Introduction to Data Science
CS 5963 / Math 3900

Lecture 2: Introduction to Descriptive Statistics

Required Reading: Grus, Ch.5
Available digitally from library: link

THE UNIVERSITY OF UTAH

Statistics, Descriptive Statistics, and Data

Statistics is a branch of mathematics that is used to analyze data.

Descriptive statistics quantitatively describes or summarizes features of a dataset.

For the purposes of this lecture, we'll think of a dataset consisting of a number of “items” each of which has a number of associated “variables” or “attributes”.

Example: As part of homework 0, you filled out a survey.

The “items” are each student and the “variables” are the question responses.

	first name	last name	major	…	gender	age
student 1	Braxton	Osting	math		M	33
student 2	Alex	Lex	CS		M	35
…						
student n	Science	Cat	hunting		F	2

Variable Types

Nominal: Unordered categorical variables

Ordinal: There is an ordering but no implication of equal distance between the different points of the scale.

Interval: There are equal differences between successive points on the scale but the position of zero is arbitrary.

Ratio: The relative magnitudes of scores *and* the differences between them matter. The position of zero is fixed.

Nominal Variables

Unordered categorical variables

Examples:

Survey responses: sex (M/F), true or false (T/F), yes or no (Y/N)

color

Ordinal Variables

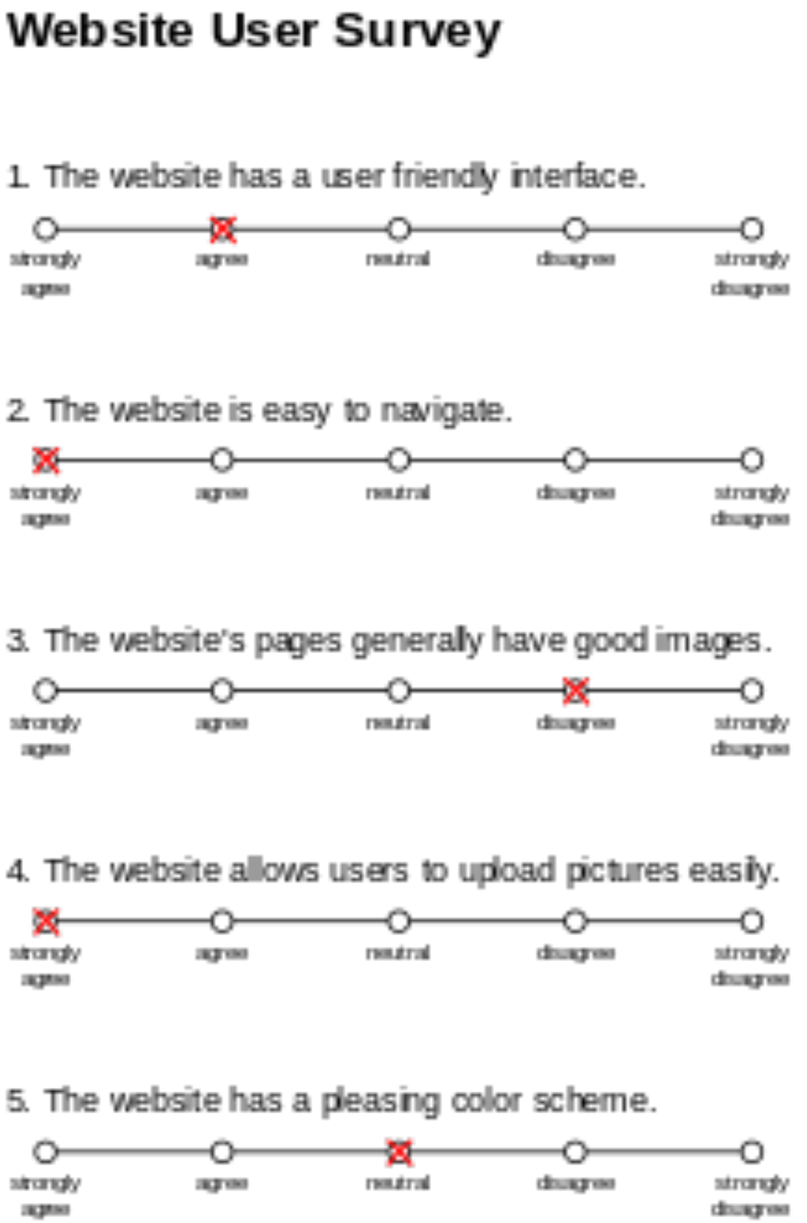
There is an ordering but no implication of equal distance between the different points of the scale.

Examples:

on Likert scale of 1 to 5, how comfortable are you with programming?

educational level (high school, some college, degree, graduate…)

size: S/M/L/XL



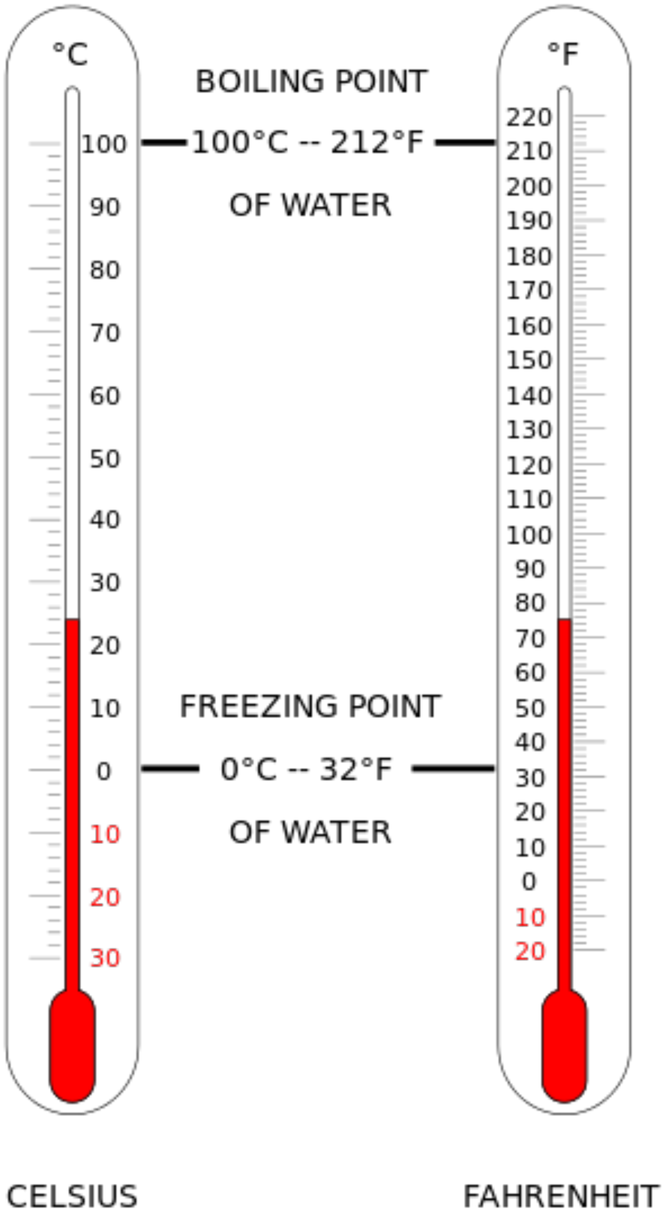
Source: Wikipedia

Interval Variables

There are equal differences between successive points on the scale but the position of zero is arbitrary.

Examples:

Measurement of temperature using the Celsius or Fahrenheit scales.



Source: Wikipedia

Ratio Variables

The relative magnitudes of scores *and* the differences between them matter. The position of zero is fixed.

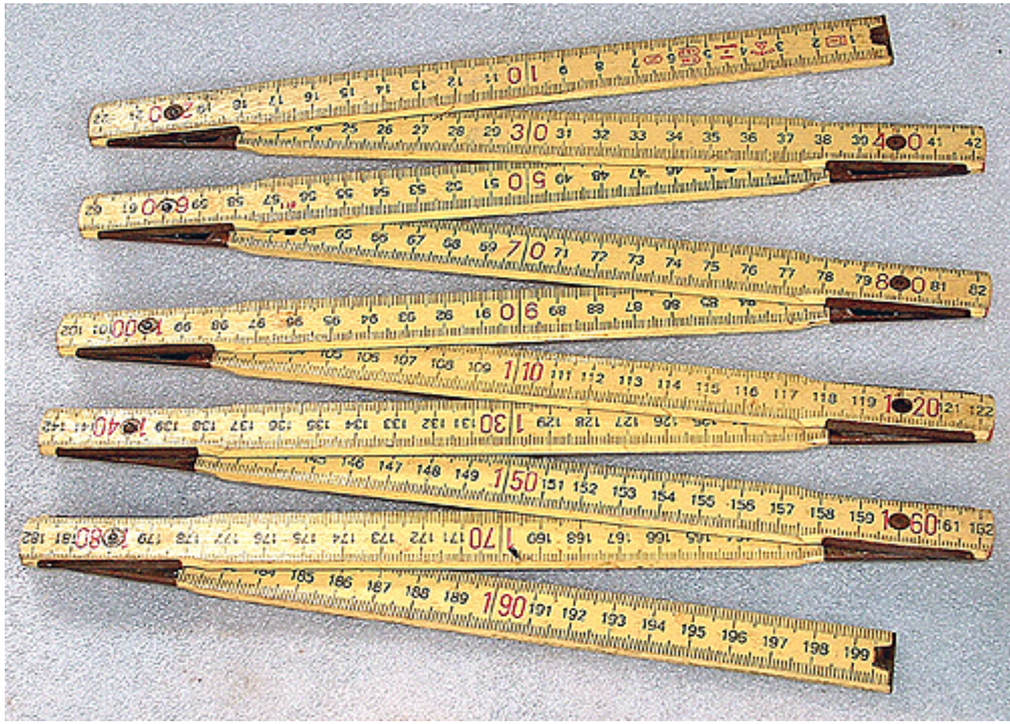
Examples:

Absolute measure of temperature (Kelvin scale)

Age

Weight

Length



Source: Wikipedia

Quiz!

What type of variable (Nominal, Ordinal, Interval, or Ratio) are the following:

1. Olympic 50 meter race times
2. College major
3. Amazon rating for a product
4. Olympic high jump
5. Olympic floor gymnastics score

Can you think of an example of an interval variable?

Descriptive or summary statistics

The goal is to describe a dataset with a small number of statistics or figures

Suppose we are given a “sample” or collection of variables,

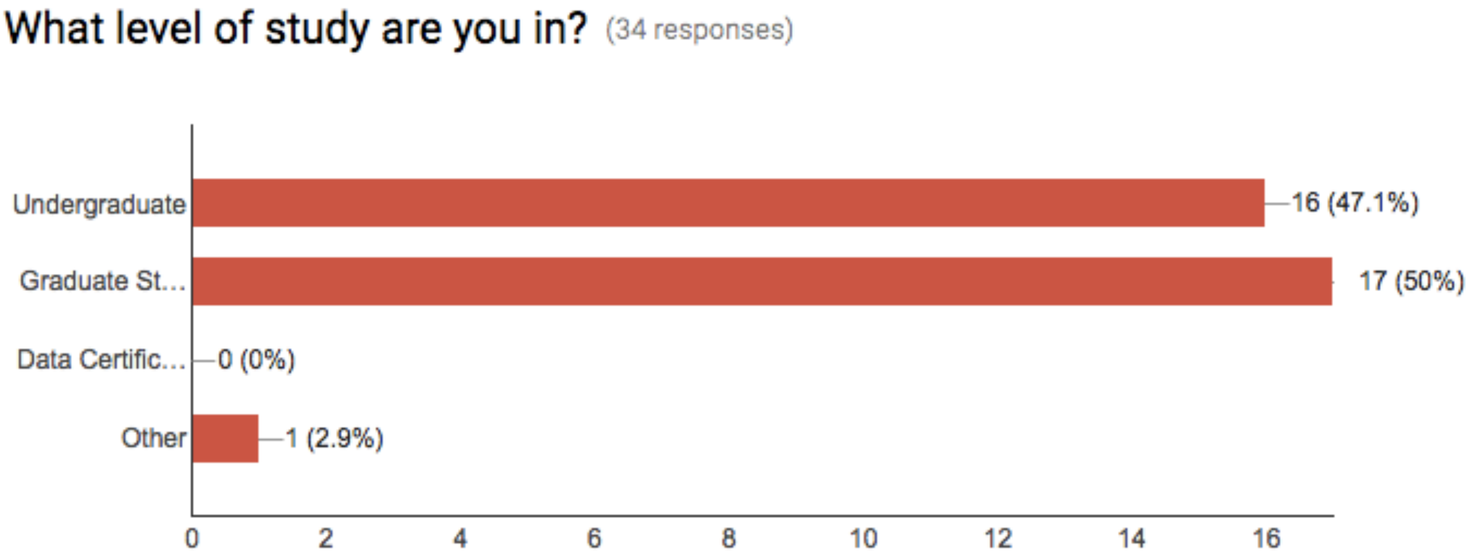
$$x_1, x_2, \ldots, x_n$$

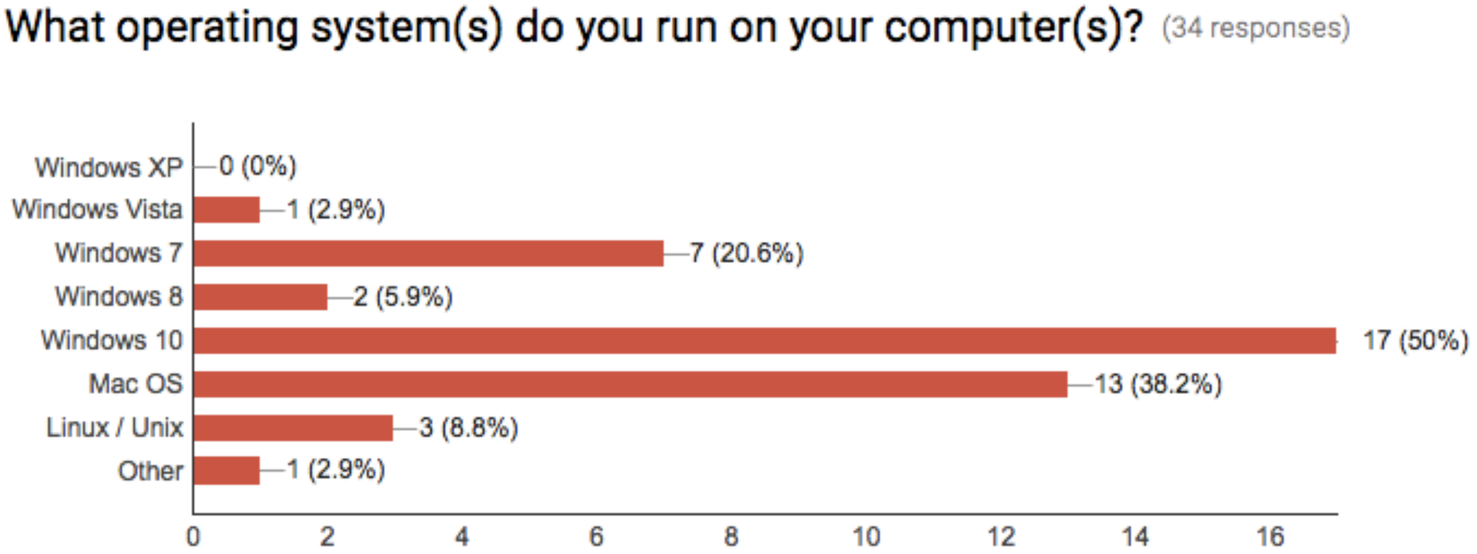
To describe the sample, we might give the sample size (n), max, min, median, or mean

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

Figures include histograms, pie charts, boxplots, scatter plots, …

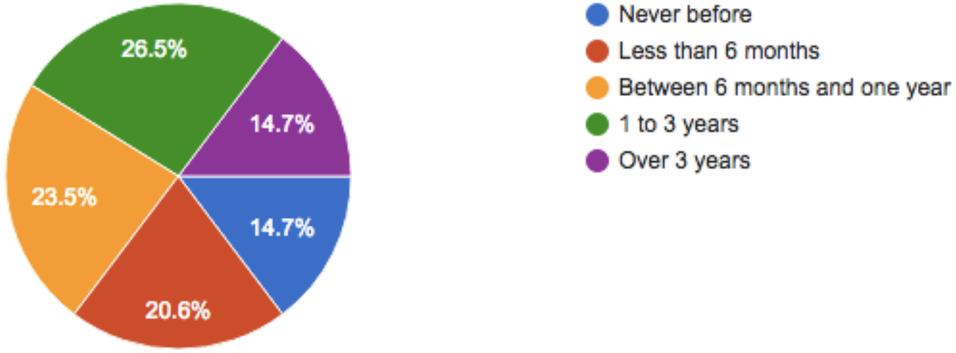
Description of hw0 survey results…



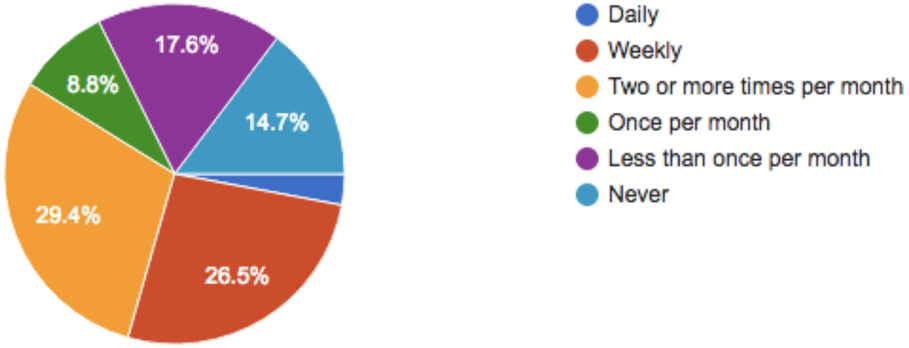


Description of hw0 survey results…

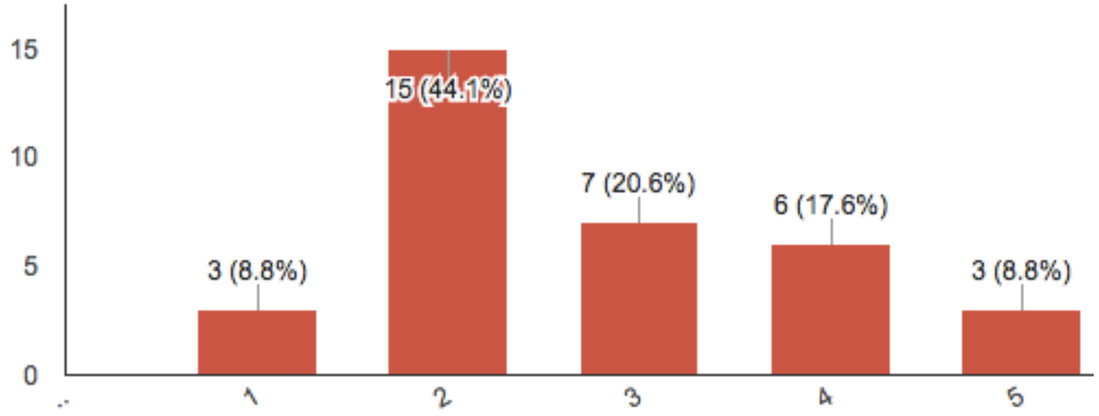
How long have you been programming? (34 responses)



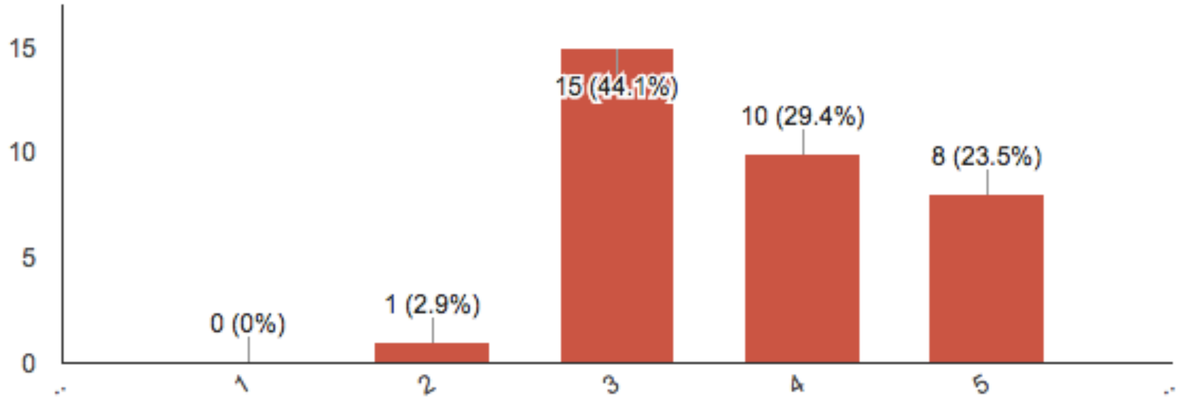
How often do you write code? (34 responses)



Overall, how comfortable are you with programming? (34 responses

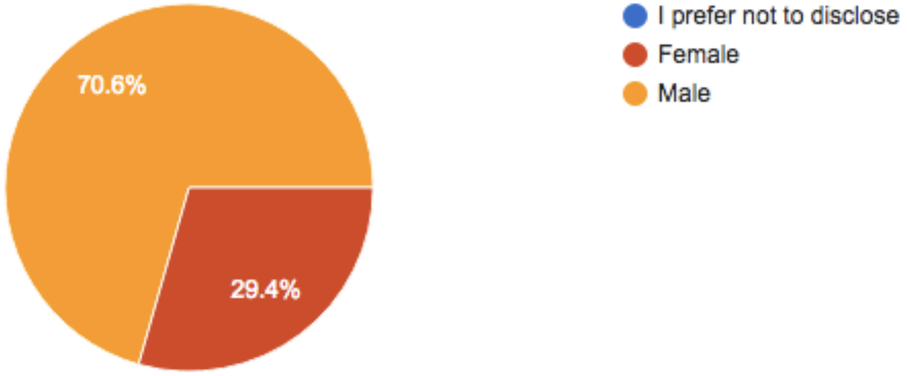


Overall, how comfortable are you with math/statistics? (34 responses)

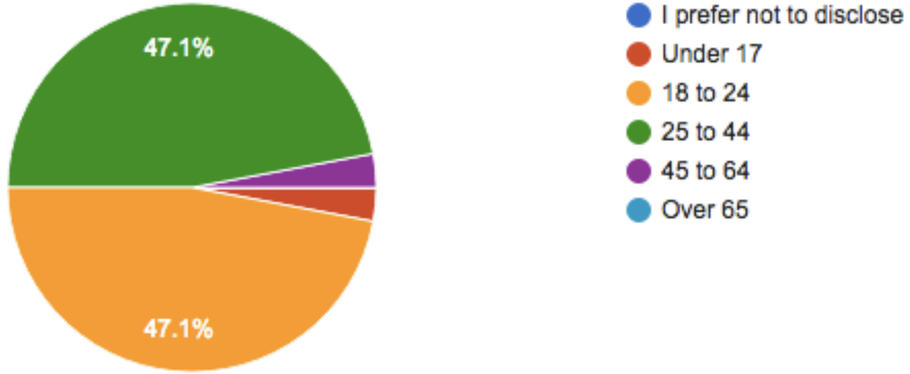


Description of hw0 survey results…

Gender (34 responses)



Age (34 responses)



Description of hw0 survey results…

Why did you decide to take the data science course?

- It is extremely relevant to the work I do as a graduate student
- I have 600 GB of data recorded and now I need to figure out what to do with it.
- To get a hands-on introduction to data science
- I want to explore a career in data science.

Statistics in python?

We'll use the following python libraries with built-in statistical functions:

- SciPy (https://www.scipy.org/); see scipy.stats
- pandas (http://pandas.pydata.org/)
- scikit-learn (http://scikit-learn.org/stable/)

Descriptive Statistics in Python

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib notebook
```

In [2]:
```
# Utah Utes football team
# 2015 offensive scores
# https://en.wikipedia.org/wiki/2015_Utah_Utes_football_team
utah_2015_off_scores = [24, 24, 45, 62, 30, 34, 24, 27, 34, 30, 9, 20, 35] # list of integers
type(utah_2015_off_scores), type(utah_2015_off_scores[0])
```

Out[2]:
```
(list, int)
```

In [3]:
```
np.min(utah_2015_off_scores), np.max(utah_2015_off_scores)
```

Out[3]:
```
(9, 62)
```

In [4]:
```
len(utah_2015_off_scores)
```

Out[4]:
```
13
```

In [5]:
```
np.mean(utah_2015_off_scores) # in python 3.x, same as sum(made_up_data)/len(made_up_data)
```

Out[5]:
```
30.615384615384617
```

* In hw 1, you will write functions to compute the mean, median and other descriptive statistics

Ages from the 1994 U.S. Census

In [8]:
```
# import data from the following dataset
# https://archive.ics.uci.edu/ml/datasets/Adult
#
# "ages" is a list containing ages of people in 1994 Census

data = pd.read_table("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data", sep=",",
                     names=("age", "type_employer", "fnlwgt", "education", "education_num", "marital",
                            "occupation", "relationship", "race","sex","capital_gain", "capital_loss",
                            "hr_per_week","country", "income"))
ages = data["age"].tolist()
```

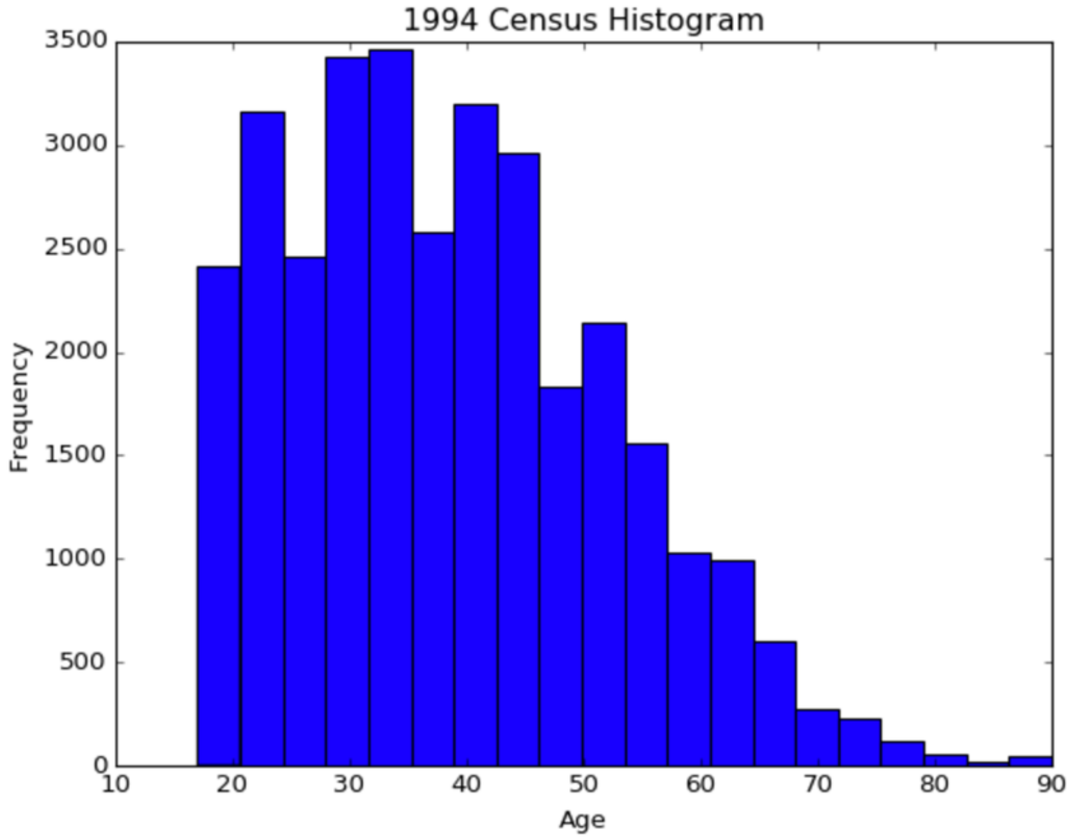
In [9]:
```
print(len(ages))
print(np.min(ages))
print(np.max(ages))
print(np.mean(ages))
print(np.median(ages))
```

```
32561
17
90
38.5816467553
37.0
```

These descriptive statistics gives us some idea of what the data looks like, but a histogram is much more … descriptive.

Histogram of data

In [11]:
```
plt.hist(ages,20)
plt.title("1994 Census Histogram")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```



Quantiles

Quantiles describe what percentage of the observations in a sample have smaller value

```
In [12]: np.percentile(ages,25), np.percentile(ages,75)
Out[12]: (28.0, 48.0)
```

For this data, 25% of the people are under 28 yr.

The middle 50% of the data (the data between the 25% and 75% quantiles) is between 28 yr. and 48 yr.

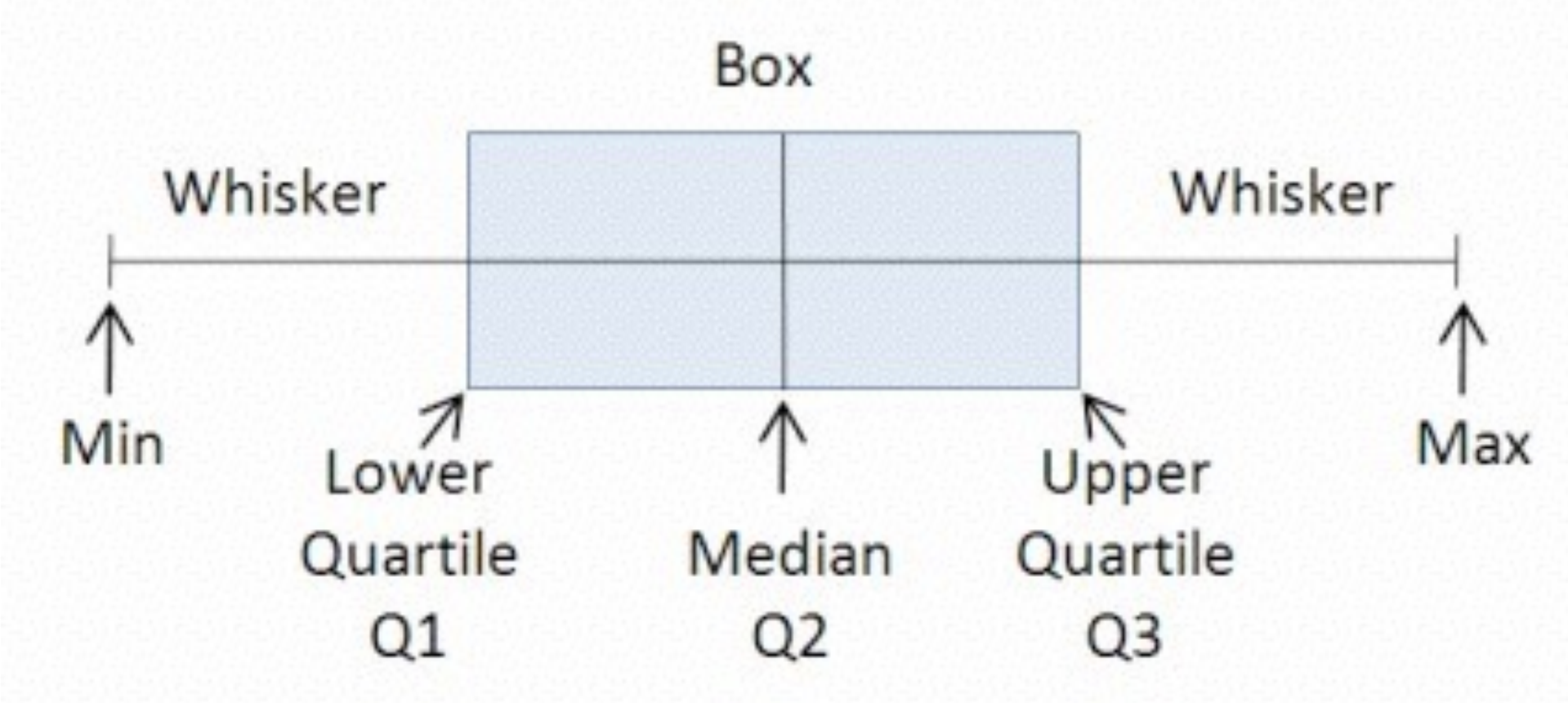
Question: how do I read off quantiles from histogram?

SAT Composite Score Range	Percentile Score
1550-1600	99+
1500-1550	98 to 99
1450-1500	97 to 98
1400-1450	94 to 97
1350-1400	91 to 94
1300-1350	86 to 91
1250-1300	80 to 86
1200-1250	72 to 80
1150-1200	64 to 72
1100-1150	55 to 64
1050-1100	44 to 55
1000-1050	34 to 44
950-1000	25 to 34
900-950	18 to 25
850-900	12 to 18
800-850	7 to 12
750-800	4 to 7
700-750	2 to 4
650-700	1 to 2
600-650	1
550-600	1
500-550	1
450-500	1
400-450	1

SAT quantiles

Boxplot

The *box plot* or *box and whisker diagram* shows several descriptive statistics: minimum, first quartile, median, third quartile, and maximum.



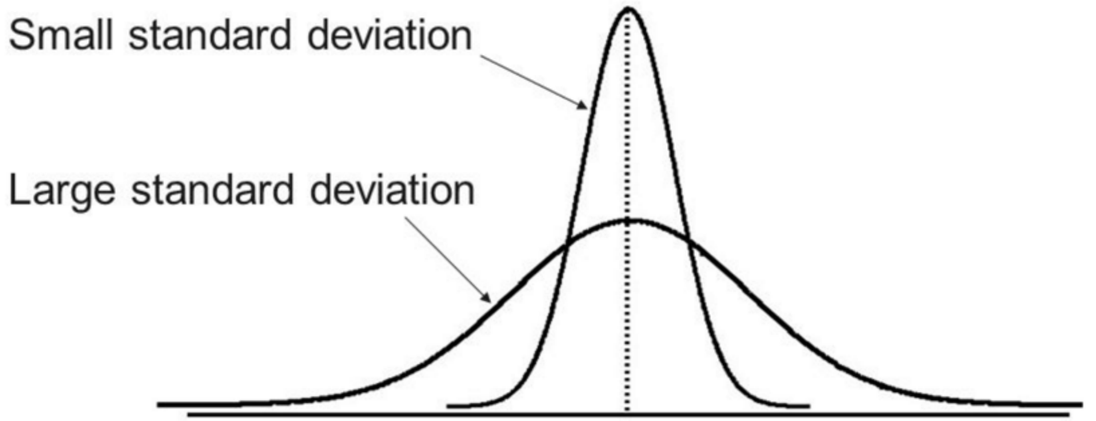
Sample Variation and Standard Deviation

Variance and standard deviation quantify the amount of variation or dispersion of a set of data values.

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

Mean = $\bar{x}$

Std. dev. = s



In terms of the histogram …

Covariance and Correlation

Covariance and correlation measure of how much two variables change together.

Suppose for each item, we collect two variables: x_i & y_i

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

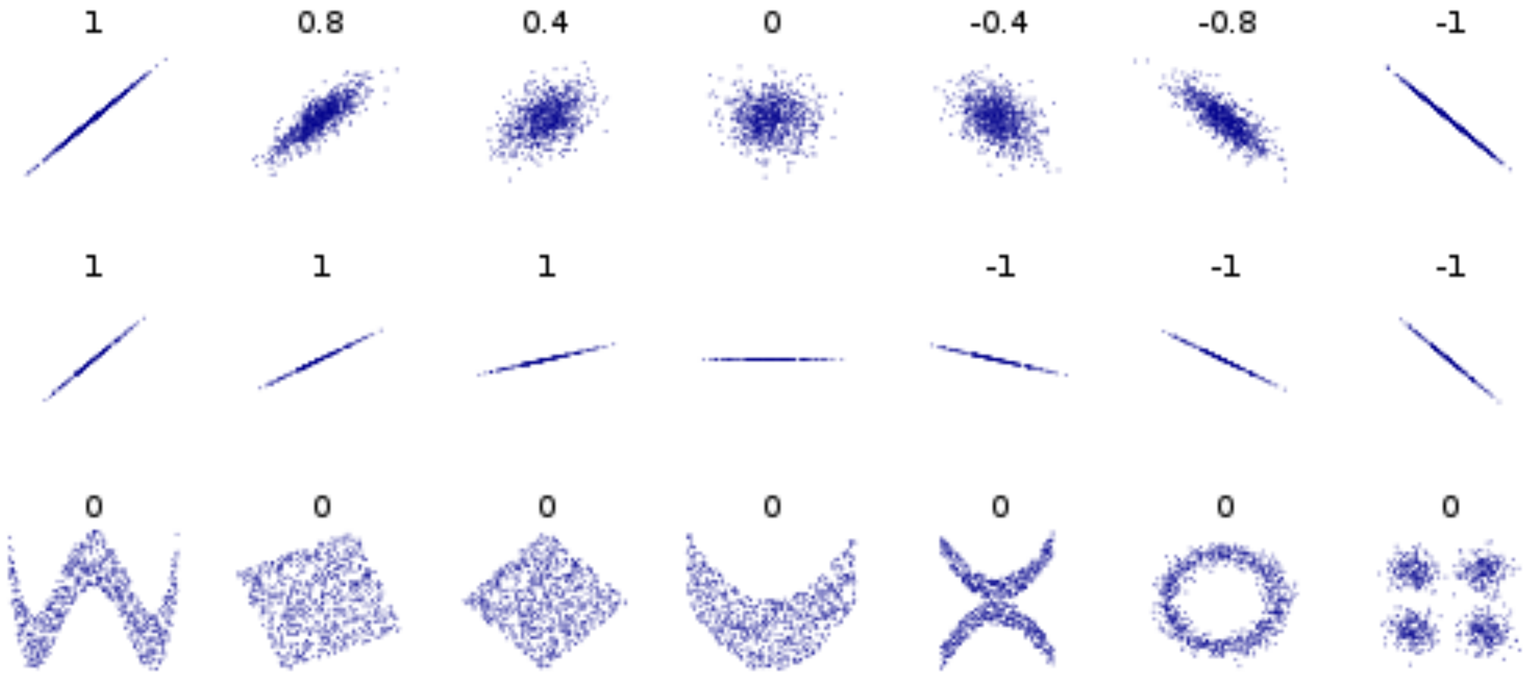
$\bar{x}$ is the mean of x_i

$\bar{y}$ is the mean of y_i

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{s_x s_y}$$

s_x is std. dev. of x_i

s_y is std. dev. of y_i

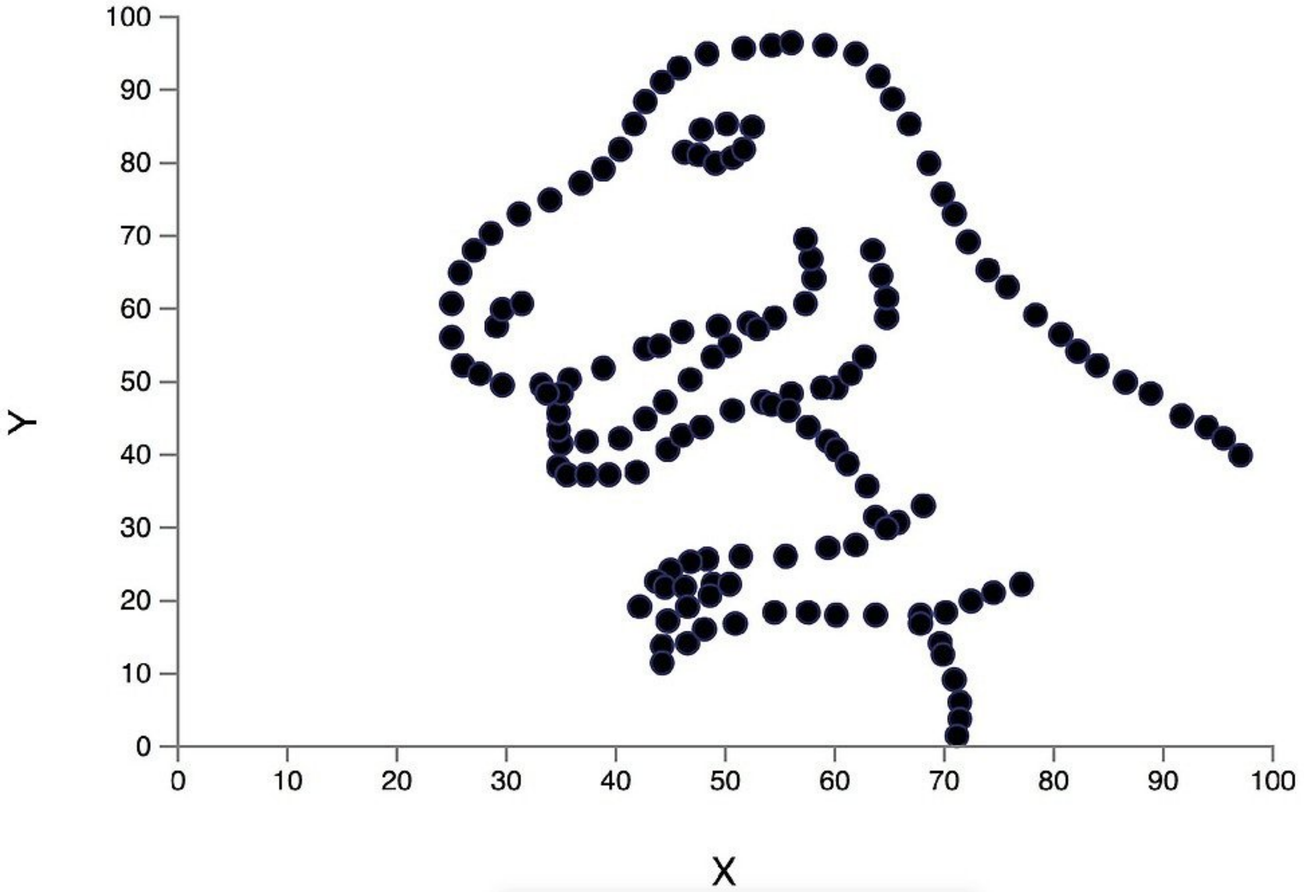


Correlations of two variables

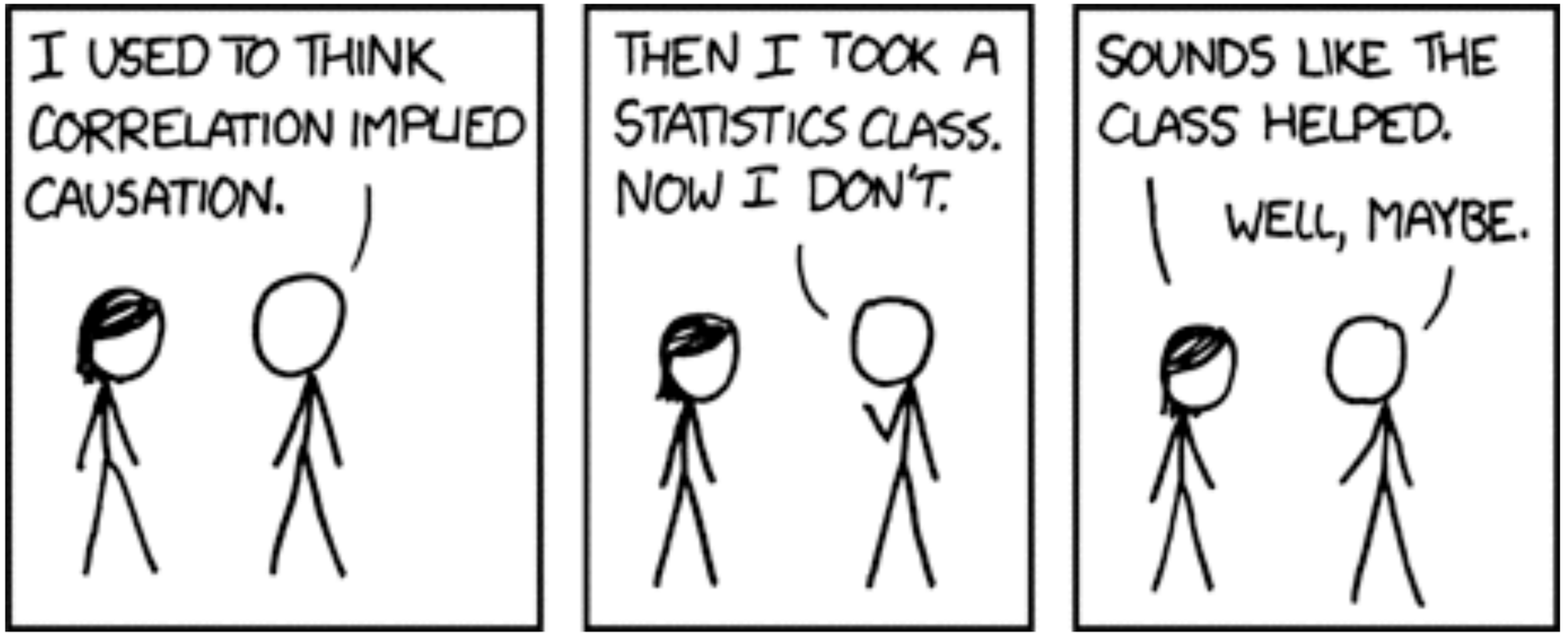
Source: Wikipedia

“Never trust summary statistics alone. Always visualize your data.”

N = 146 ; X mean = 53.7355 ; X SD = 15.7801 ; Y mean = 49.0995 ; Y SD = 24.0605 ; Pearson correlation = -0.143

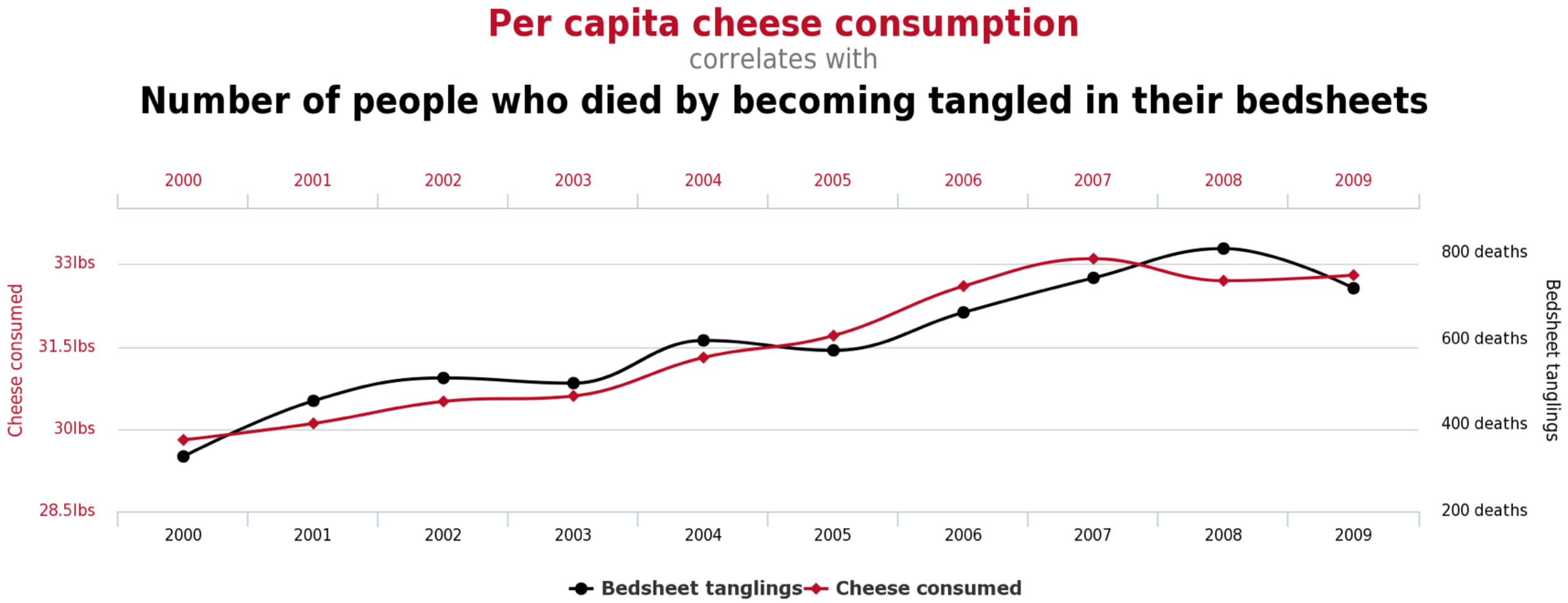


Correlation vs Causation



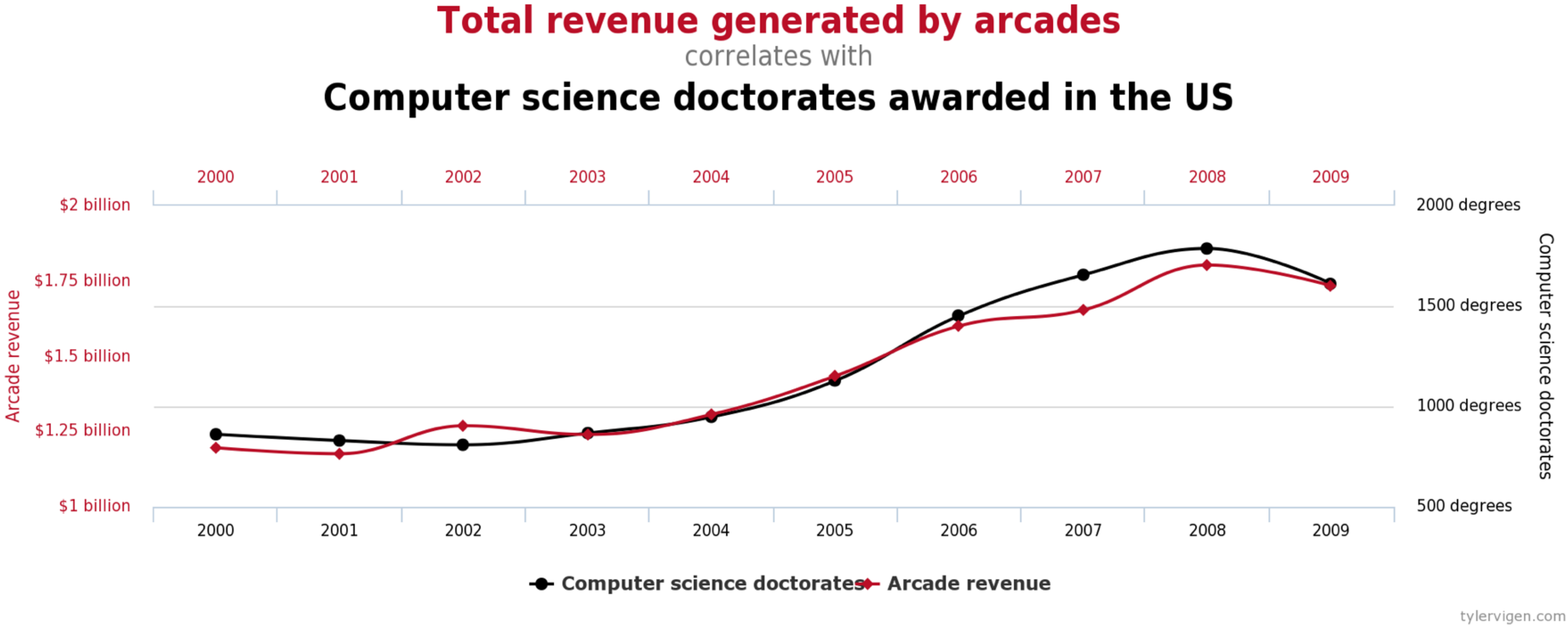
Source: XKCD Comics

Spurious Correlations



Source: http://www.tylervigen.com

Spurious Correlations



Source: http://www.tylervigen.com

Confounders: example

Suppose we are given city statistics covering a four-month summer period, and observe that swimming pool deaths tend to increase on days when more ice cream is sold.

Should we conclude that ice cream is the killer?

Source: doi:10.1371/journal.pone.0152719

Confounders: example cont.

No!

As astute analysts, we identify average daily temperature as a *confounding variable*: on hotter days, people are more likely to both buy ice cream and visit swimming pools.

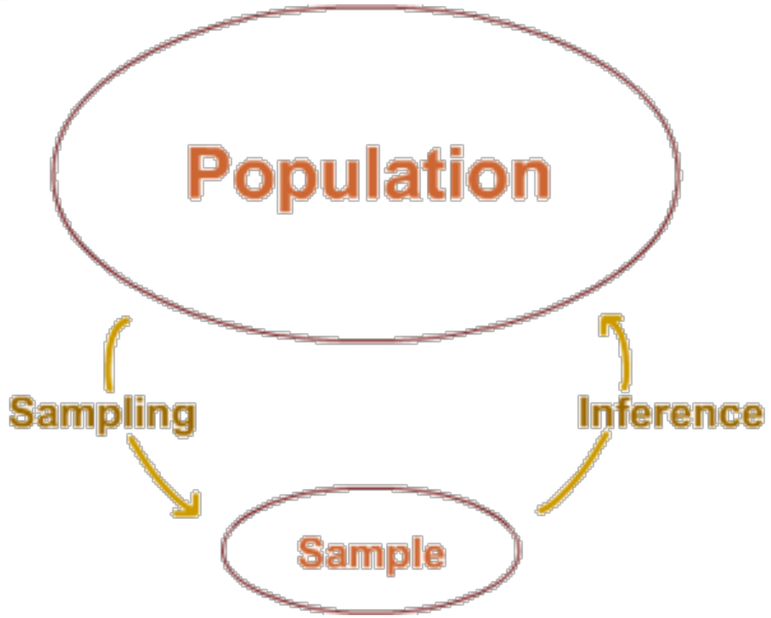
Regression methods can be used to statistically control for this confound, eliminating the direct relationship between ice cream sales and swimming pool deaths.

Source: doi:10.1371/journal.pone.0152719

Descriptive vs. Inferential Statistics

Descriptive statistics quantitatively describe or summarize features of a dataset.

Inferential statistics attempt to learn about the population that the sample of data is thought to represent.



Hypothesis testing (next lecture) uses inferential statistics.