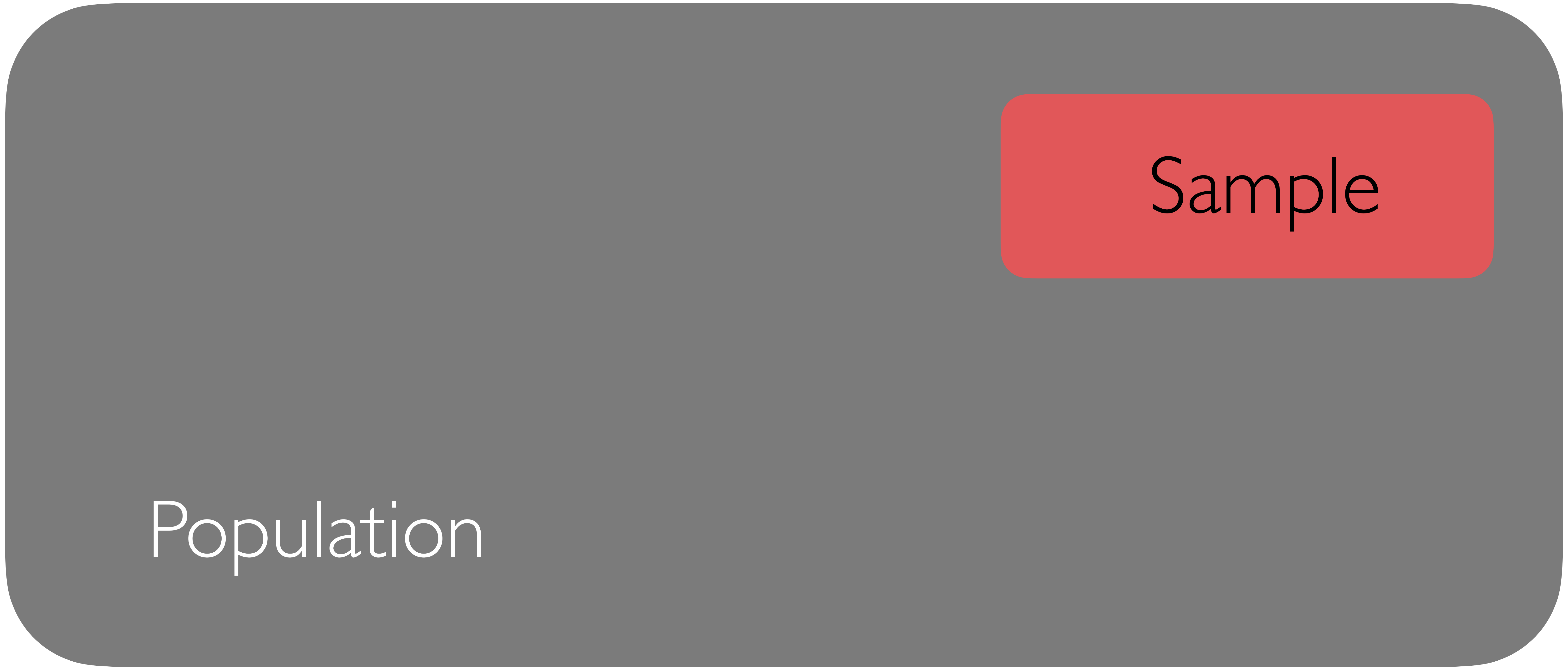
1) A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.
2) The more churches in a city, the more crime there is. Thus, churches lead to crime.
3) 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

from onlinestatbook.com

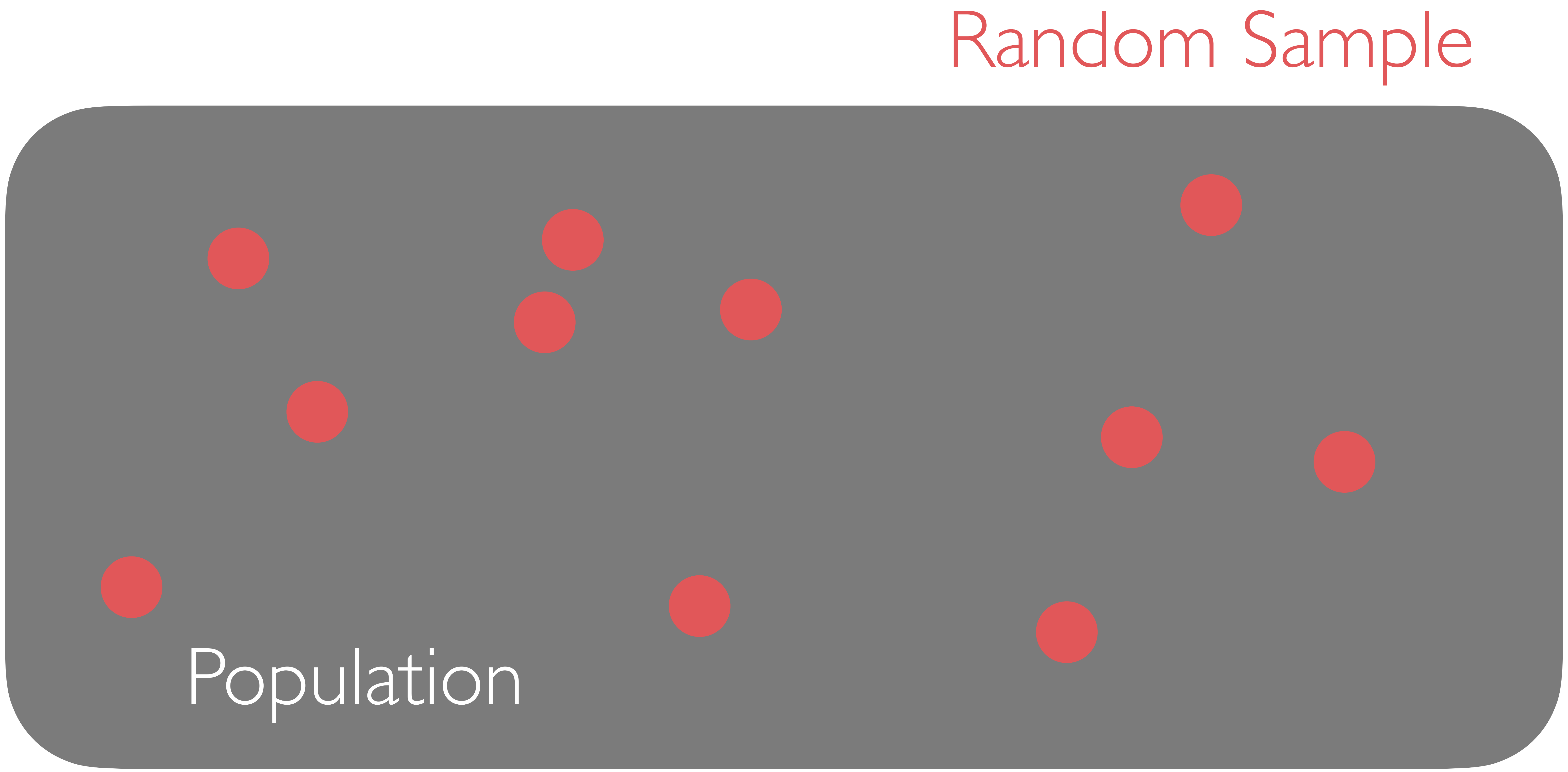
Other Examples of Misleading Statistics

https://wpdatatables.com/misleading-statistics/

Sampling



Sampling



Good Samples?

A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

Good Samples?

A quality engineer wants to inspect electronic microcircuits in order to obtain information on the proportion that are defective. She decides to draw a sample of 100 circuits from a day's production. Each hour for 5 hours, she takes the 20 most recently produced circuits and tests them.

Principles of Statistics for Engineers and Scientists, Navidi

Independence

knowing the value of one sample does not help predict the value of another

Summary Statistics

sample mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	np.mean(x)
sample variance	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$	np.var(x, ddof=1)
sample standard deviation	$s = \sqrt{s^2}$	np.std(x, ddof=1)

Summary Statistics

median

np.median(x)

Compare median vs mean:

$$x = [1, 2, 3, 4, 5, 6, 100]$$

Summary Statistics

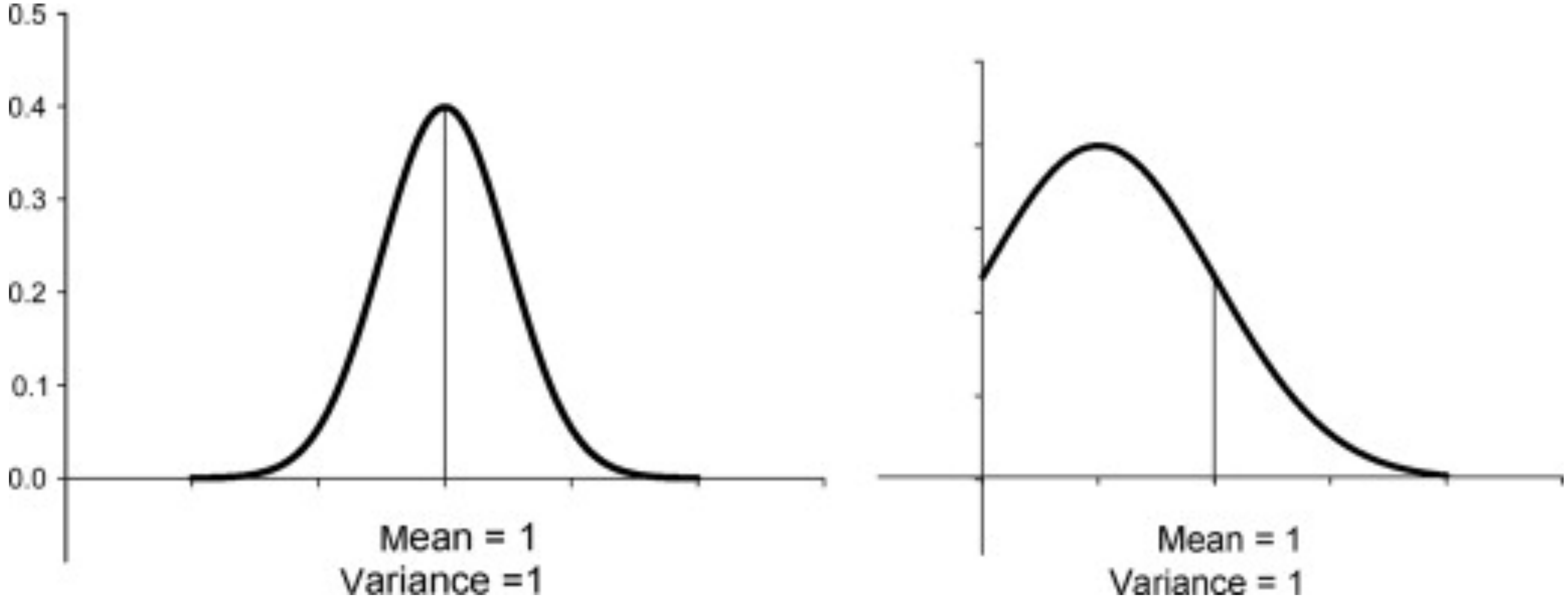
the pth *percentile*: for n data points, labeled 0 to n, the pth percentile is at the p/100*n th position (interpolate between points)

the 50th percentile is the median

```
x = np.array([1, 2, 3, 4, 5, 6, 100])
```

try computing different percentiles with `np.percentile`

But Summary Statistics are Not Enough



Kandethody M. Ramachandran, Chris P. Tsokos, Mathematical Statistics with Applications in R

Histograms (distributions of the data)

get sample data from:

```
x = np.random.normal(0, 1, 100)
```

(we'll talk more about this later, but in brief it samples 100 random data points from a distribution that has mean=0 and stdev=1)

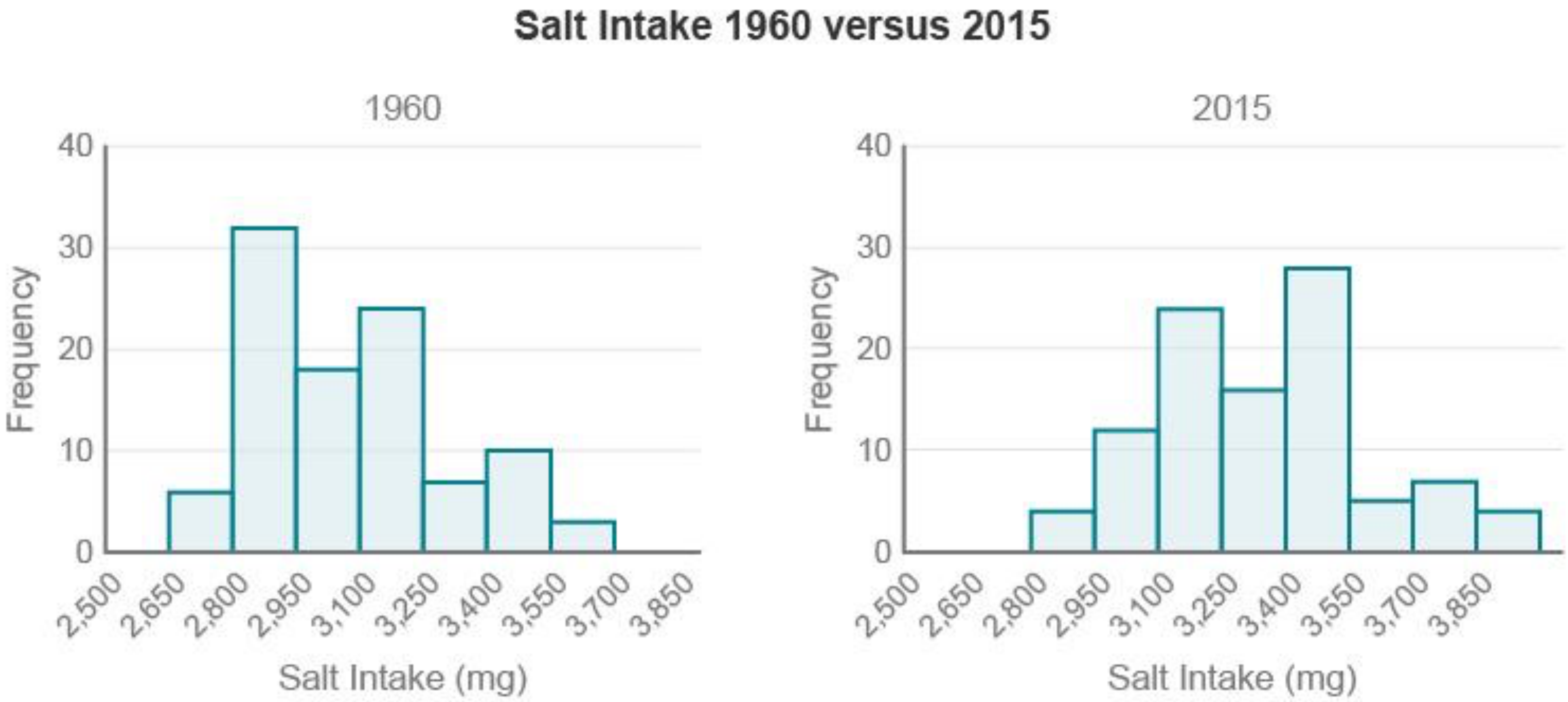
now plot a histogram.

try the pyplot.hist function.

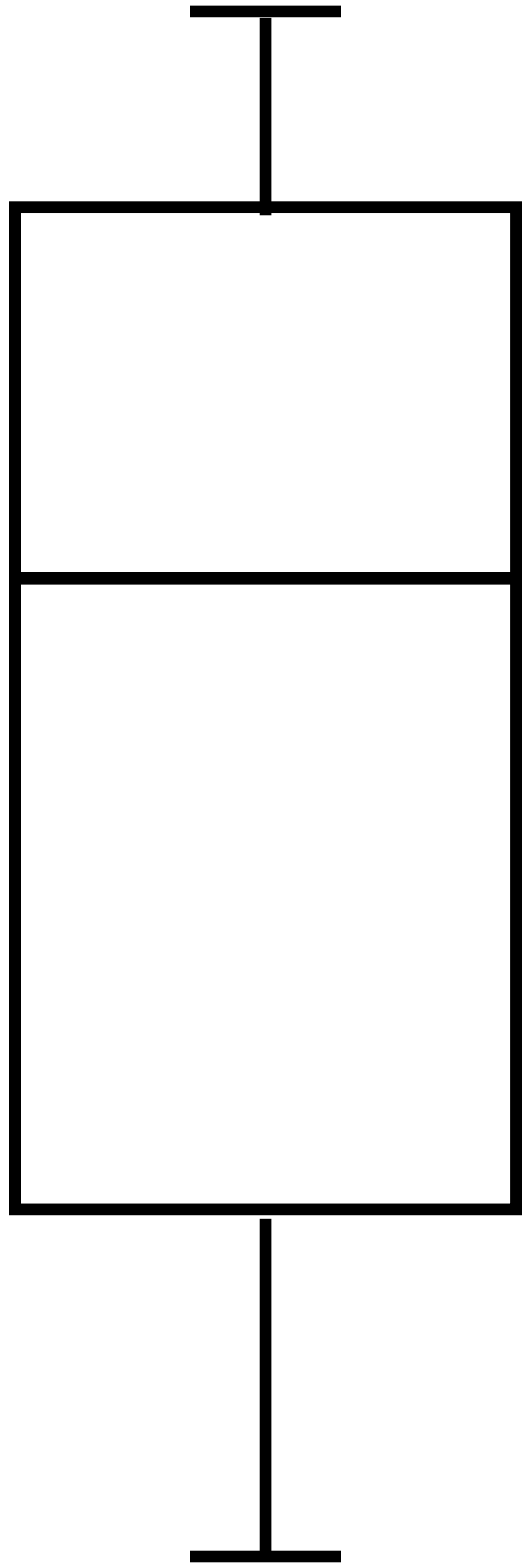
you should also try the bins keyword argument

try also increasing the sample size (change the 100 above)

Histograms


Salt Intake 1960 versus 2015
1960
2015
Frequency
0
10
20
30
40
2,500
2,650
2,800
2,950
3,100
3,250
3,400
3,550
3,700
3,850
Salt Intake (mg)

Box and Whisper Plot



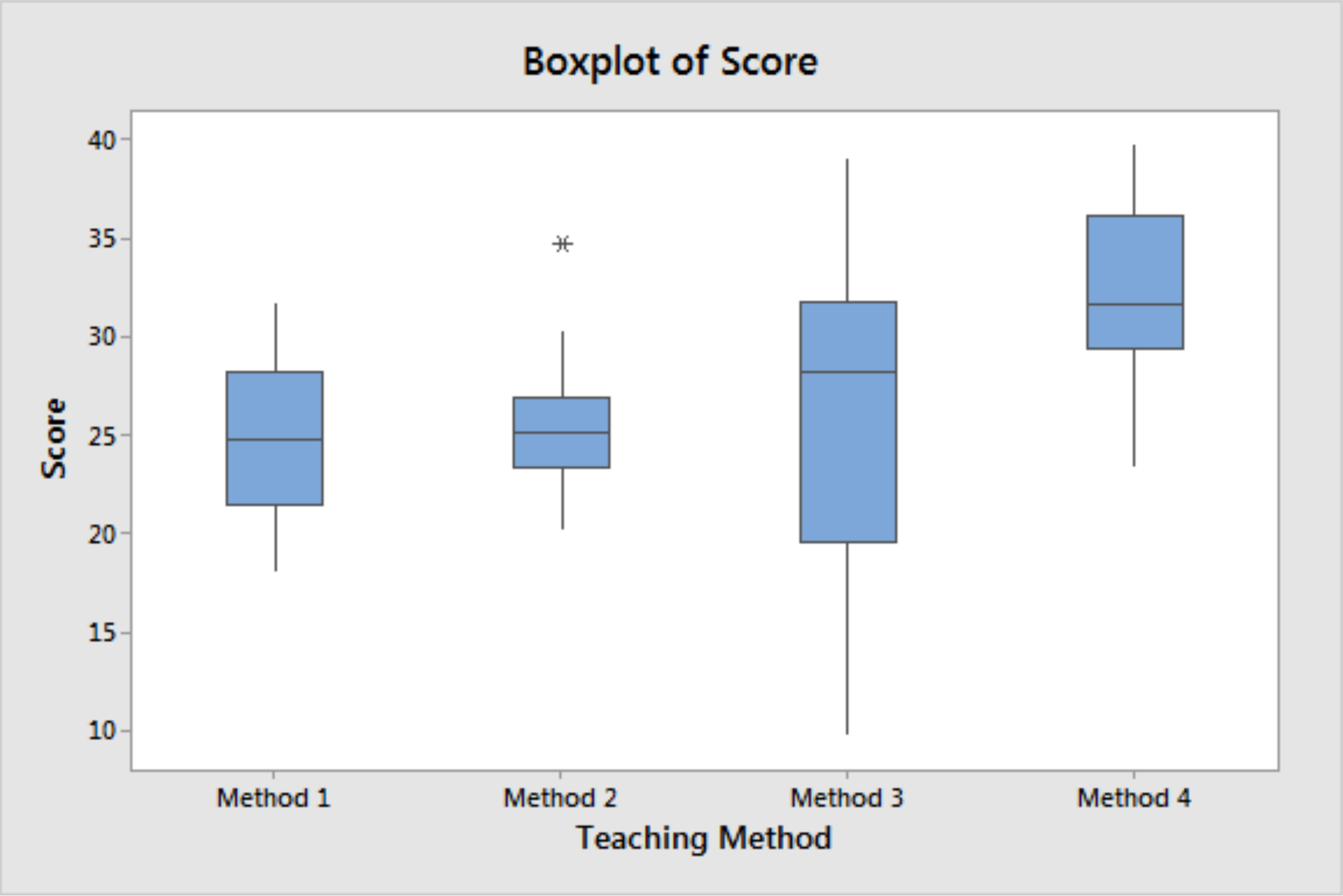
upper extreme

75th percentile

50th percentile (median)

25th percentile

lower extreme



Box Plot Explained with Examples, Statistics with Jim

Bivariate Data

(two variables)

Correlation Coefficient

-1
-0.8
-0.6
0
0.6
0.8
1

A perfect negative relationship.

A strong negative relationship.

A moderate negative relationship.

No relationship.

A moderate positive relationship.

A strong positive relationship.

A perfect positive relationship.

Olga Berezovsky, How to do linear regression and correlation analysis

Pearson’s Correlation Coefficient

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

Python

```
x = np.linspace(0, 1, 10)
y = np.array([1, 2, 4, 5, 8, 12, 18, 20, 40, 50])
```

plot scatter plot and compute correlation coefficient (r)

scipy.stats.pearsonr

Limitations

Only measures linear relationships

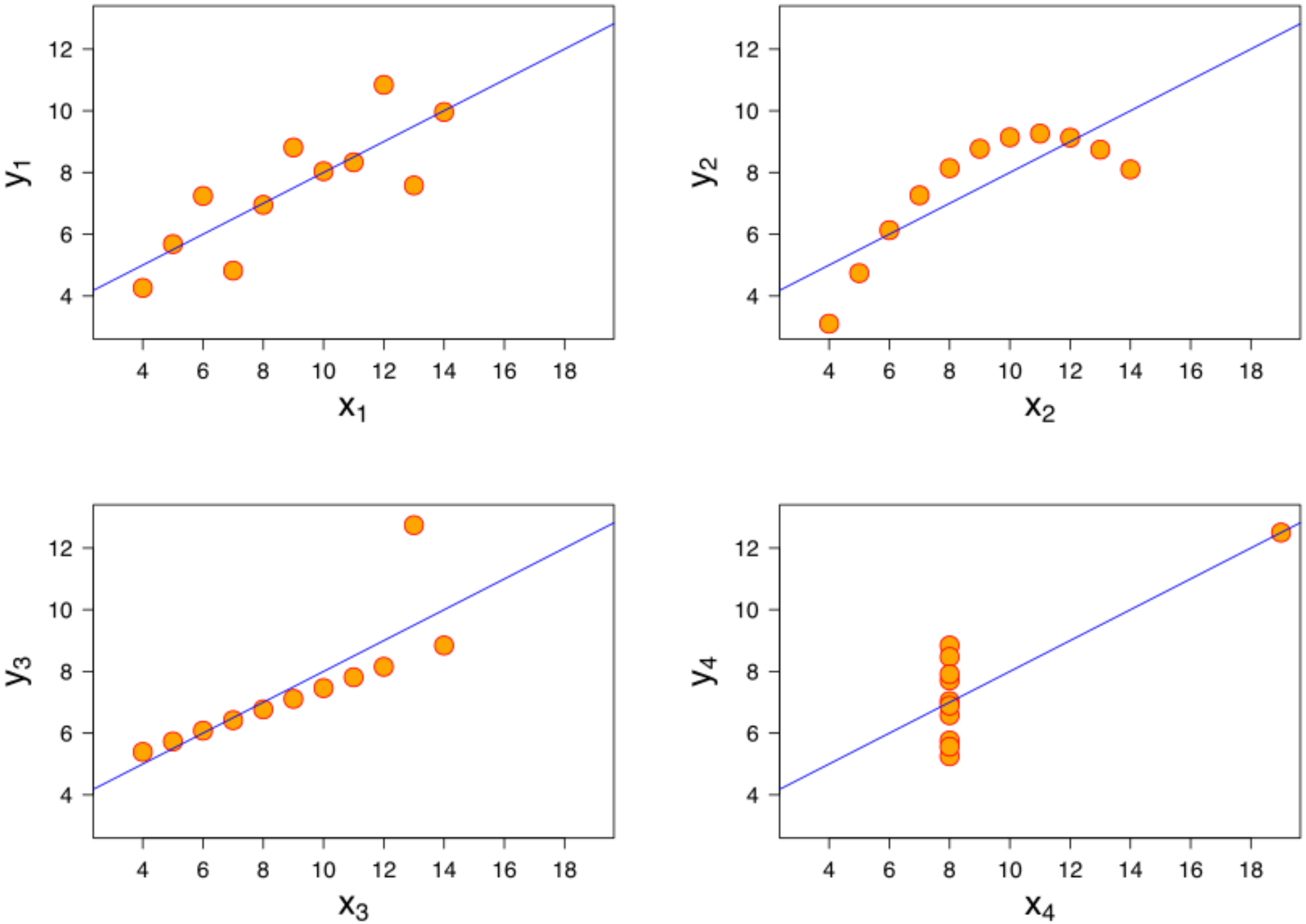
Highly sensitive to outliers

Correlation does not imply causation!

Correlation Does Not Imply Causation

https://tylervigen.com/spurious-correlations

Be careful with summary statistics - look at the data!



12
10
8
6
4
y_1
4
6
8
10
12
14
16
18
x_1
y_2
x_2
y_3
x_3
y_4
x_4