Machine Learning

Introductory Statistics

Assistant Professor of Data Science
Paul Intrevado, Ph.D.

UNIVERSITY OF SAN FRANCISCO

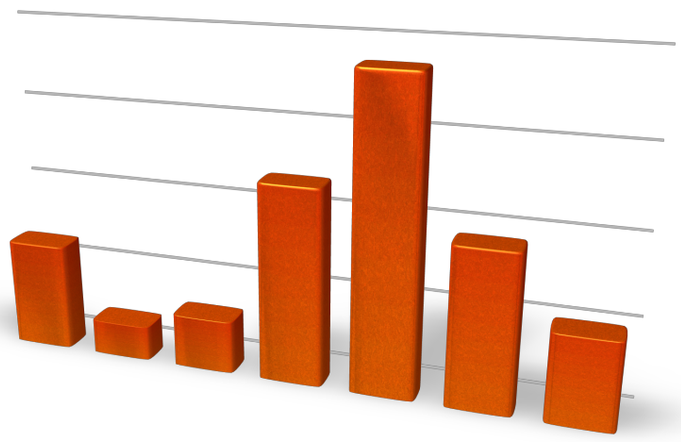
Descriptive Statistics

- Deals with methods of organizing, summarizing and presenting data

Graphical Techniques	**Numerical Techniques**

Pie Charts

Bar Graphs



- Mean
- Median
- Standard Deviation
- Variance
- etc.

Inferential Statistics

- Used to draw conclusions or *inferences* about characteristics of populations based on sample data
  - What is sample data?
    - Example: Exit polling during elections
      - What conclusions can you draw from a small sample of individuals which allows you to make an inference about the general population?

Inferential Statistics

- Four key terms

1. Population: the group of all items of interest
    - Descriptive measure of a ***population*** is a ***parameter*** (e.g., mean)
2. Sample: a set of data drawn from the studied population
    - Descriptive measure of a ***sample*** is a ***statistic*** (e.g., sample mean)

Types of Variables

Interval

- 3.2 miles
- 643 USD
- 43 widgets
- also known as Quantitative or Numerical variables
- ----
- values are numbers
- **all calculations are valid**

Ordinal

- Strongly Disagree, Disagree, Agree, Strongly Agree
- Freshmen, Sophomore, Junior, Senior
- ----
- **calculations based on an ordering process are valid**

Nominal

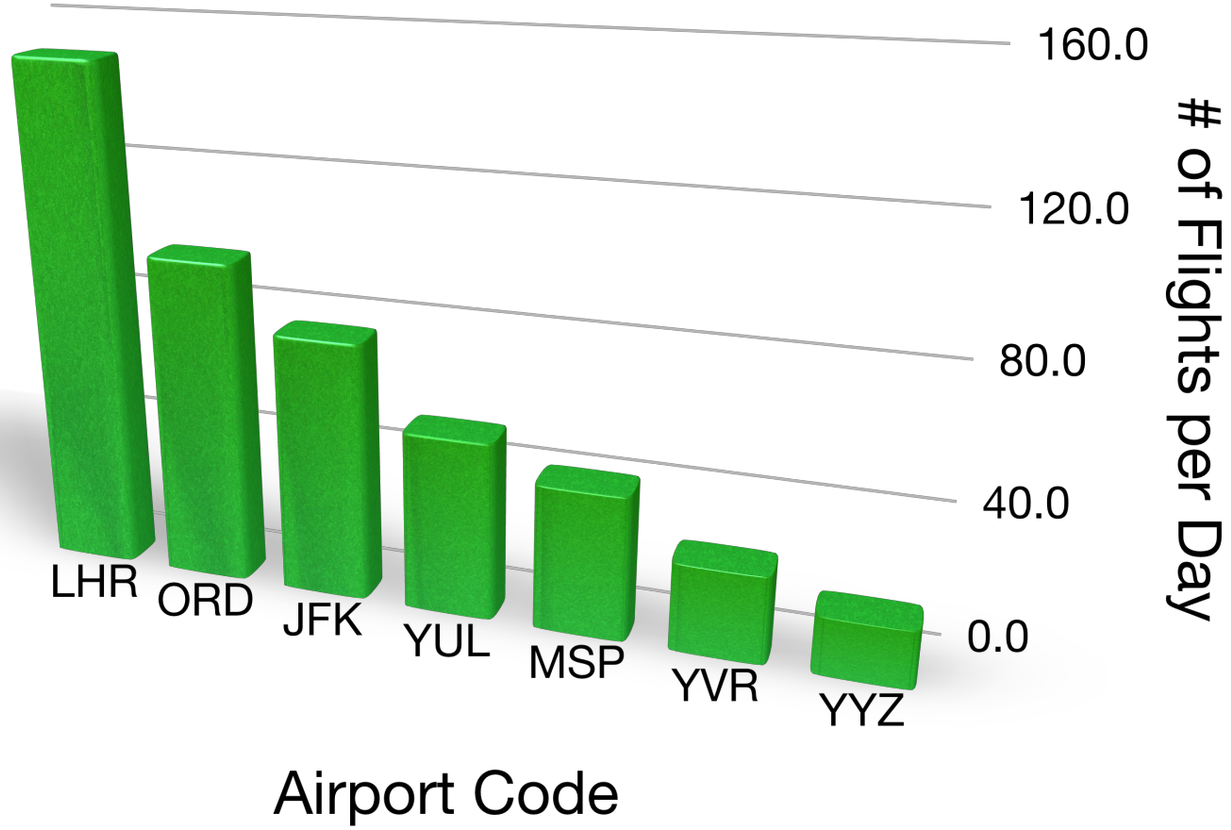
- Orange, Yellow, Brown
- Male, Female
- Finance, HR, Marketing
- also known as Qualitative or Categorical variables
- ----
- **only calculations based on the frequencies or % of occurrence are valid**

Variables *(example)*

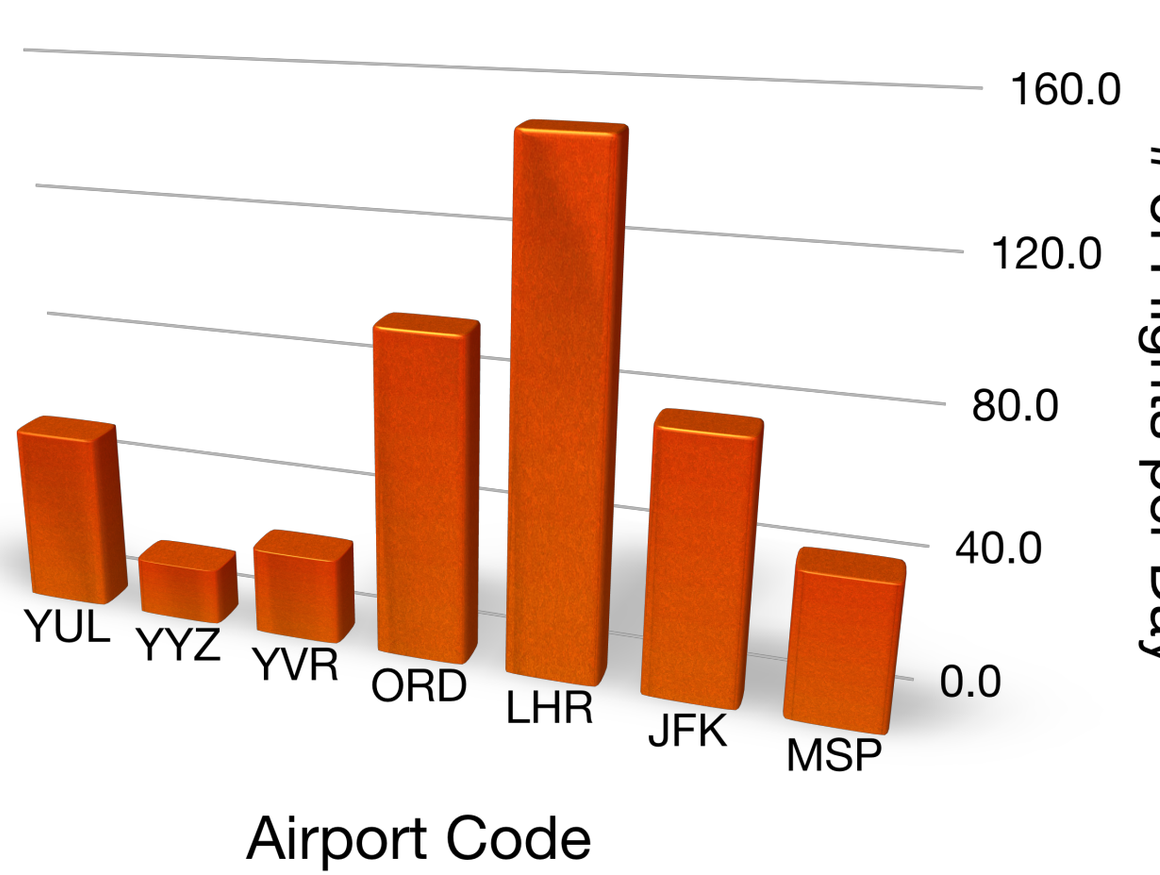
	Nominal	Nominal	Ordinal	Interval
Student	**Gender**	**Program**	**Happy?**	**GPA**
1	Male	Law	Unhappy	3.56
2	Female	Engineering	Very Happy	3.92
3	Male	Business	Happy	3.07
4	Female	Education	Somewhat Unhappy	3.21
5	Male	Medicine	Delighted	2.95
6	Male	Commerce	Depressed	4

Charting Nominal or Ordinal Data

- In bar charts or pie charts, counts (frequency) or percentages (relative frequency) can be represented
- A **Pareto chart** is sorted by rank (high to low)



Pareto Chart

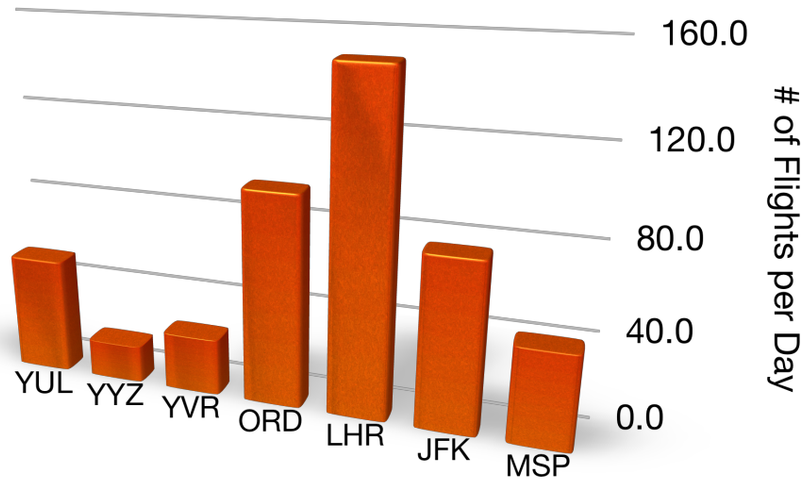


***NOT a* Pareto Chart**

Histograms

- Where as categorical data has natural “buckets...”

*# of flights **per airport***



- ...what about quantitative data?

*# of potato chips **per Lays Classic bag***

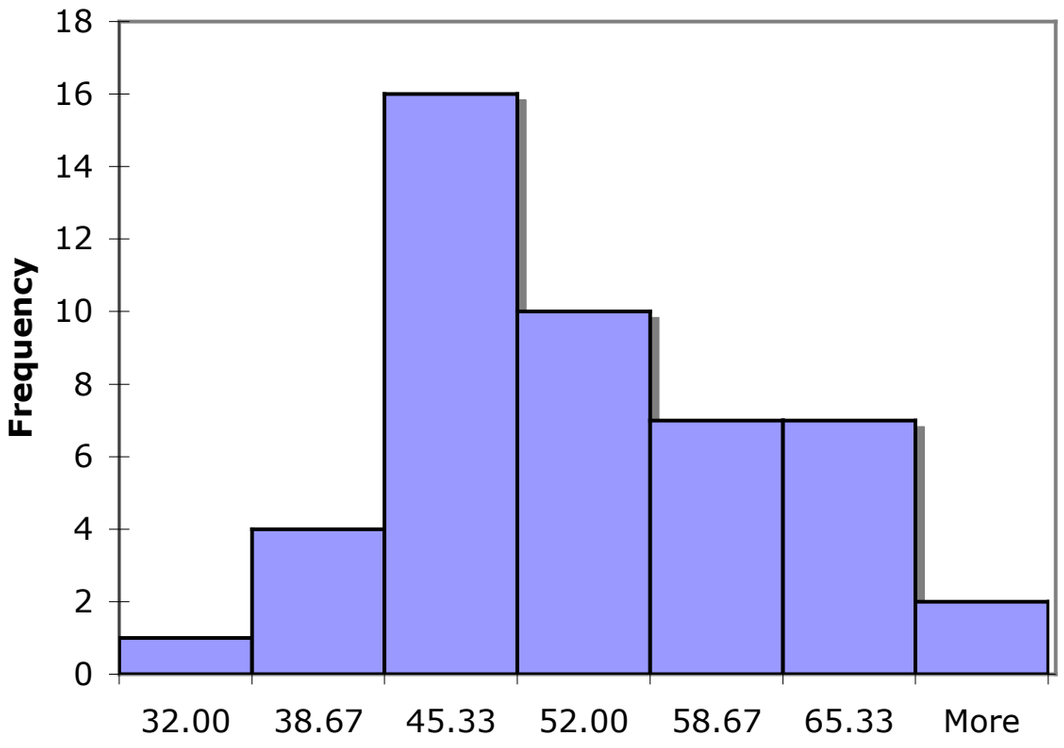
OR

Airport Code

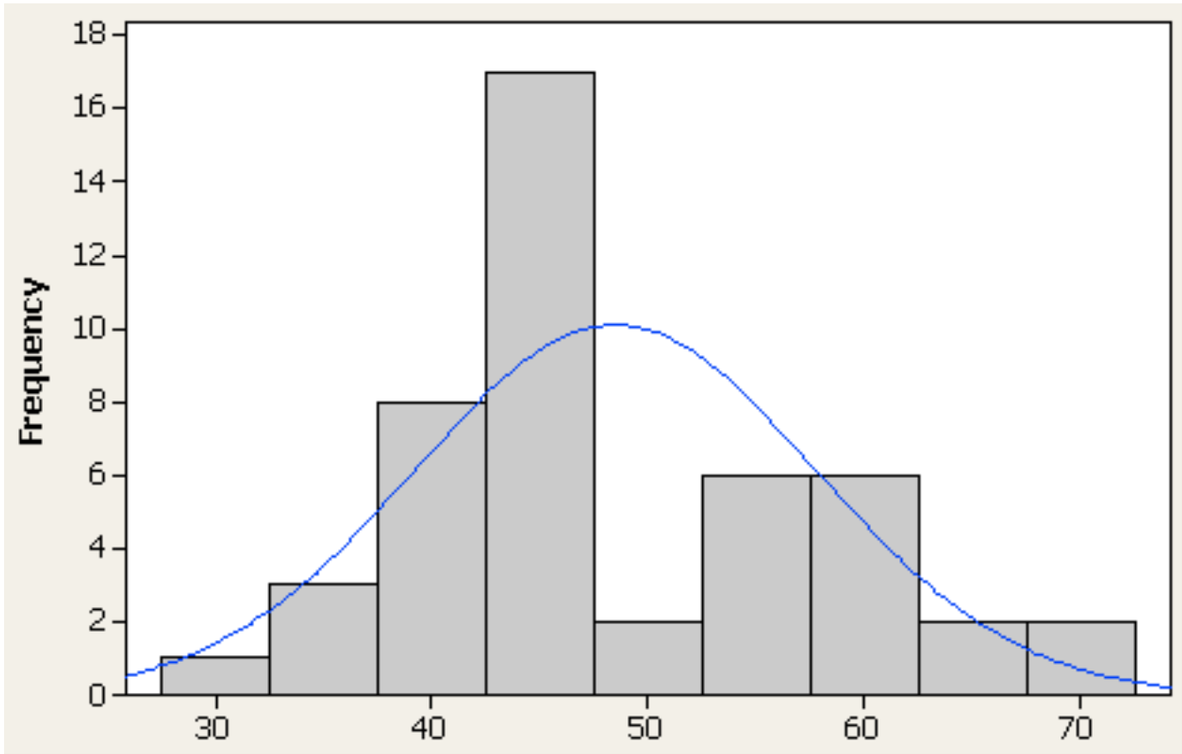
Histograms

- What bin size is the best for histograms?
- Two bar charts of the same data where the software ***automatically*** determined the size of the “buckets”

MS Excel 2004



Minitab *v*14



(bar charts above are not based on potato chip example from previous slides)

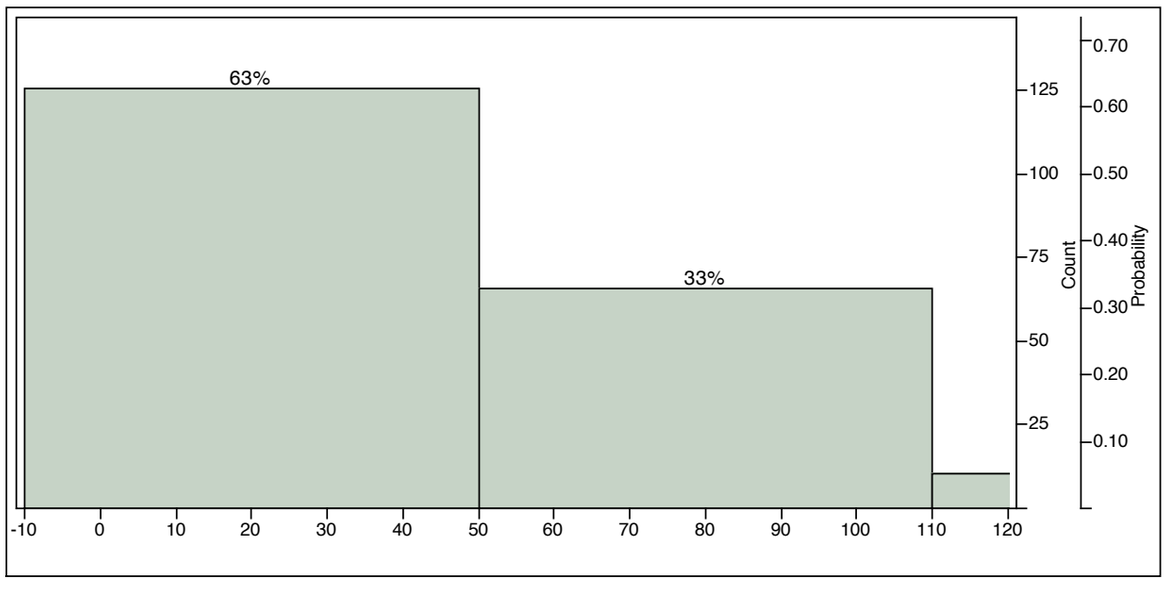
Histograms

- What bin size is the best for histograms?
  - Objective: create bins that best summarize the observed data

---

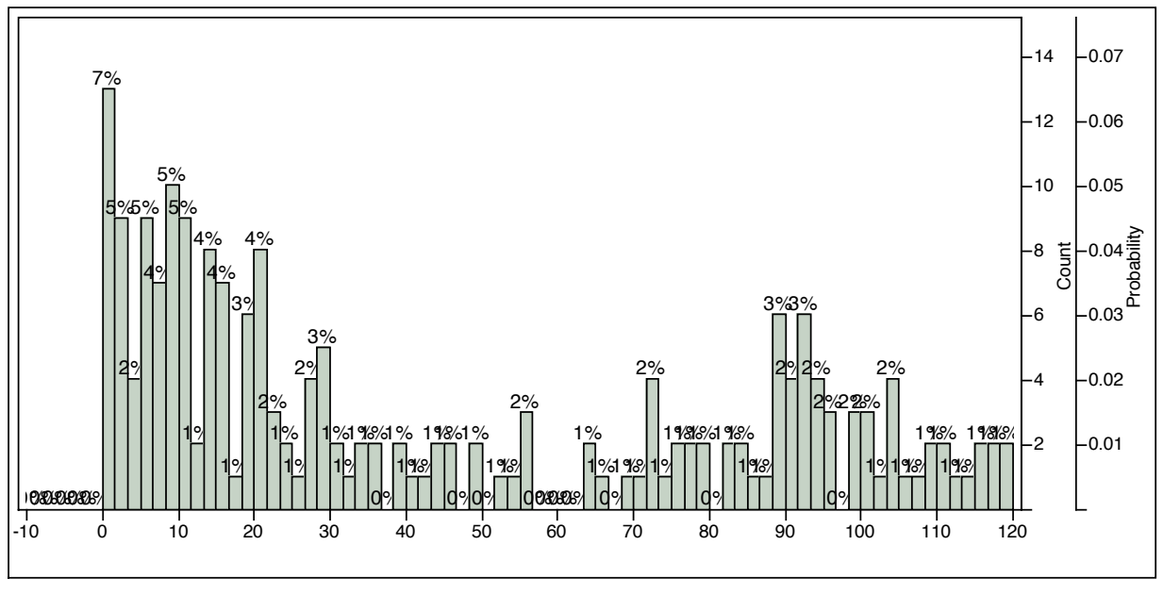
Too few bins

loose information



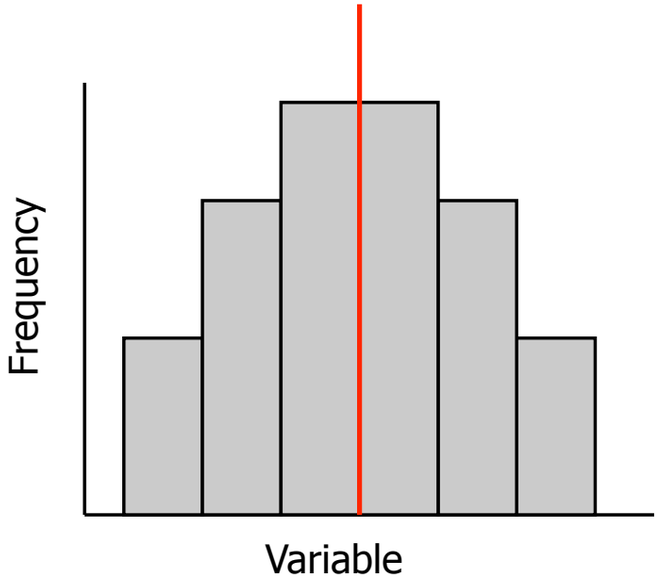
Too many bins

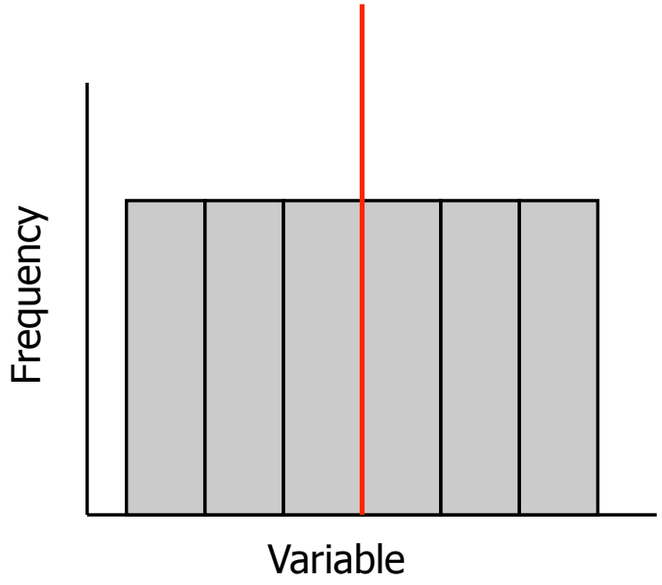
no longer a summary

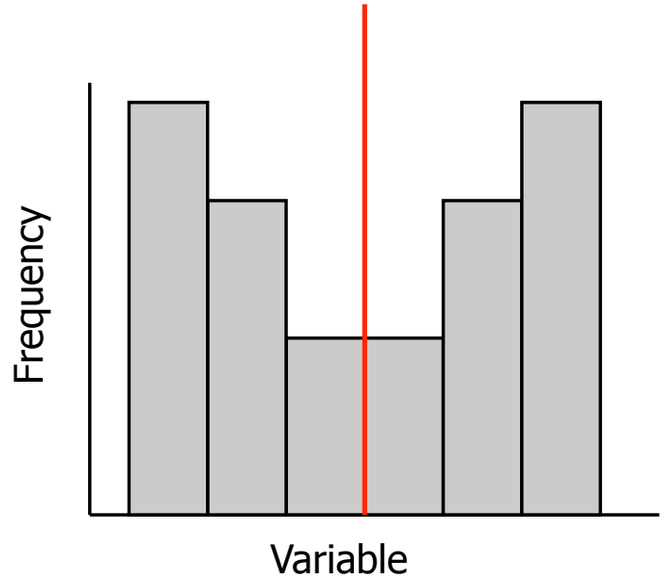


Symmetric Histograms

- Symmetric about their centers
  - Various examples of symmetric histograms

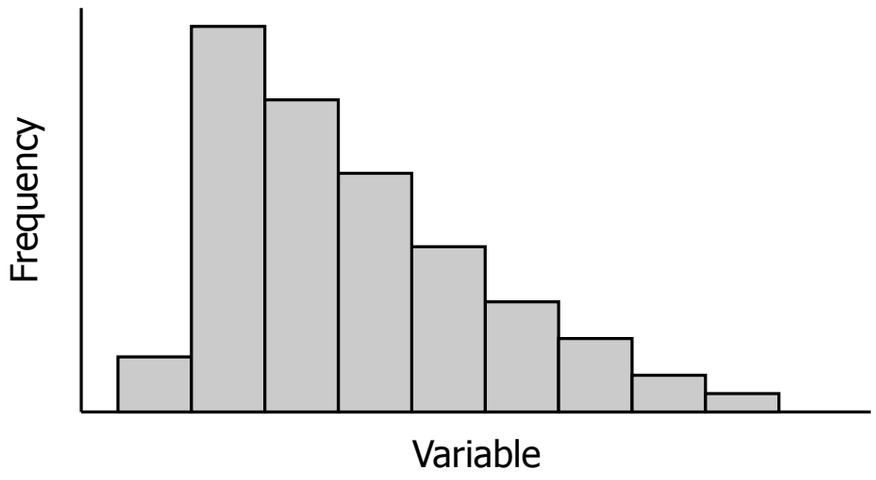




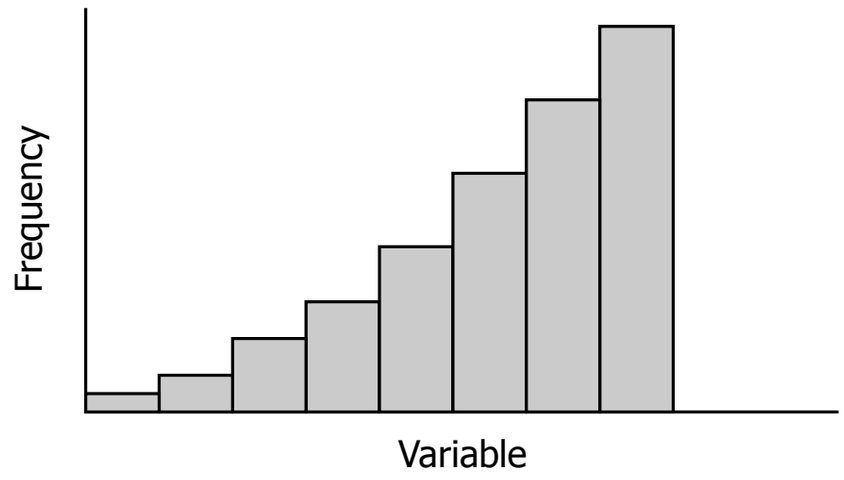


Skewed Histograms

- A histogram is skewed if it has one long tail extending in either the right of left direction
- Positive or Right-Skewed: long right tail, e.g., salaries at large corporations
- Negative or Left-Skewed: long left tail, e.g., amount of time students take to write an exam



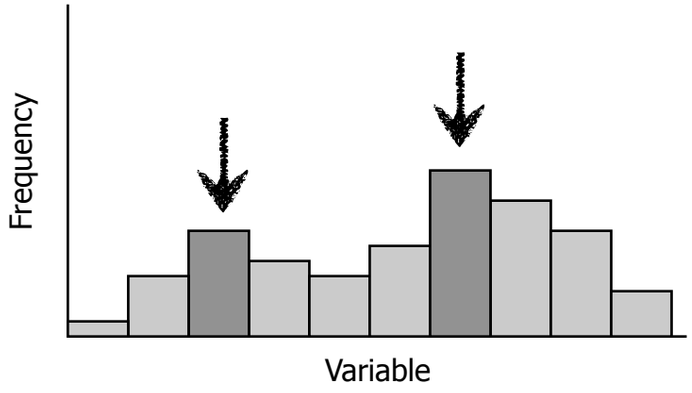
Positive or Right-Skewed



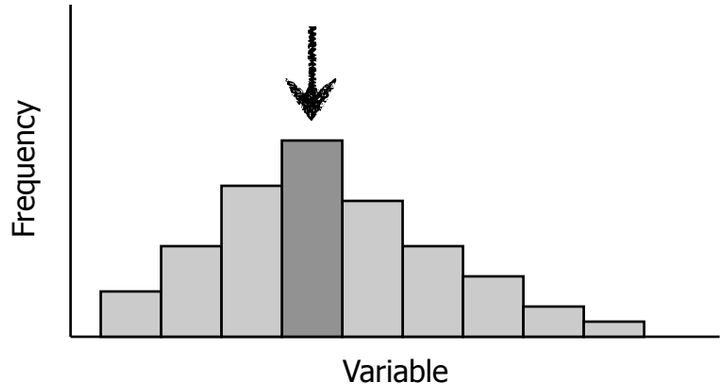
Negative or Left-Skewed

Unimodal and Bimodal Histograms

- A unimodal histogram has a single peak whereas a bi-modal histogram has two peaks (they need not be equal in height)



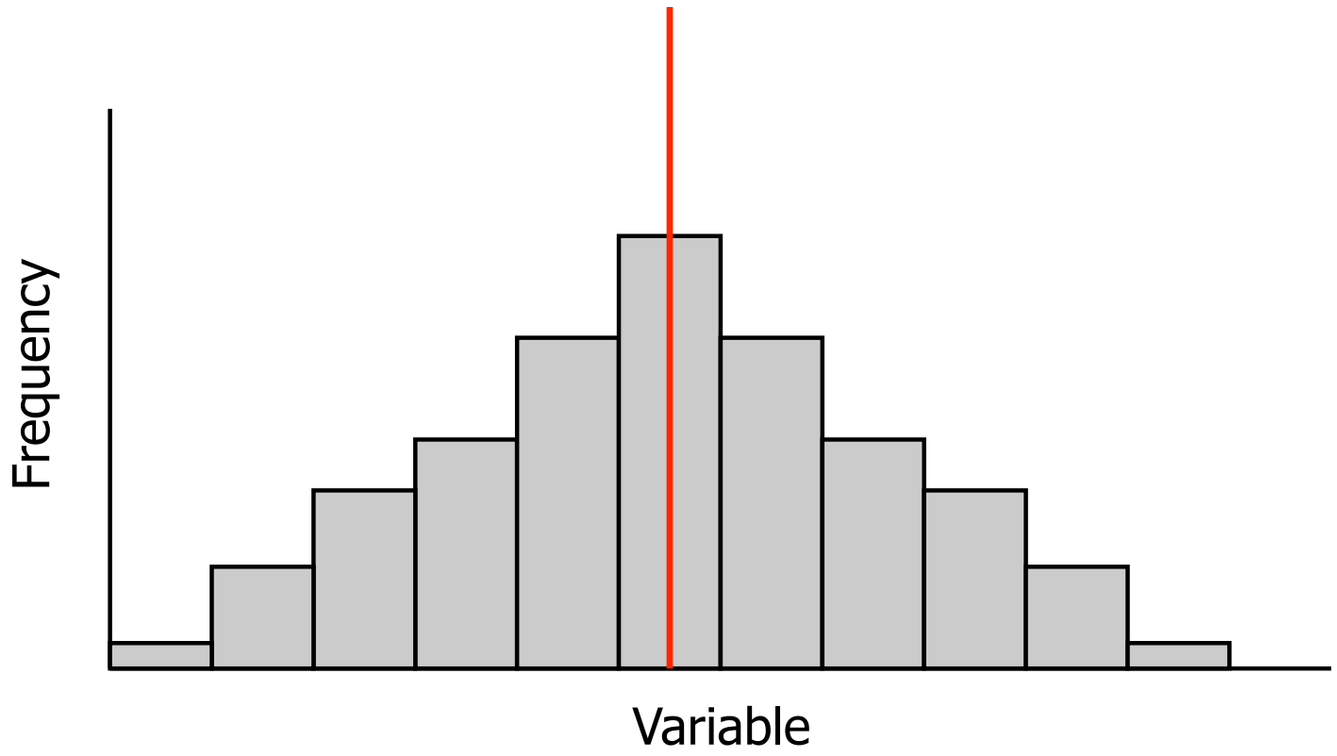
Bimodal



Unimodal

Bell-Shaped Histogram

- A symmetric, unimodal histogram



Describing the Relationship Between Two Interval Variables

Histograms vs. Scatter Diagrams

- Histogram: seeks to display summarized frequency data (in bins) of a ***single variable*** in the form of a vertical bar graph
  - Data set required for a histogram is only for a single variable
- Scatter Diagrams: seeks to display each data point in a data set individually
  - Data set required for a scatterplot contains ***two variables*** (response and explanatory)

Histograms vs. Scatter Diagrams

Sample Histogram Data

% Unemployment
6.3
1.2
8.9
2.2
2.7
4.1
0.6

Sample Scatterplot Data

% Unemployment	% of Population which is University-Educated
6.3	8.3
1.2	10.2
8.9	7.1
2.2	10.1
2.7	5.9
4.1	4.4
0.6	11.1

response variable

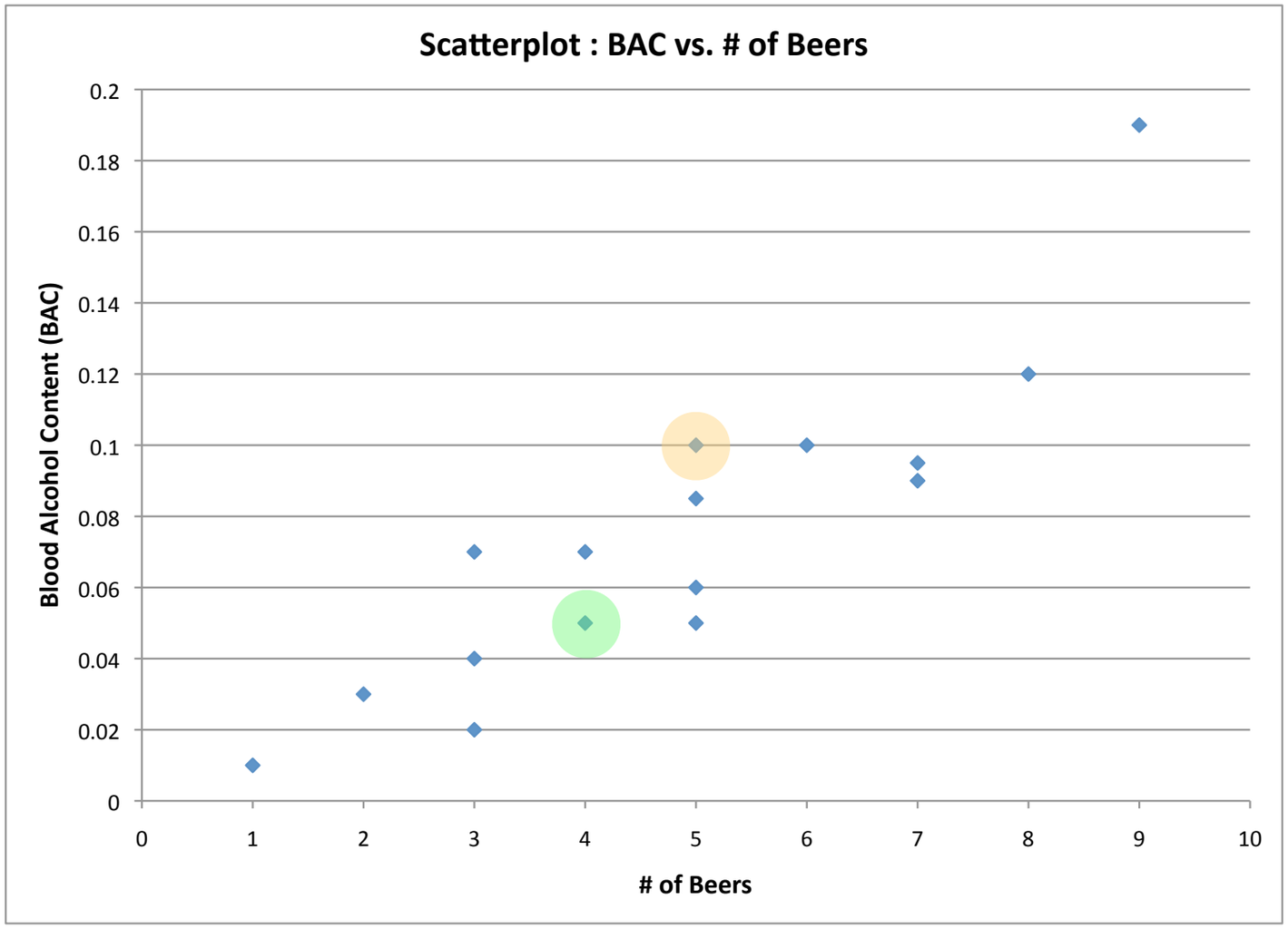
explanatory variable

Scatter Diagrams

- Plot two interval variables against each other
- It is convention to plot the explanatory variable on the x-axis (horizontal) and the response variable on the y-axis (vertical)

Scatterplot Data

Subject	# of Beers (x)	BAC (y)
1	5	0.1
2	2	0.03
3	9	0.19
4	7	0.095
...	...	...
15	1	0.01
16	4	0.05



Measures of Central Location

Arithmetic Mean

- Also commonly referred to as the average
- Measures the center of a distribution
- Often referred to by a variable with a bar overtop

- Formula:
  - $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}$$

  where $n = \text{total \# of observations}$

Arithmetic Mean

- *Example:*
  - *Observe that n = 5*
  - *Recall the formula:*

Input Data

	Class	Enrollment
1	MATH 106 - 79173	52
2	MATH 106 - 79174	24
3	MATH 106 - 79175	47
4	MATH 106 - 79448	12
5	MATH 106 - 79449	7

All MATH 106 Courses

- *Compute:* $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\bar{x} = \frac{52 + 24 + 47 + 12 + 7}{5} = 28.4$$

- *Interpret: The average enrollment for a MATH 106 class is 28.4 students*

Arithmetic Mean

- Observe
  - *is it possible to have a fractional number of students (28.4) in a class?*
- Conclude
  - *even though each data point (i.e., the number of students enrolled in each class) is an integer, the arithmetic mean can be a non-integer value*

Arithmetic Mean = Mean

- For the rest of this class, and more generally speaking, references to a ***mean*** will imply a reference to the arithmetic mean
  - For this class, I will specifically mention the term *Geometric Mean* if I intend for you to compute or interpret it

Median

- The midpoint of a distribution is a number such that half of the observations are smaller or equal to that number and the other half are larger or equal
- How to find the median of a data set

1. Order all data points in either ascending or descending order
2. Find the number for which half of the data points are smaller and half are larger
3. If there is no natural dividing number, take the arithmetic mean of the two middle numbers

Median

Input Data

	Class	Enrollment
1	MATH 106 - 79173	52
2	MATH 106 - 79174	24
3	MATH 106 - 79175	47
4	MATH 106 - 79448	12
5	MATH 106 - 79449	7

All MATH 106 Courses

- *Example 1*

1. Order all data points in the data set above in increasing order

7 12 24 47 52

2. The middle number, and hence the median is 24 (2 numbers are smaller and two numbers are larger than the median value of 24)

Median

- *Example 2*

1. If the ordered data set were the following

2. The is no middle number for which half of the numbers are smaller and the other half are larger

3. Take the arithmetic mean of the two middle numbers

$$\frac{9+17}{2} = \frac{26}{2} = 13 = median$$

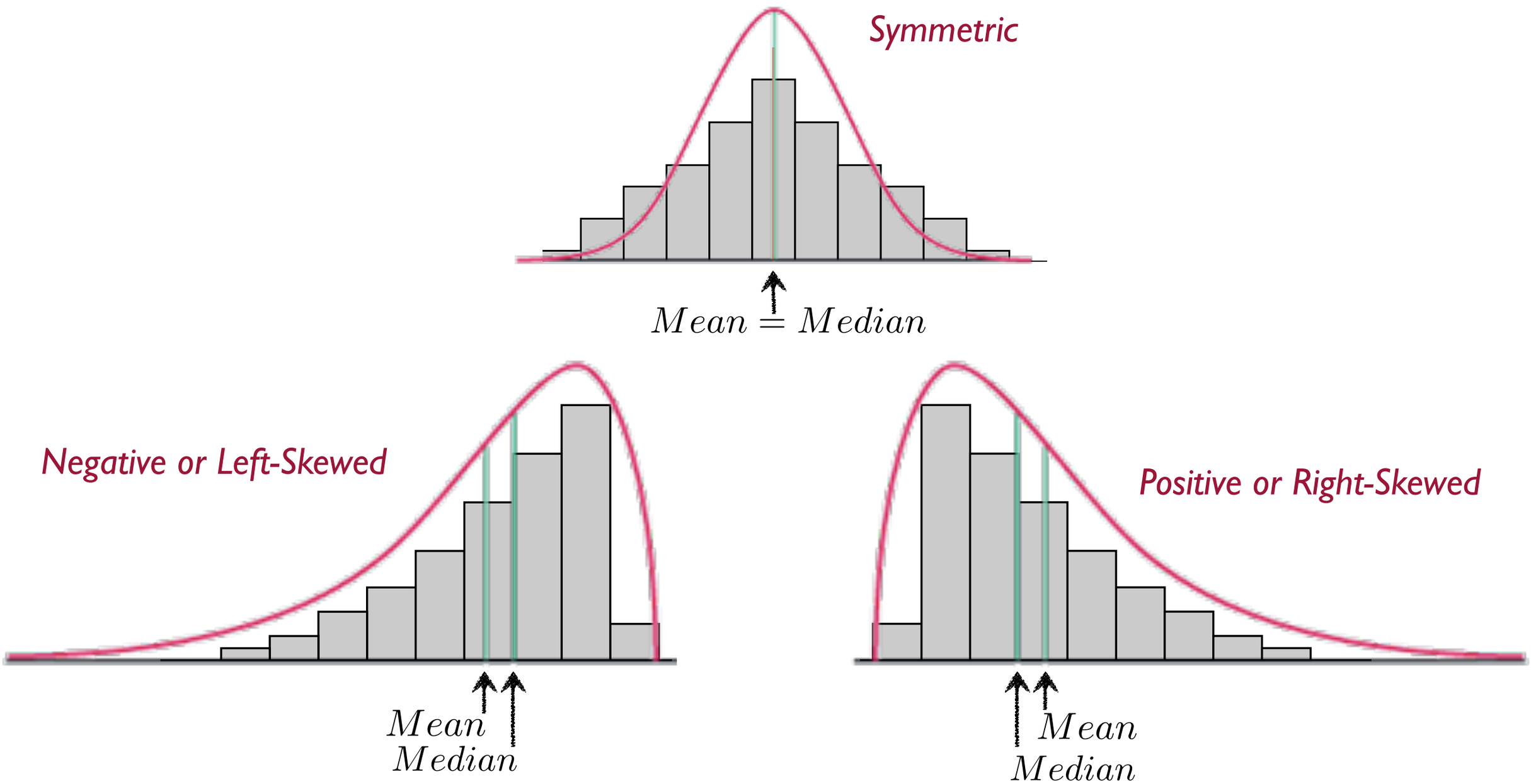
Comparison: Mean vs. Median

A ***mean*** is much more sensitive to individual data points than a ***median*** is

Data Set	Data							Mean	Median
1	11	12	23	32	43	53	100	39	32
2	11	12	23	32	43	53	10,000	1,453	32
3	0	0	0	32	43	10,000	10,000	2,873	32

Comparison: Mean vs. Median

- Rule of Thumb for *Skewness*

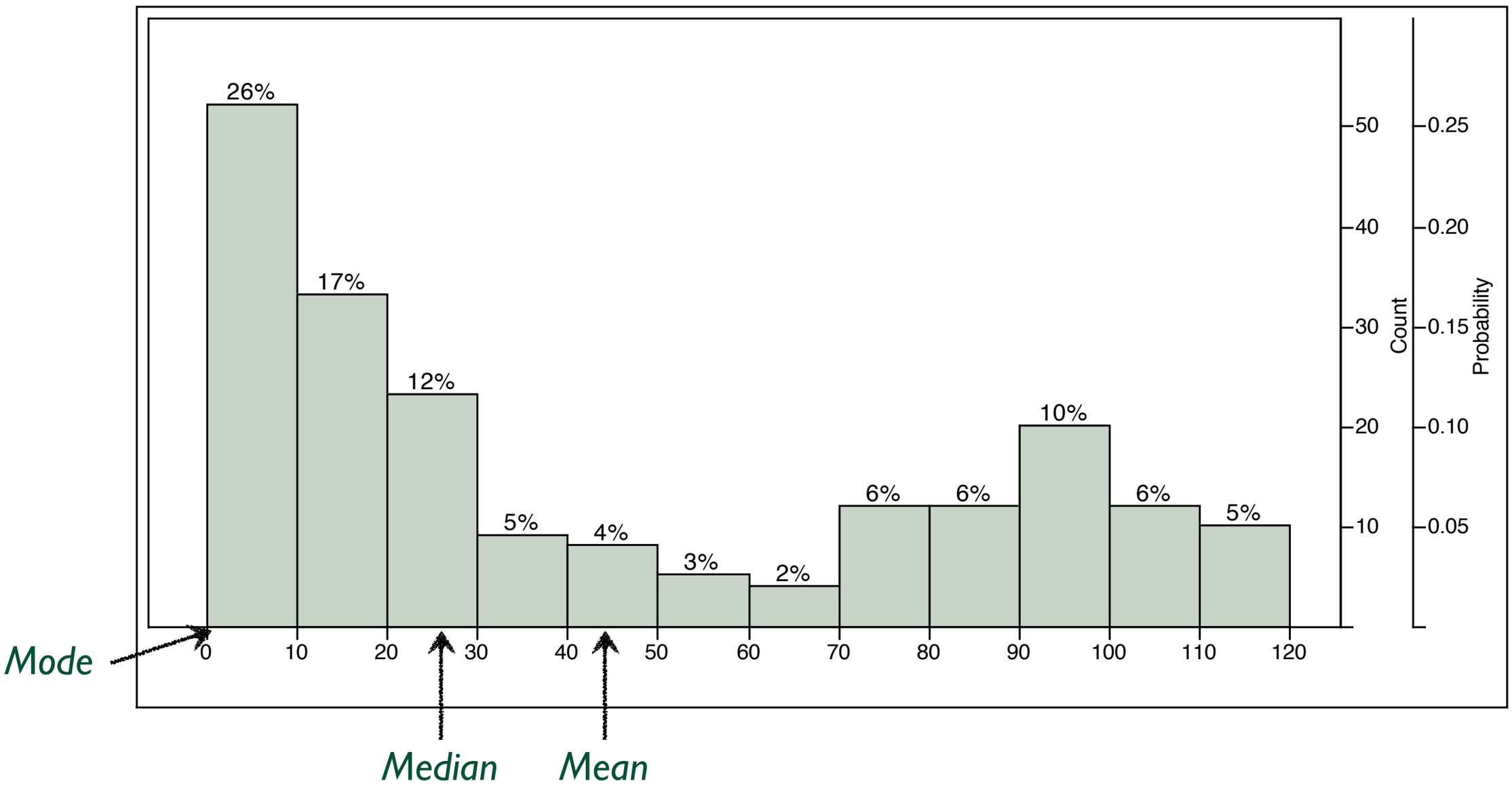


Mode

- The observation (or observations) in a data set that occur with the greatest frequency
- Possible issues with using the mode as a measure of central tendency
  - not particularly useful in describing small data sets
  - will not necessarily be unique

Mode

- Where do you think the mean and median are? What about the mode?



Advertisement

Median Healing Time: 4.1 Days

Geometric Mean

- Much less common than the Arithmetic Mean
- Also measures the center of a distribution

- Formula: $\bar{a} = \sqrt[n]{a_1 a_2 \dots a_n} = \sqrt[n]{\prod_{i=1}^{n} a_i}$
- *where*
  - $n =$ total # of observations
- *note: all* a_i's *must be positive numbers*

Geometric Mean

- Application
  - The geometric mean is often used for numbers which are meant to multiplied together or are exponential in nature, e.g., bacterial population, interest rates, etc.

- *Example:*

$$\bar{a} = \sqrt[5]{1,037 \cdot 9,449 \cdot 41,000 \cdot 99,003 \cdot 221,955}$$

$$\bar{a} = \sqrt[5]{8.8280 \times 10^{21}}$$

$$\bar{a} = 24,500$$

Input Data

Day	Bacterial Population
1	1,037
2	9,449
3	41,000
4	99,003
5	221,955

Are Means Enough?

Mean Yearly Temperature in San Francisco and New York city (in Fahrenheit, from Weather.com)

57.4 55.0

Which is which?

Measures of Variability

Range

- The difference between the largest and smallest observations
- Easy to compute, but doesn't tell of much about the data in between the largest and smallest value

Example 1 4 4 4 4 4 50

$$range = 50 - 4 = 46$$

Example 2 4 8 15 24 39 50

$$range = 50 - 4 = 46$$

same range for both examples which have very different data

Variance

- The most common measure of the spread of the data points within a data set is variance (s^2)
  - Formula

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  - *where*
    - $\bar{x}$ *is the sample mean of the data set*
    - $n = \text{total } \# \text{ of observations}$
- Interpretation: the *variance* grows as the spread of the data increases

Shortcut Calculation for Variance

- With a little algebraic manipulation, we can rewrite the formula for variance as follows, making it easier to compute (manually)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i^2\right)}{n}\right]$$

- *where*
  - x_i *are individual values of data*
  - $n = \text{total } \#\text{ of observations}$

Standard Deviation

- Often times, instead of *variance*, the *standard deviation* is cited, which is simply the square root of Variance
  - Formula

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

  - *where*
    - $\bar{x}$ *is the mean of the data set*
    - n *is the number of observations*
- Interpretation: the *standard deviation* grows as the spread of the data increases

Why we need Variance

- Both graphics depict a distribution of data which share the same *mean* but very different levels of *variance*

High Variance

Low Variance

Variance

- *Example:*
  - *Observe* $n = 5$, $\bar{x} = 28.4$
  - *Recall the formula*

Input Data

	Class	Enrollment
1	MATH 106 - 79173	52
2	MATH 106 - 79174	24
3	MATH 106 - 79175	47
4	MATH 106 - 79448	12
5	MATH 106 - 79449	7

All MATH 106 Courses

- *Compute*

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{(52 - 28.4)^2 + (24 - 28.4)^2 + (47 - 28.4)^2 + (12 - 28.4)^2 + (7 - 28.4)^2}{5-1}$$

$$s^2 = \frac{1649.2}{4} = 412.3 \Rightarrow s = 20.3$$

Variance

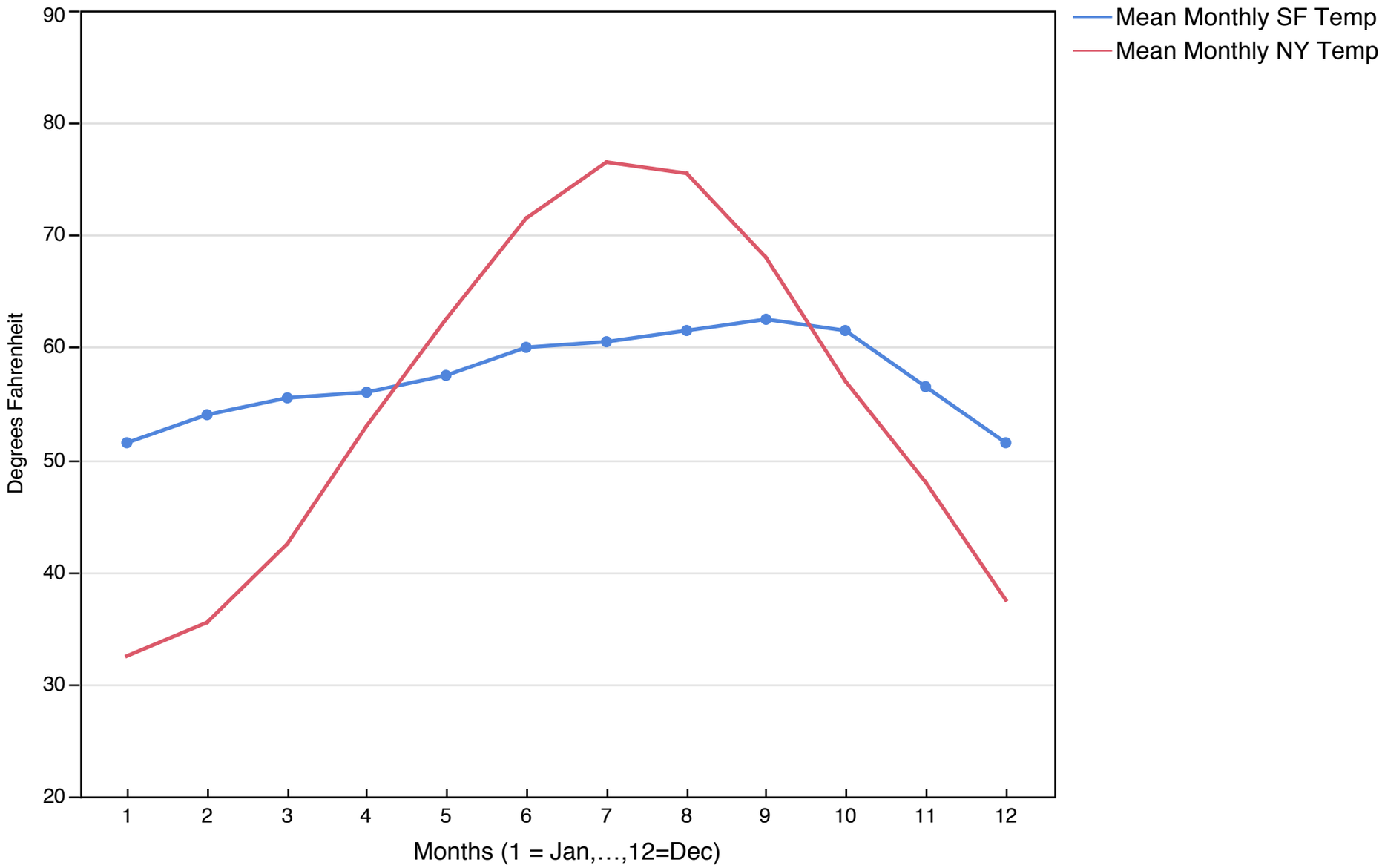
- *Example: (cont'd)*
  - *<u>Interpret</u>: The variance of the average enrollment for an OMIS 40 class at SCU is 412.3 students squared, and the standard deviation is 20.3 students*

- Note that when stating variance, the units are squared, therefore, very often, Standard Deviation is cited because the units are the same as those of the mean (to the power of 1)

- Low variance is good, high variance is bad

The Value of Variance

	Mean	Variance	Standard Deviation
San Francisco	57.4	14.8	3.85
New York	55	253.3	15.91



Measures of Relative Standing and Box Plots

Percentile

- A measure of the spread of the data
- Recall: the ***median*** is the value for which half of the data points in a data set are less than or equal to that value and the other half are larger or equal. *Is there a number for which one third of the data points are smaller or equal in value and two thirds are greater or equal in value?*
- If we divide a data set into four evenly-spaced intervals, we could call these quantiles *quartiles*
- 100 evenly spaced intervals are called *percentiles*

Percentiles

- The *Pth* percentile is the value for which *P*% are less than that value and *(100–P)*% are greater than that value
- You can locate the *P*th percentile using the following equation

$$L_p = (n+1)\frac{P}{100}$$

  where L_p is the location of the *P*th percentile

- Don’t be fooled and just plug in values for *P* and *n* and solve: ***this is a multistep process***

Quartiles

- How many quartiles are there?
- What fraction of the data points from a given data set are smaller than its first quartile?
- What is the second quartile? Is there a relationship between the second quartile and the median? The mean?
- Are the second quartile and the median always equal?
- Quartiles are denoted by the letter Q with a subscript representing the quartile #: Q_1, Q_2, Q_3

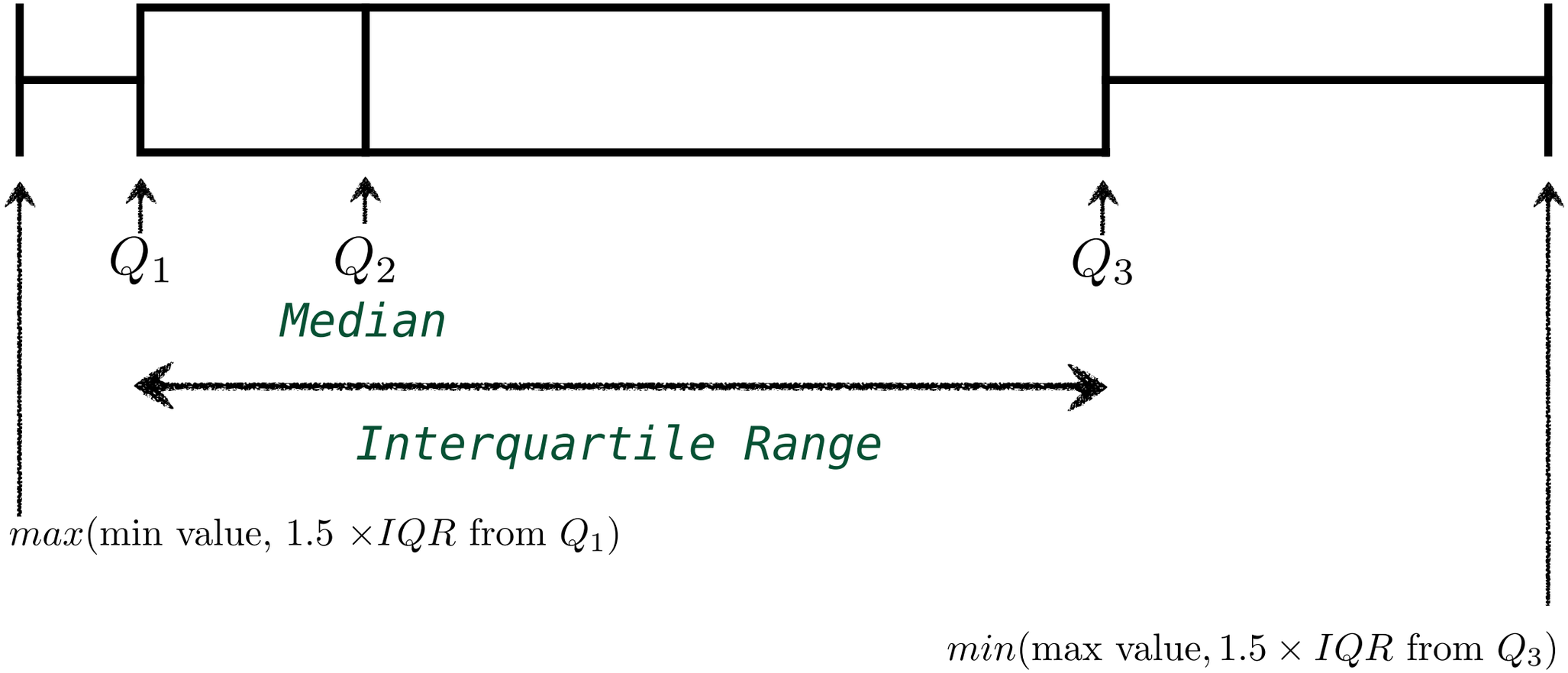
Quartiles & Interquartile Range

- The Range (maximum value minus the minimum value) develops a sense of the spread of the data
- The **Interquartile Range** (IQR) is $Q_3 - Q_1$
  - What percentage of values lie within the IQR?

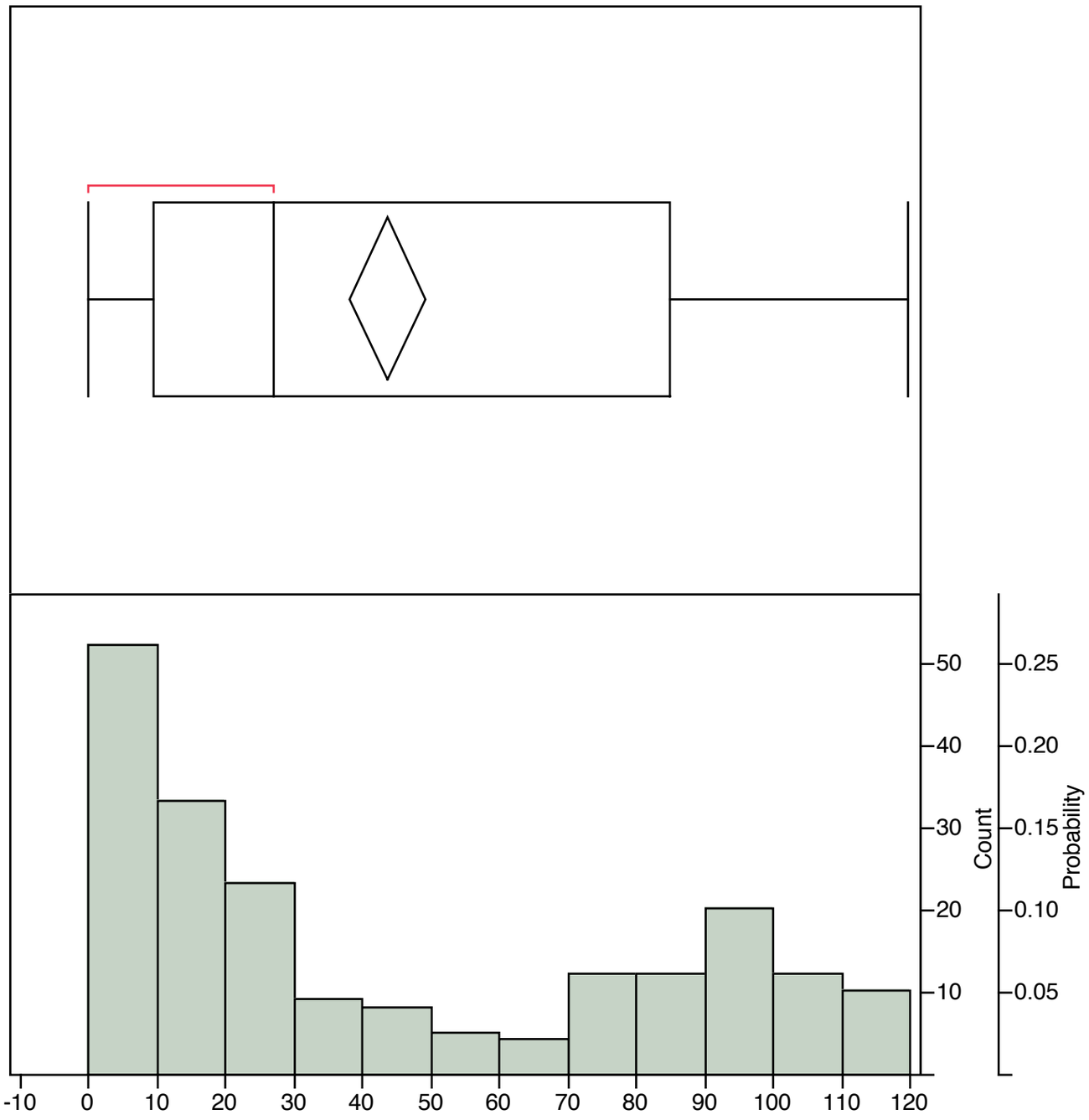
- **Important Observation**
  - Quartiles (and by association) the median divide a data set according the number of data points **and not** the values associated with each of those data points

Box Plot

- A useful graphical interpretation which depicts quartile values together with the minimum and maximum values of the data set

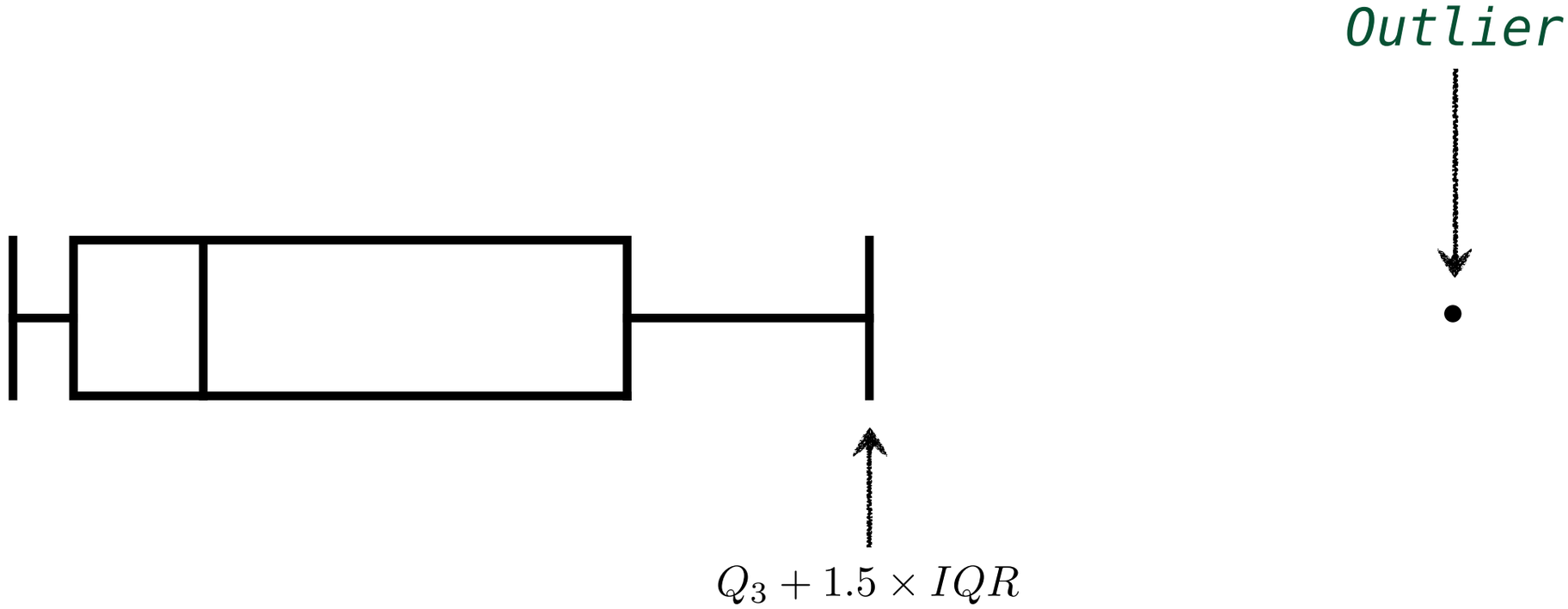


Box Plot & Histogram



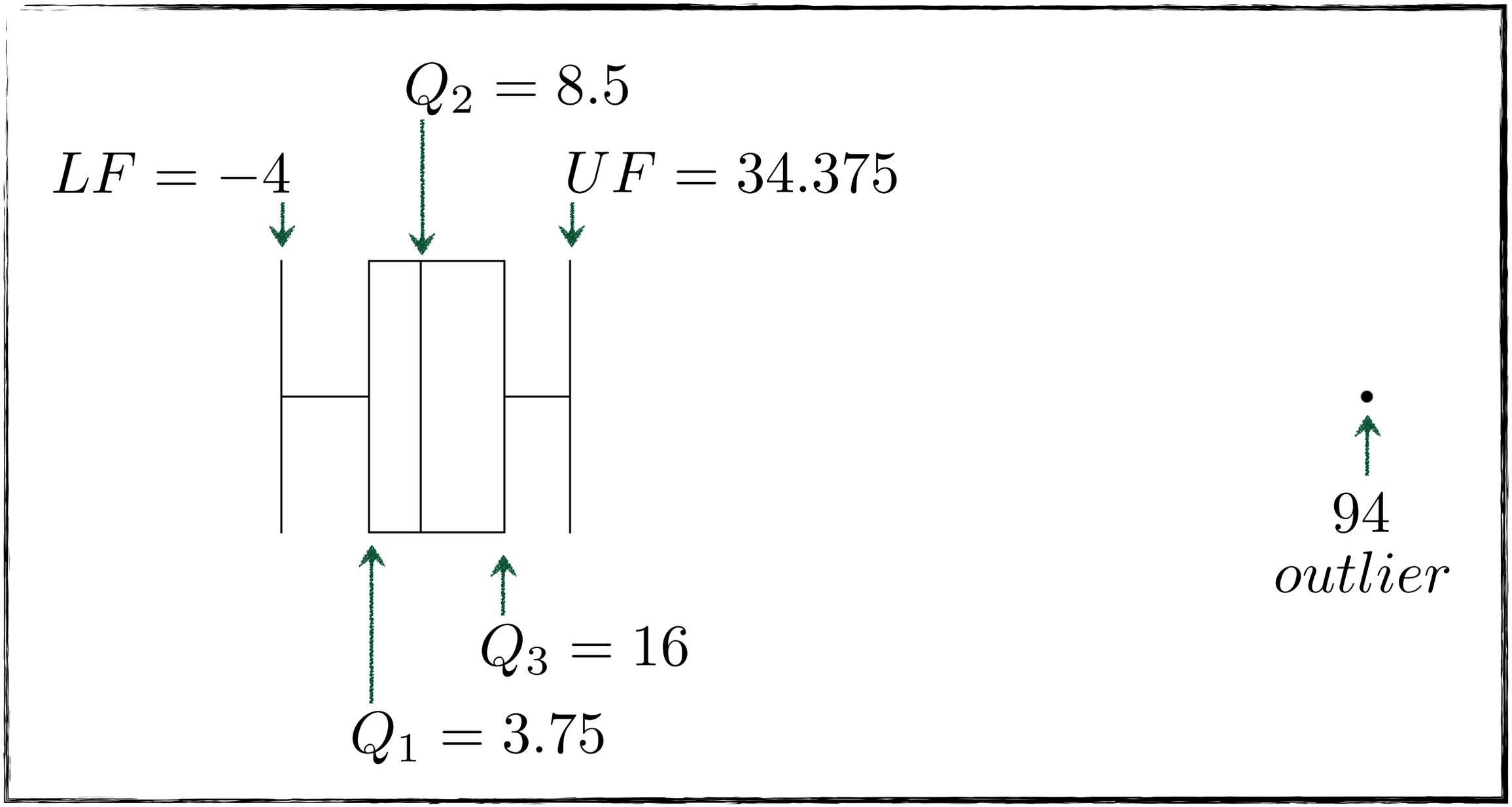
Box Plot with an Outlier

- If certain data exceed the *1st* or *3rd* quartile by more than *1.5∗IQR* in either direction, they are considered outliers and are represented in a box plot as points
- In this example there is one outlier



Box-Plot: *Example*

-4 0 5 7 8 9 12 14 22 94 $\rightarrow n = 10$



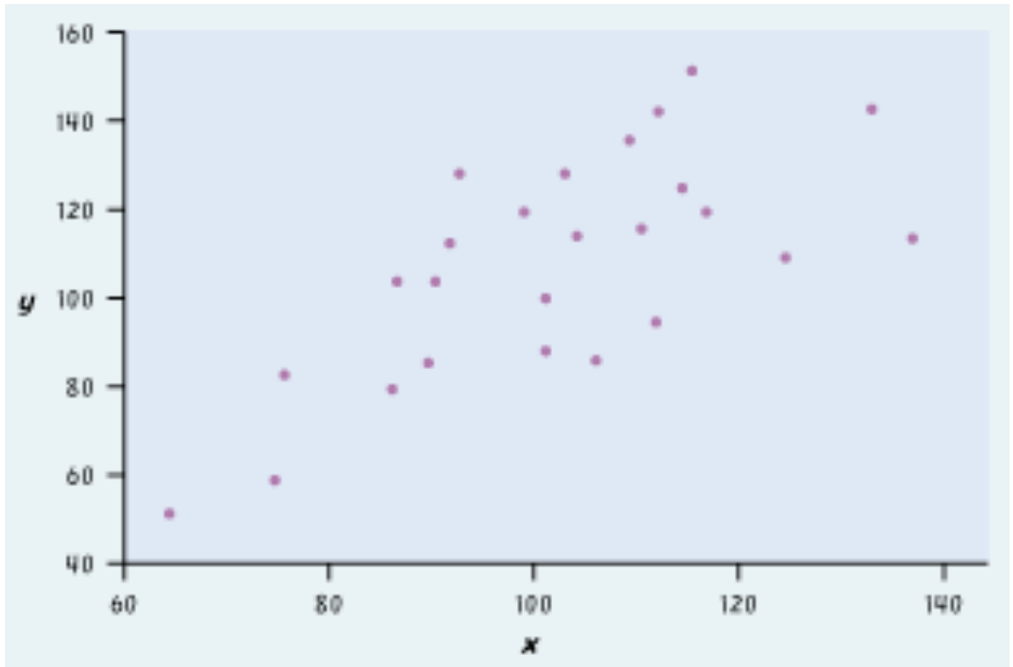
Describing Distributions

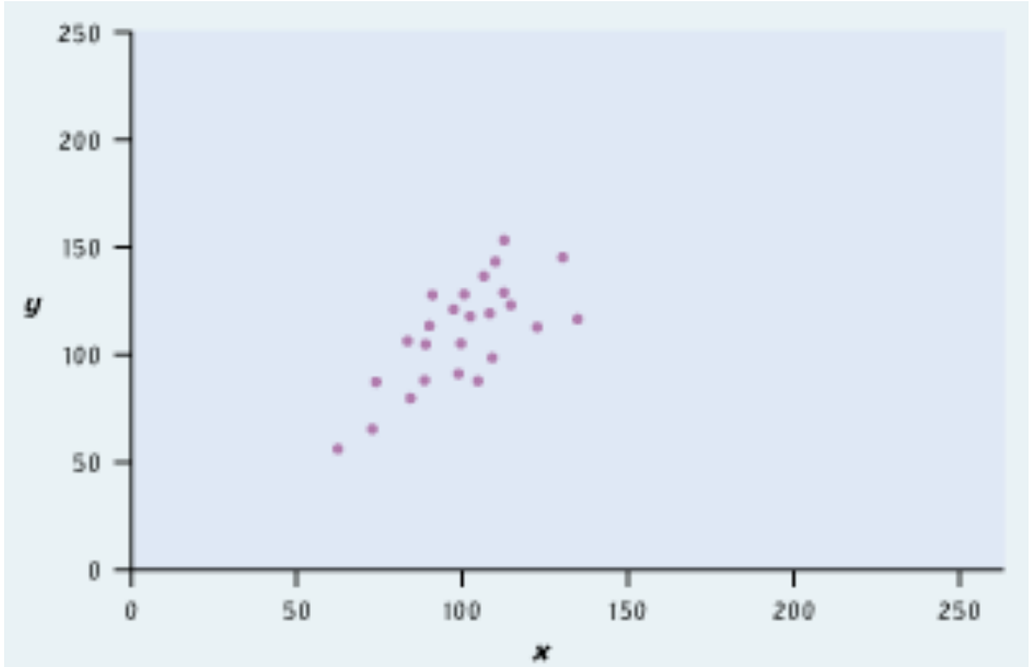
Measures of Central Tendency	Measures of Spread
Arithmetic Mean	Range
Geometric Mean	Variance
Median	Quantiles/ Quartiles / Percentiles
Mode*	

Measures of Linear Relationship

Linear Relationships

- Which graph exhibits a stronger linear positive relationship?





Interpreting Scatter Diagrams

- Our previous interpretation of these diagrams were visual in nature
- We will now examine three numerical measures of ***linear*** relationship that provide this information
  - Covariance
  - Coefficient of Correlation

Covariance

- A measure of how two data sets linearly vary with respect to each other

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Equivalent

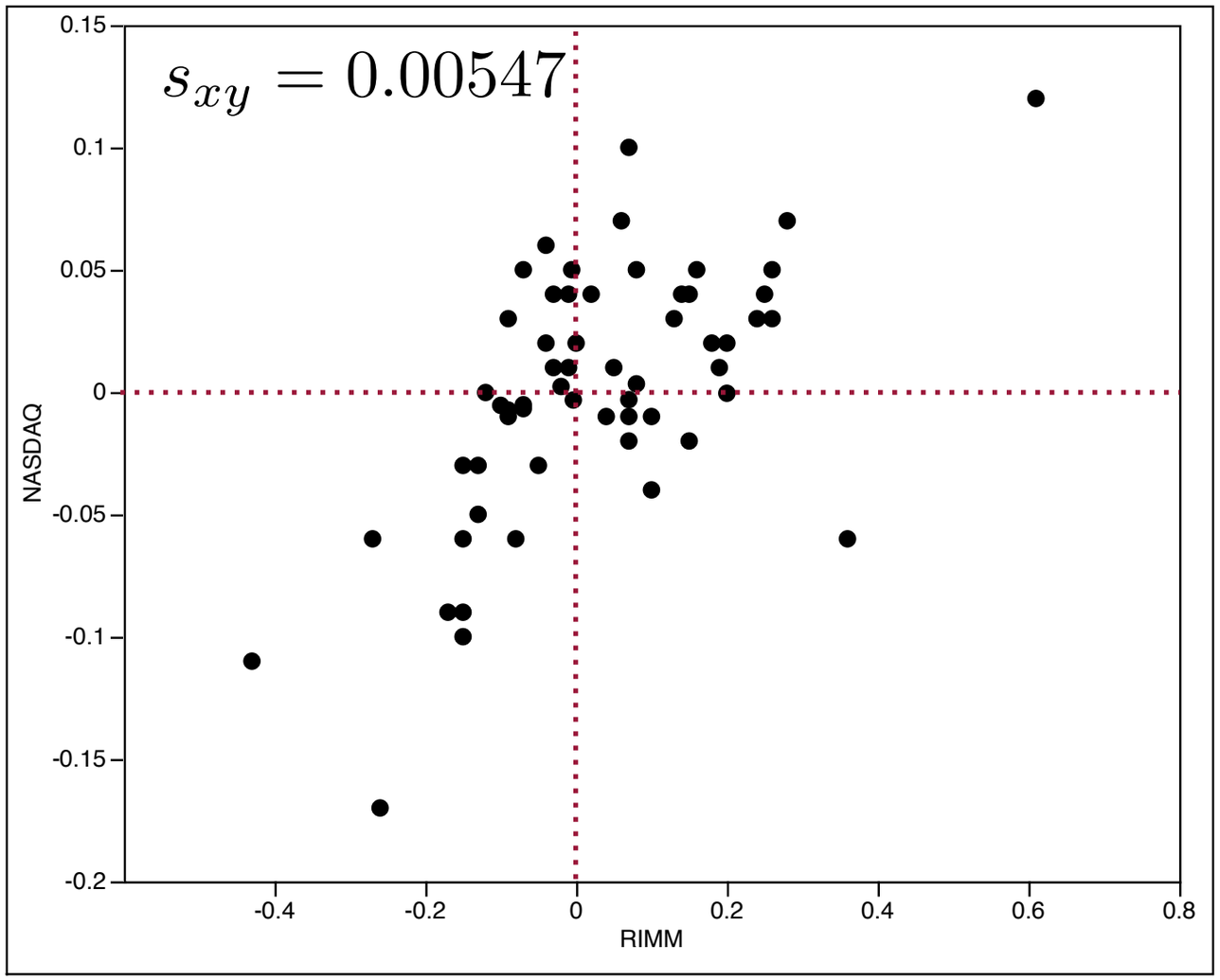
$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

Shortcut for Manual Calculations

- When we obtain a result for the covariance (s_{xy}), we are looking for two pieces of information
  - Sign: is it positive or negative?
  - Value: is it a large number or a small number?

Covariance

- A scatter diagram of the monthly rates of return of RIMM against the NASDAQ for Jan–05 to Dec–09



We visually conclude that there is positive covariance between the monthly rates of return between RIMM and NASDAQ.

We compute the covariance which confirms that covariance is indeed positive, although it seems quite small at 0.00547. Does this mean that they covary poorly? What number is high enough? How do we compare this value of covariance against other covariance results from other data set?

We need a method to standardize the magnitudes of covariance values so that they can be more easily interpreted and compared.

Coefficient of Correlation

- Formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} = \frac{s_{xy}}{s_x s_y}$$

Covariance

Standard Deviation of y

Standard Deviation of x

  - *where*
    - $r = correlation$
    - $n = \#\ of\ data\ points\ in\ data\ set$
    - $\bar{x}, \bar{y} = mean\ of\ variables\ x\ and\ y$
      - **recall:** $s_x, s_y = standard\ deviation\ of\ variables\ x\ and\ y$

$$s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

Coefficient of Correlation

- By dividing the covariance by standard deviation of both variables, we obtain the coefficient of correlation, a much easier number to interpret

$$-1 \leq r \leq +1$$

- $r = +1 \longrightarrow$ perfect positive linear relationship
- $r = -1 \longrightarrow$ perfect negative linear relationship
- $r = 0 \longrightarrow$ no linear relationship between variables

Correlation: *Example*

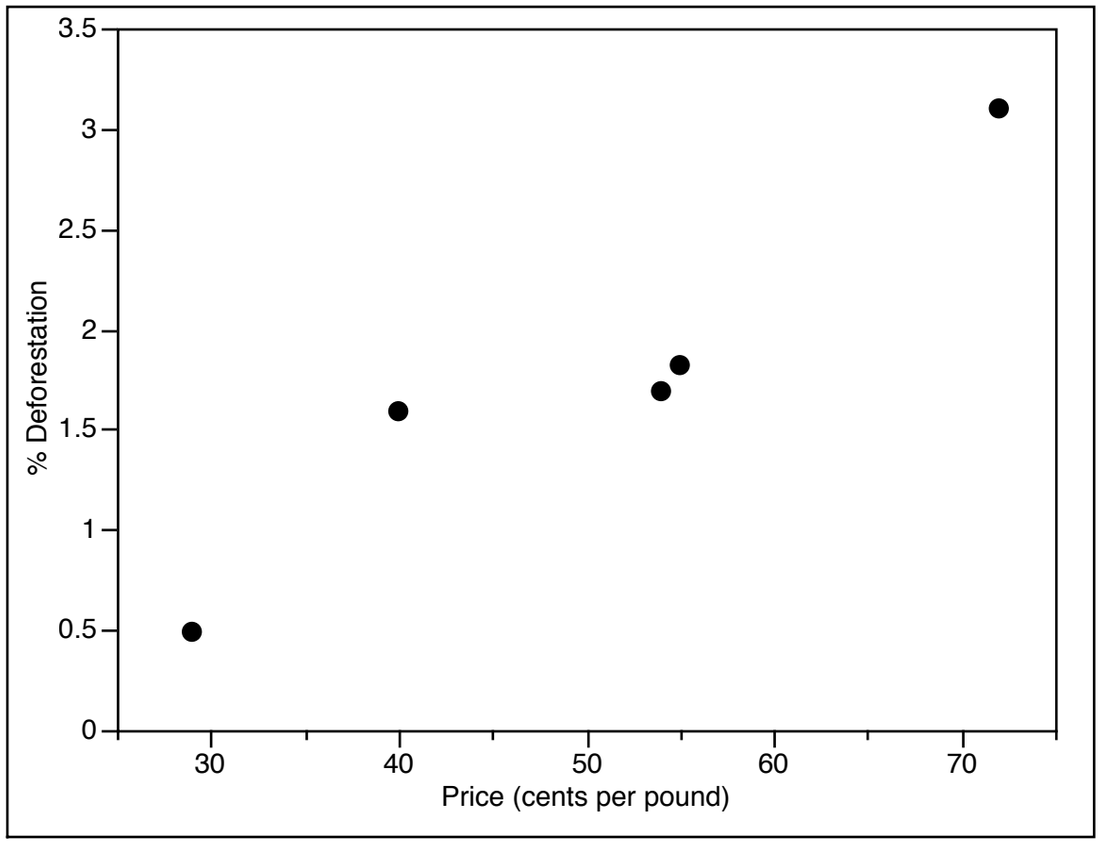
- Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forest to plant more coffee trees. Here data for five years on prices paid to coffee growers in Indonesia and the rate of deforestation in a national park that lies in a coffee producing region.

Price (cents per pound)	Deforestation %
29	0.49
40	1.59
54	1.69
55	1.82
72	3.1

Correlation: *Example*

a. What is the explanatory variable?

b. What pattern can be observed?



Correlation: *Example*

c. Find the correlation (r) and relate it to the scatterplot.

$$recall \quad r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$observe\ n = 5$$

Price (cents per pound)	Deforestation %
29	0.49
40	1.59
54	1.69
55	1.82
72	3.1

Step 1

recall

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

compute

$$\bar{x} = \frac{29 + 40 + 54 + 55 + 72}{5} = 50$$

$$\bar{y} = \frac{0.49 + 1.59 + 1.69 + 1.82 + 3.10}{5} = 1.738$$

Correlation: *Example*

c. Find the correlation (r) and relate it to the scatterplot.

$$recall \quad r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$observe\ n = 5$$

Price (cents per pound)	Deforestation %
29	0.49
40	1.59
54	1.69
55	1.82
72	3.1

Step 2

$$recall\ s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

compute

$$s_x = \sqrt{\frac{(29-50)^2 + (40-50)^2 + (54-50)^2 + (55-50)^2 + (72-50)^2}{4}} = 16.32$$

$$s_y = \sqrt{\frac{(0.49-1.738)^2 + (1.59-1.738)^2 + (1.69-1.738)^2 + (1.82-1.738)^2 + (3.10-1.738)^2}{4}} = 0.928$$

Correlation: *Example*

c. Find the correlation (r) and relate it to the scatterplot.

$$recall \quad r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$n = 5, \bar{x} = 50, \bar{y} = 1.738, s_x = 16.32, s_y = 0.928$$

Price (cents per pound)	Deforestation %
29	0.49
40	1.59
54	1.69
55	1.82
72	3.1

Step 3

compute

$$r = \frac{1}{4}\left[\left(\frac{29-50}{16.32}\right)\left(\frac{0.49-1.738}{0.928}\right) + \left(\frac{40-50}{16.32}\right)\left(\frac{1.59-1.738}{0.928}\right) + \ldots + \left(\frac{72-50}{16.32}\right)\left(\frac{3.10-1.738}{0.928}\right)\right]$$

$$r = 0.955$$

Facts about Correlation

- Correlation (r) does not differentiate between explanatory and response variables
- Correlation (r) has no units of measurement

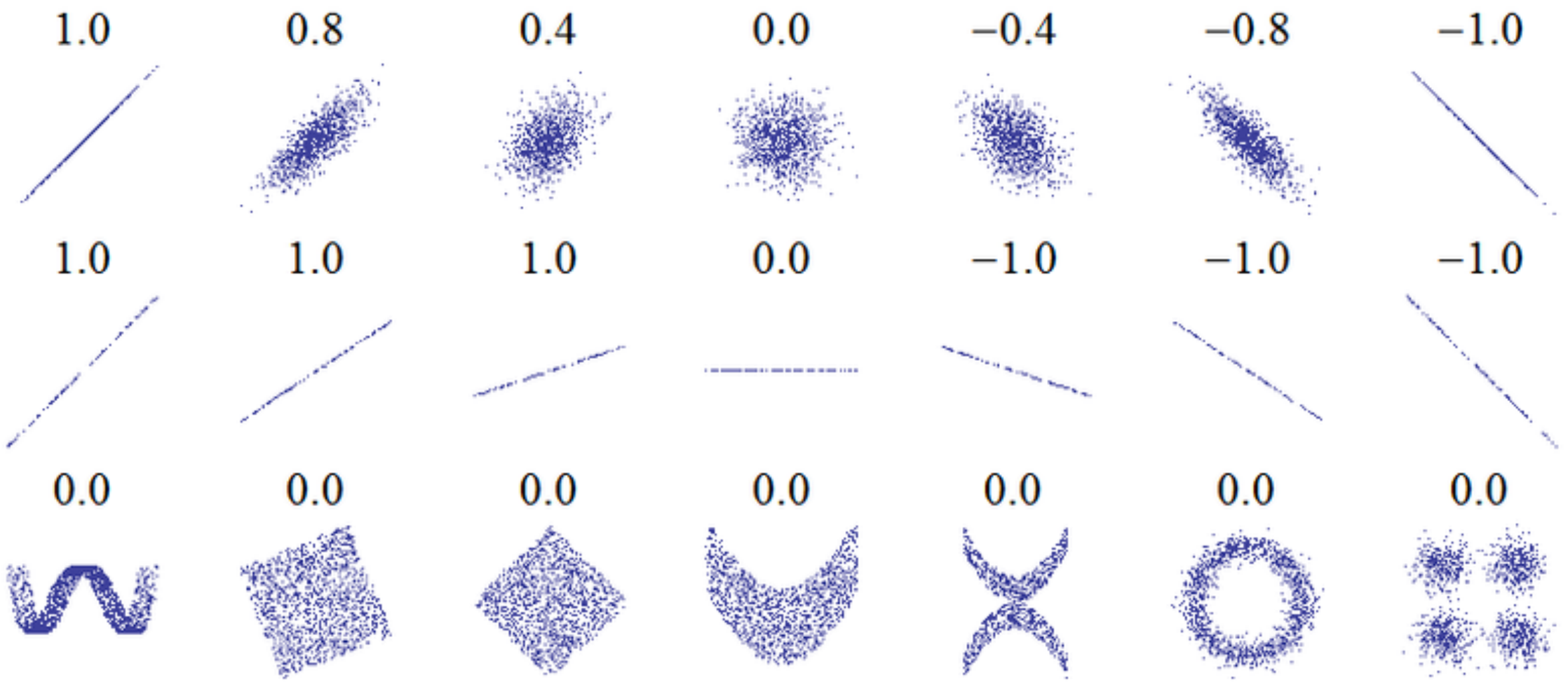
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$\frac{units - units}{units} = unitless$$

- Correlation is strongly affected by outliers

Coefficient of Correlation

$$-1 \leq r \leq 1$$



graphic obtained from Wikipedia (2009) http://en.wikipedia.org/wiki/Correlation

Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

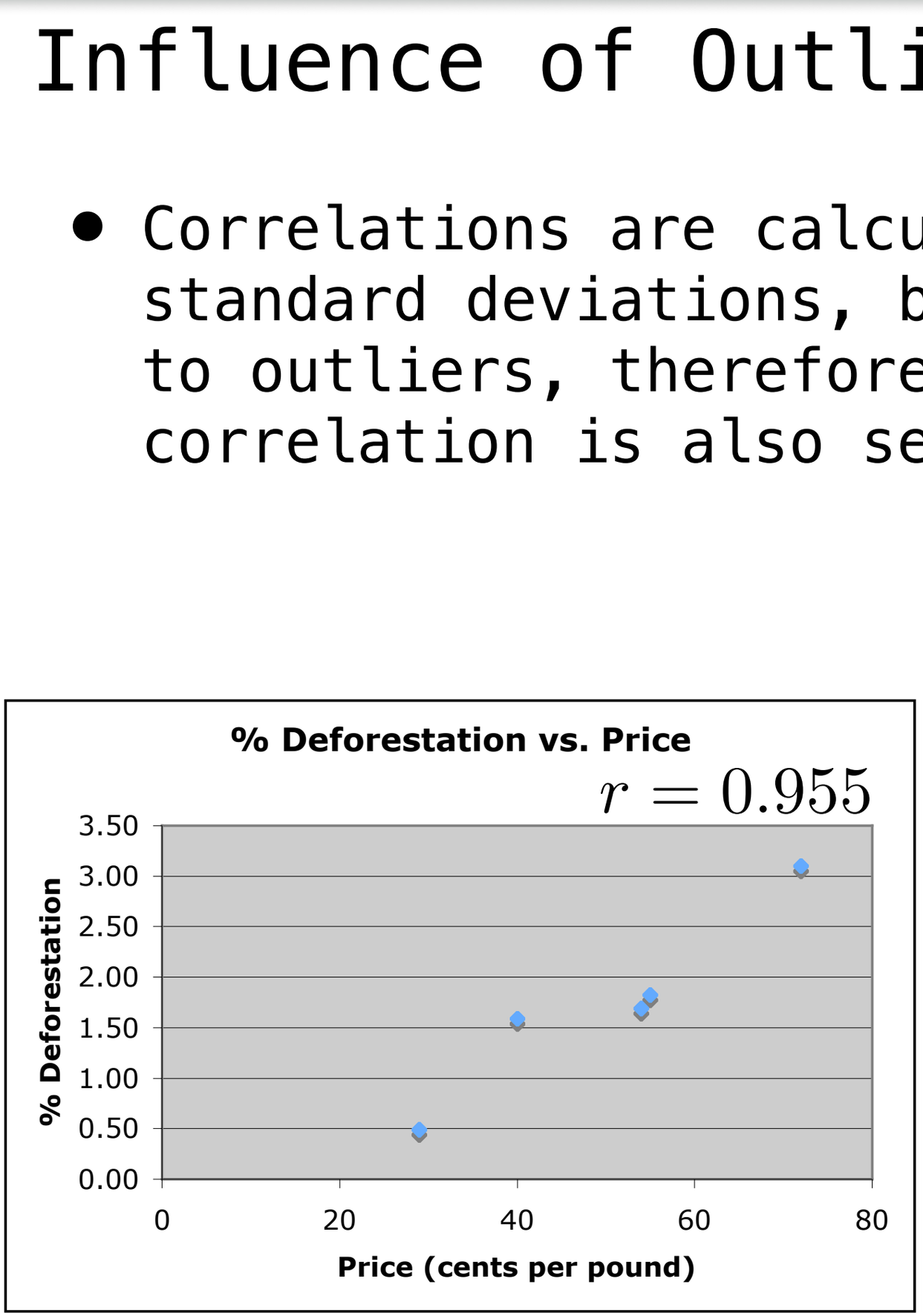
$\Rightarrow$ = *implies that*

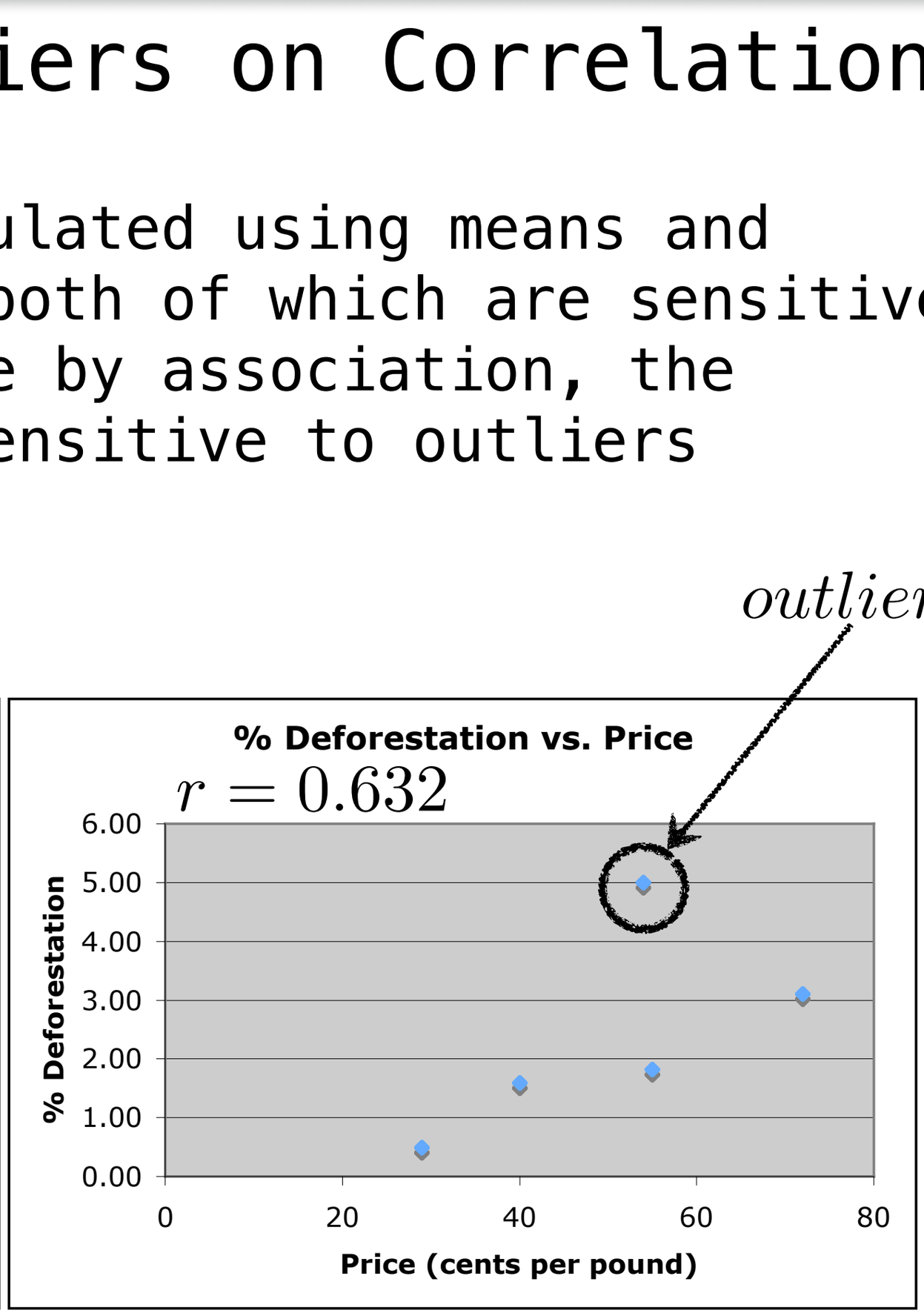
As $s_x \downarrow \Rightarrow \left(\frac{x_i - \bar{x}}{s_x}\right)\uparrow \Rightarrow r\uparrow$

- Interpretation: as the standard deviation (and by association the variance) of the data points (scatter) is reduced, the correlation increases in strength

Influence of Outliers on Correlation

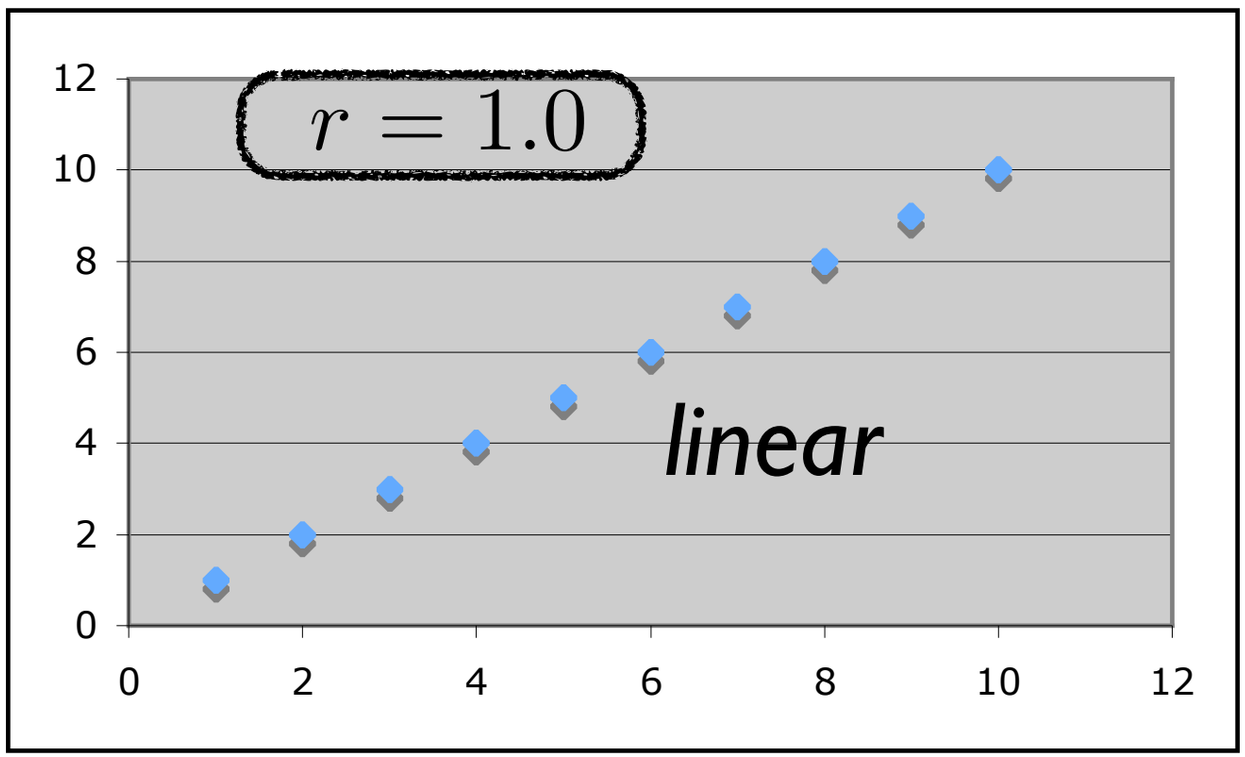
- Correlations are calculated using means and standard deviations, both of which are sensitive to outliers, therefore by association, the correlation is also sensitive to outliers



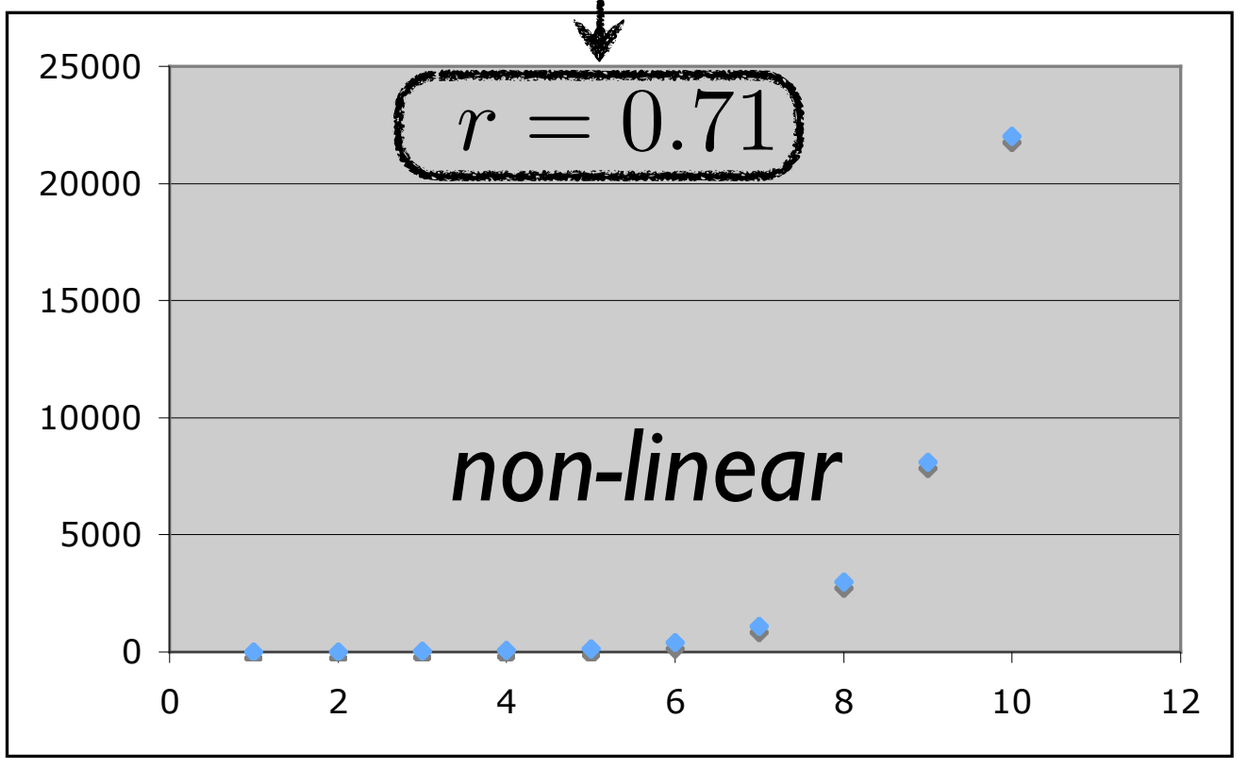


Linear vs. Non-Linear

- Correlation describes ***only linear relationships***
- Never use correlation to describe non-linear relationships



Meaningless and Misleading!



Summary Classification of Measures

Relative Measures of Data	Absolute Measures of Data	Measures of Linear Relationships
Mean	Median	Covariance
Standard Deviation	Range	Coefficient of Correlation
Variance	Min / Max	
Coefficient of Variation	Interquartile Range	

Assigning Probability to Events

Sets

- A set is a collection of items
- If a set *A* has members *x*, *y*, *z*, this is represented in the following manner
  - *Example* $A = \{x, y, z\}$
    - *the collection of all possible outcomes from tossing a coin consists of the set* $S = \{H, T\}$
    - *if we defined the event A as Heads coming up, then the set* $A = \{H\}$
    - *the empty set, a set with no members is represented with a* $\emptyset$ *or* $\{\}$

Random Experiment

- A *random experiment* is a procedure or an operation whose outcome is uncertain and cannot be predicted in advance, even if we may intuitively know the probabilities associated with each outcome
- A *sample space* (S) is the collection of **all** possible *outcomes* of a random experiment

Random Experiment	Sample Space (S)
Tossing a coin once	S = {H, T}
Tossing a coin twice	S = {HH, HT, TH, TT}
Rolling a die once	S = {1, 2, 3, 4, 5, 6}
Rolling two dice	S = {(1,1), (1,2), (1,3), ... , (6,5), (6,6)}
Your grade in this class	S = {A, A–, B+, B, B–, C+, C, C–, D+, D, D–, F}

Events

- An **event** is a set of one or more outcomes of a random experiment (a subset of the sample space *S*)

Random Experiment	Sample Space (*S*)	Event
Tossing a coin once	*S* = { H, T }	Heads = { H }
Tossing a coin twice	*S* = { HH, HT, TH, TT }	Getting exactly one head = { (H,T), (T,H) }
Rolling a dice once	*S* = {1, 2, 3, 4, 5, 6 }	Outcome is even = { 2, 4, 6 }
Rolling two dice	*S* = {(1,1), (1,2), (1,3), ... , (6,5), (6,6)}	Sum equals 5 = { (1,4), (2,3), (3,2), (4,1) }
Your grade in this class	*S* = { A, A−, B+, B, B−, …, F }	Pass = { A, A− ,B+, B, B−, C+, C, C−}

Axioms of Probability

- *Axiom*: a statement or proposition that is regarded as being established, accepted, or self-evident
- Let S denote the sample space of an experiment, let us define A as an event in S, and define as the probability of A, $P(A)$, occurring such that the following two axioms hold
- Axiom 1: $0 \leq P(A) \leq 1$ *(between zero and one)*
- Axiom 2: $P(S) = 1$ *(all possible events are contained in the sample space)*

Axioms of Probability

- For example, when tossing a coin
  - The sample space $S = \{H, T\}$
  - Define event A as heads coming up $A = \{H\}$
  - Define event B as tails coming up $B = \{T\}$
    - We know that $P(A) = 0.50$
- Which satisfy the axioms
  - Axiom 1: $0 \leq P(A) = 0.50 \leq 1$
  - Axiom 2: $P(S) = P(A) + P(B) = 0.50 + 0.50 = 1.0$

Bayes's Theorem

Baye's Theorem

$$P(A_1|B_1) = \frac{P(B_1|A_1)P(A_1)}{P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)}$$

Baye's Theorem

*Baye's Theorem is particularly useful when you are **given conditional and unconditional probabilities** and asked to **find conditional probabilities***

Baye's Theorem

- The types of questions are very interesting when presented in a medical context.
  - Why not screen everyone for HIV/AIDS, cancer, or other diseases on a regular basis?
  - Medical tests are not perfect, and the possibility of a false-positive test can be very stressful as well as costly, especially when testing for chronic or potentially fatal diseases.

Baye's Theorem

- *Example*: HIV/AIDS testing
  - Assume that 99% of the general population is not infected with the HIV/AIDS virus. Of those people who are not infected with HIV/AIDS and are tested for the virus, 3% test positive. 98% of people infected with HIV/AIDS test positive for the virus. What is the probability of not having HIV/AIDS given you test positive for the virus?

2010 Global Stats: 0.8% adults are infected with w/ HIV
2010 USA Stats for the general population:
false-positive rates of 0.0004 to 0.0007, false-negative rates of 0.003

HIV / AIDS Testing Primer

- In increasing order of accuracy:
  - ELISA Test
  - Western Blot Test
  - PCR Test
  - All have different false-positive and false-negative rates, e.g., false-positive rates for the ELISA test are approximately 0.002, but false-negatives are much lower

Baye's Theorem

- *Example cont'd*
  - Define the events
    - A not infected with the HIV/AIDS virus
    - A^c infected with the HIV/AIDS virus
    - B test positive for the HIV/AIDS virus
    - B^c test negative for the HIV/AIDS virus

Baye's Theorem

- *Example cont'd*
  - What information is given to us?
    - $P(A) = 0.99$, *which implies* $P(A^c) = 0.01$
    - $P(B|A) = 0.03$ *(false-positive rate)*
    - $P(B|A^c) = 0.98$
- What are we looking for?
  - $P(A|B)$

Baye's Theorem

- *Example cont'd*
  - Using Baye's Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

$$P(A|B) = \frac{(0.03)(0.99)}{(0.03)(0.99) + (0.98)(0.01)} = 0.7519$$

- *Interpret:* Of the people who test positive for HIV/AIDS, over 75% of them aren't actually infected. What do you think of this probability?

False-Positives & False-Negatives

IMPORTANT

- False-Positive
  - given that a person does not have a disease, what is the probability that the test generates a positive result?
  - $P(positive|not\ infected)$
- False-Negative
  - given that a person has a disease, what is the probability that the test generates a negative result?
  - $P(negative|infected)$

The Monty Hall Problem

Random Variables & Probability Distributions

Random Variables

- A random variable (r.v.) is a function which assigns a unique number or result to each outcome in the sample space of an experiment
- A r.v. is traditionally denoted by upper-case letter X

When flipping a coin

$$X = \begin{cases} heads \\ tails \end{cases}$$

When rolling a die

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$

Random Variables and Probability

- Random variables are assigned a unique value associated with an outcome in the sample space of an experiment, and associated with that unique value is a probability
- *Example: Rolling a die*

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$

Random Variable

Outcomes

and

Random Variable

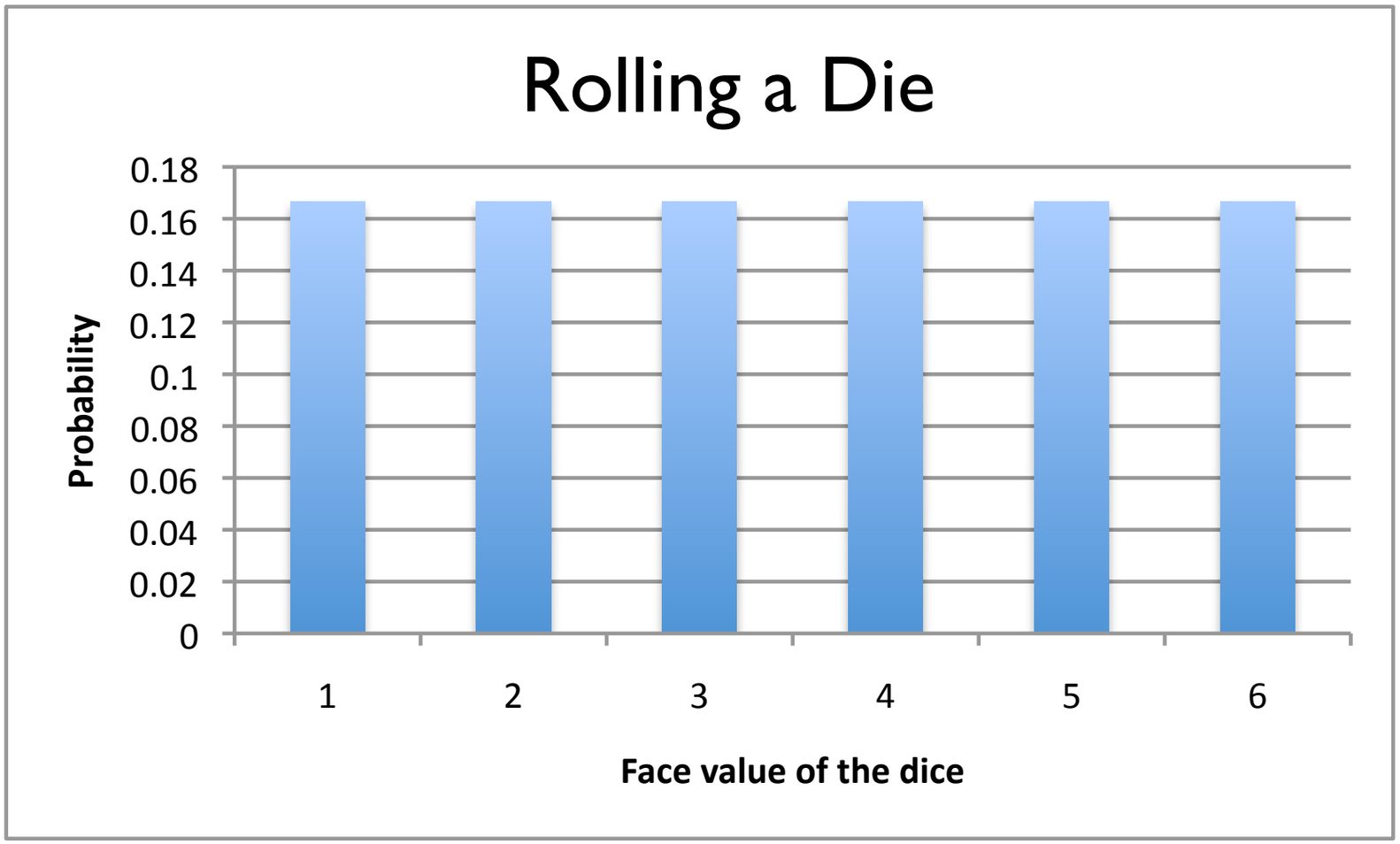
Outcome

Probability

$$P(X = 1) = 1/6$$
$$P(X = 2) = 1/6$$
$$P(X = 3) = 1/6$$
$$P(X = 4) = 1/6$$
$$P(X = 5) = 1/6$$
$$P(X = 6) = 1/6$$

Probability Mass Function

- The collection of probabilities that a *discrete* random variable can take can be represented graphically in a *probability **mass** function* (histogram)



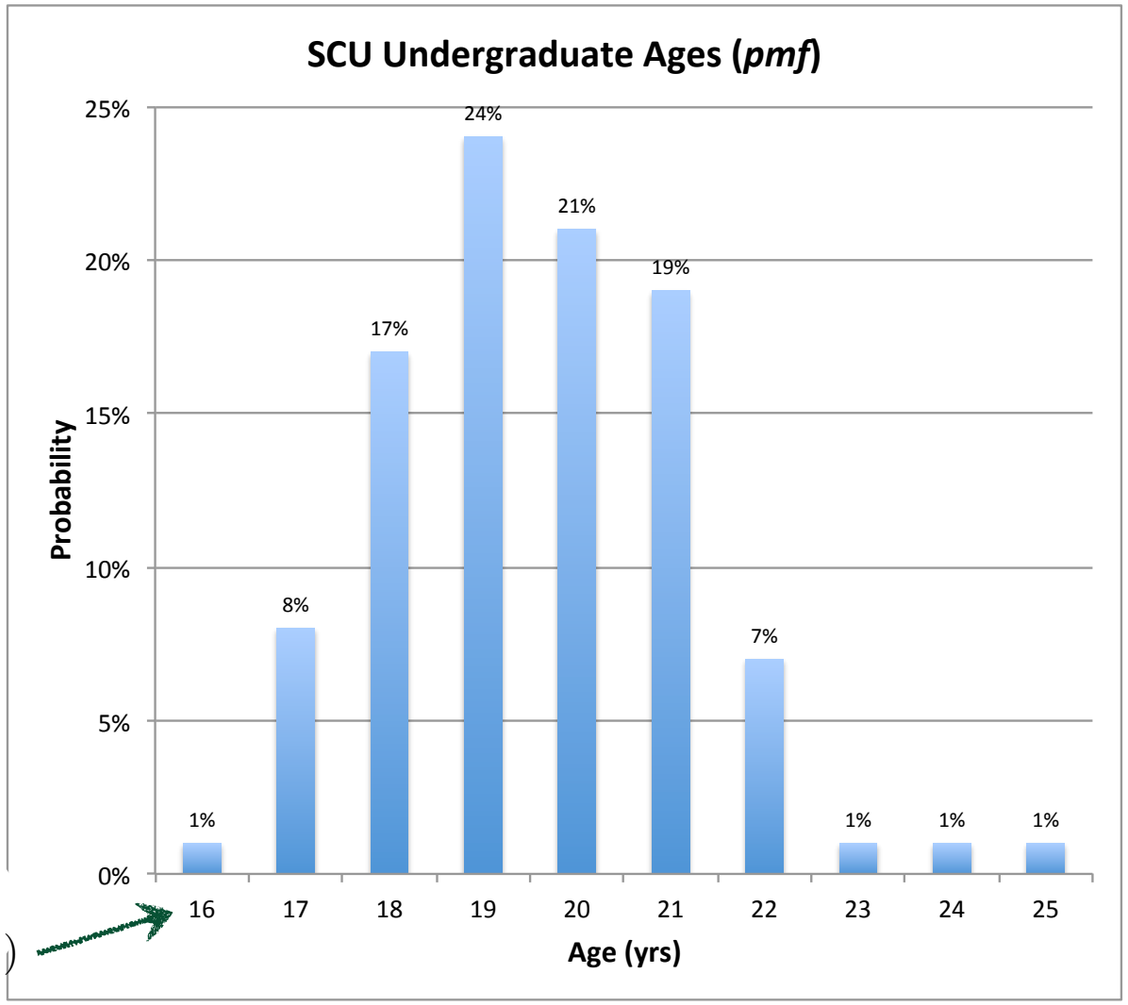
$$P(X = 1) = 1/6$$
$$P(X = 2) = 1/6$$
$$P(X = 3) = 1/6$$
$$P(X = 4) = 1/6$$
$$P(X = 5) = 1/6$$
$$P(X = 6) = 1/6$$

Probability Mass Function

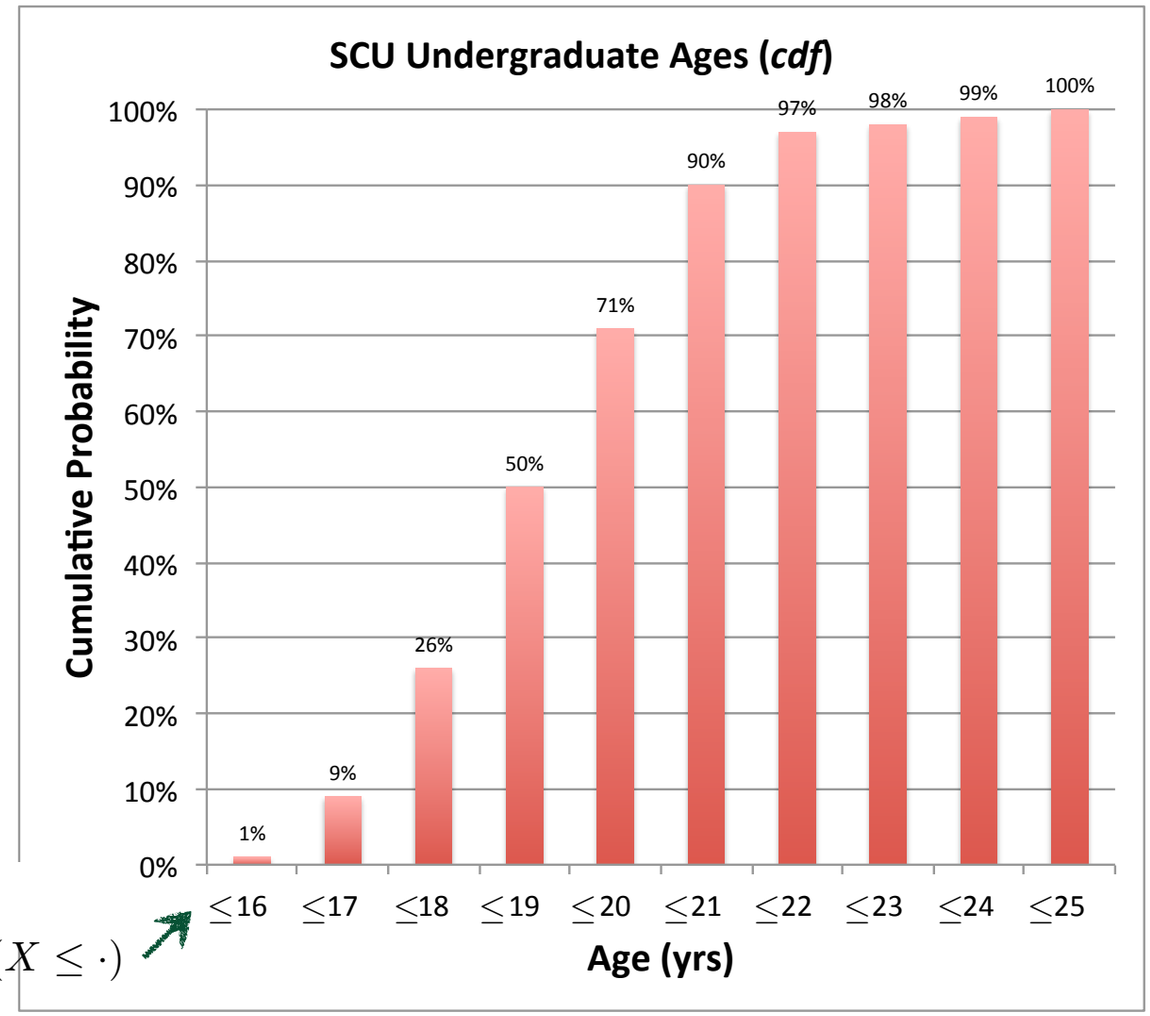
- Given a *pmf*, what questions can we answer?
- E.g., if we are looking at rolling a die, we can ask
  - *what is the probability of rolling a 2?*
  - *what is the probability of rolling an even #?*
  - *what is the probability of rolling a number > 1?*
- This last question is a segue into the cumulative distribution function (*cdf*)

Cumulative Distribution Function

- As its name implies, the *cdf* is the cumulative sum of probabilities (remember ogives?)
- Observe the direct link between the *pmf* and the *cdf*



$P(X = \cdot)$



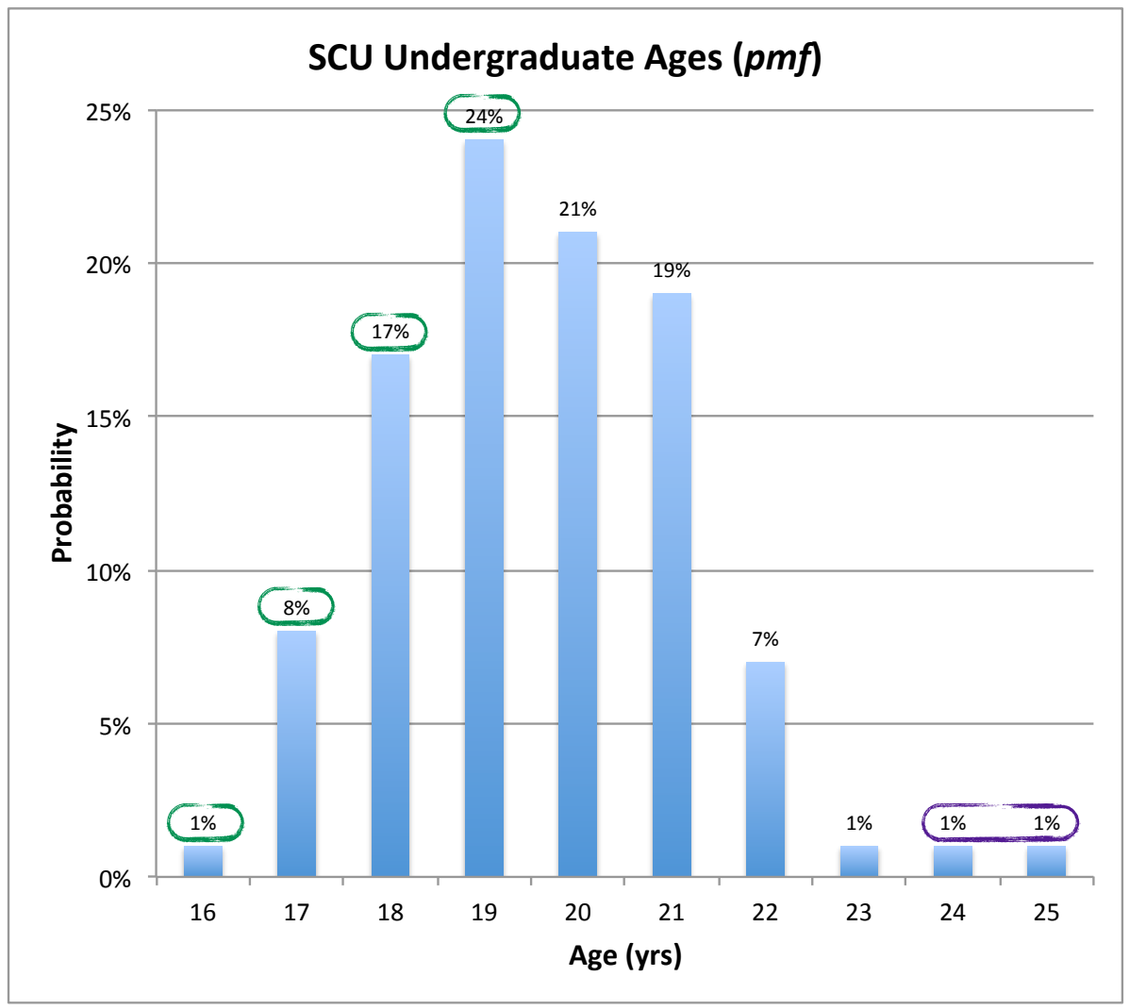
$P(X \leq \cdot)$

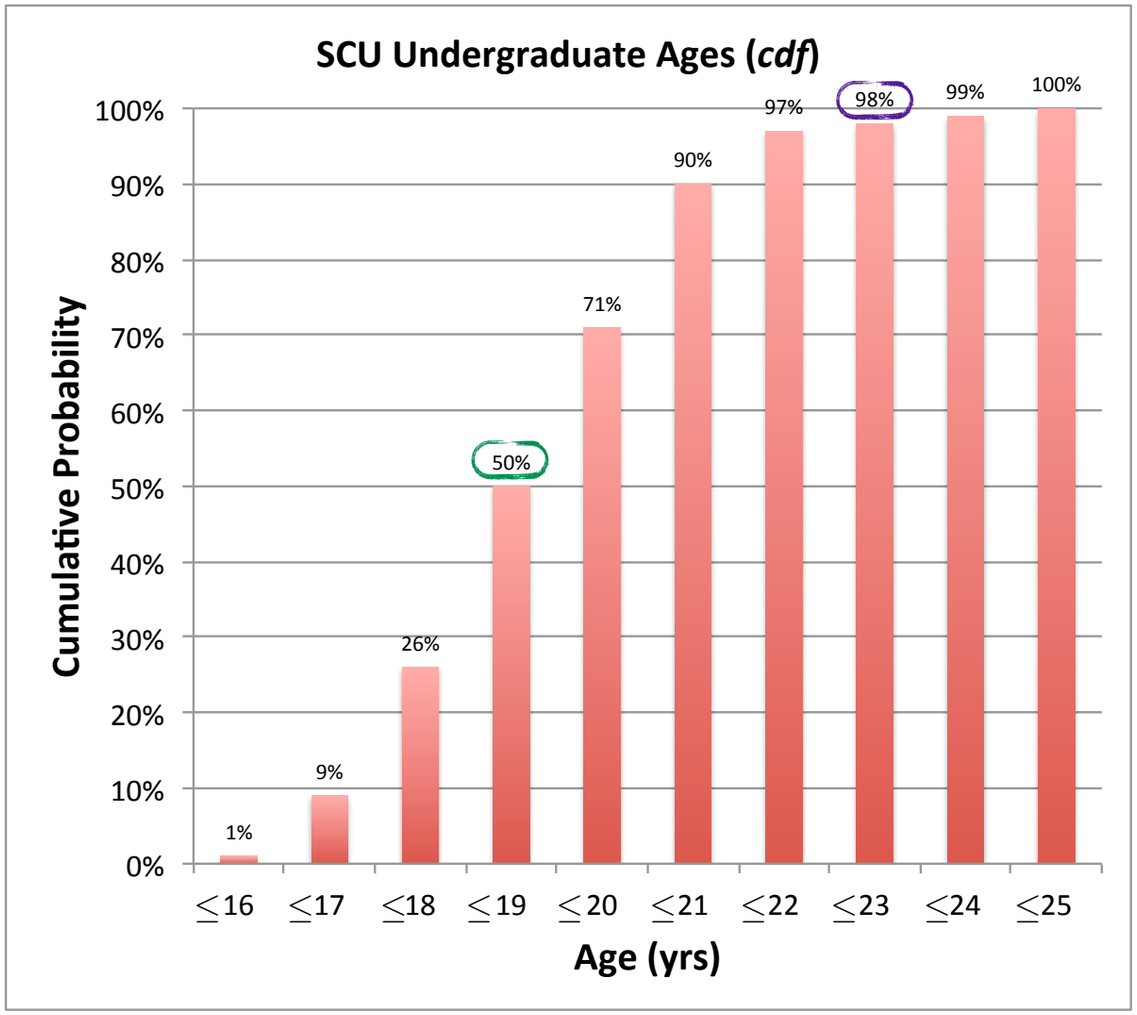
Cumulative Distribution Function

$$P(X \leq 19) = 0.01 + 0.08 + 0.17 + 0.24 = 0.50$$

$$P(X > 23) = 0.01 + 0.01 = 0.02$$

$$1 - P(X \leq 23) = 1 - 0.98 = 0.02$$





Probability Density Functions

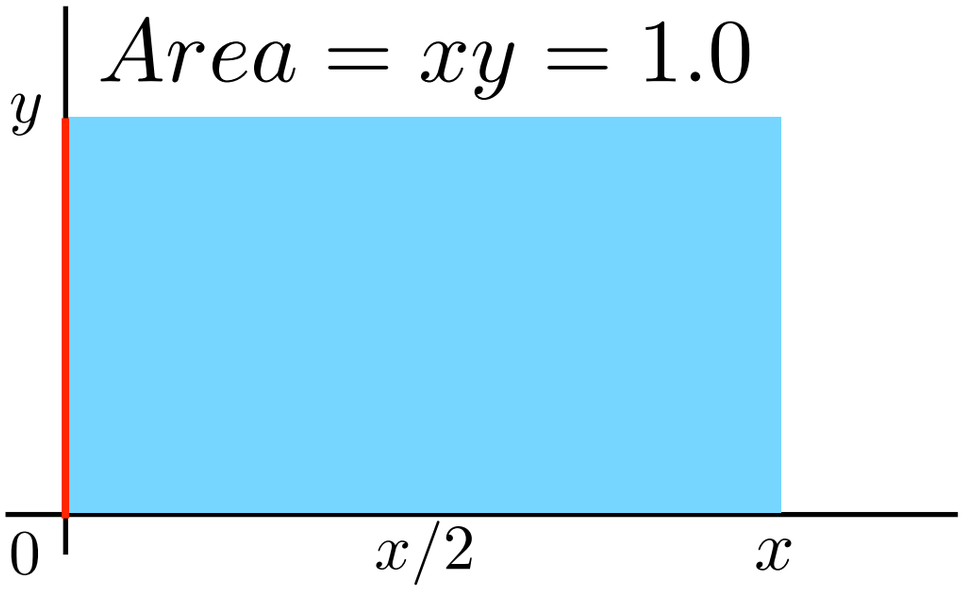
Uniform Distribution

Continuous Random Variables

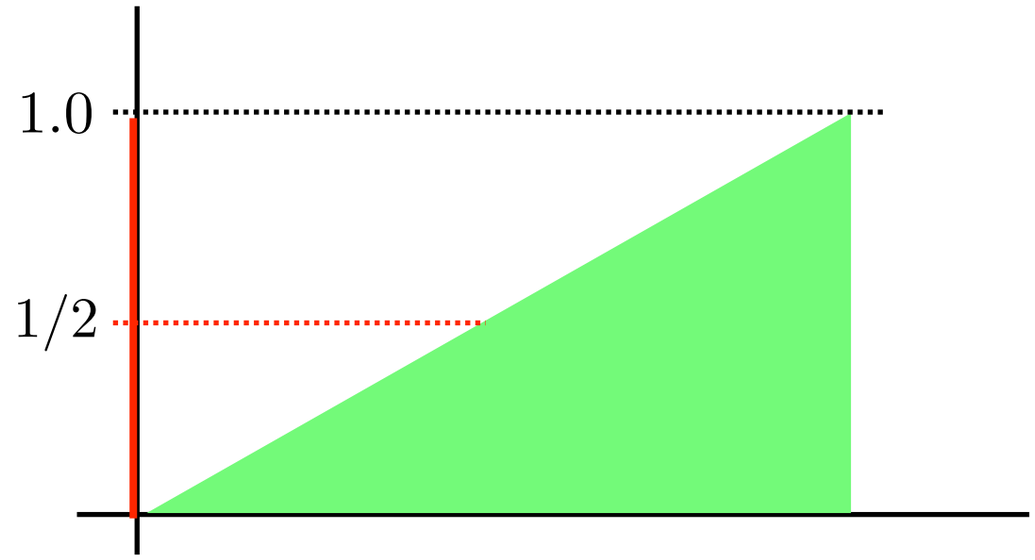
- If a random variable can take on any real-numbered value (e.g., π , $\sqrt{2}$), we define that random variable to be ***continuous***
- We are not able to represent the probability distribution of the random variable X with a *probability mass function* because its possible values are infinite; instead we will define a ***probability density function (pdf)*** such that **areas and not heights** represent probabilities
- The cumulative counterpart to the pdf remains the *cumulative distribution function (cdf)*

Area Under a Continuous Distribution

- The area under a probability density function *is equal to one*, which we intuitively understand to be the maximal value a probability can take
- *Observe the link between the pdf and cdf*



pdf

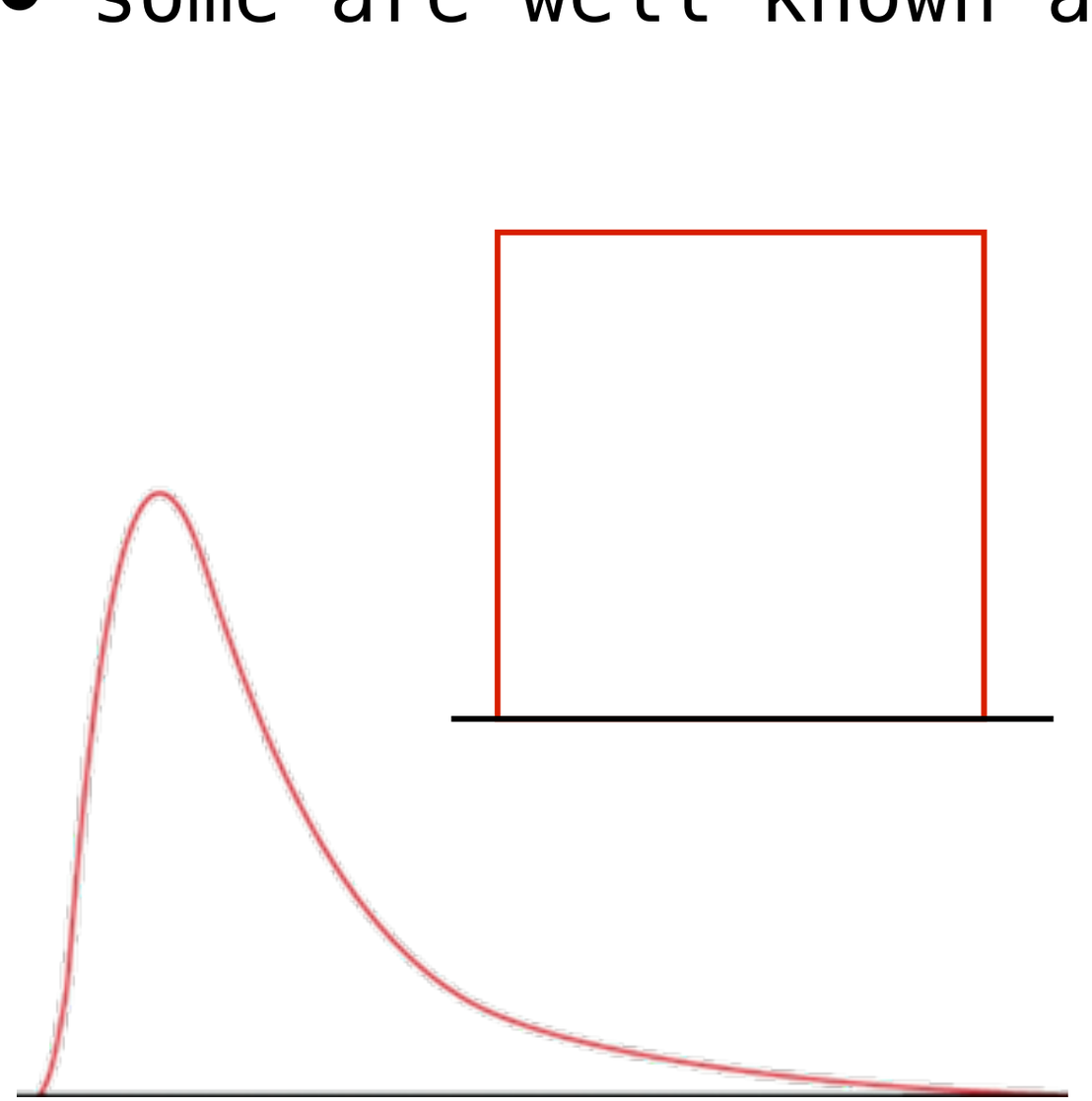


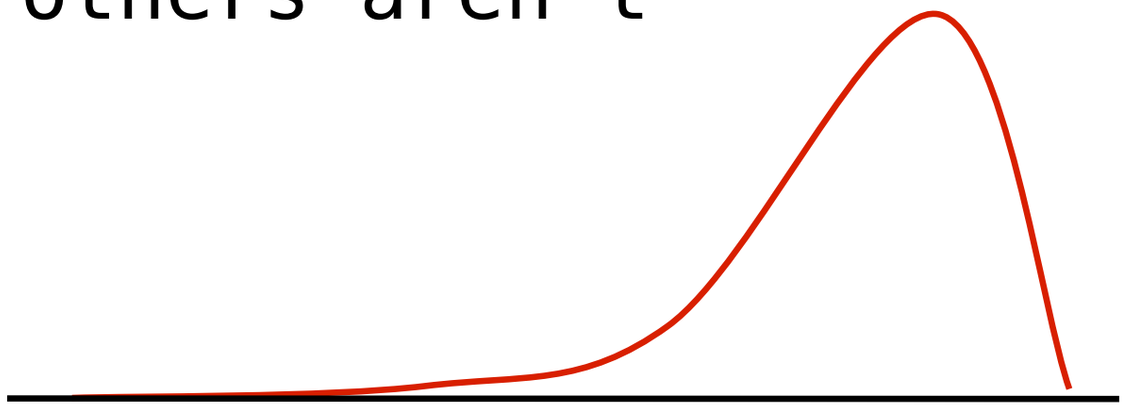
cdf

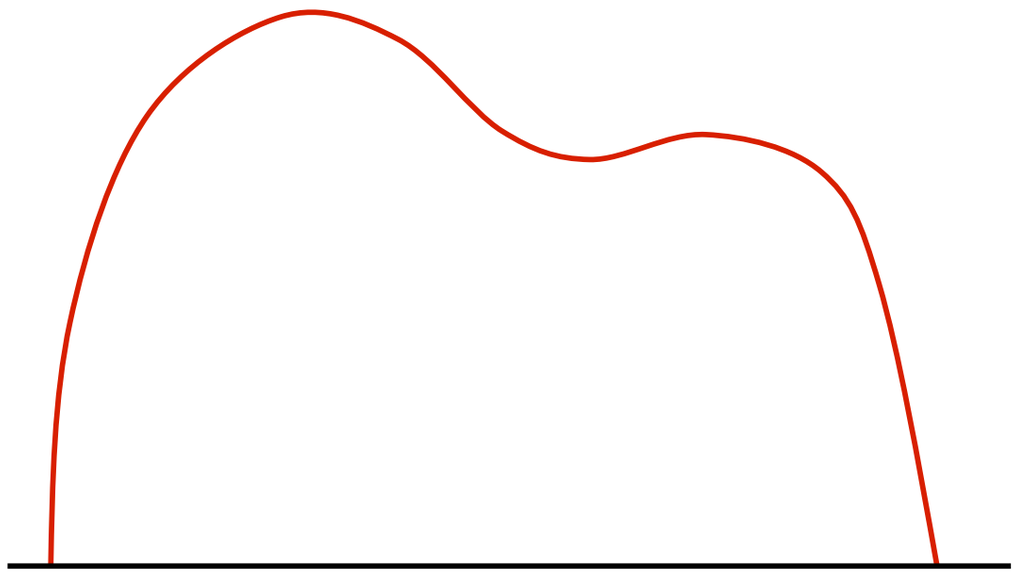
Observe that the cdf is the integral (the area) of the pdf

Probability Density Functions

- pdf's come in any imaginable shape
- some are well known and others aren't





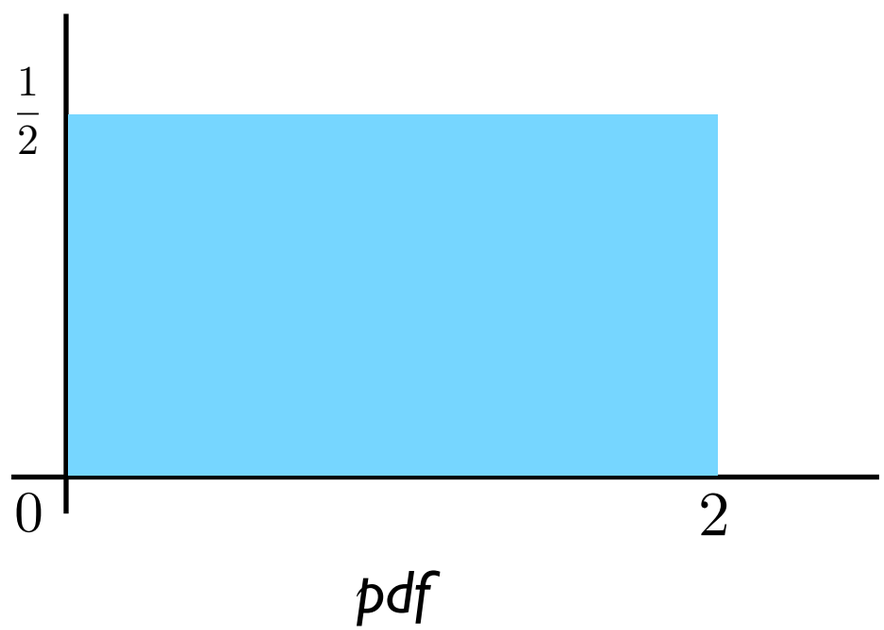


Do we Need to Compute Integrals?

- ***NO, you will never be asked to compute an integral for an assignment or a test***
  - Normally, we need to integrate over continuous probability distributions to determine probabilities, but it turns out that *we are going to examine continuous distributions that have easy shortcuts which do not involve us taking integrals*

Uniform Distribution

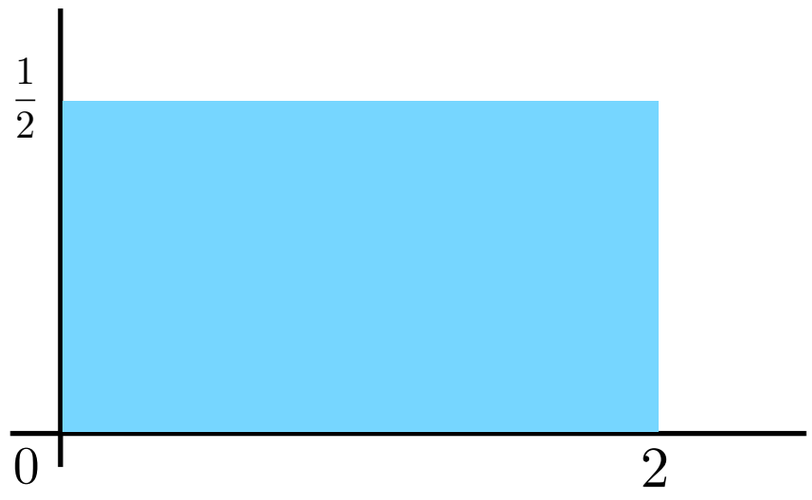
- A uniform distribution – as its name implies – assigns equal probability to every possible value over which the distribution spans
- In the example uniform distribution displayed below, every number from 0 to 2 has equal probability of being selected when sampling from the distribution



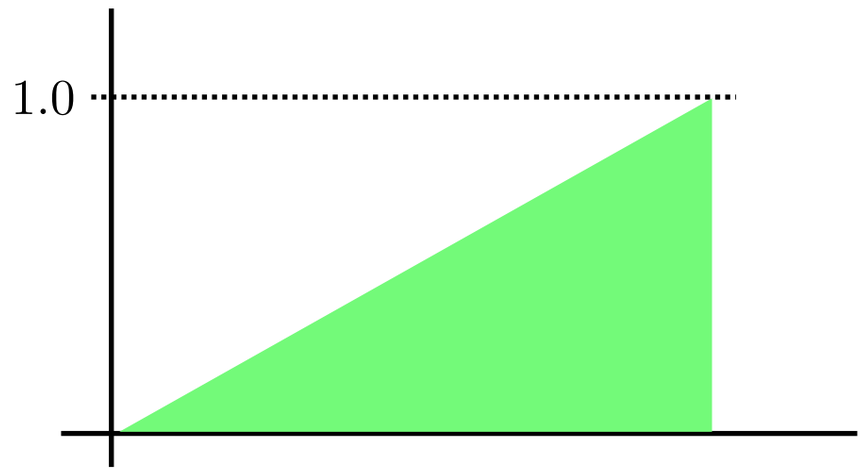
What do you think a discrete Uniform pdf would look like?

Uniform Distribution

- A very nice property about the Uniform Distribution is that is it always rectangular in shape, and computing the area of a rectangle is trivial
- *<u>Example</u>: what is the probability that a random variable sampled from the uniform distribution below is between 1.81 and 0.63?*



pdf

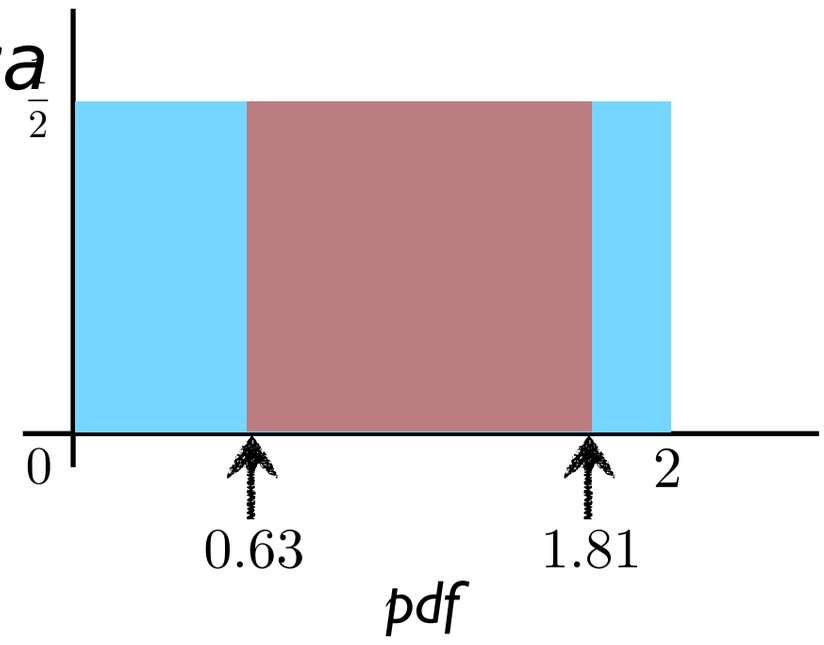


cdf

Uniform Distribution

- *Example (cont'd)*
  - *Using integrals* $\int_{0.63}^{1.81} \frac{1}{2} dx = \frac{1.81}{2} - \frac{0.63}{2} = 0.59$
  - *More simply using basic algebra*
    - *compute the area*

$$(1.81 - 0.63) \times (1/2) = 0.59$$

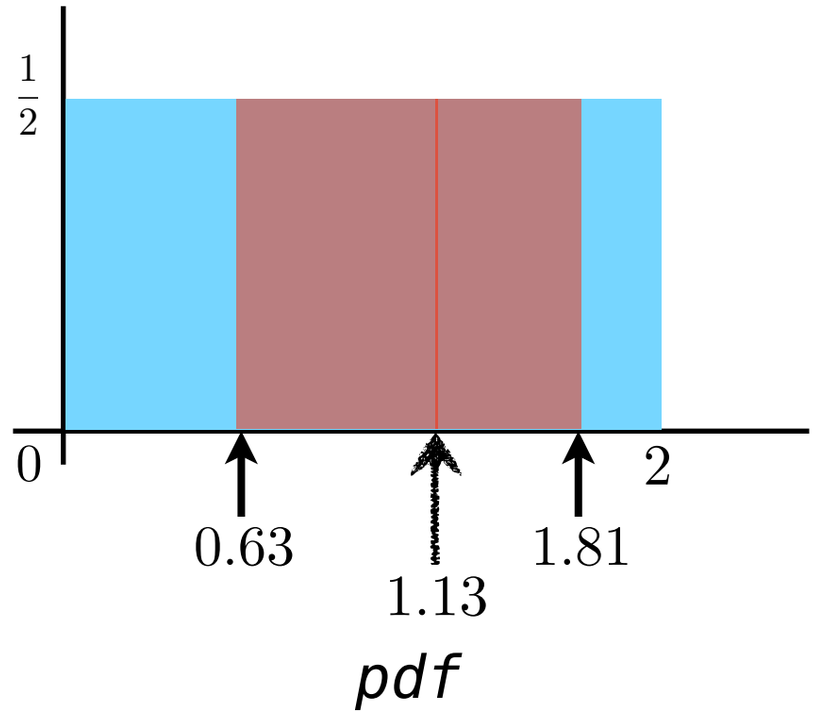


- Both methods arrive at the same conclusion that 59% of random samples from this uniform distribution will fall between 1.81 and 0.63

Uniform Distribution

- *Another Example*
  - What is the probability that a random variable sampled from the uniform distribution below is equal to 1.13?
  - What are the integration limits?

$$\int_{1.13}^{1.13} \frac{1}{2} dx = \frac{1.13}{2} - \frac{1.13}{2} = 0$$



  - What is the area of a line?

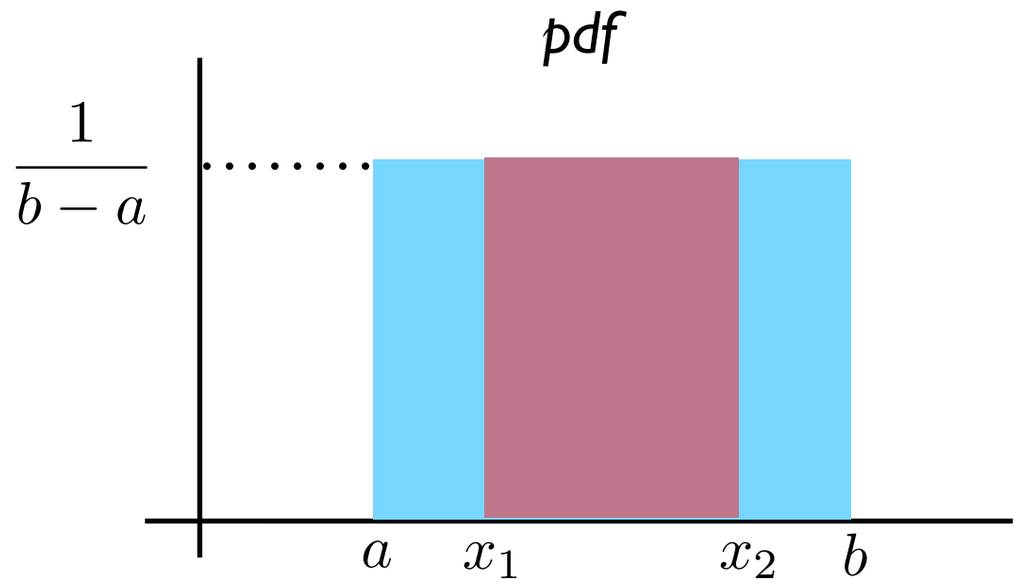
$$(1.13 - 1.13) \times (1/2) = 0$$

Very Important Observation

- For ***ALL*** continuous distributions, the probability of a random variable X being equal to a single value is always equal to 0

Uniform Distribution

- More generally, for any uniform distribution with the following pdf



$$P(x_1 < X < x_2) = \frac{x_2 - x_1}{b - a}$$

Uniform Distribution: *Example*

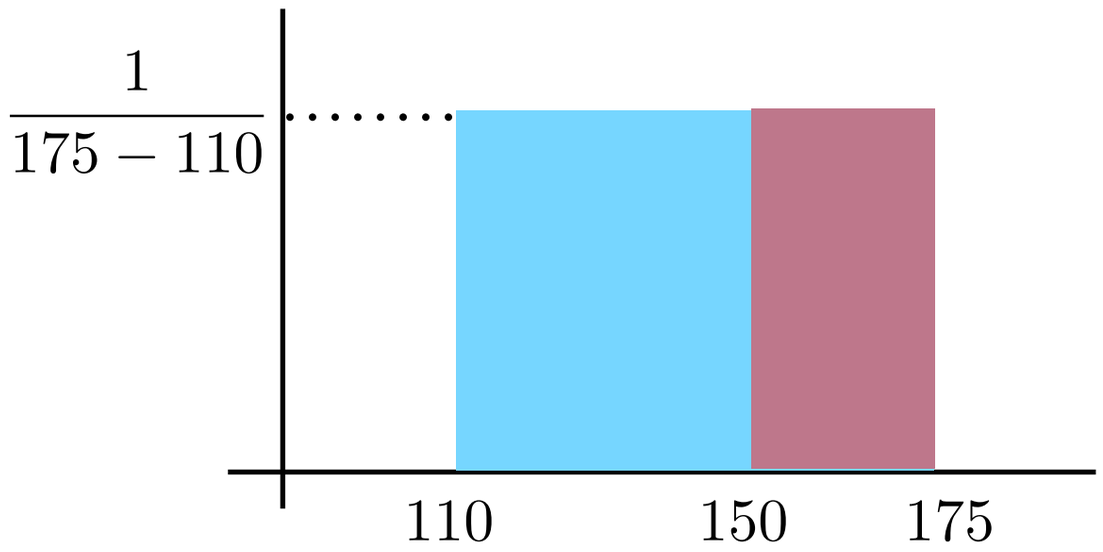
- The weekly output of a steel mill is a uniformly distributed random variable that lies between 110 and 175 metric tons.

a. Compute the probability that the steel mill will produce more than 150 metric tons this week.

b. Determine the probability that the steel mill will produce between 120 and 160 metric tons next week.

Uniform Distribution: *Example*

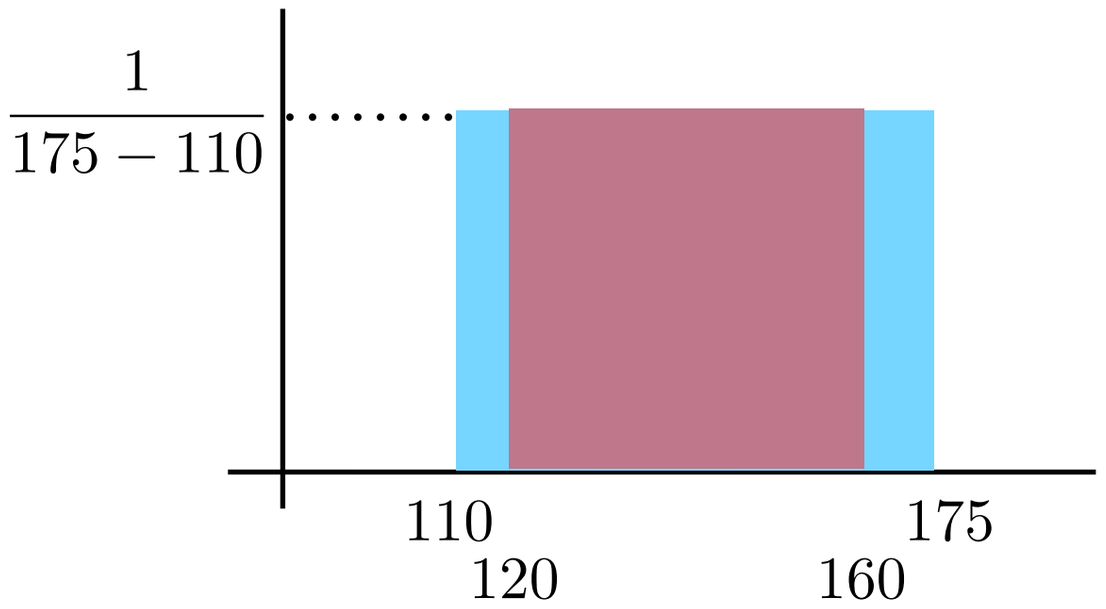
a. Compute the probability that the steel mill will produce more than 150 metric tons this week.



$$P(150 < X < 175) = \frac{175 - 150}{175 - 110} = \frac{25}{65} = 0.385$$

Uniform Distribution: *Example*

b. Determine the probability that the steel mill will produce between 120 and 160 metric tons next week.



$$P(120 < X < 160) = \frac{160 - 120}{175 - 110} = \frac{40}{65} = 0.615$$

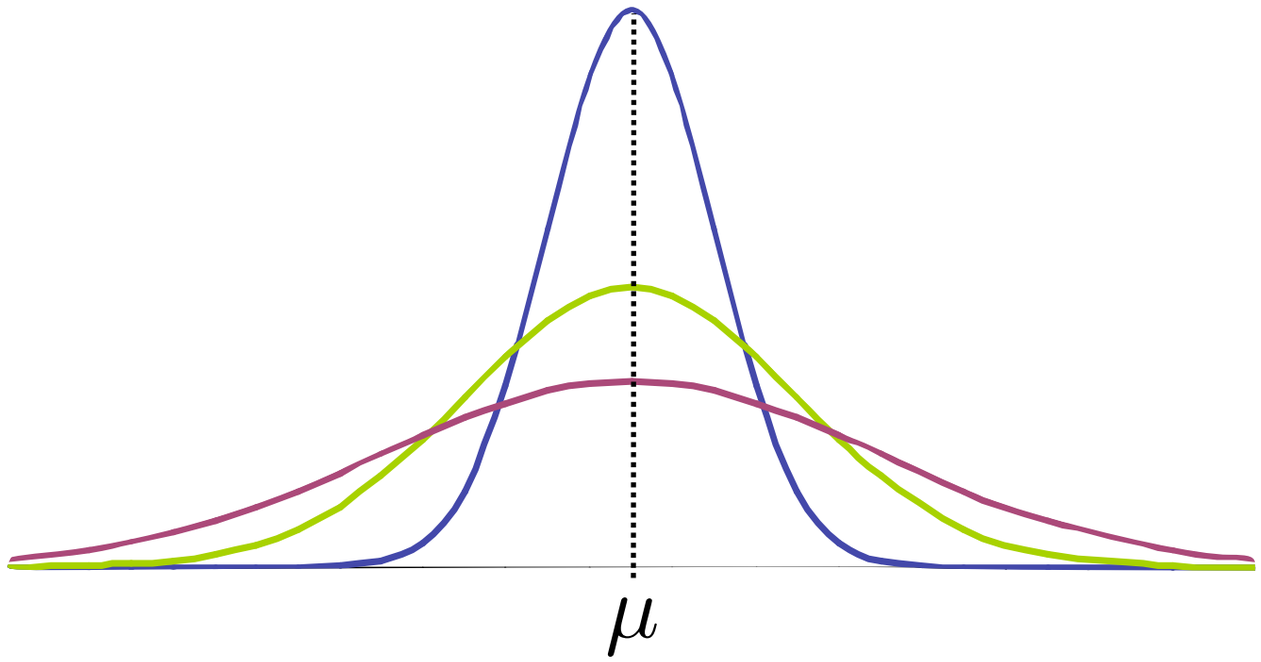
Normal Distribution

Normal Distribution

- The other continuous distribution we will examine in this class is the ***Normal Distribution***
- A defining characteristic is the *symmetric clustering of data around the mean of the distribution*
- This distribution tends to occur very often in various disciplines, e.g.,
  - people's heights, blood pressure, weight (natural sciences)

Normal Distribution: Characteristics

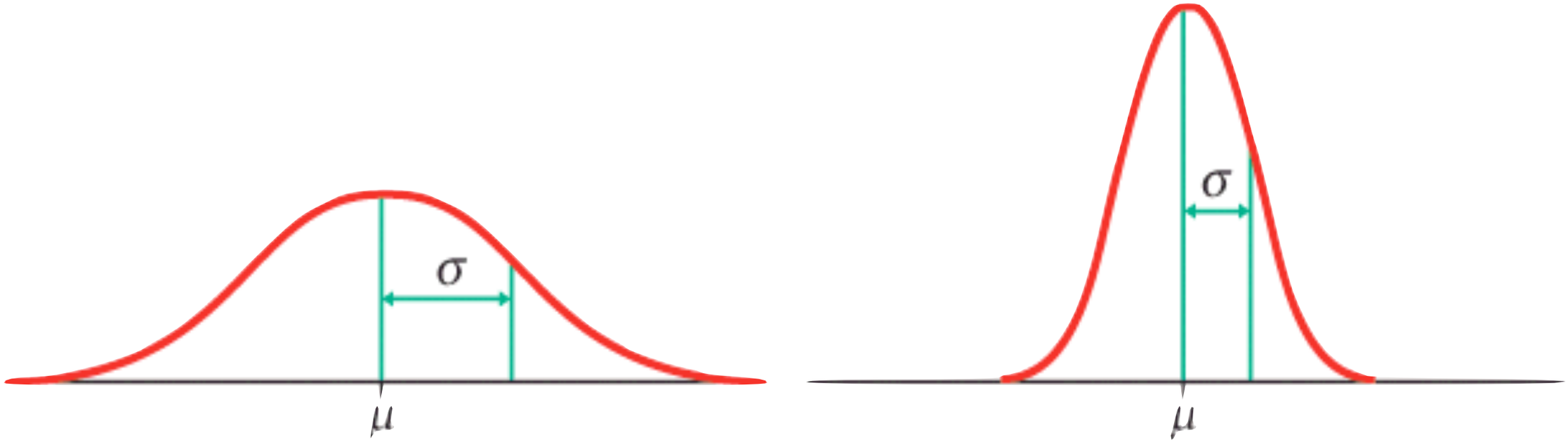
- Symmetric about the mean
- Infinitely long tails
- Can be *wholly described using only two parameters*, *mean and variance*



various normal pdf's

Normal Distribution: Parameters

- Whereas the mean dictates where the center of the distribution lies, the variance dictates the spread and ultimately shape of the curve (*narrow versus broad)*

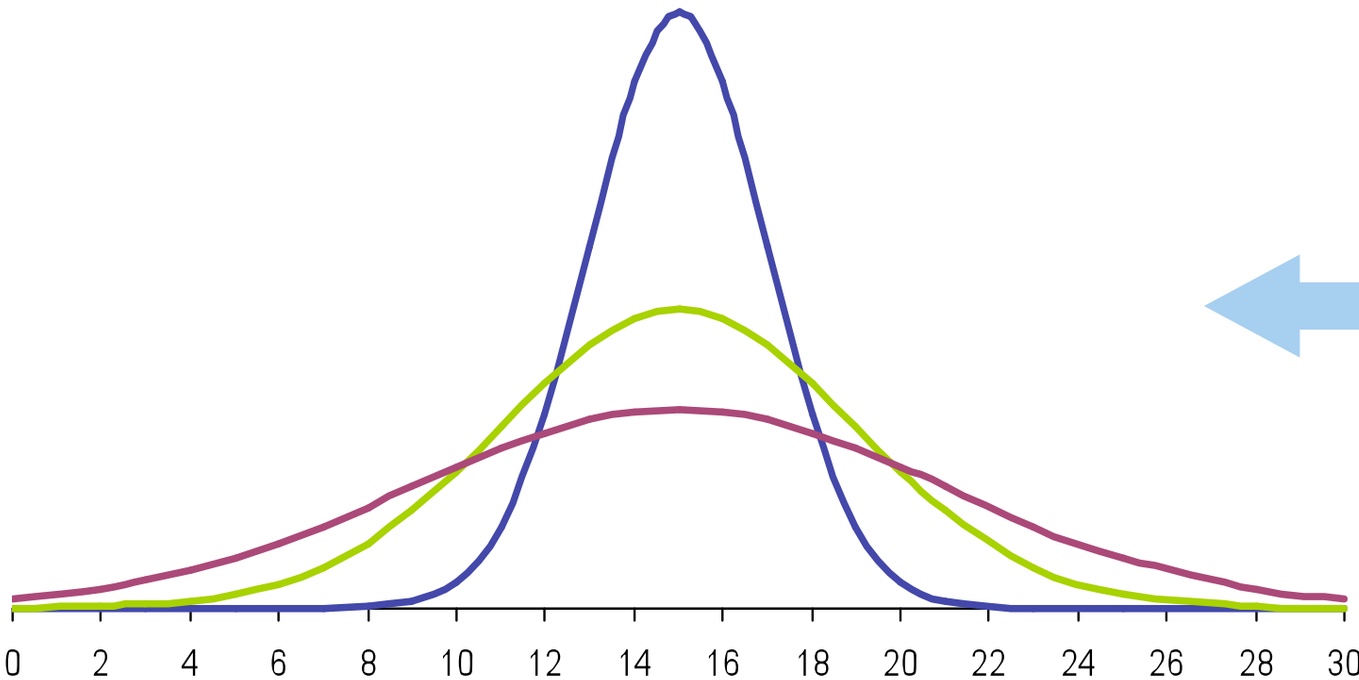


Nomenclature

$$\sim N(\mu, \sigma^2)$$

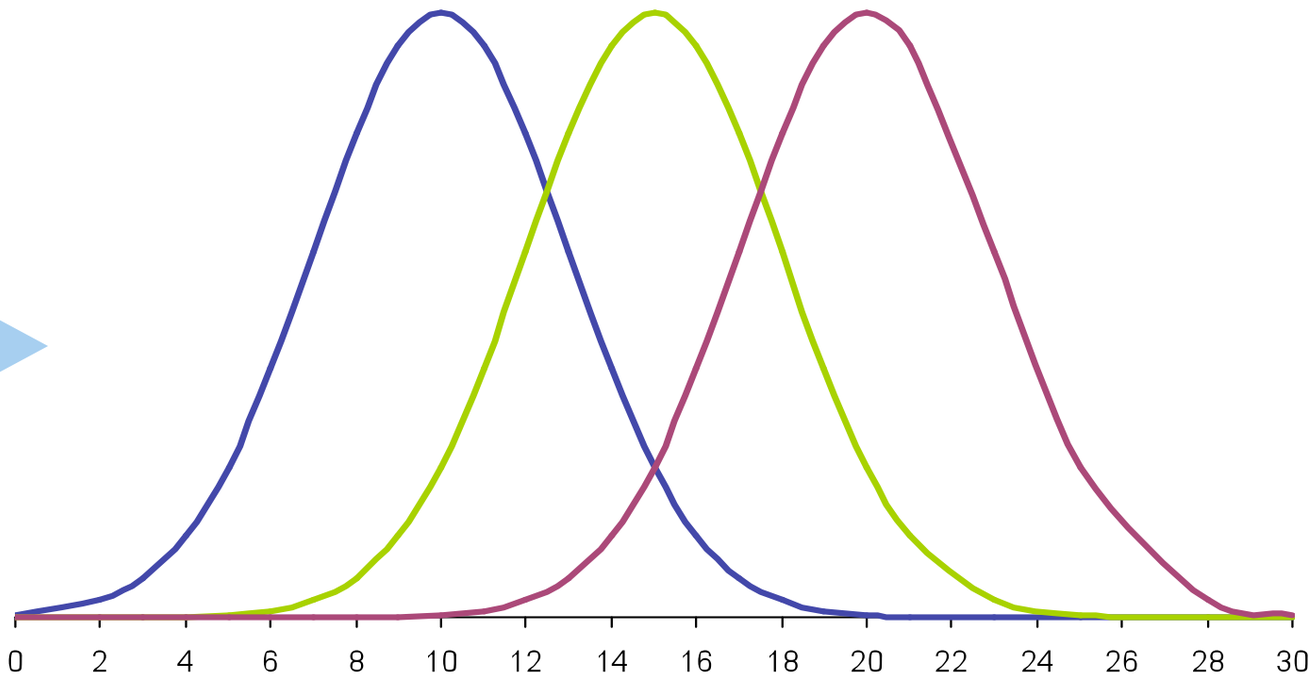
distributed normally mean variance

Normal Distribution: Parameters



Same means
Different
variances

Different means,
Same variance



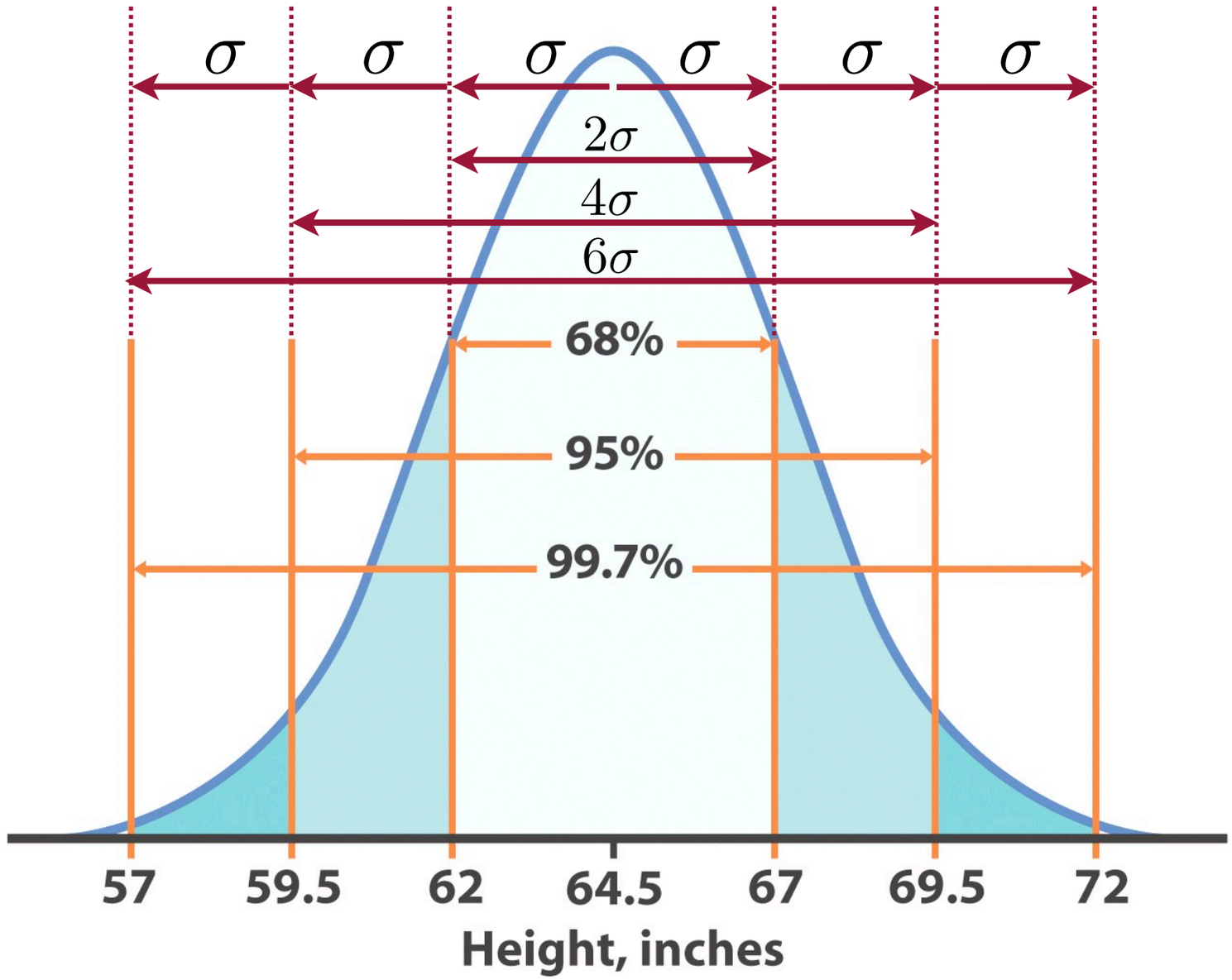
Normal Distribution: Sigma Relationship

- A very unique and defining feature of the normal distribution is the relationship with the standard deviation (σ)
- Moving from the mean (or *center because the normal distribution is symmetric*), if we compute the area under the curve at a distance of one sigma in each direction, we will have captured 68% of the area, regardless of the values of the mean (μ) and the variance (σ^2)
- At a distance of two and three sigma, the area under the curve increases to 95% and 99.7% respectively

Normal Distribution: Sigma Relationship

- *Example*: We determine (*somehow*) that the heights (in inches) students is

$$\sim N(64.5, 6.25) \Rightarrow \mu = 64.5, \sigma^2 = 6.25, \sigma = 2.5$$



Area Under the Curve

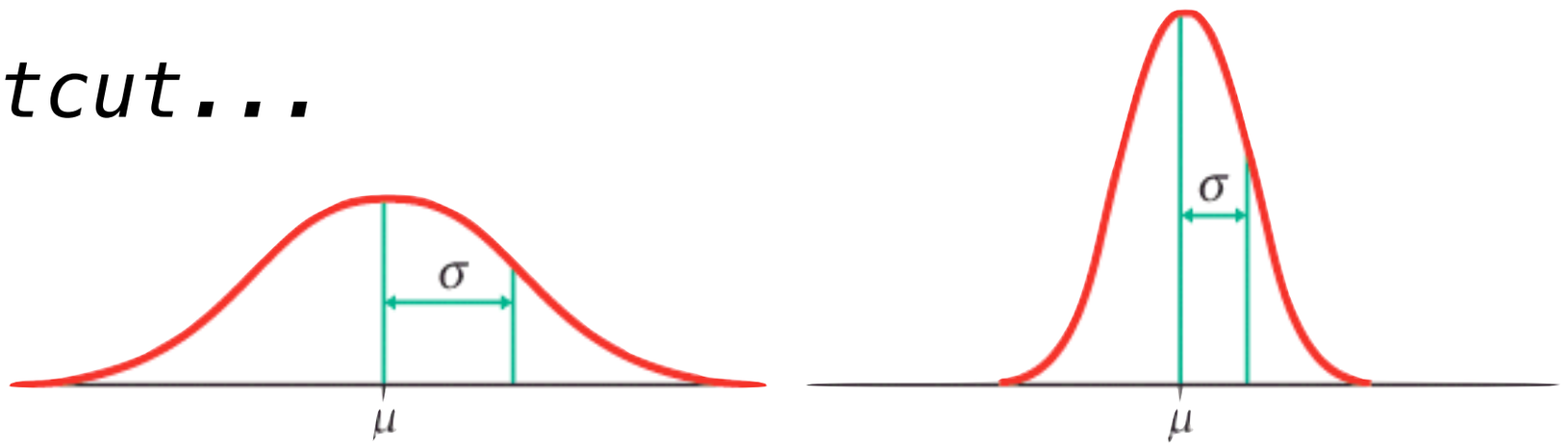
- Recall from our previous slides that determining the probability of a random variable X from a continuous distribution involved integrating to compute an area
- This was easy enough for the Uniform Distribution which was rectangular (we integrated a constant), but for the normal distribution we would have to integrate the following:

$$\int_{l_1}^{l_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Area Under the Curve

- Look impossible? (*It is*)
- There is no closed form (*analytical*) solution to this integral
- Good numerical approximations have been computed using computer packages

- Although numerical approximations algorithms exist, do we want to have to run an approximation algorithm every time we want a probability?
- *There is a shortcut...*

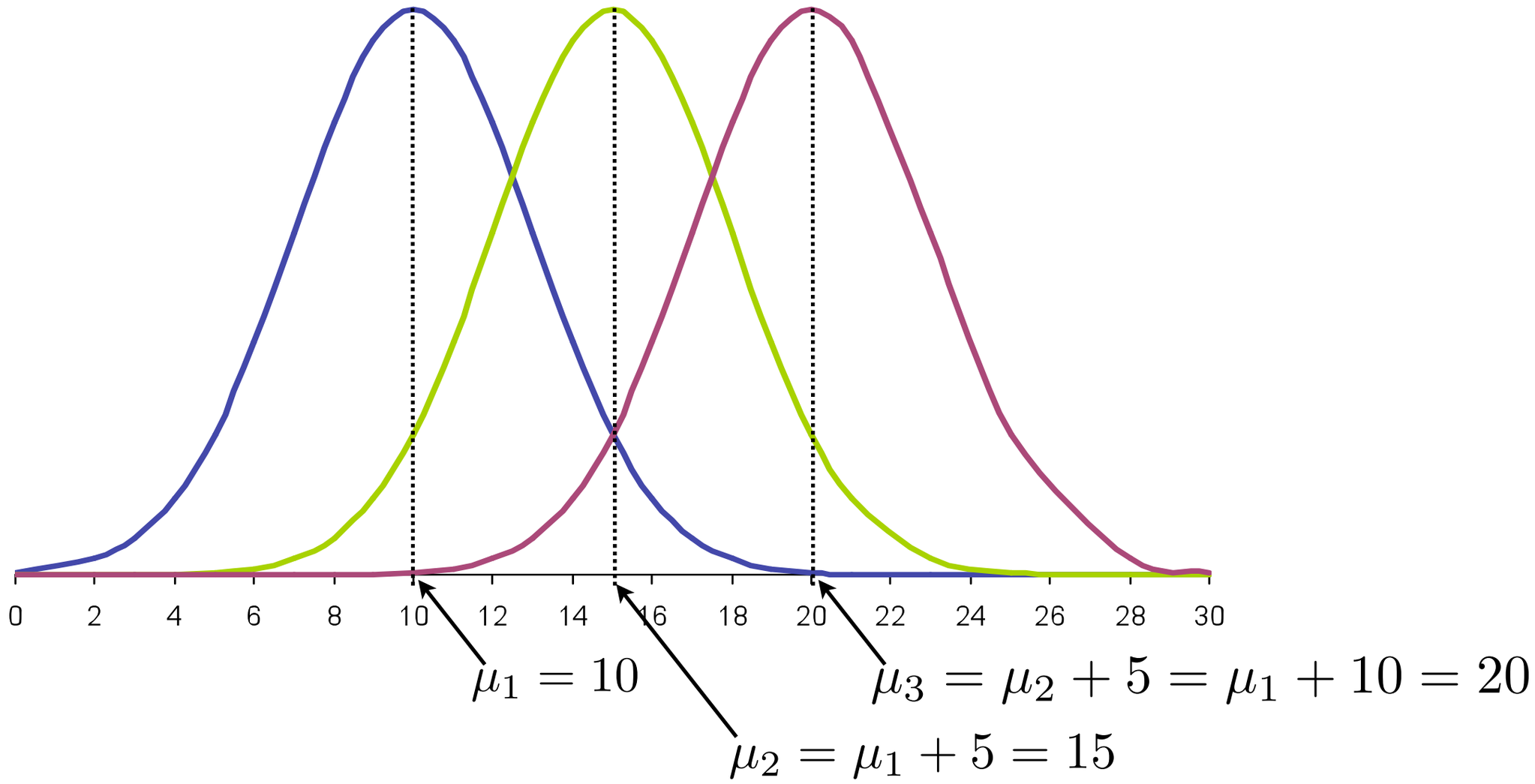


Standardizing the Normal Distribution

- *Recall* that all Normal distributions share identical characteristics: *infinite tails, centered and symmetric about the mean*
- Leveraging these characteristics, we can transform any Normal distribution into a common, standard version
- For this standardized version of the Normal distribution, we then run the numerical approximation algorithm once, record the values, stick them in the back of a book, and never worry about this problem again (*which is exactly what is done*)

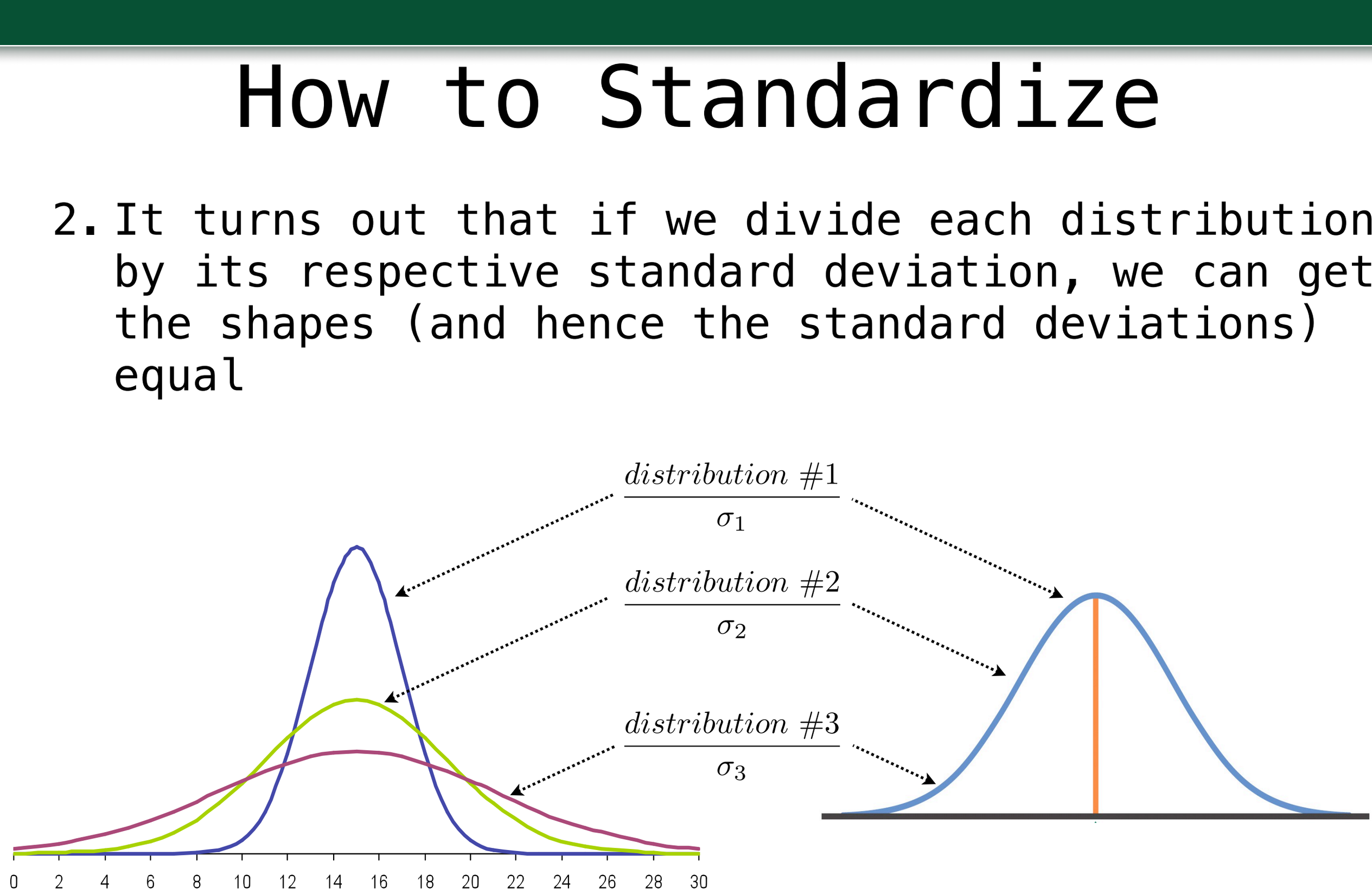
How to Standardize

1. Recall from a previous graphic (*repeated here*) that we can shift a Normal distribution without changing its shape just by shifting the mean



How to Standardize

2. It turns out that if we divide each distribution by its respective standard deviation, we can get the shapes (and hence the standard deviations) equal



How to Standardize

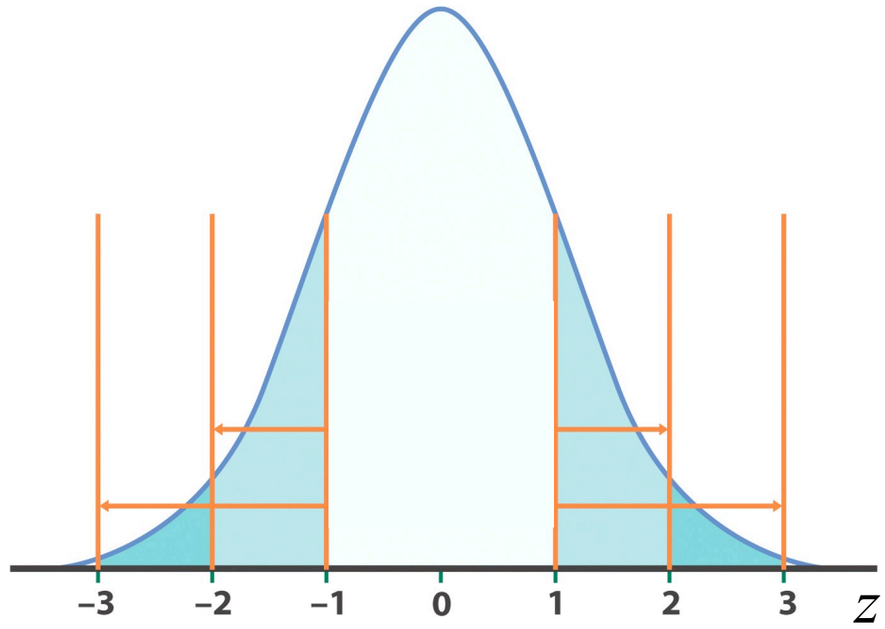
- Combining steps 1 and 2, we obtain a formula for standardizing any Normal distribution

1. Subtract the mean (μ) to shift the distribution such that it is centered about zero
2. Divide by the standard deviation (σ) to standardize the shape of the curve

$$given\ X \sim N(\mu, \sigma^2)$$

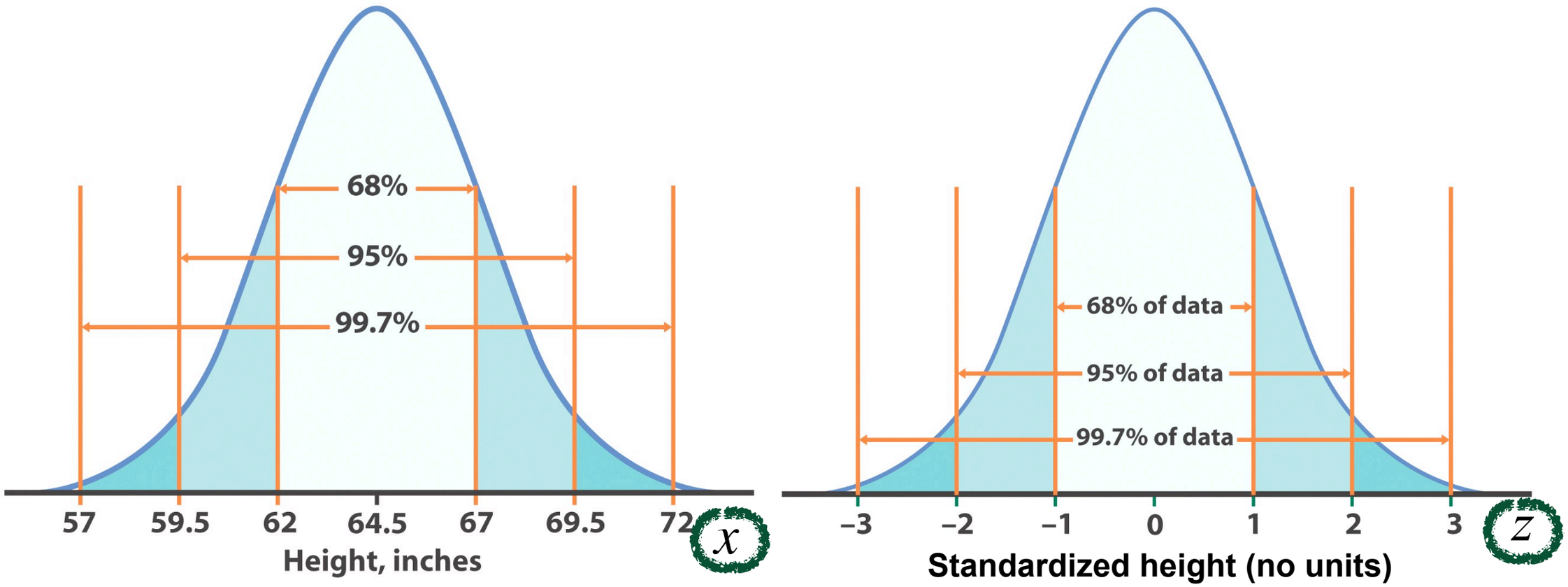
we can compute

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



Standardized Normal Distribution

- Z represents the standardized version of the random variable X
- Z has a mean $\mu = 0$ and a variance $\sigma^2 = 1$
- Observe the standardized sigma relationship...



Z-Table

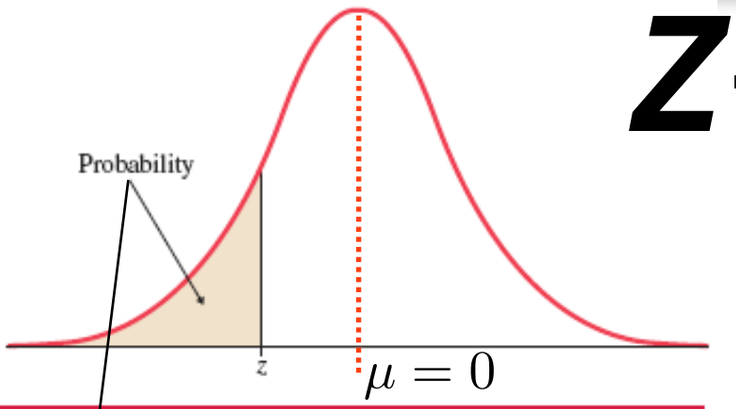


Table entry for z is the area under the standard normal curve to the left of z.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

# of σ's from the mean (μ)

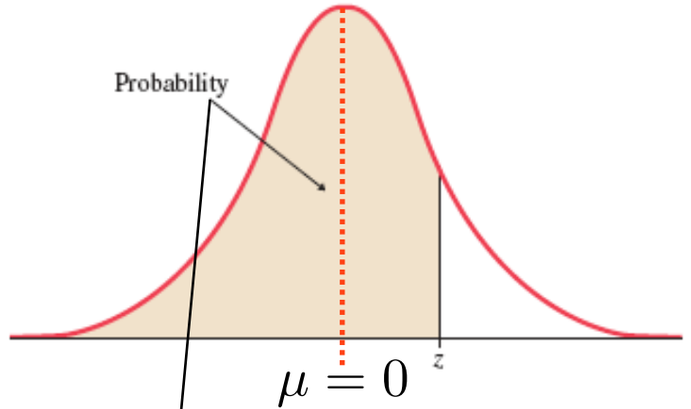


Table entry for z is the area under the standard normal curve to the left of z.

TABLE A Standard normal probabilities *(continued)*

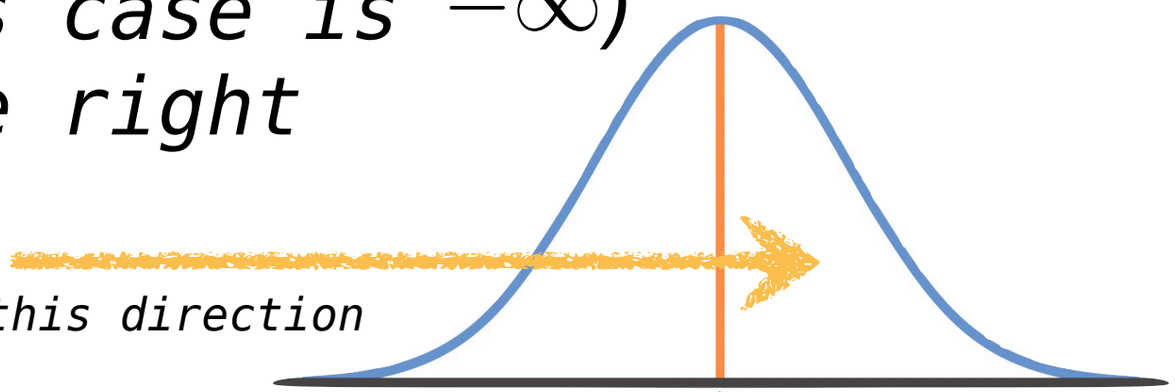
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Interpretation of the *Z*-Table

- The Z-values in the top and left margins are the number of standard deviations (where $\sigma = 1$) you are away from the mean (where $\mu = 0$)
- The values in the middle of the table are probabilities, namely, the probability that lies between $-\infty$ and the z-value that corresponds with the cell you have selected

Elementary Observations

- On Z-Table, we notice that at a value $Z = 0$, the area under the curve (probability) is equal to 0.50
  - *Interpretation*: $Z = 0$ implies that we are 0 $\sigma's$ above the mean (μ), and since the Normal distribution is symmetric about the mean, the area under the curve from $Z = -\infty$ to $Z = 0$ is 0.50
- *note that when integrating to find the area under the curve, we begin from the left-most point (which in this case is $-\infty$) and integrate to the right*



Elementary Observations

- We claimed earlier that when we were 1, 2, and 3 σ around the mean, that the area under the curve was 68%, 95%, and 99.7% respectively; let's verify that on the Z-Table
- $Z = 1.00$ implies we are 1σ above the mean, and from the table, we observe that at $Z = 1.00$, the area under the curve (or the probability) is equal to 0.8413. Why doesn't that match up with the 68% we claimed?
  - When looking up $Z = 1.00$ in the Z-Table, we obtain the probability (area under the curve) from $Z = 1.00$ to $Z = -\infty$

Elementary Observations

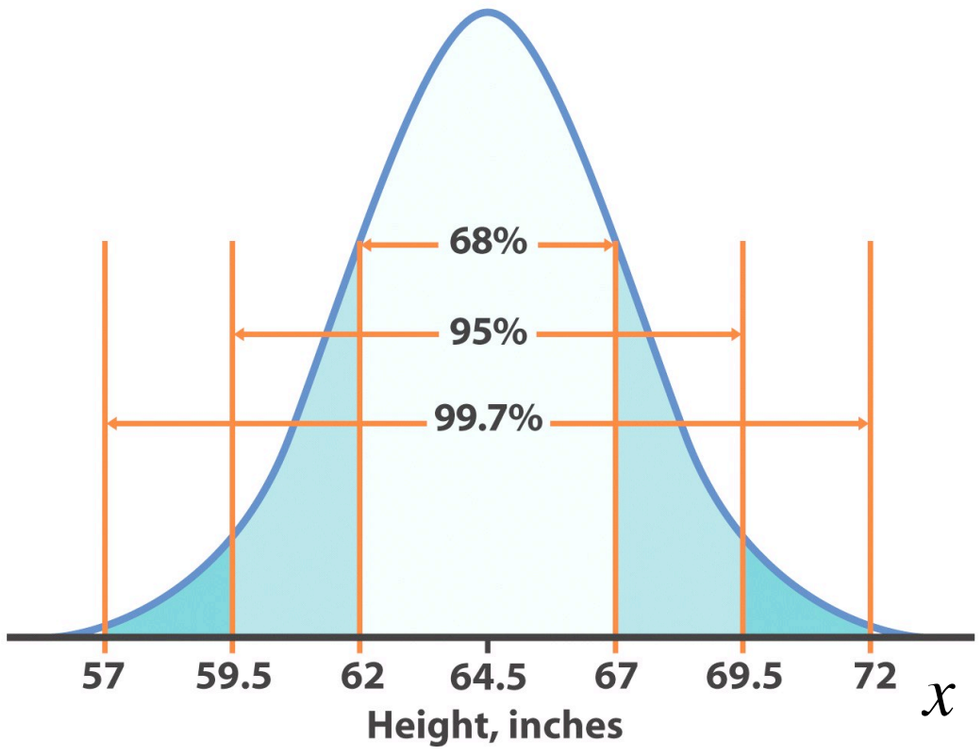
- To obtain the 68% we are seeking, we must take the value from the Z-Table at $Z = 1.00$ and then subtract the part we don't want which is the left part of the tail from $Z = -\infty$ to $Z = -1.00$

$$Z_{(1.00)} - Z_{(-1.00)}$$

$$0.8413 - 0.1587 = 0.6828$$

Normal Probabilities

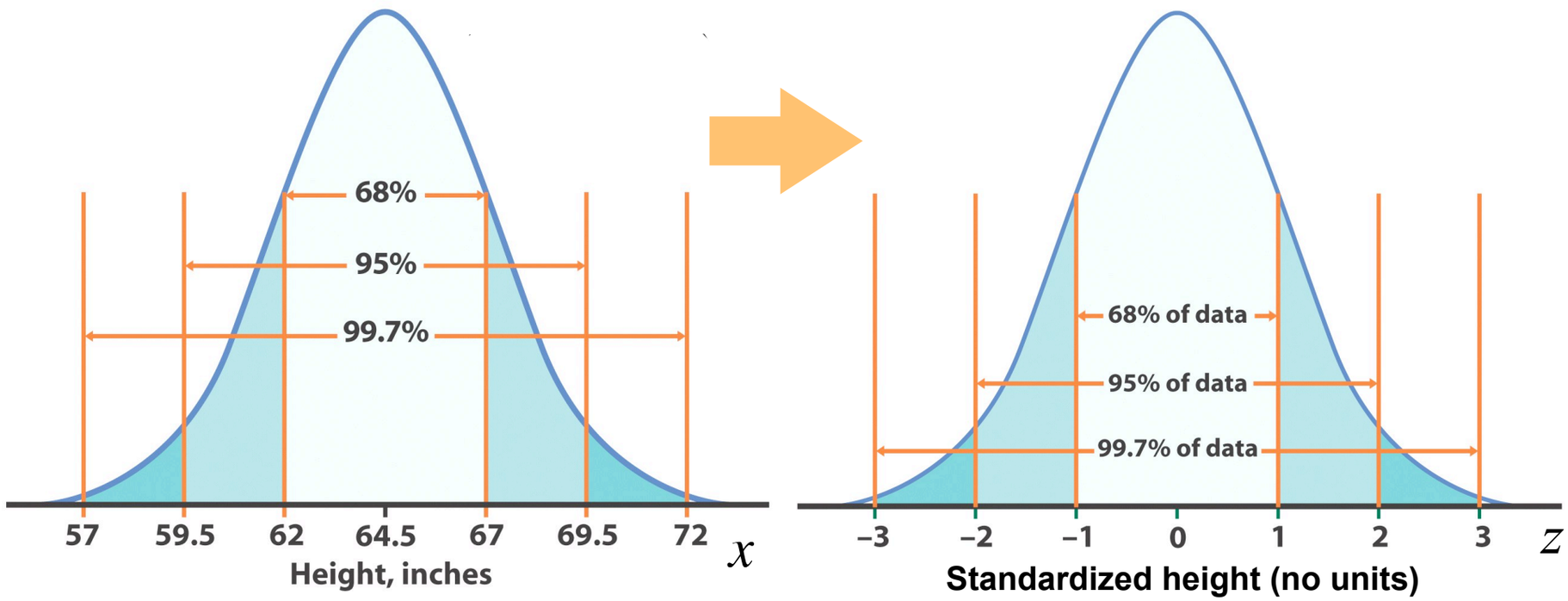
- *Example (cont'd)*
  - Given the graph below, what are the parameters of this Normal distribution? $\sim N(64.5, 6.25)$
  - What is the probability, when sampling from this distribution, of someone's height being between 68.35 and 60.22 inches?



Normal Probabilities

- *Example (cont'd)*
  - *Firstly, we need to standardize both Normal distribution values so we can use the Z-Table*

$$\frac{68.35 - 64.5}{2.5} = 1.54 \qquad \frac{60.22 - 64.5}{2.5} = -1.712$$



Normal Probabilities

- *Example (cont'd)*
  - *Now we can restate the question:*
    - *What is the probability, when sampling from the standardized Normal distribution, that a random variable Z will be between −1.712 and 1.54*
  - *Look in the Z-Table to determine the values for $Z = -1.712$ and $Z = 1.54$*

z-table only contains z-values in the margins accurate to two decimal places – round numbers up or down at your discretion

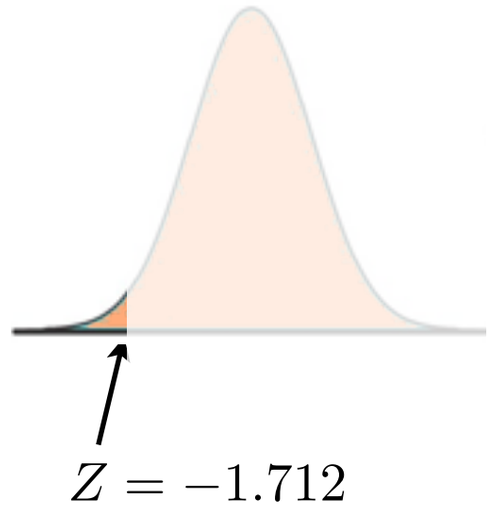
Normal Probabilities

- *Example*

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

$$Z_{(-1.712)} = 0.0436$$



Normal Probabilities

- *Example*

TABLE A Standard normal probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$$Z_{(1.54)} = 0.9382$$

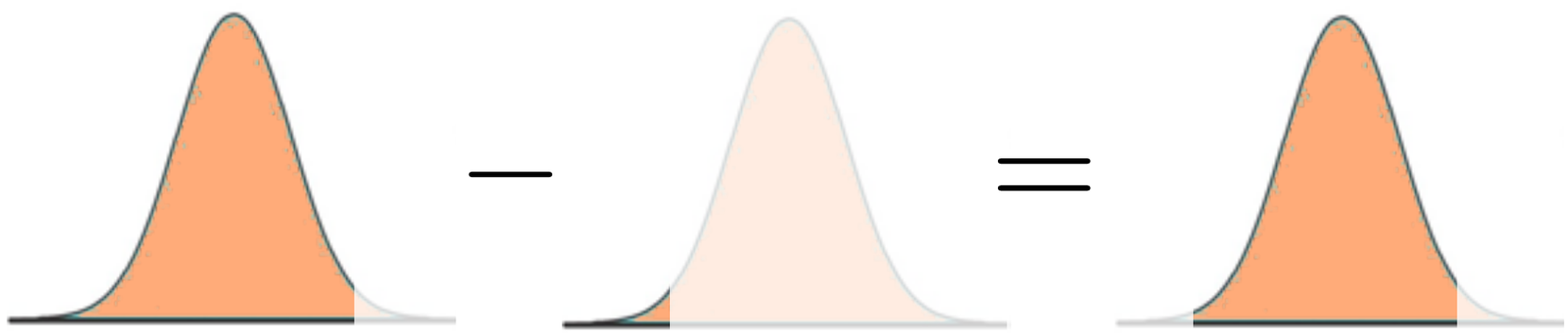
$$Z = 1.54$$

Normal Probabilities

- *Example (cont'd)*
  - *We can finally compute the probability*

$$Z_{(1.54)} - Z_{(-1.712)}$$

$$0.9382 - 0.0436 = 0.8946$$



Normal Probabilities

- **Conclusion and Interpretation**
  - We can conclude that a random variable Z sampled from a standard Normal distribution has a 89.46% chance of being between −1.712 and 1.54

$$P(-1.712 \leq Z \leq 1.54) = 0.8946$$

  - We can conclude by association that sampling from the original Normal distribution of heights that the probability that a random variable X is between 60.22 and 68.35 inches is 89.46%

$$P(60.22 \leq X \leq 68.35) = 0.8946$$

Sampling Distribution of the Mean

The Law of Large Numbers

- Recall that for a distribution, a ***parameter*** (e.g., μ) is a fixed and (usually) unknown number
- A ***statistic***, e.g., $\bar{x}$, is a random variable
- A Simple Random Sample (SRS) should fairly represent a population, hence $\bar{x}$ should be fairly close to μ
- The Law of Large Numbers (LLN) ensures that as our sample size increases, the statistic $\bar{x}$ is guaranteed to get closer and closer to the parameter μ

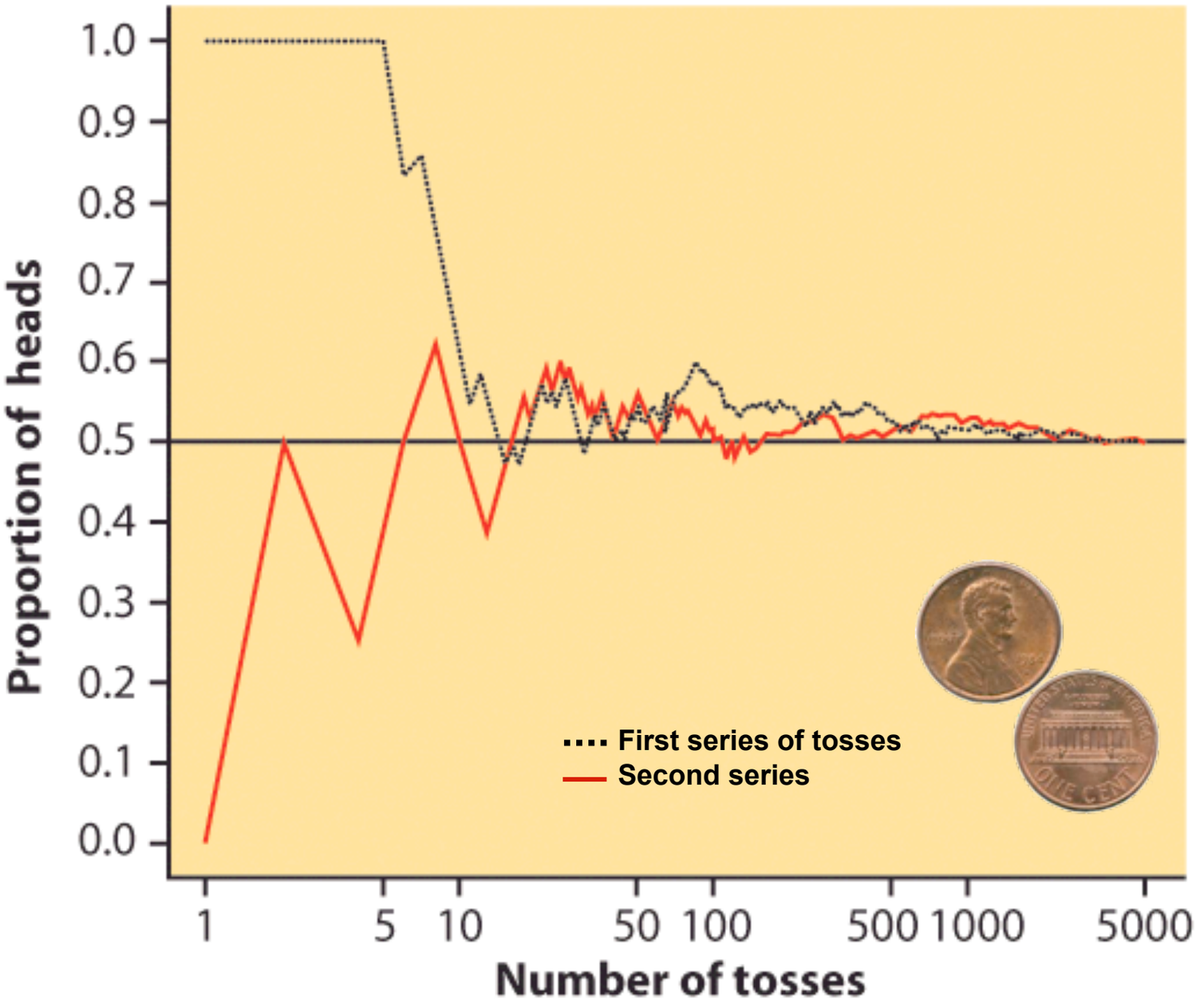
The Law of Large Numbers

- Formal Definition
  - Draw independent observations at random from any population with a finite mean μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values gets closer and closer to the mean μ of the population.

- *Less formally, this is also referred to as a "Regression to the mean"*

LLN: Flipping a Coin

- Gambling and the Law of Large Numbers



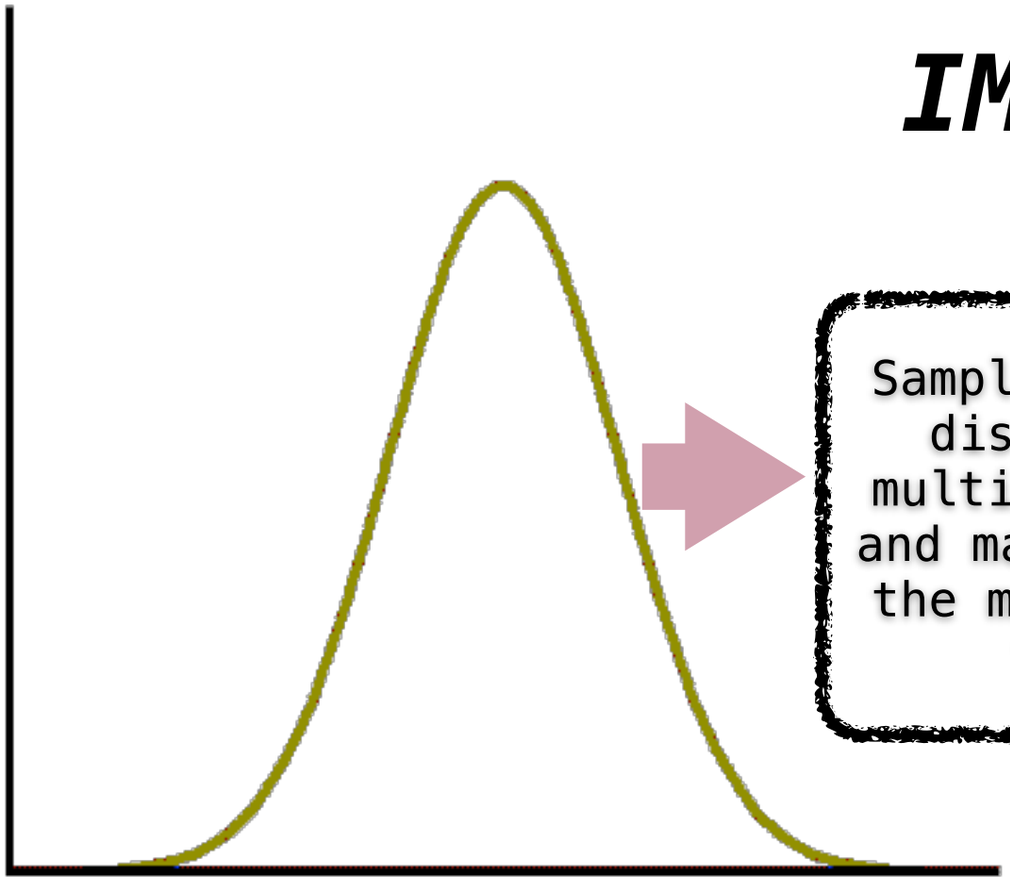
Mean & Std Dev of a ***Sample Mean***

- Suppose that $\bar{x}$ is the *sample mean* of an SRS of size n drawn from a large population with *population mean* μ and a population standard deviation σ . Then the mean of the sampling distribution is μ and its sample standard deviation is $\frac{\sigma}{\sqrt{n}}$

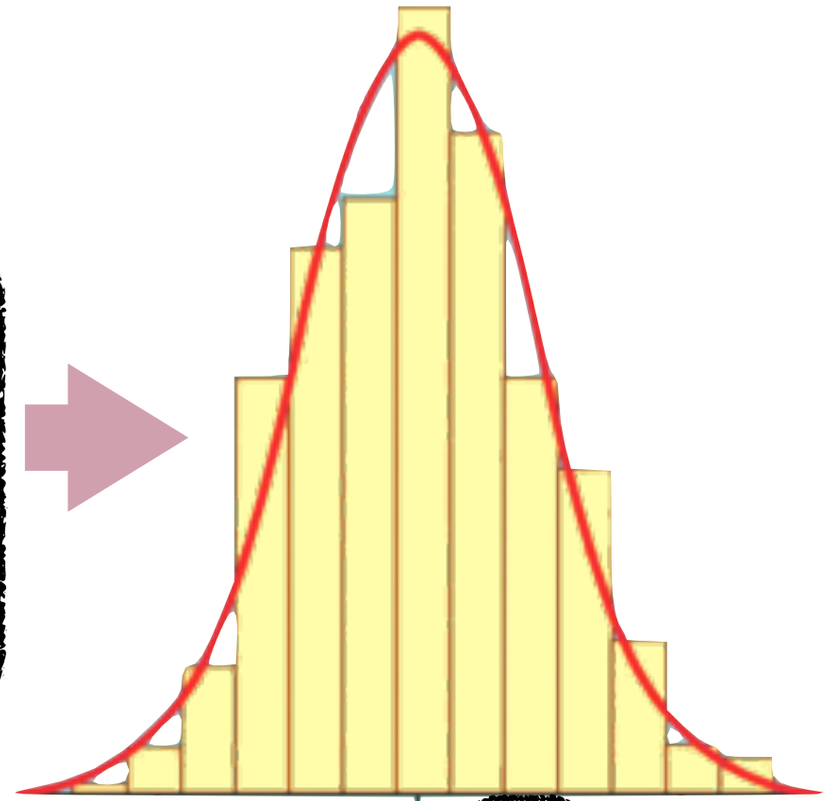
- *Interpretation of the Standard Deviation*: if we wish to shrink the sample standard deviation, all we need to do is increase our sample size n

Mean & Std Dev of a ***Sample Mean***

IMPORTANT



Sampling from a Normal distribution, take multiple SRS of size *n* and make a histogram of the means of those SRS to obtain...



$\sim N(\mu, \sigma^2)$

$\sim N(\mu, \frac{\sigma^2}{n})$

Mean of the Sampling Distribution

Variance of the Sampling Distribution ***of the Sample Mean***

Sample Means: *Example*

- The Dean of a business school claims that graduates from her program make an average of $800 a week with a standard deviation of $100. To verify this claim, a current student takes a sample of 25 graduates and discovers their sample mean weekly salary to be $750. What is the probability that a sample of 25 former students would have a mean weekly salary of $750 if the true (population) mean weekly salary is $800 with a standard deviation of $100?

Sample Means: *Example*

- Recall that when we **know** the **population mean and variance,** we can solve probabilities for a single random variable X by standardizing and obtain the z-score

$$Z = \frac{X - \mu}{\sigma}$$

- Now that we are calculating the probability of a **sample statistic** $\bar{X}$ **and not a single random variable** X, the formula is modified as follows

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Means: *Example*

- With the modified formula for sample means, we can now compute the probability that a sample of 25 former students would have a mean weekly salary of $750 if the true (population) mean weekly salary is $800 with a standard deviation of $100?

$$P(\bar{X} < 750) = P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}) < P(\frac{750 - 800}{\frac{100}{\sqrt{25}}}) = P(Z < -2.5)$$

- We can now check the standard normal tables to see what is the probability (area under the curve) to the left of $Z = -2.5$

Sample Means: *Example*

We conclude the probability of observing a sample mean of $750 or lower assuming that the population mean is actually $800 with a standard deviation of $100 is 0.62%.

This probability is extremely small and might call into question the validity of the dean's claim.

Table entry for z is the area under the standard normal curve to the left of z.

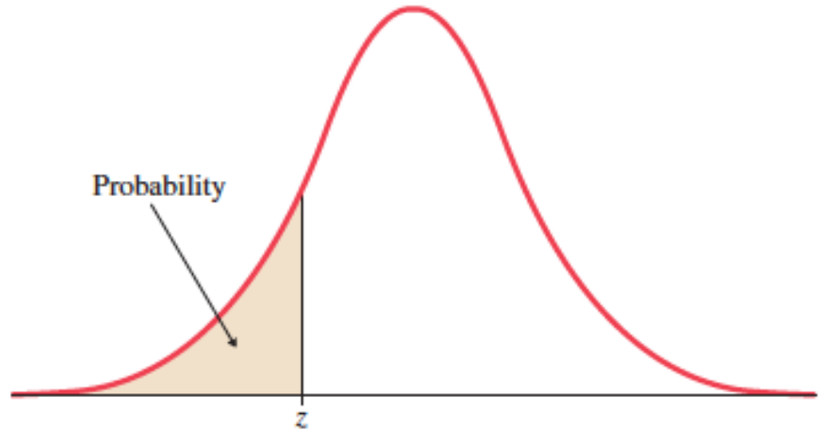


TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

Problem Structure

Note the structure of the interpretation of the previous question (let's break it down into parts):

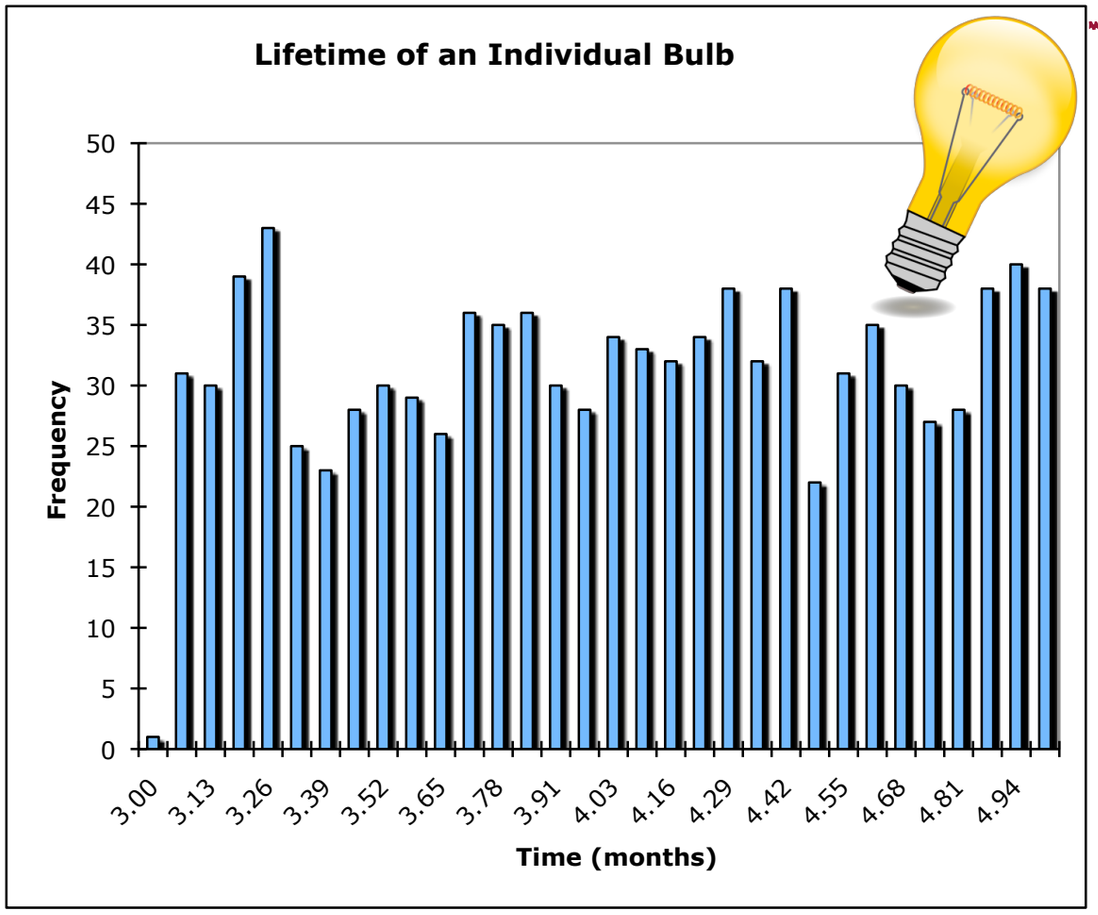
1. If the population mean (parameter) is in fact μ and it's standard deviation is σ ...
2. ...then the probability that we observe a sample mean $\bar{X}$ is... *(compute probability)*
3. Conclude: *IF* the claim about the population parameters is indeed true, the probability of observing the sample statistic we observed is...

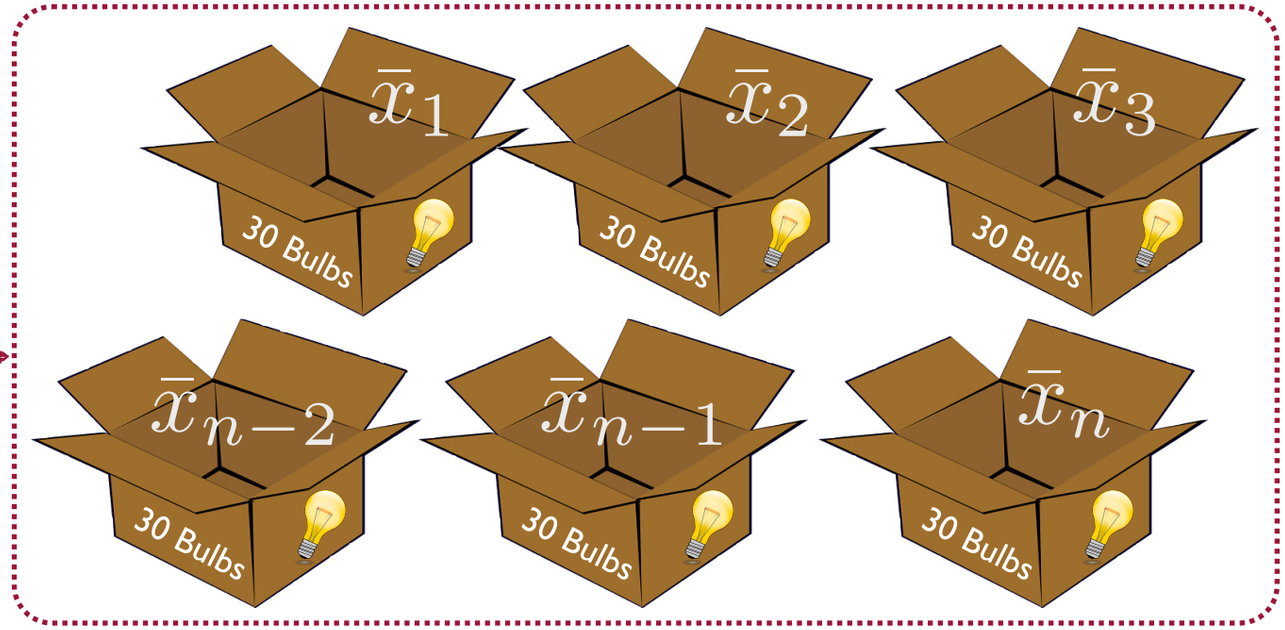
Central Limit Theorem

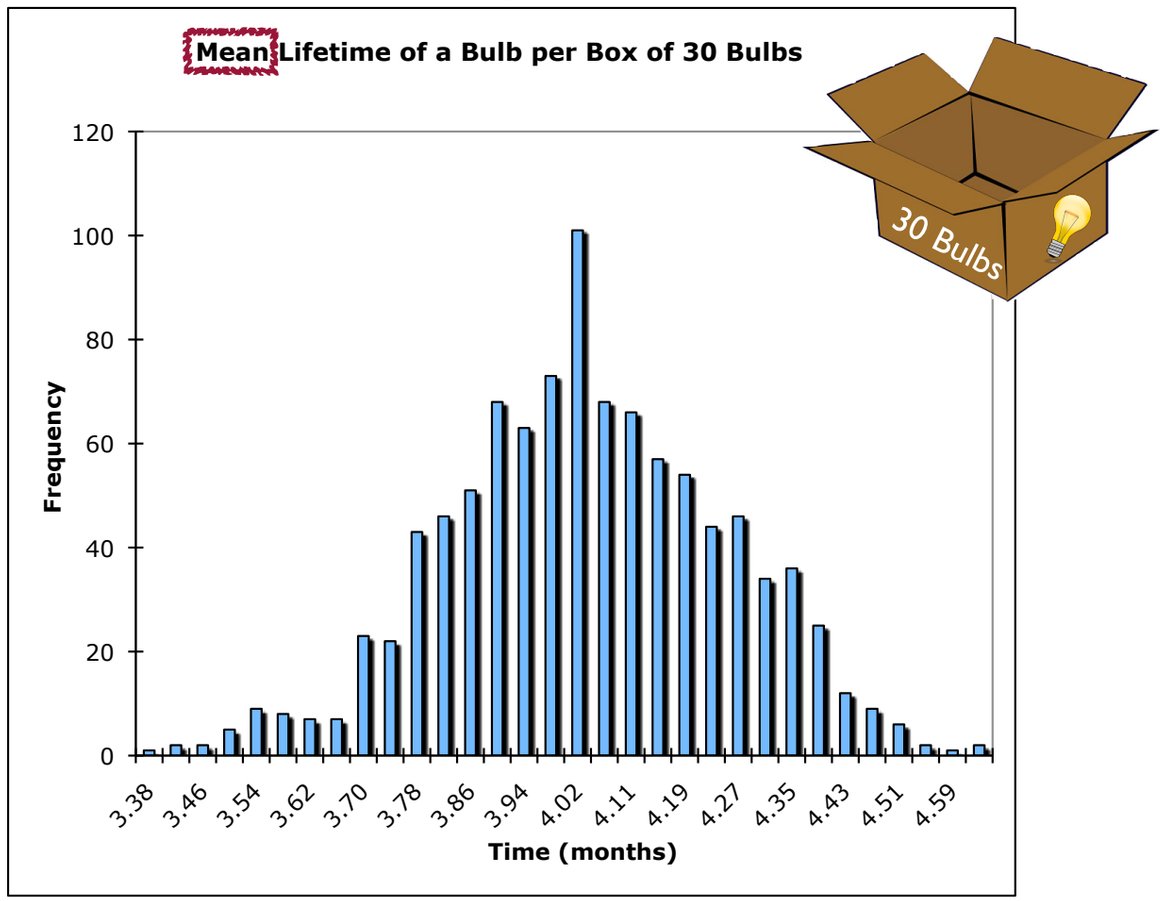
- The Normal distribution is of great interest to us not only because it frequently occurs in nature, but it has many other nice features, one of them being its relation to the Central Limit Theorem (CLT)
- The CLT is a *very powerful* theorem which states that, when sampling from ***ANY*** probability distribution, for sufficiently large *n* the sample mean $\bar{x}$ of a distribution has an approximately normal distribution.
- As a rule of thumb, *n* should be at least 30, but we will see examples where an *n* of even 3 begins to exhibit Normal tendencies

Central Limit Theorem

Original Distribution







Distribution of the Sample Mean

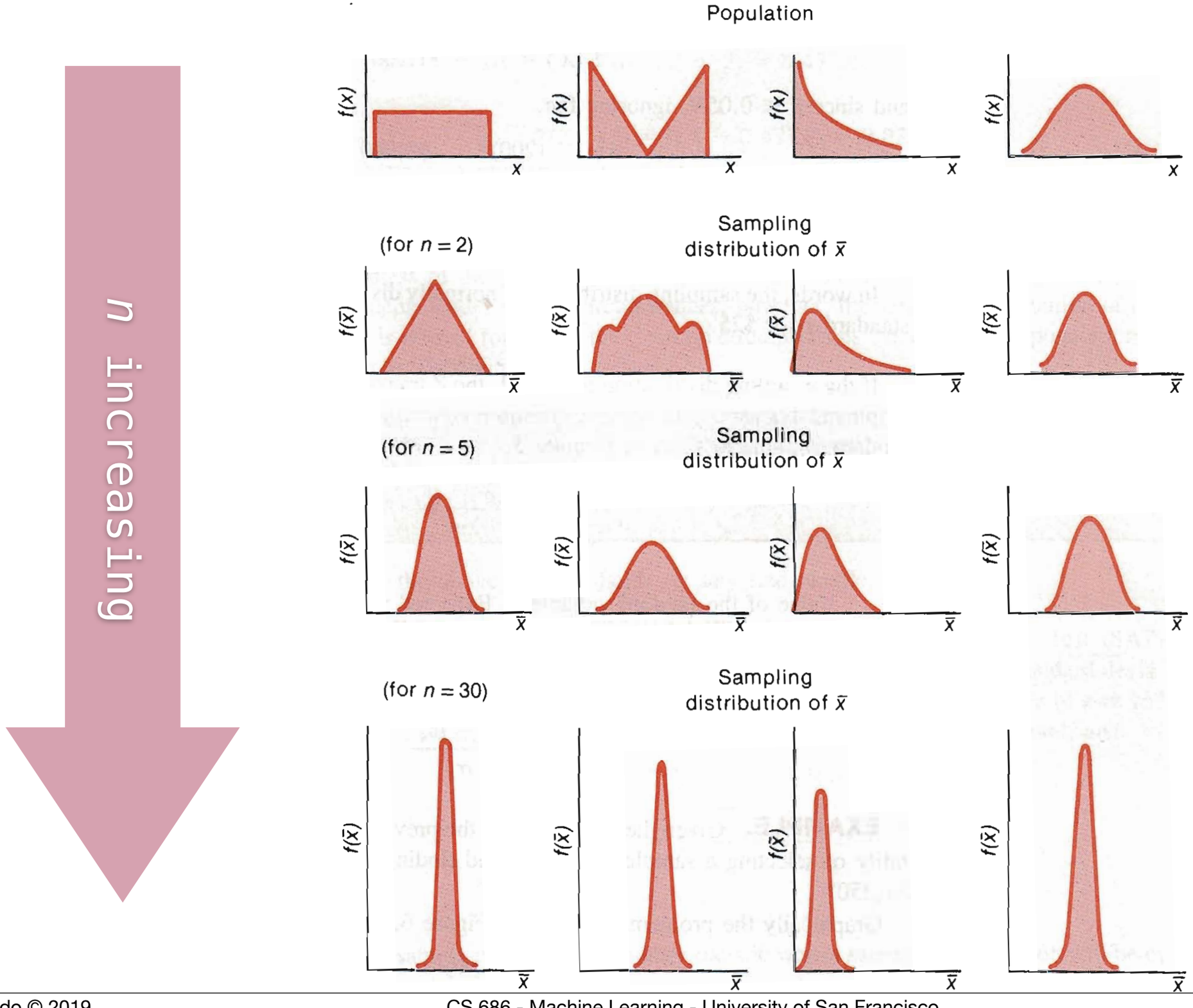
Misinterpretations of the CLT

- Often, the CLT is misrepresented or misunderstood to imply that, when sampling from any distribution, taking any 30 numbers from this distribution and then generating a histogram based on those numbers will yield a normal distribution - *THIS IS INCORRECT*

- Sampling from any distribution and then generating a histogram based on that sample should give you back the exact same distribution you sampled from!

- The operative word when dealing with the CLT is ***mean***

CLT Holds for any Distribution

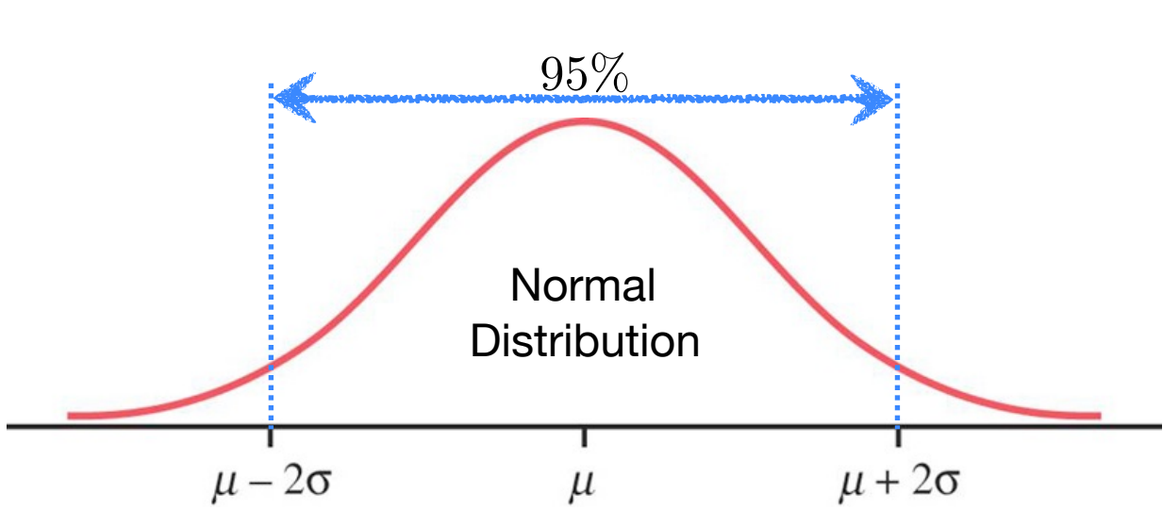


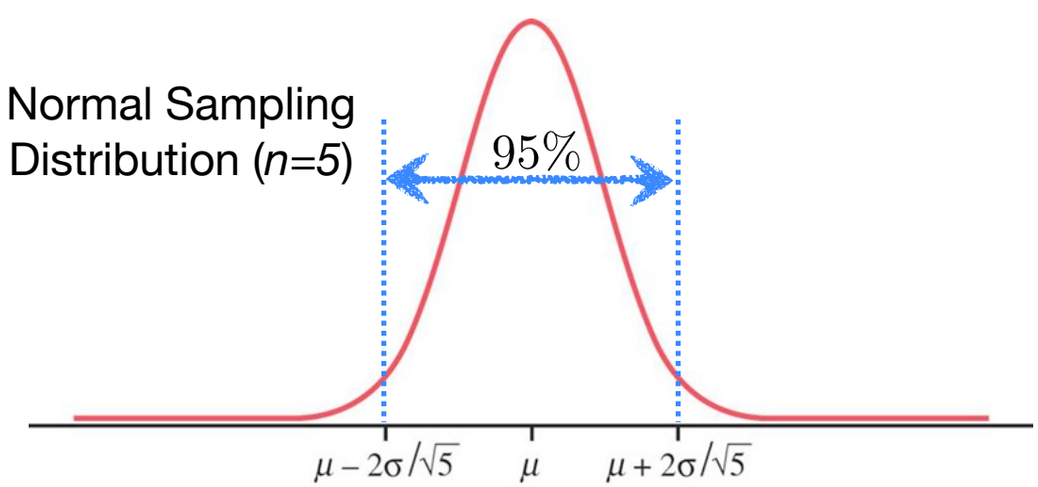
Introduction to Estimation

Estimating the Population Mean When the Population Standard Deviation is Known

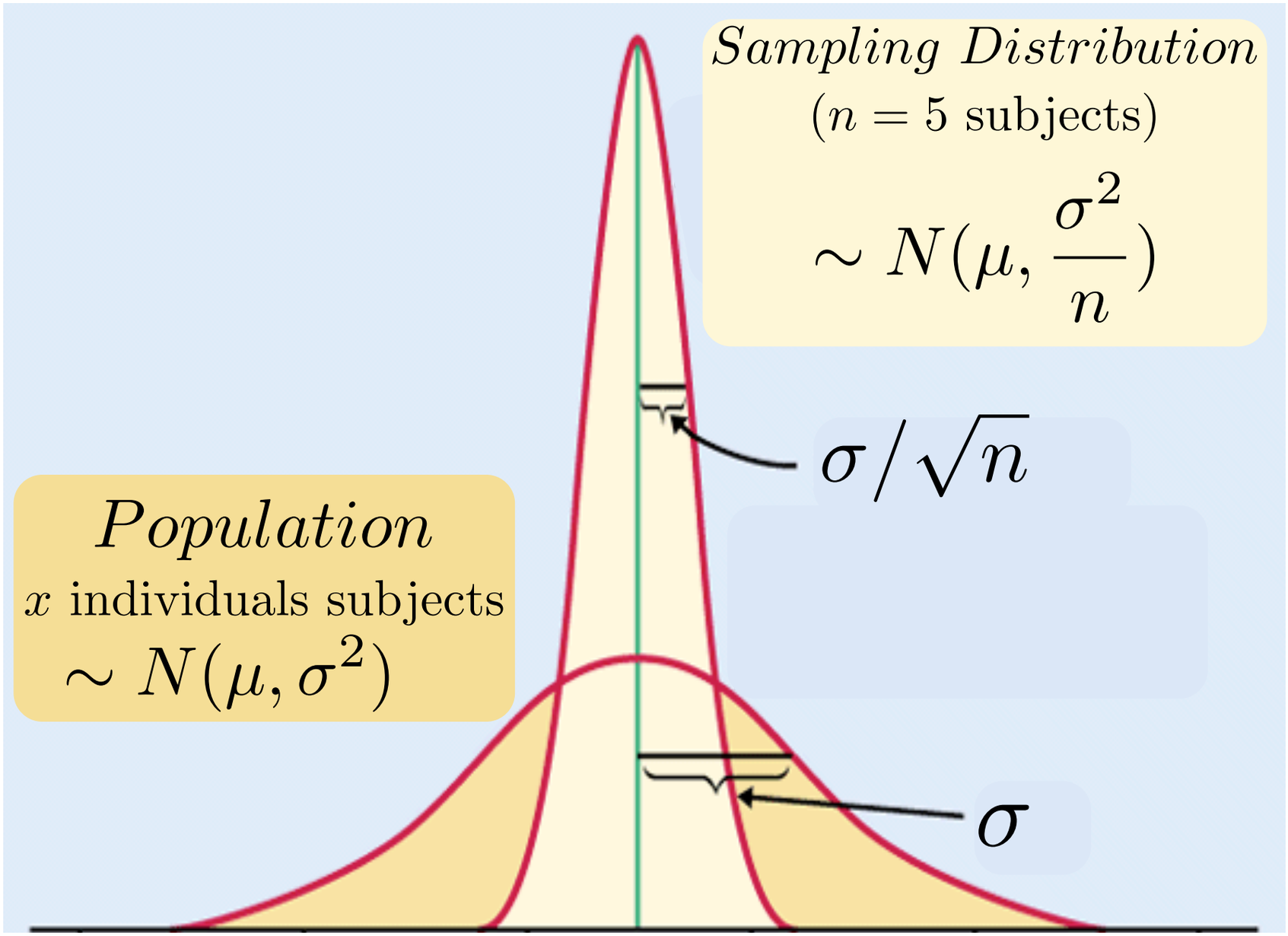
Recall...

- A Normal distribution is a distribution defined by two **parameters,** the mean (μ) and the variance (σ^2)
- When sampling (*randomly*) from a Normal distribution, e.g., take sample sizes such that $n = 5$, recall that the ***mean*** of this sampling distribution is μ and the standard deviation is $\sigma/\sqrt{n} = \sigma/\sqrt{5}$ for our sample size of $n = 5$





Sampling Distribution



...to Summarize

- If we are sampling from a population of individuals that is $\sim N(\mu, \sigma^2)$, then...
  - The sampling distribution of sample means (with sample sizes equal to *n*) is $\sim N(\mu, \sigma^2/n)$ where s^2 and $\bar{x}$ are **unbiased estimators** of μ and σ^2 respectively

- Note from the previous graphic that the variances in the population and sampling distribution are unequal. Does this matter? Why or why not?

Locating Ourselves in Statistical Theory

1. We are sampling from a population
2. The purpose for sampling from the population?
   - 2.1. To (hopefully) get a good estimate ($\bar{x}$) of the population mean (μ)
3. What do we know about the population?
   - 3.1. Do we know the distribution of the population? Does it matter?
   - 3.2. We only ***assume one thing about the population***, that σ is given (known).

Condition

A Brief Example

- If we have a population with a mean μ and variance σ^2 then in repeated samples of size $n = 100$, the sample mean $\bar{x}$ is approximately

  $\sim N(\mu, \sigma^2/100)$

- What conclusions can we draw from this information?
- Recalling the 68% – 95% – 99.7% rule...
  - there is a 68% probability that the population mean μ lies within $\sigma/10$ (plus or minus) units from the sample mean $\bar{x}$
    - note that $\frac{\sigma}{10} = \sqrt{\frac{\sigma^2}{100}}$

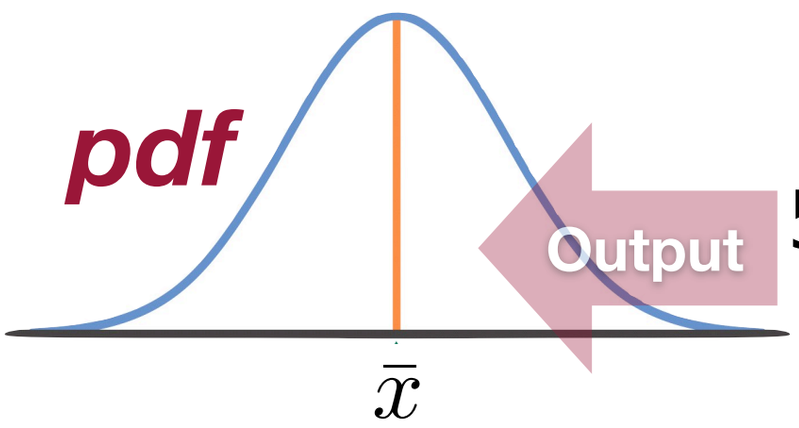
Confidence vs. Probability

- We will now introduce the word ***Confidence*** as means for representing the level of certainty we have about a claim
- Note from the previous slide we said that there was a given ***probability*** that the mean lies within a certain amount (plus or minus) the sample mean
- We explicitly say ***probability*** here because the interval is stated in a random fashion

random (statistic) $\bar{x} \pm \frac{\sigma}{\sqrt{n}}$ ***constant (parameter) WHY?***

A quick review...

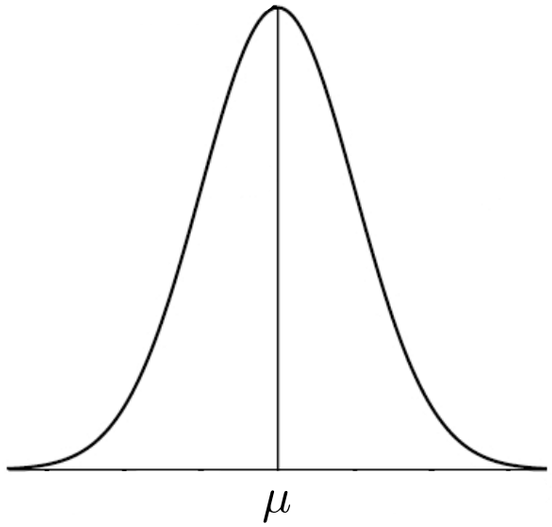
1. Select a sample size n
2. Take a simple random sample of size n from the probability density function



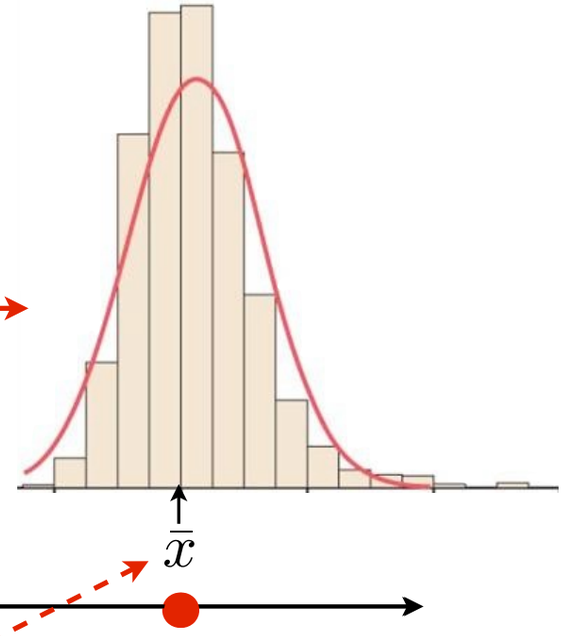
3. Take the mean of that sample ($\bar{x}_1$)
4. Repeat steps 2 and 3 until you have a sufficient amount of samples
5. Plot a histogram of all of the sample means ($\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$) and you obtain a Normal distribution (by the *CLT*)

Interpreting Confidence

- If you can imagine executing the steps indicated on the previous slide many time over...



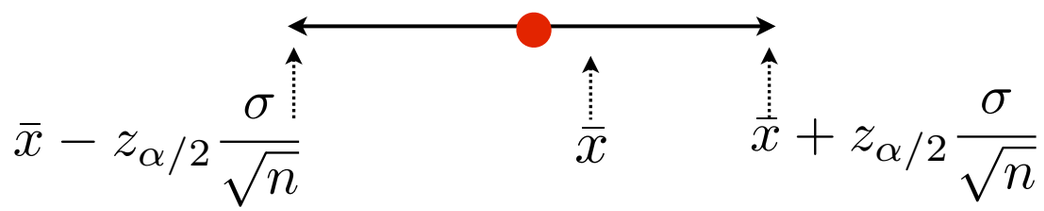
Execute the steps once to obtain...



...and recalling that $\pm\frac{\sigma}{\sqrt{n}}$ *is a constant*

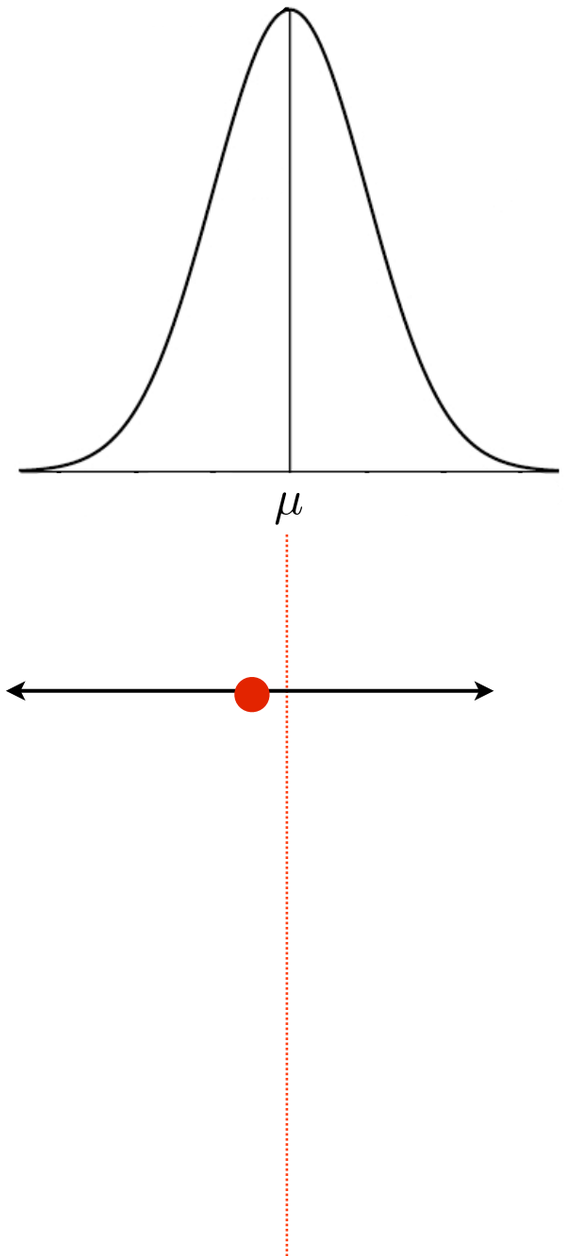
Note
This $\bar{x}$ is no longer random because we have sampled

Diagram Legend



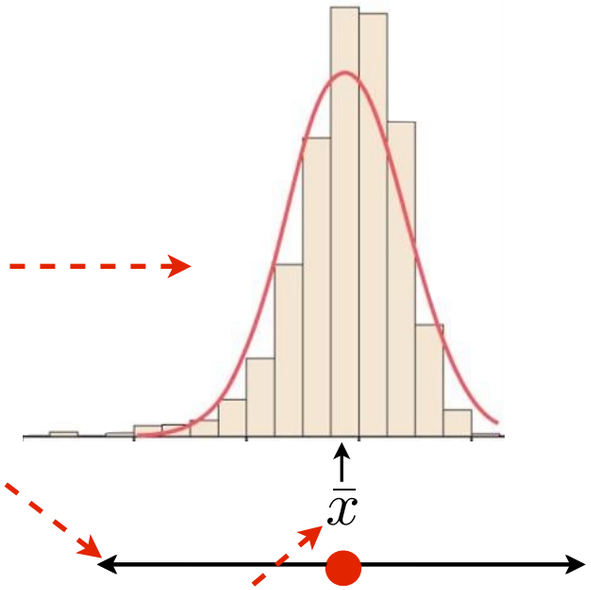
Interpreting Confidence

- If you can imagine executing the steps indicated two slides previous to this one many time over...



Execute the steps a <u>SECOND</u> time to obtain...

...and recalling that $\pm\frac{\sigma}{\sqrt{n}}$ *is a constant*

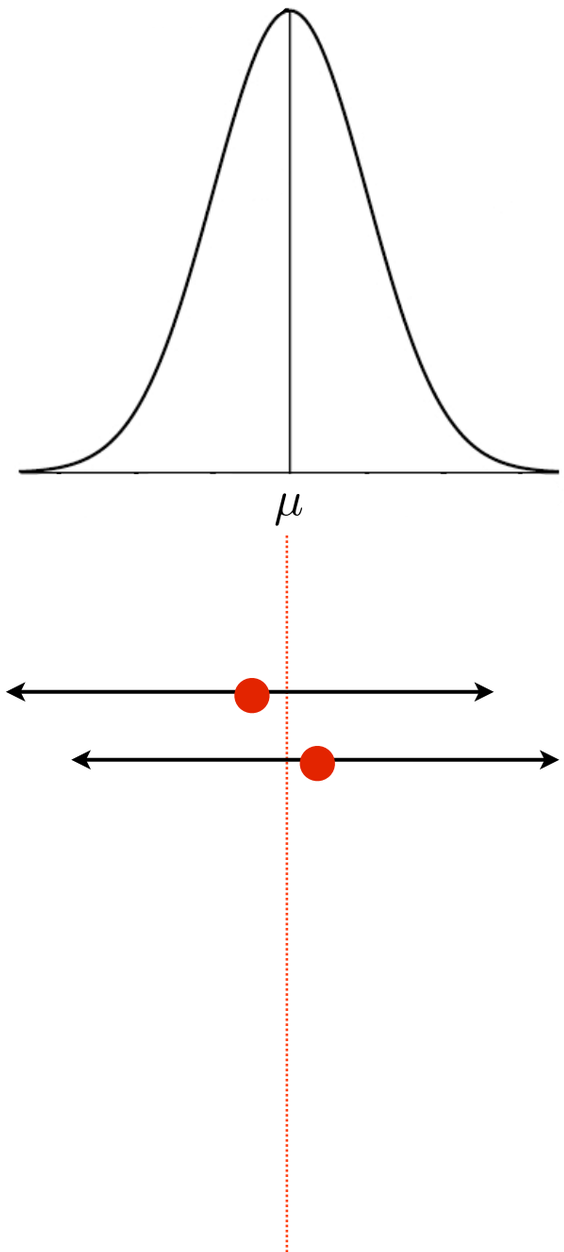


Note

This $\bar{x}$ is different from the one in the previous slide, and is also no longer random because we have sampled

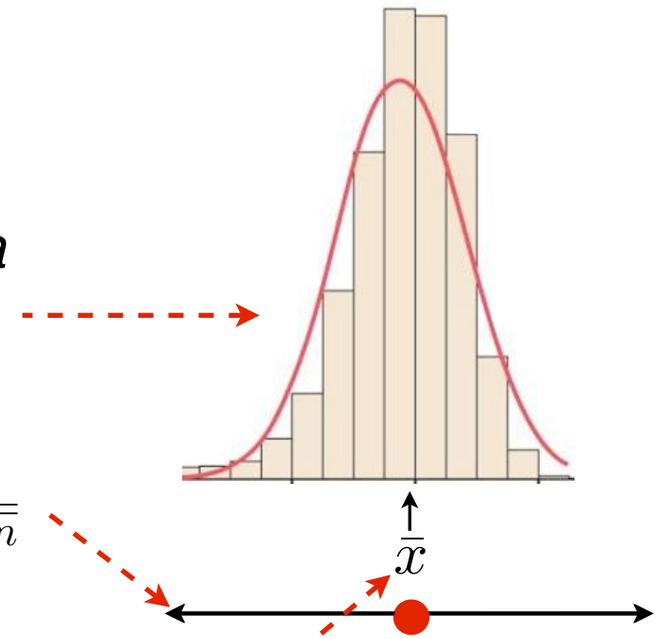
Interpreting Confidence

- If you can imagine executing the steps indicated two slides previous to this one many time over...



Execute the steps a <u>THIRD</u> time to obtain...

...and recalling that $\pm\frac{\sigma}{\sqrt{n}}$ *is a constant*

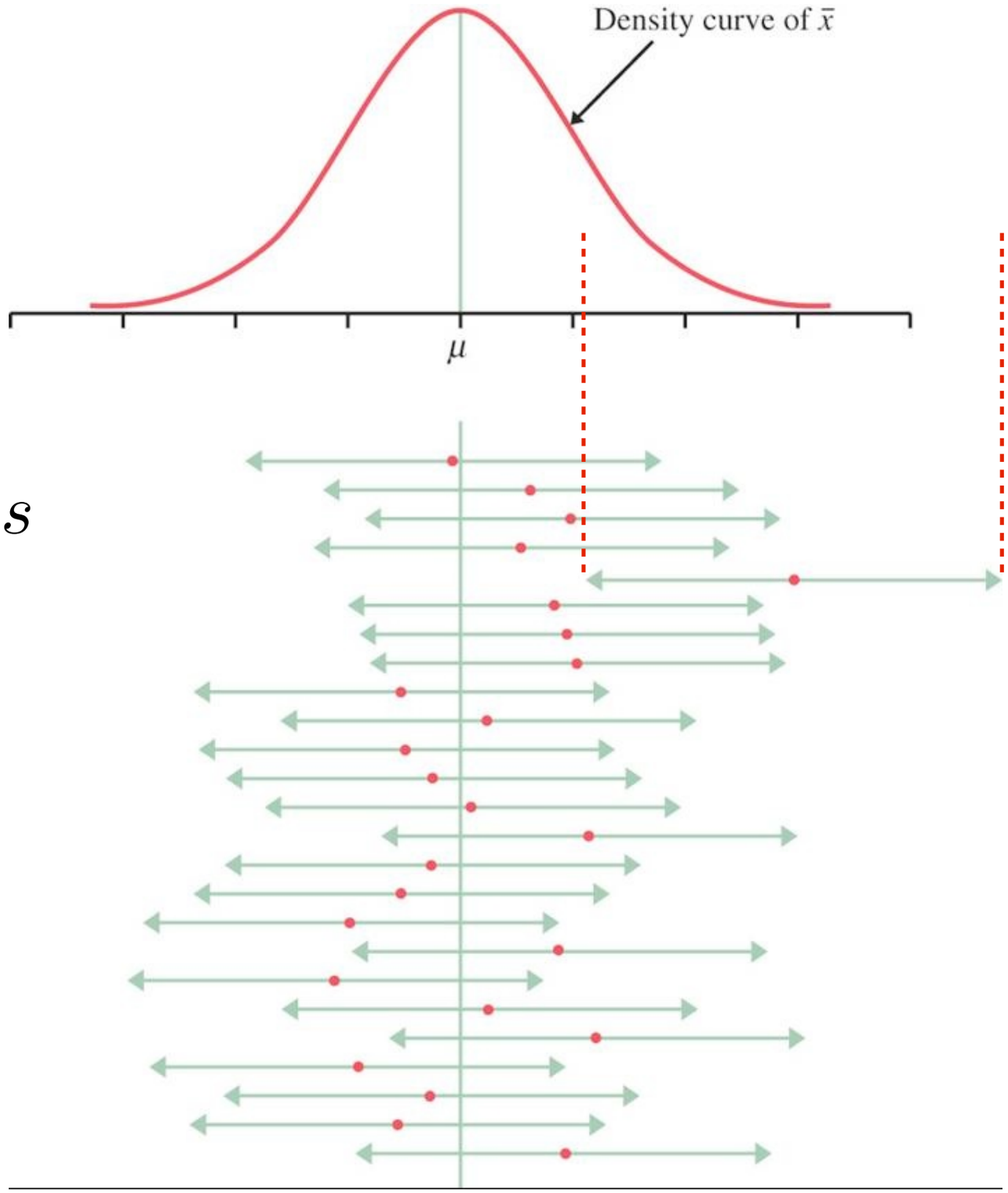


Note

This $\bar{x}$ is different from the one in the previous slide, and is also no longer random because we have sampled

Interpreting Confidence

- If you did this 100 times, you would obtain 100 intervals of identical length centered about their respective $\bar{x}'s$



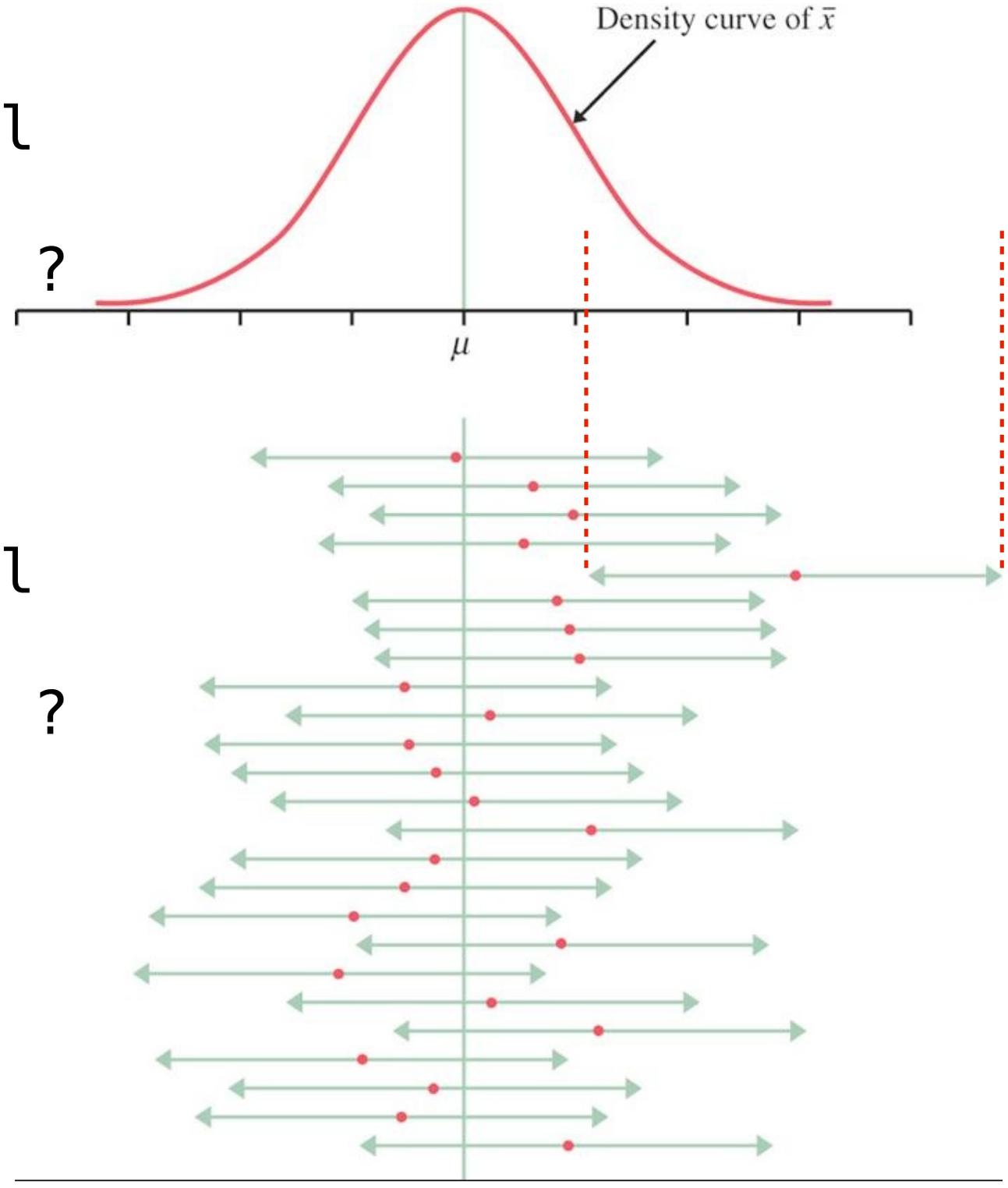
- What percentage of these intervals will contain μ if the interval is $\bar{x} \pm \frac{\sigma}{\sqrt{n}}$?

Interpreting Confidence

- What percentage of these intervals will contain μ if the interval is $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$?

- What percentage of these intervals will contain μ if the interval is $\bar{x} \pm 3\frac{\sigma}{\sqrt{n}}$?

- Each interval is called a ***Confidence Interval***



Interpreting Confidence

- Of course, we don't (can't) generate and infinite number of confidence intervals and then claim that 95% of them contain the population mean. Rather, we create one confidence interval and claim that we are $1 - \alpha$ % confident that it contains the population mean μ, where α is the total error, and $\alpha/2$ is the error in each tail

Conclusions

When referring to any given confidence interval, it is ***correct*** to say that this confidence interval has a $1-\alpha$% ***probability*** that the sample mean $\bar{x}$ will be equal to a value such that the confidence interval will contain the population mean μ

Conclusions

- Once it is computed, any given confidence interval either does or does not contain the true population parameter μ , hence it is ***incorrect*** to say that the probability is 95% that the true population mean μ lies within a given (specific) confidence interval
- If we generate 100 different confidence intervals, 95% of those intervals are expected to contain the parameter μ , however, for any particular confidence interval, it is not known whether or not the confidence interval contains μ

Confidence Intervals: *Example*

from Statistics & Data Analysis, Tamhane & Dunlop, 2000, p.204

- Airlines use sampling to estimate mean the revenue for passengers. Suppose that the revenues for a certain airlines are normally distributed with $\sigma = 50$. To estimate mean share per ticket, airlines uses a sample of 400 tickets resulting a sample mean of $\bar{x} = 175.60\$$

- Calculate a 95% Confidence Interval (CI) for the true mean revenue μ

Confidence Intervals: *Example*

$$\bar{x} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

$$175.60 \pm 2 \times \frac{50}{\sqrt{400}}$$

$$[170.6, 180.6]$$

- We are 95% confident that the mean revenue μ is between *$ 170.60* and *$180.60*

Confidence Intervals: *Example*

- Calculate an 80% Confidence Interval (CI) for the true mean share

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

- Get standardized value from the Z-Table for 80%

TABLE A Standard normal probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

Confidence in the *Z*-Tables

- When we look up 80% probability in the Z-Table, we obtain $z \approx 0.845$ which implies that we are 0.845 standard deviations above the mean
  - But we know that being one standard deviation above the mean implies that we are covering 68% of the area under the curve (and hence have a probability of 68%)
- Why do we have this conflict of information?

Lack of Symmetry

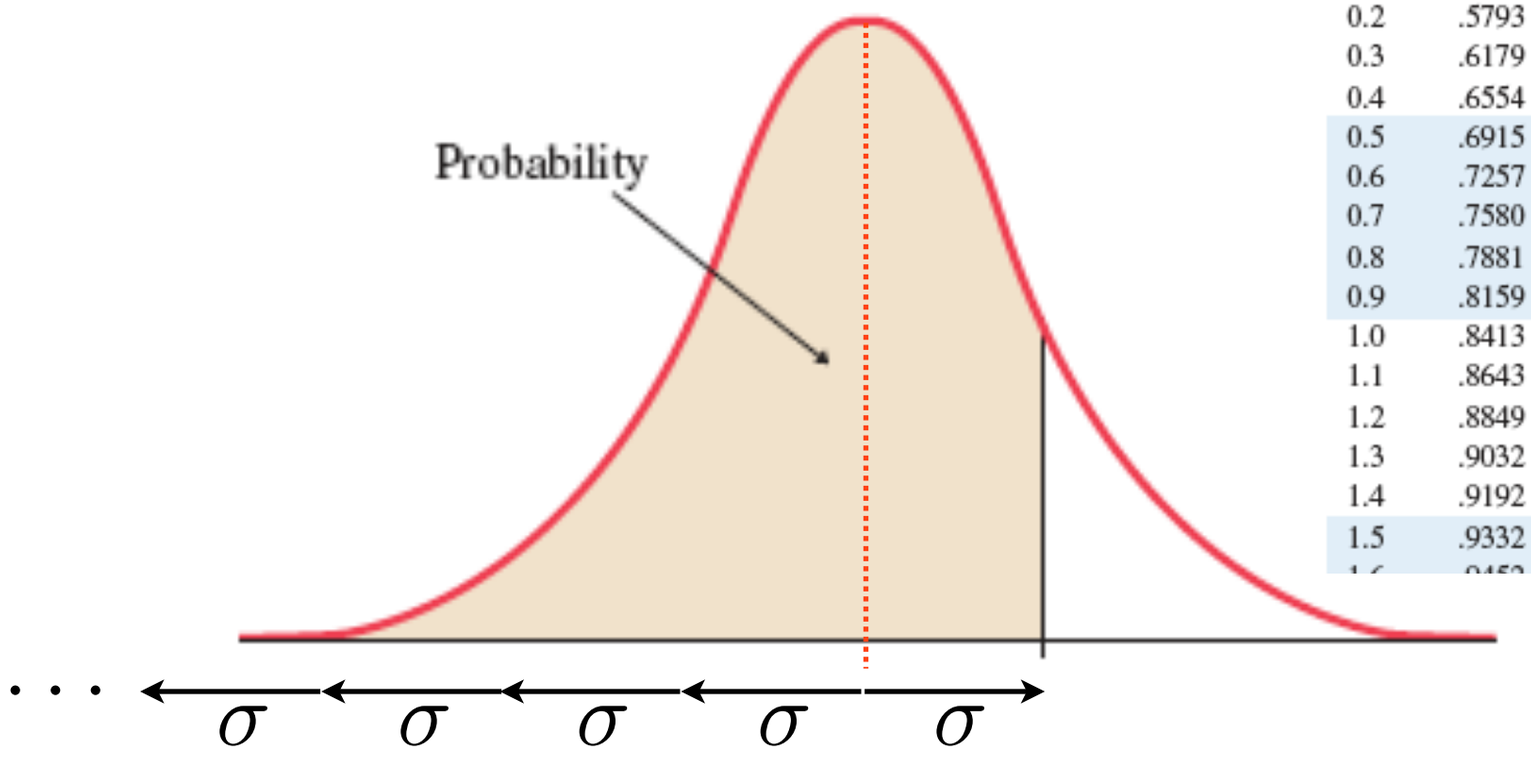


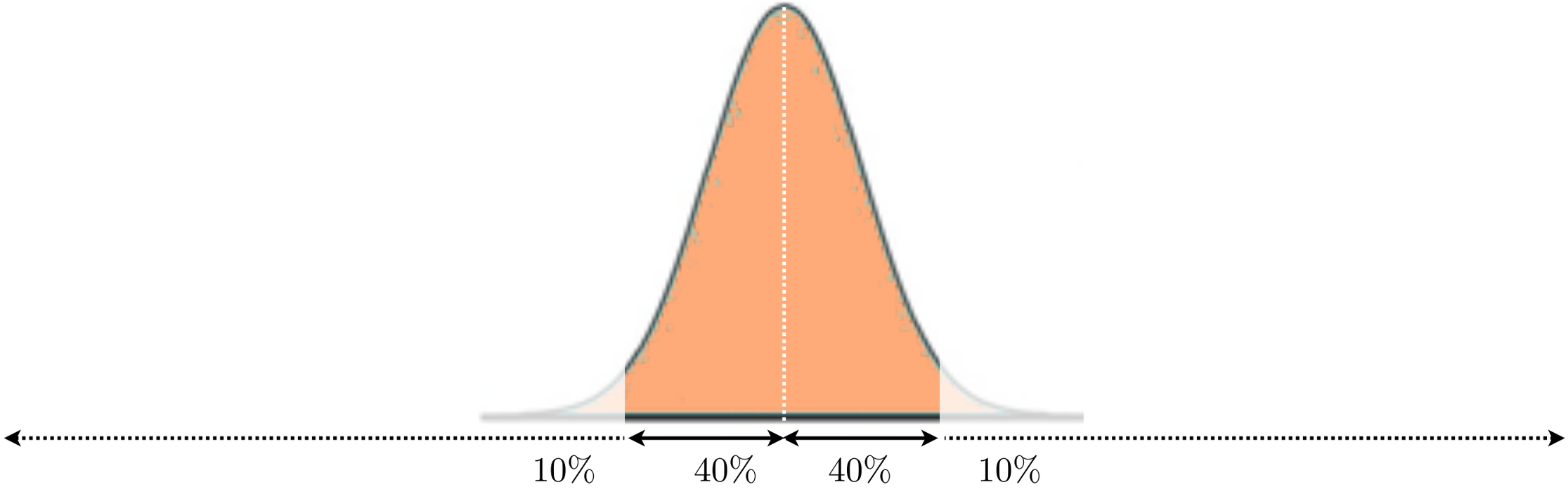
TABLE A Standard normal probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

- How many sigma are we above the mean? **1**
- How many sigma are we below the mean? ***infinity***

The Correct Value

- If we want to obtain 80% confidence, then we must observe that the Normal Distribution is symmetric and that each tail should share the same amount of “unaccounted” confidence or risk



Confidence Values

- Therefore, if we want 80% confidence – which implies 20% risk – we observe the fact that this requires us to have 10% risk in each tail, therefore in the standardized Normal table, we look up the probability value of 90%, which will give us the number of standard deviations above the mean we need to get 10% in the upper tail
- Once we get this value, we can compute our confidence interval

...back to the Example

TABLE A Standard normal probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$175.60 \pm 1.285 \times \frac{50}{\sqrt{400}}$$

$$[172.39, 178.81]$$

A Note on *Z*

Note that $z_{\alpha/2}$ is a single term. The $\alpha/2$ subscript is aesthetic, a reminder to divide your error by 2 before looking up the *z*-value. *Z is <u>not</u> multiplied by $\alpha/2$ nor is it in any way numerically part of the equation.*

Finding the Common $z_{\alpha/2}$ Values

- Three very common levels of confidence used in practice are 90%, 95%, and 99%
  - What are the z values associated with each?

$z_{\alpha/2}$			
$Confidence$	90%	95%	99%

These values are for 2-tailed tests, hence $\alpha/2$

Finding the Common $z_{\alpha/2}$ Values

TABLE A Standard normal probabilities *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974

Finding the Common $z_{\alpha/2}$ Values

- Three very common levels of confidence used in practice are 90%, 95%, and 99%
  - What are the z values associated with each?

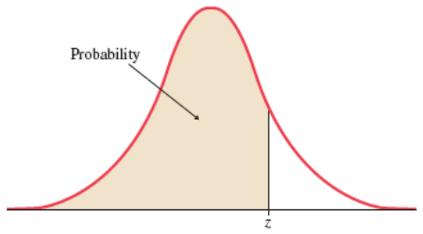
$z_{\alpha/2}$	1.645	1.96	2.576
$Confidence$	90%	95%	99%
α	10%	5%	1%
$\alpha/2$	5%	2.5%	0.5%

These values are for 2-tailed Confidence Intervals

Correction

- Recall from the previous airline example problem that we used 2 as the $z_{\alpha/2}$ value when creating a 95% confidence interval
- This was slightly incorrect as we now know that the exact number is 1.96, as calculated on the previously slide

Probability vs. Standard Deviations



1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50

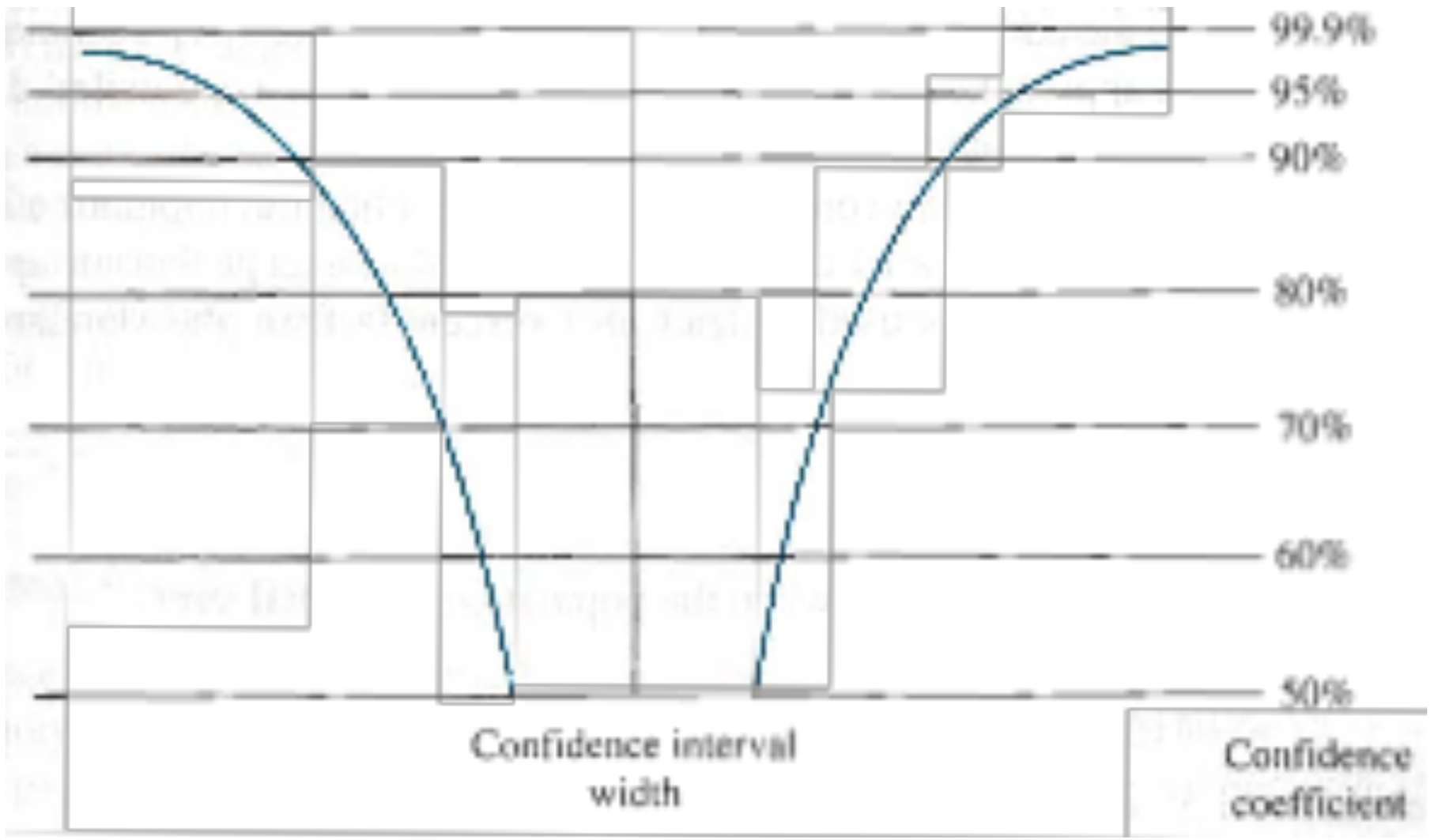
Probability

0 0.46 0.92 1.38 1.84 2.3 2.76 3.22 3.68 4.14 4.6

Standard Deviations above the Mean

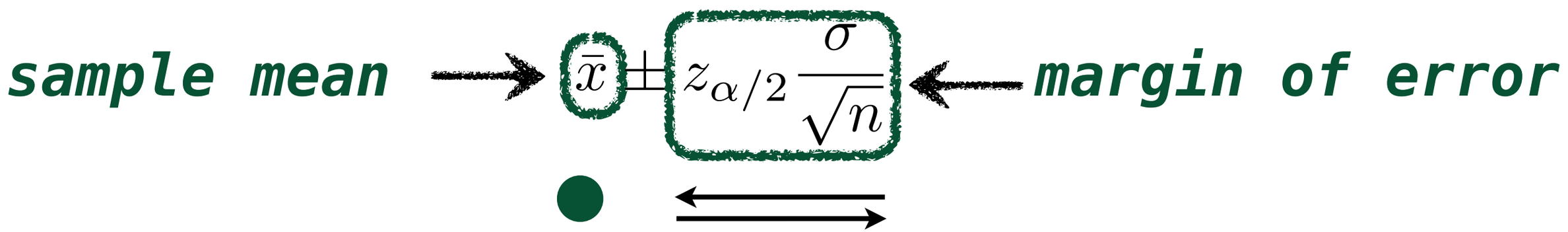
CI: *Observations*

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



CI: *Observations*

- If the margin of error is too large, how can we reduce it?
  - Use a lower level of confidence (hence smaller z)
  - Reduce the variance (σ^2) *(not really...why?)*
  - Increase the sample size (n)



Final Notes on Inference

- Recalling that $\bar{x}$ is a measure which is sensitive to outliers, an outlier can have a large effect on the variance and by association the width of the confidence interval
- If the sample size is too small and it is sampled from a non-Normal population, confidence levels may be affected; simply aware of this
- **Remember that in this context, *we need to know the population variance* (σ^2)**

Inference About a Population Mean When the Standard Deviation is Unknown

Inference for the Mean of a Population

- Recall the conditions and objective for inference from the previous chapter
  - We were trying to estimate the mean of a population μ with a sample mean $\bar{x}$ obtained from an SRS, and we were given the population standard deviation
- In reality, **we usually don't have information about the population standard deviation** σ , so this section will present tools, very similar to those previously discussed, which account for this difference

Standard Error

- Firstly, some redefinitions
  - Recall that the distribution of the sample mean $\bar{x}$ (*as defined in the previous topic*) is

$$\sim N(\mu, \frac{\sigma^2}{n})$$

- When in a scenario where the ***standard deviation*** σ ***is unknown***, we estimate it by the sample standard deviation s, and then we estimate the standard deviation of $\bar{x}$ with $s/\sqrt{n}$
- The quantity $s/\sqrt{n}$ is referred to as the standard error of the sample mean $\bar{x}$ (SE)

...the Switch...

- When substituting the value $\sigma/\sqrt{n}$ with $s/\sqrt{n}$, the distribution of the sample mean $\bar{x}$ is no longer Normal; instead, it follows a *Student's t-Distribution*

if σ known	*if σ unknown*
$\sigma/\sqrt{n}$	$s/\sqrt{n}$
$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
Normal Distribution	**Student's *t* Distribution**

recall

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

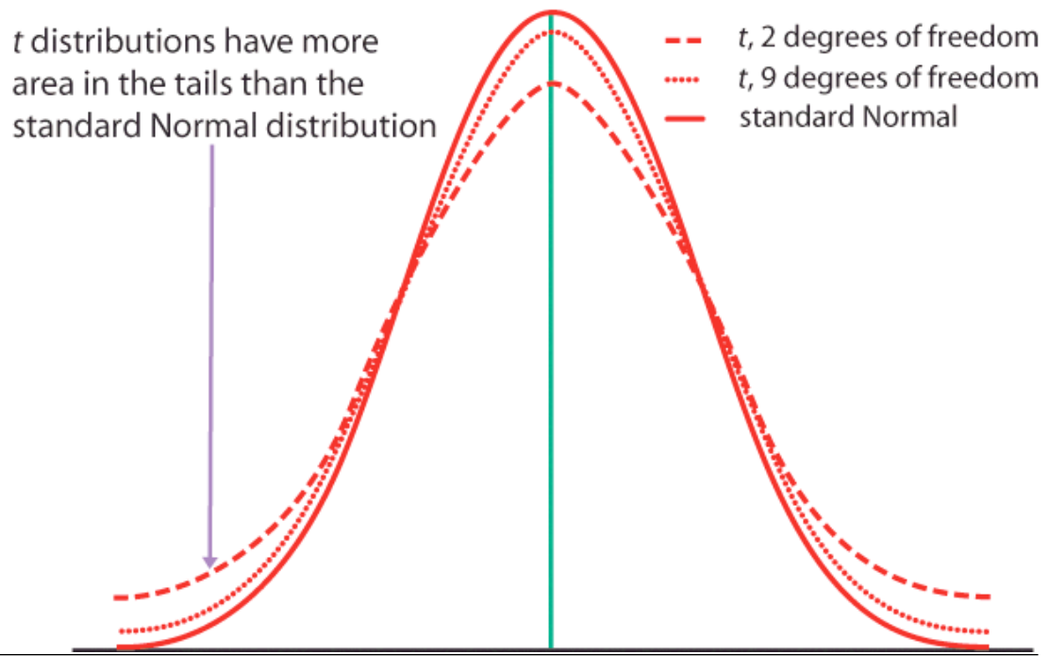
Why Student's *t* Distribution?

This distribution was investigated by the Irish chemist William Sealy Gosset (1876–1937), who worked for Guiness Breweries and published his statistical research under the pseudonym "Student." The distribution of t is named the Student's t–Distribution in his honor.

from Statistics & Data Analysis, Tamhane & Dunlop, 2000, p.179

t-Distribution: A Few Notes

- The *t*-Distribution is almost Normal in appearance
- The *t*-Distribution has slightly heavier tails than the Normal distribution (*see below*)
- **There is a different distribution associated with each different degree of freedom; as the degrees of freedom increase, the *t*-Distribution gets closer and closer to a Normal distribution**



t-Distribution: A Few Notes

- **Why does the *t* Distribution have heavier tails that the Normal Distribution?**
- When obtaining the *z*-value from the Standard Normal table, the only statistic we have is $\bar{x}$
- When obtaining the *t*-value from the *t*-Distribution table, we have two statistics, $\bar{x}$ and s
  - *because of the greater uncertainty associated with having two statistics instead of one, it makes sense that the t Distribution has higher variability; this translates in to a distribution with heavier tails!*

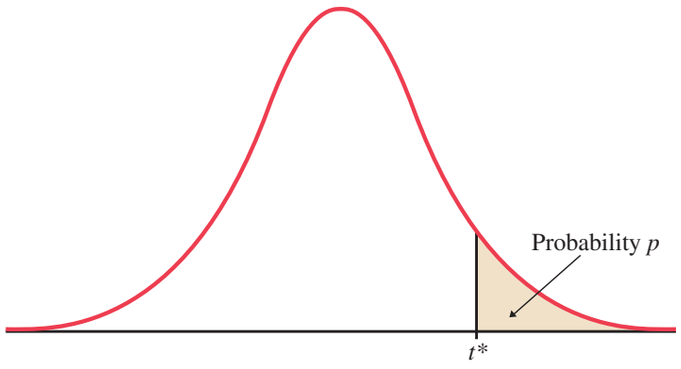
Degrees of Freedom for Limiting t Values

TABLE D *t* distribution critical values

$\left(\frac{\alpha}{2}\right)$

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

$n-1$



$$t\left(n-1, \frac{\alpha}{2}\right)$$

degrees of freedom

single-tail risk

z-Table vs. *t*-Table

z-Table

- the *z*-values are located in margins and the cumulative probabilities are located inside the table

t-Table

- the *t*-values are inside the table, the degrees-of-freedom are in the left-hand margin, and the one-sided (one-tailed) probabilities are in the top margin
- **n.b.** there is a confidence-interval cheat sheet at the bottom of the *t*-Table

Comparison of *CI* Procedures

Confidence Interval for Population Mean

if σ known	*if σ unknown*
$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{(n-1, \frac{\alpha}{2})} \frac{s}{\sqrt{n}}$
Normal Distribution	**Student's *t* Distribution**

Population Mean w/ Sample Variance

- *Example*
  - Of the 135 million tax returns filed in the United States in 2009, auditors selected 1% to review to confirm they were correctly submitted. We are curious to know how much additional income is generated by this auditing practice, so an SRS of the audited files is selected. Compute a 95% confidence interval of the mean additional income generated by auditing files.

Population Mean w/ Sample Variance

- *Example* *(cont'd)*
  - We are given the following information

$$\bar{x} = 11,343$$

$$s = 4,400$$

  - To compute the confidence interval, we are required to look up the appropriate t value in the table $n = 184$
  - ***Is the value we are searching for one- or two-tailed?***

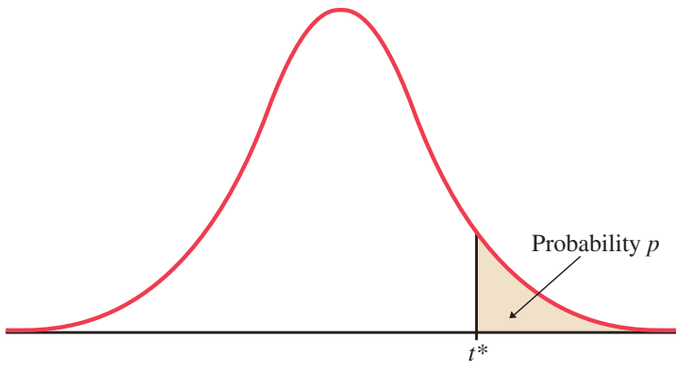
Population Mean w/ Sample Variance

TABLE D *t* distribution critical values

$\left(\frac{\alpha}{2}\right)$

	Upper tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

$n-1$



$$t_{\left(n-1,\frac{\alpha}{2}\right)}$$

degrees of freedom

single-tail risk

Population Mean w/ Sample Variance

- *Example* *(cont'd)*
  - *Recall* $\bar{x} = 11,343 \quad s = 4,400 \quad n = 184$
  - *Therefore the 95% Confidence Interval is*

$$\bar{x} \pm t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}} = 11,343 \pm 1.984 \left(\frac{4,400}{\sqrt{184}} \right) = 11,343 \pm 643$$

$$[\$10,700 \ , \ \$11,986]$$

Answer

Hypothesis Testing

"…Doing statistics really is easier now than doing plumbing, but unfortunately errors are much better hidden – there is no statistical equivalent of a leaky pipe."

Hypothesis Testing

- Another inferential approach is hypothesis testing, in which a claim or *hypothesis* is put forward and subsequently either accepted or rejected with a certain level of confidence using statistical tools
- *Examples of hypotheses*
  - Mean revenue for a company is 6,000,000
  - Mean age of student in this class is at least 23.2
  - Fraction of defective parts in a box is less than or equal to 1%

Salk Vaccine Trial

- In 1954, the Salk vaccine trial was initiated, where 400,000 grade school students were randomly and separated into two different groups, each with 200,000 students to test the quality of a Polio vaccine
- In medical trials like this one, usually one group is called the “control group,” the group who is given a placebo, and the “treatment group” who is given the actual medication (neither group knows *apriori* whether they are getting the treatment or the placebo)

Salk Vaccine Trial

- The vaccine trial hypothesized that the "treatment group" would have a lower incidence of Polio than the control group
- After the vaccination, the incidence rates for Polio were 2.68 children in the "treatment group" and 7.06 in the "control group" per 10,000 students
- How do we interpret these numbers? Does this result prove anything? Can we conclude that all children should be vaccinated or could this difference have happened by pure chance? Is the difference in incidence rate ***statistically significant***?

Constructing a Hypothesis Test

- A hypothesis test is constructed by making both a claim (hypothesis) and a counter-claim
  - The claim is called the ***null hypothesis*** and is denoted H_0
  - The counter-claim is called the ***alternative hypothesis*** and is denoted H_1 or H_a
- The alternative hypothesis is what drives a hypothesis test

Constructing a Hypothesis Test

Conceptually Important

In the classical approach, we begin with the assumption that H_0 is true. If the data fail to contradict H_0 beyond a reasonable doubt, then H_0 is not rejected. However, failing to reject H_0 does not mean that we accept it as true; it simply means that H_0 cannot be ruled out as a possible explanation for the observed data. Only when the data strongly contradict H_0 is it rejected and H_1 accepted. Thus, proof of H_1 is by contradiction of H_0

from Statistics & Data Analysis, Tamhane & Dunlop, 2000, p.205-206

Constructing a Hypothesis Test

Conceptually Important

The Canadian justice system provides an analogy to the logic of hypothesis testing. An accused person is presumed innocent until proven guilty. The burden of proof is on the prosecution to show that the accused is guilty. Thus the hypotheses are H_0 : *Accused person is not guilty* vs. H_1 : *Accused person is guilty*. The evidence plays the role of the data. Only when the evidence strongly contradicts the person's innocence is a "guilty" verdict rendered. A "not guilty" verdict does not prove that the accused person is innocent, it simply confirms that there is not enough evidence to prove guilt.

adapted from Statistics & Data Analysis, Tamhane & Dunlop, 2000, p.206

Constructing a Hypothesis Test

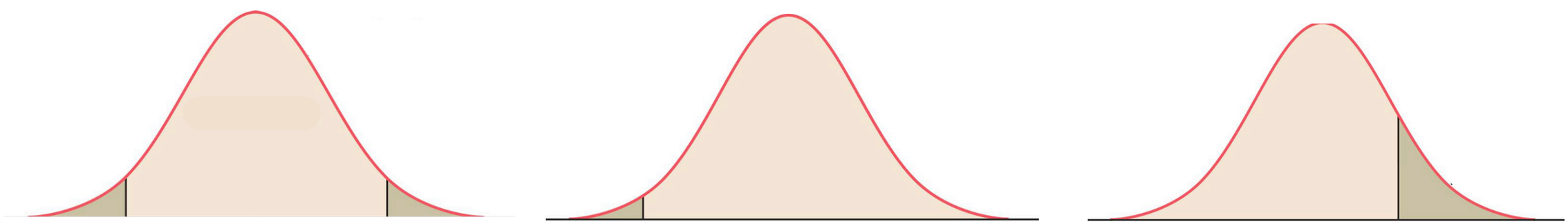
- Observe from the previous slide's interpretation that what we are looking to prove, "that a person is guilty" is set as the alternative hypothesis H_1
- Reprising the example of the Salk's Polio vaccine test, the hypotheses would be as follows
  - H_0 : incidence rate in the control group is less than or equal to the incidence rate in the treatment group
  - H_1 : incidence rate in the control group is greater than the incidence rate in the treatment group

Comments on Hypotheses

- We establish our null and alternative hypotheses such that they are mutually exclusive and collectively exhaustive. What does this imply?

3 Styles of Hypothesis Tests

H_0 : $H_a : \mu \neq a$	H_0 : $H_a : \mu < a$	H_0 : $H_a : \mu > a$
$2 - sided$	$1 - sided$	$1 - sided$



Comments on Hypotheses

- It really doesn't matter what sign (equality or inequality) is written in the null hypothesis because we never make a statement about the null hypothesis
  - Recall from the jury example, if H_a , what we are trying to prove, is guilt, and, after some analysis, we decide that the accused is not guilty, we do not prove that the accused is innocent, rather, we simply state that we don't have enough evidence to prove he or she is guilty; we are not making any statement about the null hypothesis H_0

Comments on Hypotheses

- *Hypotheses always refer to some pollution or model, not to a particular outcome. For this reason, we always state H_0 and H_1 in terms of population parameters (note in the previous slide that all hypothesis tests we claims made about μ , the population mean).*

- Note that in all of the null hypotheses, an "equals" sign always makes an appearance, whether it be as an $\leq$, as a $\geq$, or as a $=$. Conversely observe that there are no "equal" signs in the alternative hypotheses (*reasons for this will become obvious shortly*).

Constructing Hypothesis Tests

- To recapitulate, the purpose of constructing the null and alternative hypotheses is to make claims as to whether or not there is *sufficiently significant statistical proof to either accept or reject the alternative hypothesis*
  - Pay attention to the underlined wording
  - It is incorrect to “accept” the null hypothesis – you cannot accept the null because you are not testing the null, *you are testing the alternative*
  - *All conclusions should be worded in terms of accepting or rejecting the alternative hypothesis*

Test Statistic

- Now that we have concluded with the structure of the test, let us compute the number which will serve as the basis for our test
- The Test Statistic, denoted as z^* , is simply the Normal standardizing formula which we saw previously

$$z^* = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Test Statistic

- At this point, we now have all of the tools compute can draw conclusions from a hypothesis test

1. There exists a population with an unknown mean μ (parameter) and a known variance σ^2 , and we are trying to estimate the mean
2. We take multiple SRS from the population, taking the average of each SRS and generate a frequency distribution (histogram) of these values, obtaining an unbiased estimate of the mean called the sample mean and denoted $\bar{x}$

Test Statistic

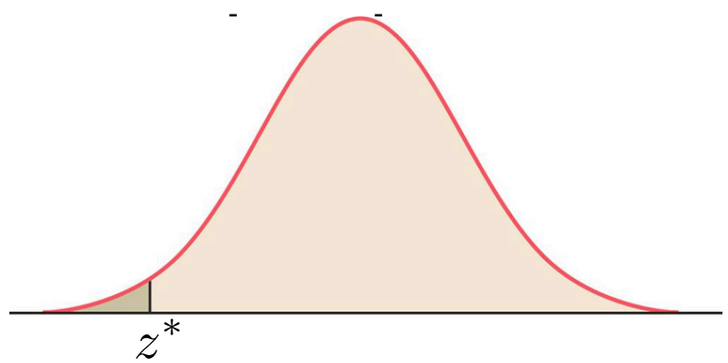
3. We standardize $\bar{x}$ using the Normal standardizing formula to obtain z^*, where

$$z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

4. We can now look up z^* in our standard Normal Table (*Z-Table*) and obtain a probability. What does this probability tell us? Our interpretation depends on whether or not our original hypotheses were stated as 1 or 2-sided tests

Test Statistic for 1-Sided Lower Tail Test

- If this is our hypothesis test: $H_0 : \mu \geq a \quad , H_a : \mu < a$
  - we have z^* so we know how many deviations below the mean it is



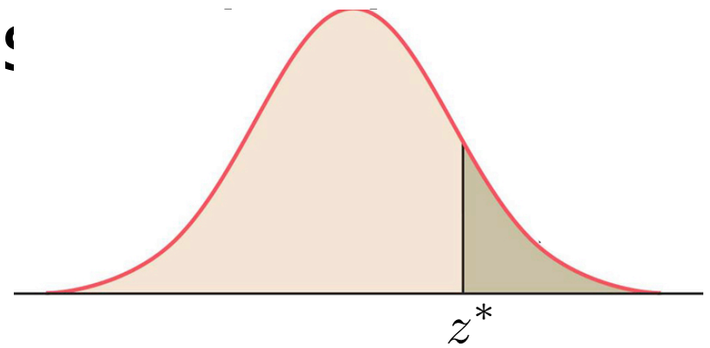
- We can now make the following statement:
  - *If $\mu \geq a$ then the probability of obtaining a value z^* is*

$$P(X \leq z^*)$$

which can be obtained from the Z-Table

Test Statistic for 1-Sided Upper Tail Test

- If this is our hypothesis test: $H_0 : \mu \leq a \quad , \quad H_a : \mu > a$
  - we have z^* so we know how many s deviations above the mean it is



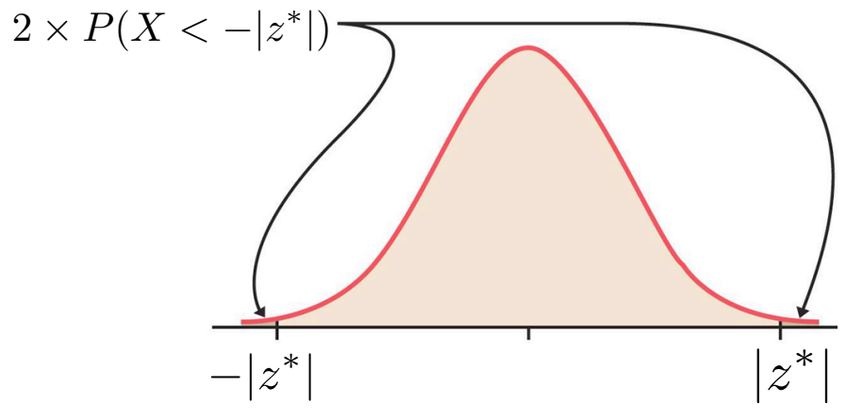
- We can now make the following statement:
  - *If $\mu \leq a$ then the probability of obtaining a value z^* is*

$$1 - P(X \leq z^*)$$

which can be obtained from the Z-Table

Test Statistic for 2-Sided (Tailed) Test

- If this is our hypothesis $H_0 : \mu = a \quad H_a : \mu \neq a$ test:
  ,
  - we have z^* so we know how many standard deviations above the mean it is



- We can now make the following statement:
  - *If $\mu = a$ then the probability of obtaining a value z^* is*

$$2[1 - P(X \leq |z^*|)] \text{ or } 2P(X \leq -|z^*|)$$

which can be obtained from the Z-Table

Interpretation of the Test Statistic

- What were are effectively asking when we design and execute a *2-tail* hypothesis test is the following:
  - *assuming that the true population mean* μ *is in fact equal to* a *(some value you choose to believe the population mean is equal to), what is the probability when taking a Simple Random Sample (SRS) that you could have obtained a sample mean of* $\bar{x}$ *?*
  - *if the probability of getting that sample mean from a population with an assumed* μ *equal to* a *is high, then we might conclude that our assumption about the true value of* $\bar{x}$ *is* μ *relatively accurate*

Interpretation of the Test Statistic

- *but if the probability of getting that sample mean* $\bar{x}$ *from a population with an assumed* μ *equal to* a *is low, then we might conclude that our assumption about the true value of* μ *was inaccurate*

Hypothesis Test: *Example*

- The McGill Health clinic is worried about how stressed Ph.D. students are. According to the clinic, the mean blood pressure of students is 128 with a standard deviation of 15. When taking a sample of 72 Ph.D. students, the mean blood pressure is 129.93? Does this value imply that Ph.D. students are more stressed than other students?

1. Observe that $\sigma = 15$ *(given)*

Hypothesis Test: *Example*

2. Set up the hypothesis test as follows: we have no reason (yet) to believe that Ph.D. students have higher blood pressure than other students, so we make the assumption (null hypothesis) than the (population) mean blood pressure of all Ph.D. Students is equal to 128

$$H_0 : \mu = 128$$

3. For the alternative hypothesis (what we are trying to prove or test), we formulate the hypothesis that Ph.D. students have a mean blood pressure greater than 128

$$H_a : \mu > 128$$

Hypothesis Test: *Example*

4. Compute the test statistic (which is simply the standardized Normal value of $\bar{x}$)

If H_0 didn't contain an equality statement, what value would we put in place of μ

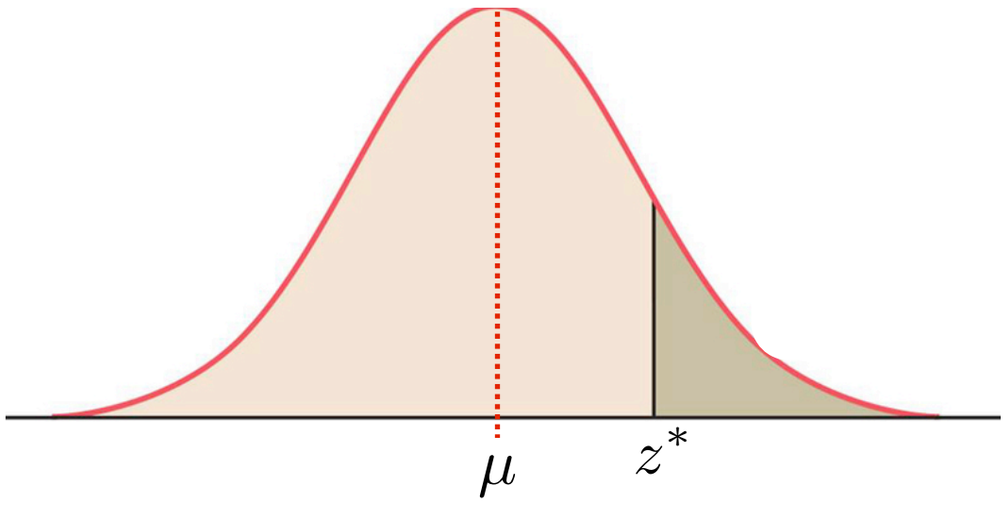
$$z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{129.93 - 128}{15/\sqrt{72}} = 1.09$$

if we stop for a moment to try an interpret this value, we can say that if the true mean blood pressure μ of Ph.D. students is in fact 128 or less, then the sample mean $\bar{x}$ of our SRS (129.93) implies that we are 1.09 standard deviations above our assumed population mean

Hypothesis Test: *Example*

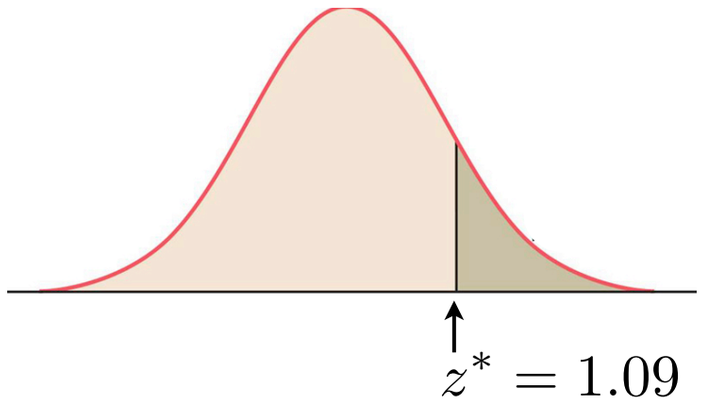
5. The question we are now trying to answer is, if the true mean blood pressure μ of Ph.D. students is in fact equal to 128, what is the probability that in a simple random sample, we would have obtained a value $\bar{x}$ equal to 129.93?

   Observing that we are preforming a 1-tailed (upper-tail) hypothesis test ,we compute (*see next page*)...



Hypothesis Test: *Example*

1. *(cont'd)*



p-value

$$P(X > z^*) = P(X > 1.09) = 1 - P(X \leq 1.09) = 1 - 0.8621 = 0.1379$$

Rewrite this as a subtraction because the Z-Tables only provide us with probabilities which are in the form of $\leq$

6. About 14% of the time an SRS of size 72 from the general Ph.D. population would have a mean blood pressure as high as ($>$) that of the sample mean $\bar{x}$

Post-Example Analysis

- We have walked our way through an example of hypothesis testing and obtained a probability of an event happening...but the question we originally asked was: is the sample mean $\bar{x}$ significantly statistically different from μ , the assumed population mean...*we never really got around to answering that*

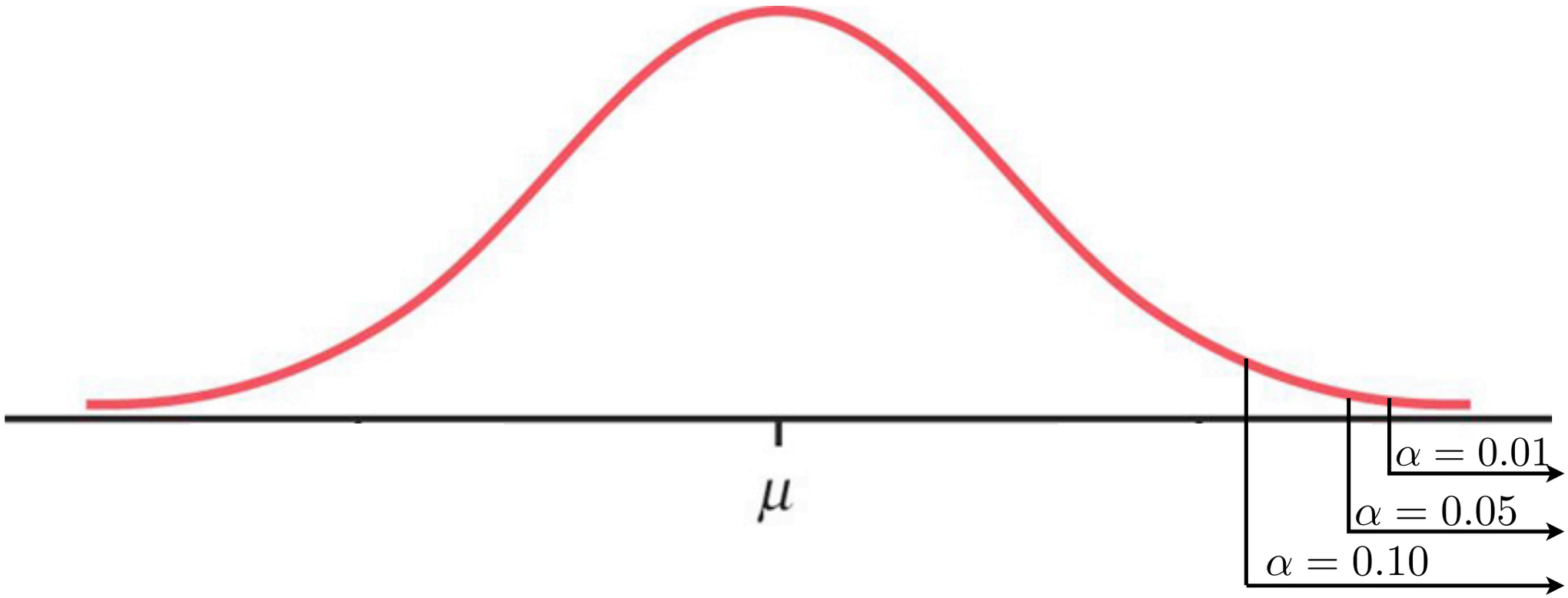
- The answer is subjective: what level confidence do you think is important to have? Let's explore this idea for a moment...

Level of Significance

- If you the repeat an SRS 100 times, what percentage of “rarity” do you think is significant?
- In the previous example, 100 SRS’s of 72 Ph.D. students will result in 14% of those SRS’s having sample means larger than 129.93. If an event occurs, on average, 14% of the time, would you consider that to be a significantly small % of time that would claim it is sufficiently different from the population mean? What about 10%? 5%? 1%? 0.1%?
- This percentage is referred to as in the context of hypothesis testing and is called the *level of significance*

Level of Significance

- Note that as α (the level of significance) decreases, we are requiring our samples to be farther and farther from the assumed population mean to be significant

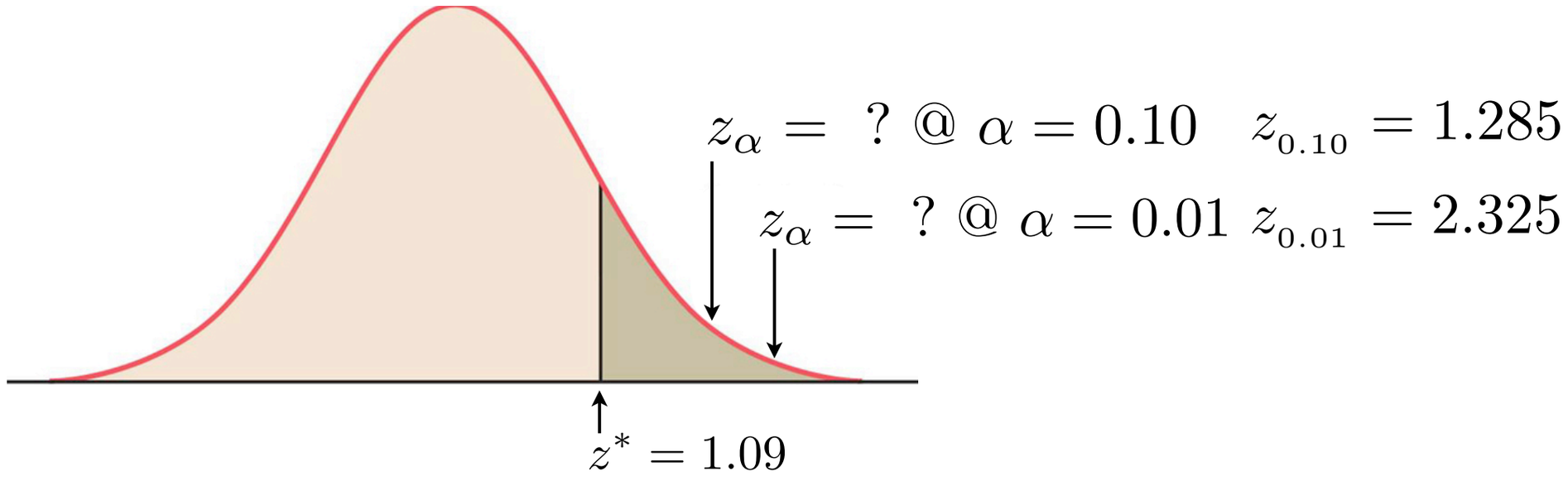


- We consider anything that happens beyond the prescribed significance level (α) to be *significant*

Level of Significance

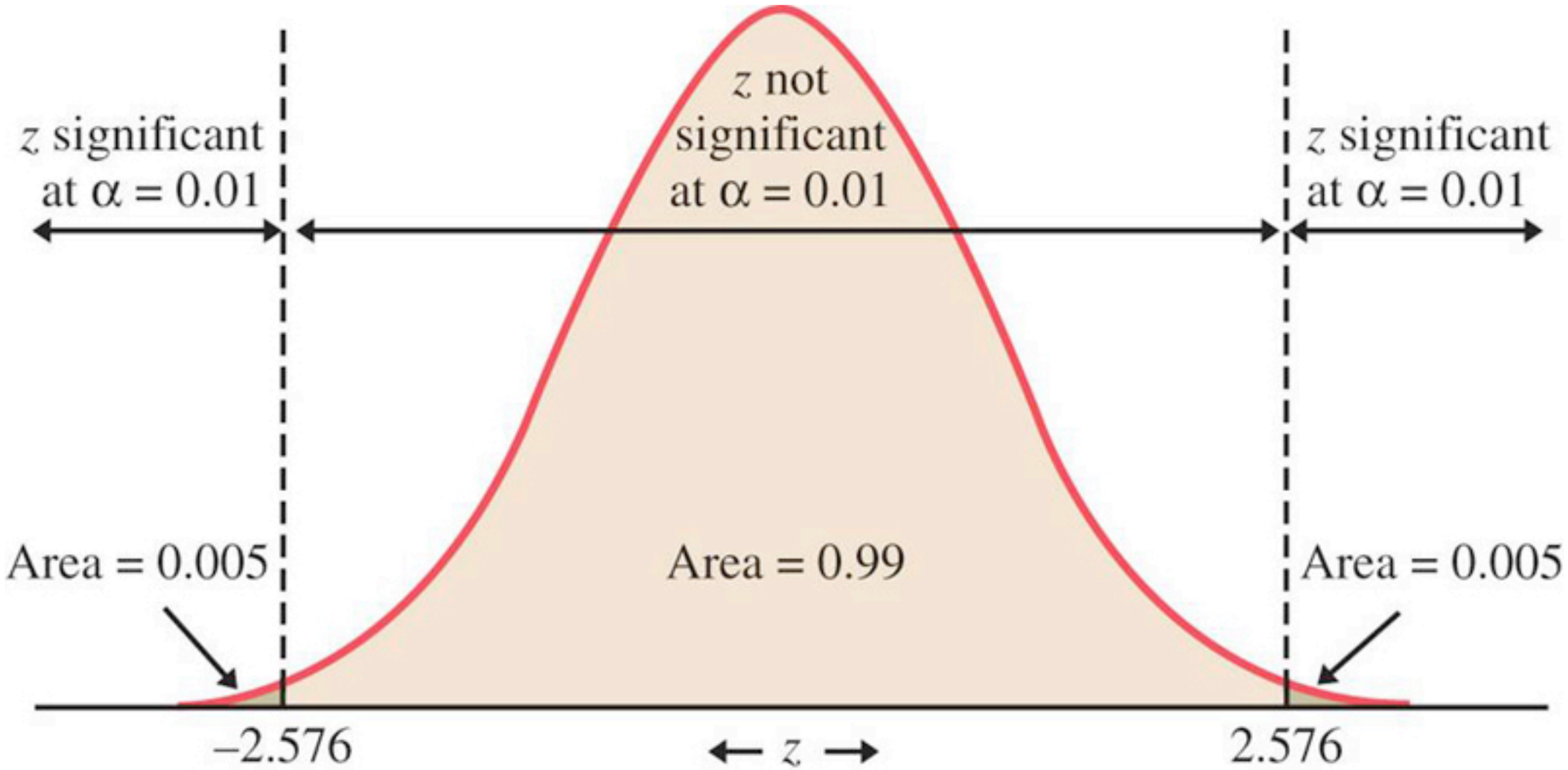
- In statistics, there are a few established levels of significance which are often referred to: is events happen with probabilities 10%, 5%, or 1% (depending on the context), we usually deem those to be significantly different from the population mean, although there is no rule keeping you from choosing any level of α you please

Example Revisited



- In a 1-tailed test, what is the equivalent z value for a level of significance $\alpha = 0.10$?
- In a 1-tailed test, what is the equivalent z value for a level of significance $\alpha = 0.01$?
- In this case we would reject H_a for any $z^* < z_\alpha$ and not reject H_a if $z^* > z_\alpha$

Level of Significance: 2-Tailed Tests



$$\alpha = 1\% \text{ significance}$$

Hypothesis Tests & Confidence Intervals

- Observe the different approaches used to solve the same problem
  - Hypothesis testing says that we assume the population mean to be a certain value and then assess the probability of obtaining a sample mean given our initial assumption
  - Confidence intervals are centered around the sample mean, with equidistant margins of error on either side
  - They really are two sides of the same coin...

Example

- A pharmaceutical company produces a drug which is supposed to have a concentration of 0.86%. Standard deviation of the concentration is given as $\sigma = 0.0068$. Three analyses of the concentrations of test samples are performed and found to be 0.8403%, 0.8363%, and 0.8447%. Is there significant evidence at the 1% level of significance that the true concentration is not equal to 0.86%?
    - We will use both a hypothesis test approach and a CI approach, in that order

Example

- Firstly, compute

$$\bar{x} = \frac{0.8403 + 0.8363 + 0.8447}{3} = 0.8404$$

- Hypothesis Test

$$H_0 : \mu = 0.86$$

$$H_a : \mu \neq 0.86$$

Is this a rare event? $p - value \approx 0$

Example

- Firstly, compute

$$\bar{x} = \frac{0.8403 + 0.8363 + 0.8447}{3} = 0.8404$$

- Hypothesis Test

$$H_0 : \mu = 0.86$$

$$H_a : \mu \neq 0.86$$

$$z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{0.8404 - 0.86}{0.0068/\sqrt{3}} = -4.99$$

Is this a rare event? $p - value \approx 0$

Example

- Hypothesis Test *(cont'd)* *d*
  - *Just as a formality, at what is the limiting value associated with* $\alpha = 1\%$

$$z_{\alpha/2} = z_{0.01/2} = z_{0.005} = \ ?$$

  - *Recalling that this is a 2-tailed test, we look up*

2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981

$$z_{\alpha/2} = 2.575$$

Example

- Hypothesis Test *(cont'd)* *d*
  - *We therefore accept* H_a *if* $|z^*| > |z_{\alpha/2}|$ *which in our case is true* $|-4.99| > |2.575|$
  - *Conclusion*
    - *Nearly 0% of the time a sample of 3 concentrations will have a sample mean of 0.8404% if the true (population) mean concentration is 0.86%, hence we accept* $H_a : \mu \neq 0.86$ *at a 99% level of confidence.*

Example

- Confidence Intervals
  - We have already computed $\bar{x} = 0.8404$
  - Compute the CI $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$
    - What is z?

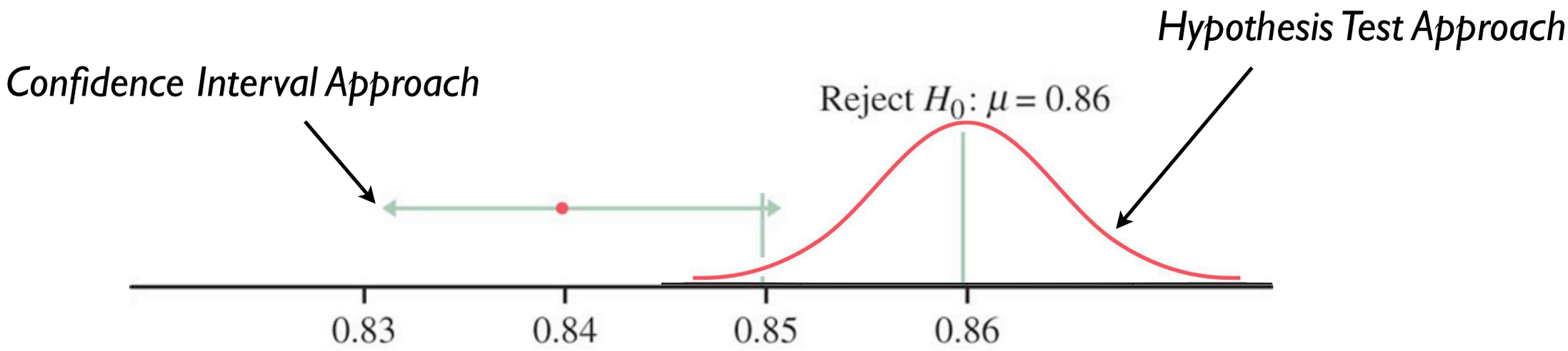
$$z_{\alpha/2} = 2.575$$

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} = 0.8404 \pm 2.575 \times \frac{0.0068}{\sqrt{3}}$$

$$0.8404 \pm 0.0101 \Rightarrow [0.8303, 0.8505]$$

- Observe that the CI does not contain our assumed population mean $\mu = 0.86\%$

Two Sides of the Same Coin *Revisited*



- Notice that when using the same level of significance α for the hypothesis test and the confidence interval, we arrive at the same conclusion
- ***Note***: one-sided CI exist for 1-tailed tests, but we will not learn that in this class, so use with caution