Introduction to Stats

MSMI Bootcamp

Lucas De Oliveira

Data Science:

Data Scientist at Capgemini

Formerly D.S. at Nextracker

Education:

M.S. in Data Science from USF

B.A. in Economics from UVA

Bootcamp Files

All files can be found in the following repository:

https://github.com/lbdeoliveira/MSMI_Bootcamp

Statistics

The practice or science of collecting and analyzing **numerical data in large quantities**, especially for the purpose of **inferring proportions in a whole** from those in a **representative sample.**

- *Google via Oxford definitions*

Distributions

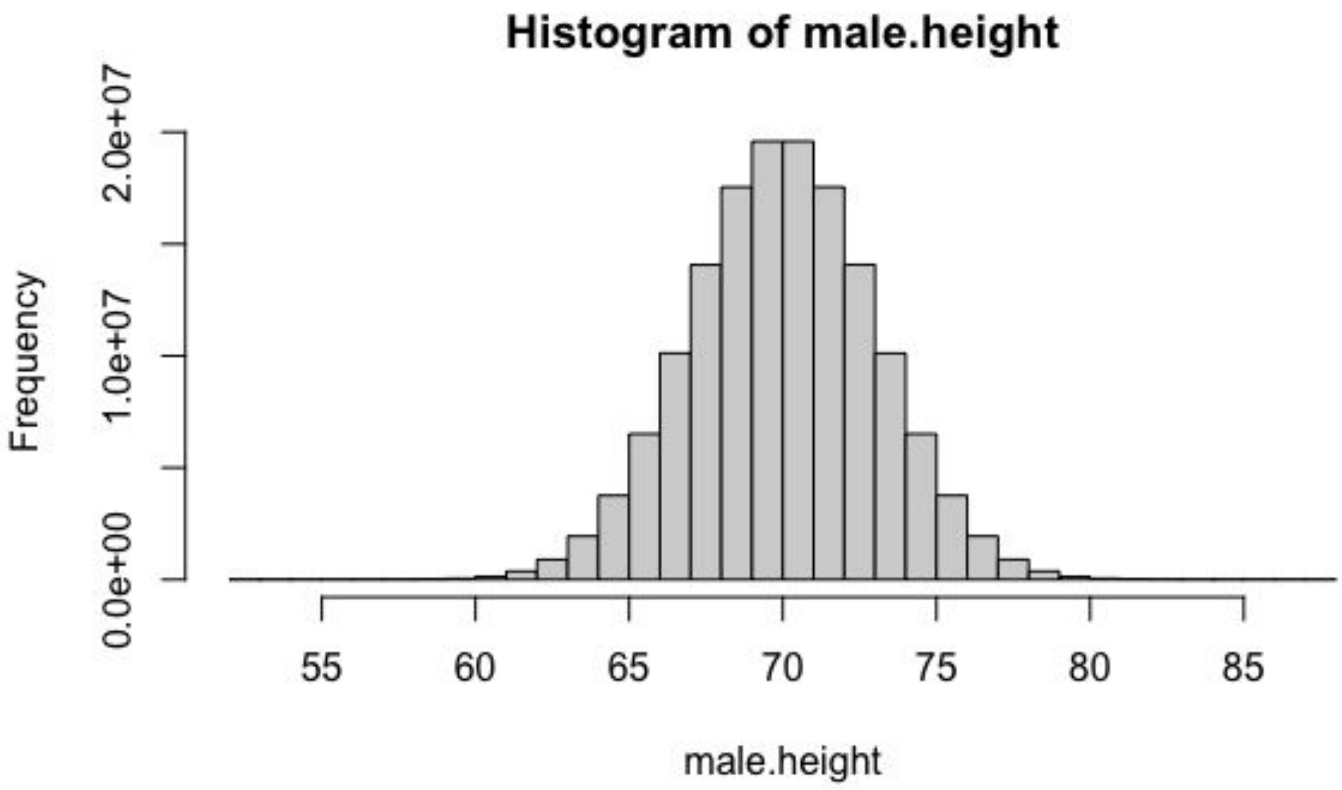
- "Large quantities of data" means we can't possibly look at every data point
- Need some way of mapping the values of our data to the frequencies with which they occur
- By looking at the distribution of the data:
    - We know the **range** of values we can expect
    - How much **variation** there can be from observation to observation
    - Which values are the **most common** and which are rare
    - If there is an error in the data (**outliers**)
    - Which tests and **models** are appropriate for analytic task
- Best tools for understanding the distributions:
    - Histograms & boxplots
    - Summary statistics

Which of the following slides represent 150M observations in the most digestible way?

The heights of adult men in the United States are approximately normally distributed with a mean of 70 inches and a standard deviation of 3 inches.

```
> male.height
  [1] 71.06754 73.11451 61.84861 71.04306 66.60774 75.33142 74.62094 72.63692
  [9] 64.34013 64.23323 61.76728 70.61079 70.57338 70.43300 72.92783 67.63482
 [17] 68.33232 70.99352 65.63758 69.36915 77.09928 70.62279 69.48665 65.95915
 [25] 66.30630 68.19937 71.53821 67.84339 67.32479 69.39501 69.63770 67.74464
 [33] 69.49204 67.01336 69.78390 64.64162 69.86481 70.43780 70.06455 73.07450
 [41] 73.60666 68.70964 65.71801 70.90021 72.76190 72.49512 71.40214 76.93740
 [49] 69.35089 66.96217 66.10697 73.27999 70.53302 69.56296 69.75982 66.63332
 [57] 66.56938 70.55239 70.04989 67.12970 78.05573 72.51517 66.59052 66.86070
 [65] 73.38959 68.49808 66.77838 71.77463 73.20678 73.49998 74.41677 70.78494
 [73] 66.86340 71.10229 70.08992 68.68983 68.80659 67.82333 69.32941 67.79487
 [81] 63.25352 67.21232 73.70053 74.33421 69.89310 72.34558 75.81894 72.83341
 [89] 70.43177 70.94397 71.70117 70.86109 69.67617 68.50836 72.90769 72.83202
 [97] 66.50942 66.82059 71.89445 71.22156 72.20907 75.86663 74.74785 71.32591
[105] 69.09332 67.03486 76.53890 67.75556 68.84377 69.61181 70.35590 60.40628
[113] 67.93395 69.47696 70.98217 68.35632 73.42198 64.21601 71.21808 73.28946
[121] 67.71032 65.44334 71.00470 72.37588 67.76722 70.09666 64.83666 67.23576
[129] 67.08049 70.82189 75.41077 74.14673 66.40636 67.30648 71.17763 68.42656
[137] 70.09483 73.74262 72.97166 71.95162 67.87660 66.01770 73.83886 78.20296
[145] 68.87451 73.71010 65.94345 66.40291 70.37389 67.78443 69.39418 68.62636
[153] 70.49621 71.20515 67.94184 67.64369 64.54484 73.52270 66.04990 71.88272
[161] 65.88620 70.95282 67.16262 69.80138 71.26670 69.07512 69.82020 69.88145
[169] 73.95875 65.10009 66.07519 67.79269 71.91866 63.99768 69.40658 66.78826
[177] 77.58736 70.65846 74.24414 69.49611 68.90296 65.32827 72.41118 70.15645
[185] 62.56113 76.69300 73.57599 66.17830 70.25872 70.24269 70.68406 64.01806
[193] 70.31706 67.48835 73.89678 70.42321 75.44047 70.53271 70.77638 71.26605
[201] 77.16321 69.43348 71.61483 67.81959 68.07122 73.61544 70.72869 71.33805
```

```
summary(male.height)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
52.68   67.98   70.00   70.00   72.02   87.45
```


Histogram of male.height
Frequency
0.0e+00
1.0e+07
2.0e+07
55
60
65
70
75
80
85
male.height

Visualizing distributions are super useful for understanding our data

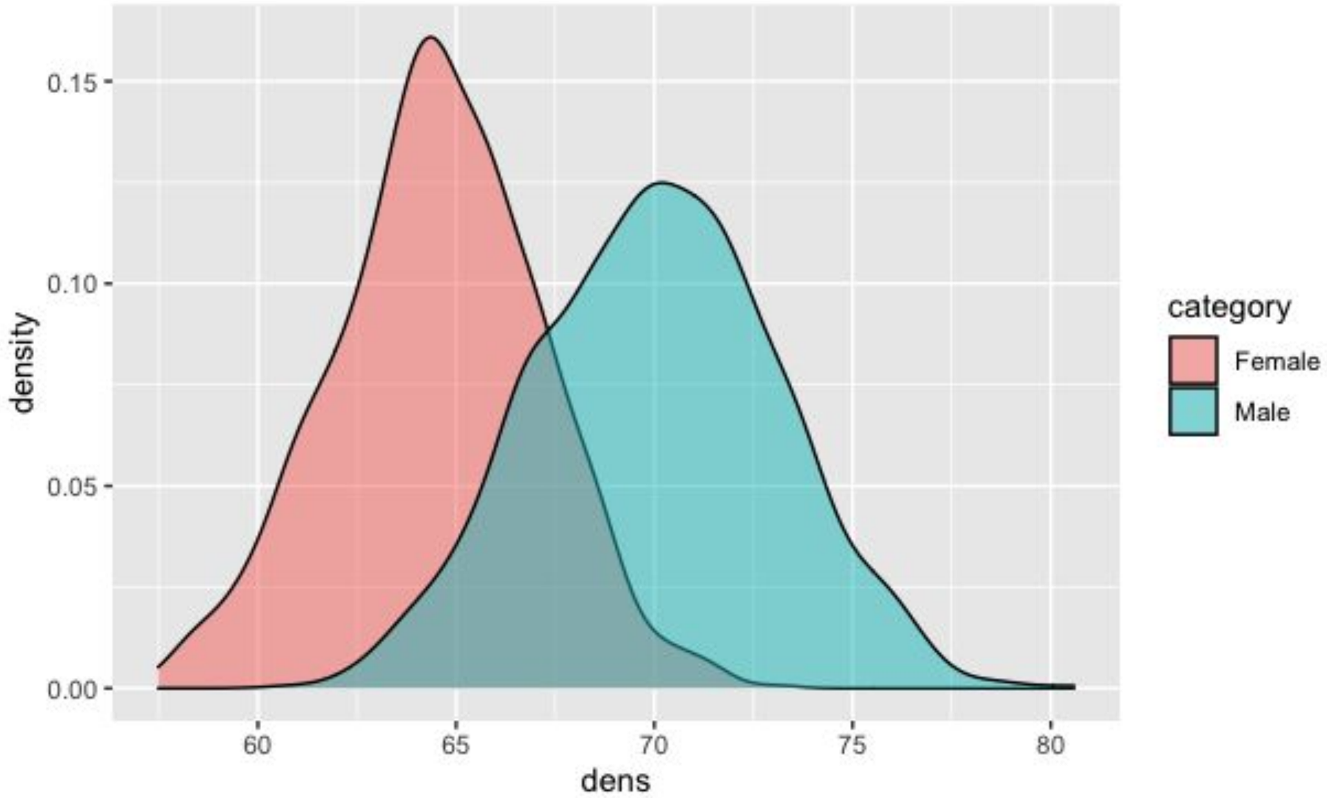
& for comparing
different populations

“The heights of adult men in the United States are approximately normally distributed with a mean of 70 inches and a standard deviation of 3 inches.

Heights of adult women are approximately normally distributed with a mean of 64.5 inches and a standard deviation of 2.5 inches.”

Source

American male v. female heights



The first step in any analysis should be to plot the distribution of your data

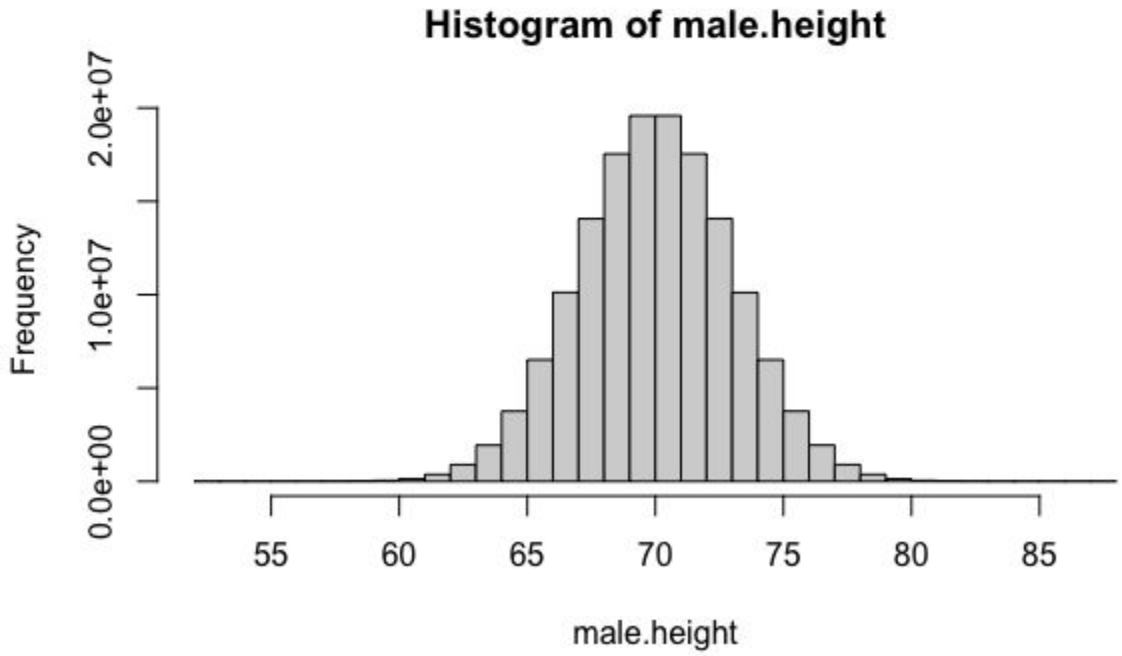
Histogram

- A histogram will build “bins” or “buckets” for your data and plot how many observations fall in those buckets
- Example:

[1, 3, 6, 12, 15, 3, 4, 8, 13, 12, 7, 2, 10, 11, 12]

Bins:	1-5,	6-10,	11-15
Counts:	5,	4,	6

Let’s look at the histogram for male height again



We can also **summarize** our data with a few numbers…

Summary statistics

- Central tendency: *mean, median,* and *mode*
  - Mean - the average value
  - Median - the “middle” value
  - Mode - the most common value
- Percentiles / quantiles: *min, Q1, median, Q3, max*
  - Min - the lowest value
  - Q1 - the 25th percentile
  - Q3 - the 75th percentile
  - Max - the highest value
- Measures of spread:
  - Variance - how “wide” is the distribution? The larger the variance, the higher the spread.
  - Standard deviation - square root of the variance

For the list of numbers:

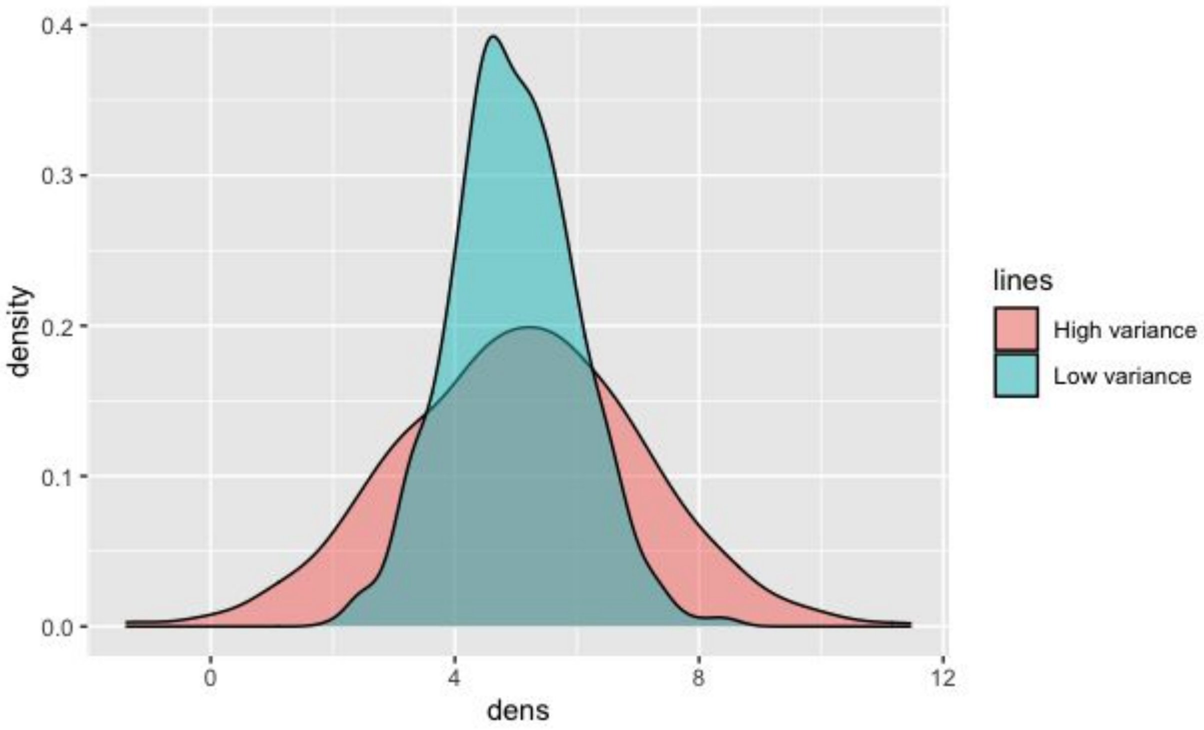
[1, 3, 6, 12, 15, 3, 4, 8, 13, 12, 7, 2, 10, 11, 12]

Calculate:

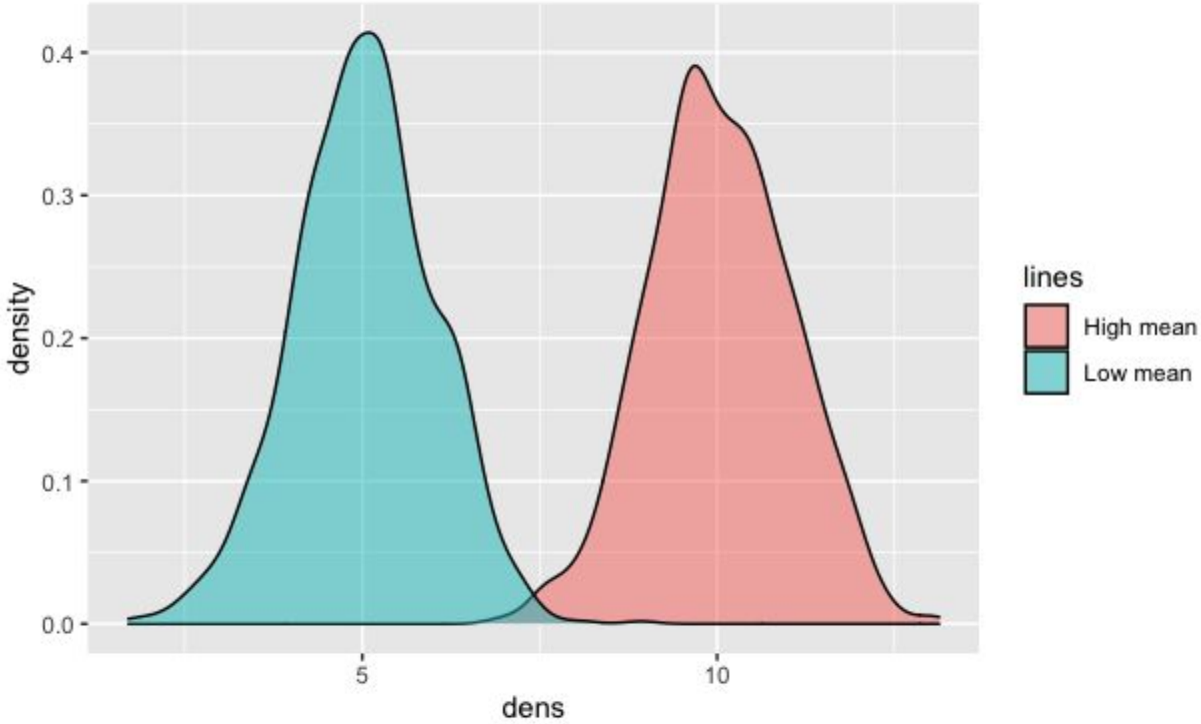
- Mean, median, mode
- Min, Q1, Q3, and max

Visualizing mean & variance

Same means, different variance

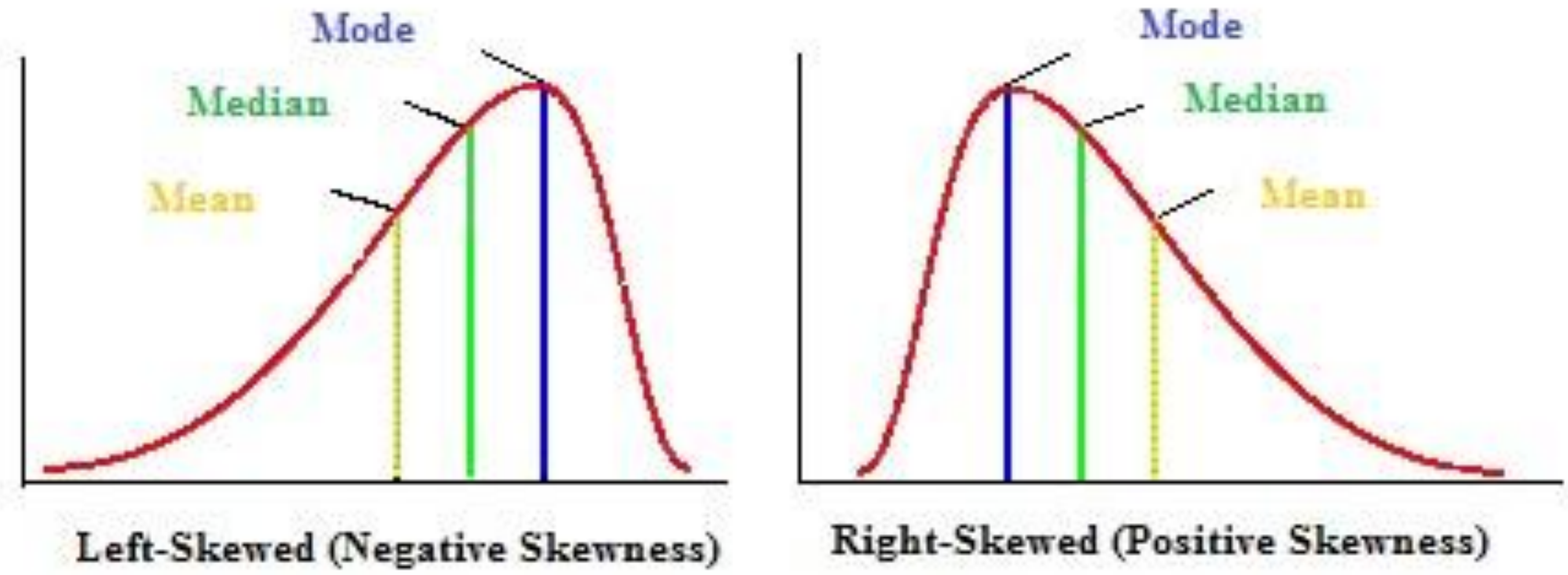


Different means, same variance



Distributions are not always symmetrical

Source


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

“inferring **proportions in a whole** from those in a **representative sample**”

The fundamental purpose of statistics is to make an inference about a **population** by computing **statistics** about one or multiple **samples.**

Population and samples

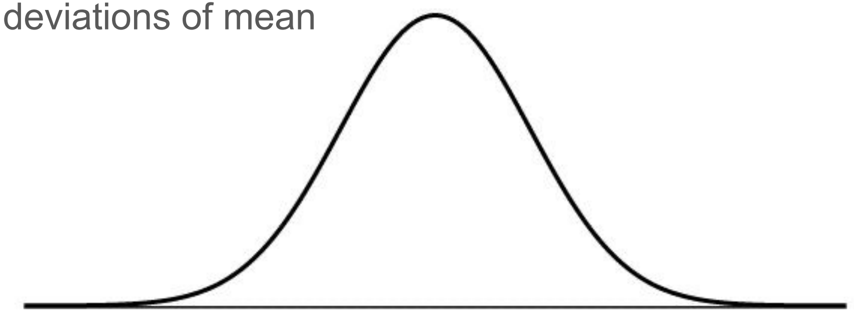
- We can never hope to actually measure all individuals or objects in a **population**
- A population is the group relevant to our study (ex: all american males, students in San Francisco, etc.)
- We want to understand:
  - The population's distributions
  - The **parameters** that describe that distribution

Recall:

The heights of adult men in the United States are approximately **normally distributed** with a **mean of 70 inches** and a **standard deviation of 3 inches**.

The normal distribution

- The most common distribution you will encounter
- “Bell curve”, symmetric
- Described by two parameters: mean and standard deviation/variance
- Empirical rule:
    - 68% of observations fall within 1 standard deviation of mean
    - 95% of observations fall within 2 standard deviations of mean
    - 99.7% of observations fall within 3 standard deviations of mean



Example: empirical rule

- Parameters:
  - Mean = 70 inches
  - Standard deviation = 3 inches
- Provide the height intervals that contain:
  - 68% of American males
  - 95% of American males
  - 99.7% of American males

Since we cannot measure all members of a population, we have to estimate the **population parameters** with **sample statistics.**

Sampling

- Randomly measuring a subset of your population of interest
- It is important that the sample observations be **independent** of each other (for instance, if you are trying to get a representative sample of the height of american men, do not survey college basketball players)
- It is also important that the observations be taken from the same distribution (**identically distributed**)
- The standard for a reliable sample is a collection of i.i.d. (independent identically distributed) observations

Estimating population parameters

- We estimate the **population mean (μ)** with the **sample mean ($\bar{x}$)**
- We estimate the **population standard deviation (σ)** with the **sample standard deviation**, also called the **standard error (s)**

How does **sample size** factor into this?

Experiment in R

Upshot: by increasing sample size, we get a better sense of the actual population distribution & better estimates for its parameters

Central Limit Theorem

As **n increases**, the **sample mean** becomes approximately **normally distributed** with mean equal to the **population mean** and variance equal to **population variance divided by n.**

Note:

- The underlying distribution of X does not matter, the distribution of $\bar{X}$ will be normally distributed if N (sample size) is sufficiently large
- If X is already normally distributed, then it does not matter

Examples in R