INTRODUCTION TO STATISTICS

Institute for Advanced Analytics

MSA Class of 2020

First things First: Explore your Data!

- BEFORE you attempt to run any models or jump towards a solution, you should **EXPLORE your data**.
  - What kind of variables do you have?
  - What do their distributions look like?
  - Do they have any interesting associations?
  - Are there any anomalies?

FUNDAMENTAL STATISTICS CONCEPTS

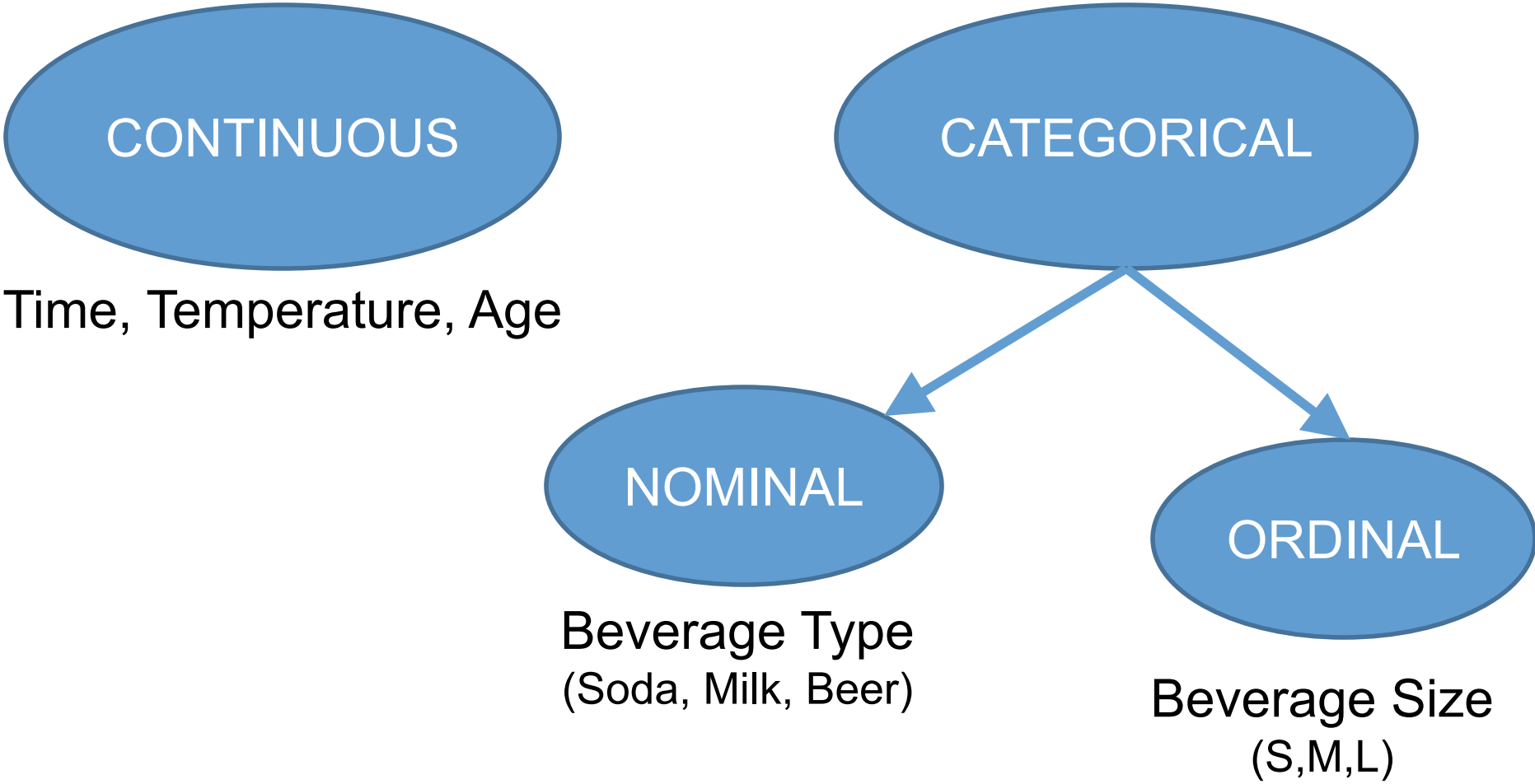
Types of Variables

Qualities and Quantities of Interest

Equivalently called:

- Variables
- Attributes
- Predictors
- Factors

Variable Type and Level of Measurement



Ordinal Variables

- A variable is **ordinal** if there are only 2 *logical* orderings:

- Size:
  - S → M → L
  - L → M → S
- Level of education:
  - No HS Diploma → HS Diploma → … → PhD
  - PhD → … → HS Diploma → No HS Diploma

- By this definition, **binary variables are ordinal!**

Ordinal Variables

- Ordinal variables treated as ***either*** continuous ***or*** categorical.

- The levels are given values if treated continuously:

Size
S
M
S
L

Size
1
2
1
3

- The levels become dummy variables if treated categorically:

Size
S
M
S
L

Small	Med.	Large
1	0	0
0	1	0
1	0	0
0	0	1

The Ames Home Sales Data Set

Printing / Listing Out Observations

```
proc print data=bootcamp.ameshousing3 (obs=10);
     title 'Listing of the Ames Housing Data Set';
run;
```

Frequencies of Categorical Variables

```
proc freq data=sasuser.ameshousing3;
     tables &categorical / plots=freqplot ;
     title "Categorical Variable Frequency Analysis";
run;
```

General form of the FREQ procedure:

```
PROC FREQ DATA=SAS-data-set;
        TABLES table-requests </ options>;
RUN;
```

FUNDAMENTAL STATISTICS CONCEPTS

Populations & Samples

Populations and Samples

- **Population** – the *entire* collection of individual members of a group of interest.
- **Sample** – a subset of a population drawn to enable inferences to the population.

- **Assumption for this course**– sample that is drawn is **representative** of the population.

Parameters and Statistics

Statistics are used to approximate population **parameters**.

	Population Parameters	**Sample Statistics**
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Poll

Quiz

DISTRIBUTIONS OF DATA

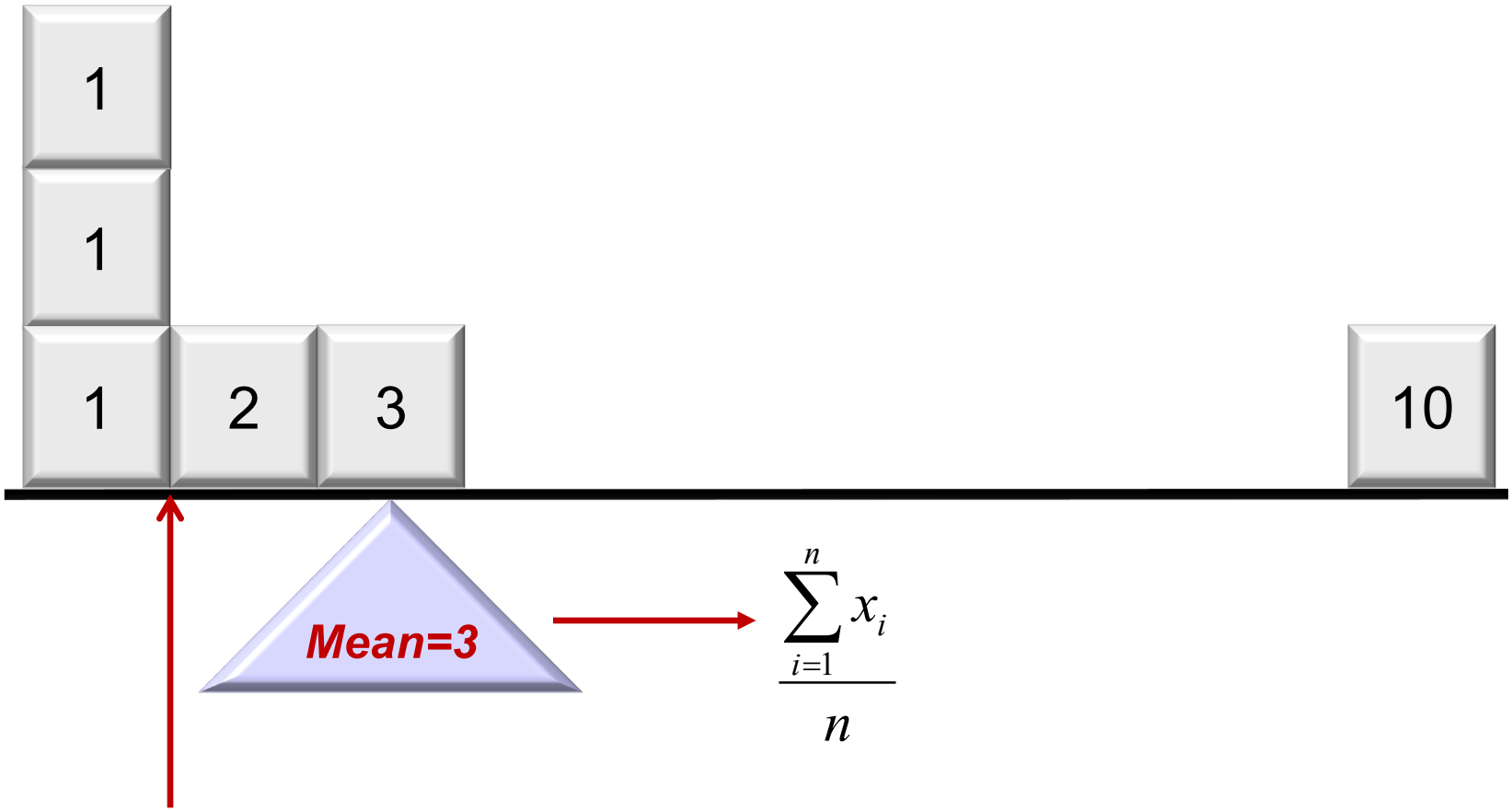
Describing Your Data

- 4 Keys Things to Look for:
  1. Center / Location
  2. Spread (Variation)
  3. Shape
  4. Anomalous Observations

Describing Your Data

- 4 Keys Things to Look for:
  1. Center / Location
  2. Spread (Variation)
  3. Shape
  4. Anomalous Observations

Central Tendency – Mean, Median, Mode



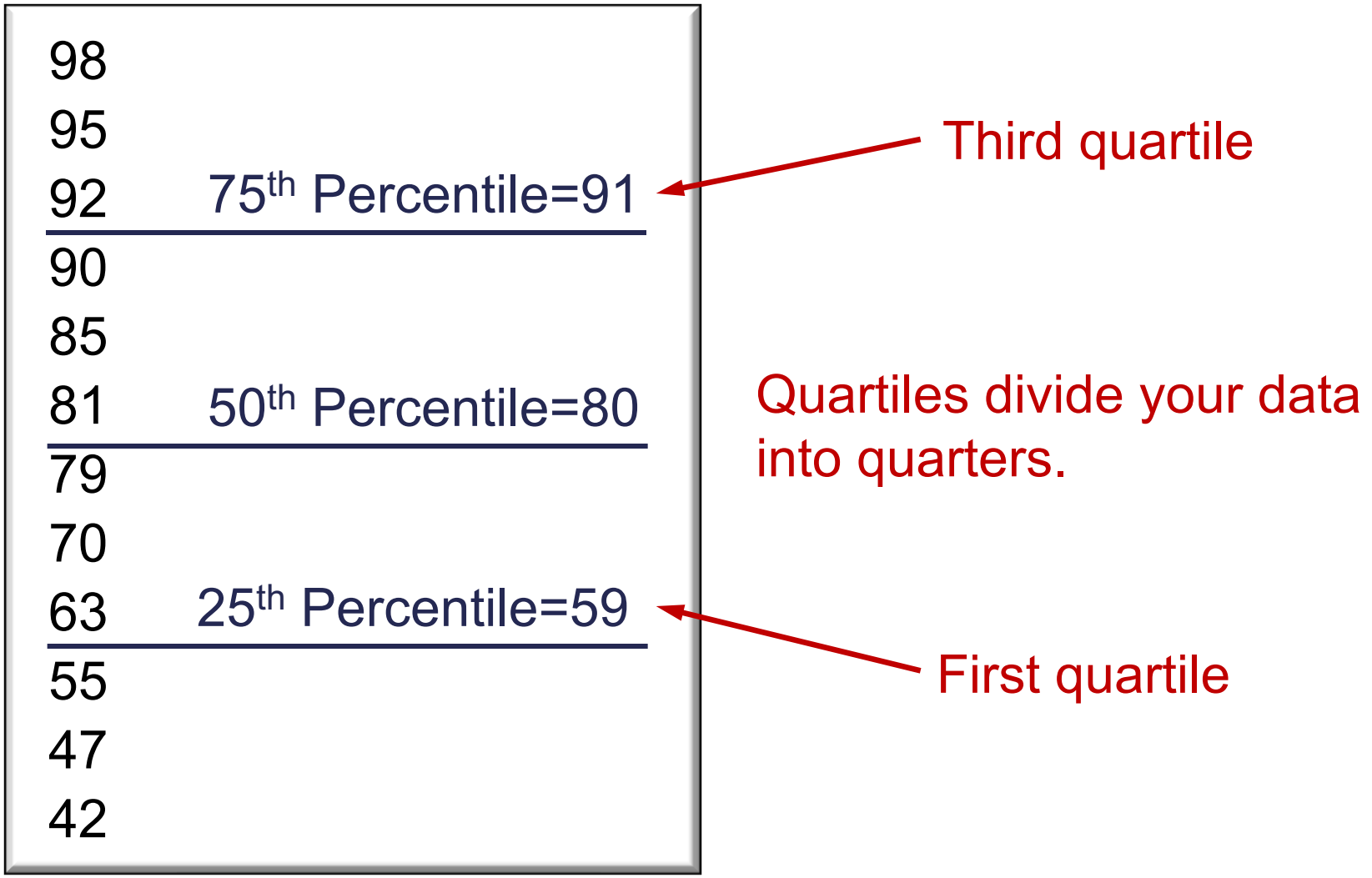
Median = 1.5 the middle value (also known as the 50th percentile)

Central Tendency – Mean, Median, Mode

Mode = 1 **(**The most frequently occurring value.)

Mode *is more commonly used to summarize* ***categorical*** *attributes*

Percentiles



Describing Your Data

- 4 Keys Things to Look for:
  1. Center / Location
  2. Spread (Variation)
  3. Shape
  4. Anomalous Observations

The Spread of a Distribution: Dispersion

Measure	Definition
Range	Difference between the maximum and minimum data values
Interquartile Range	Difference between the 25^{th} and 75^{th} percentiles
Variance	Dispersion of the data around the mean
Standard Deviation	Dispersion expressed in the same units of measurement as your data (the square root of the variance)

Basic Summary Statistics

```
proc means data=bootcamp.ameshousing3;
     var &interval;
     title 'Descriptive Statistics of Ames Data';
run;
```

Basic Summary Statistics

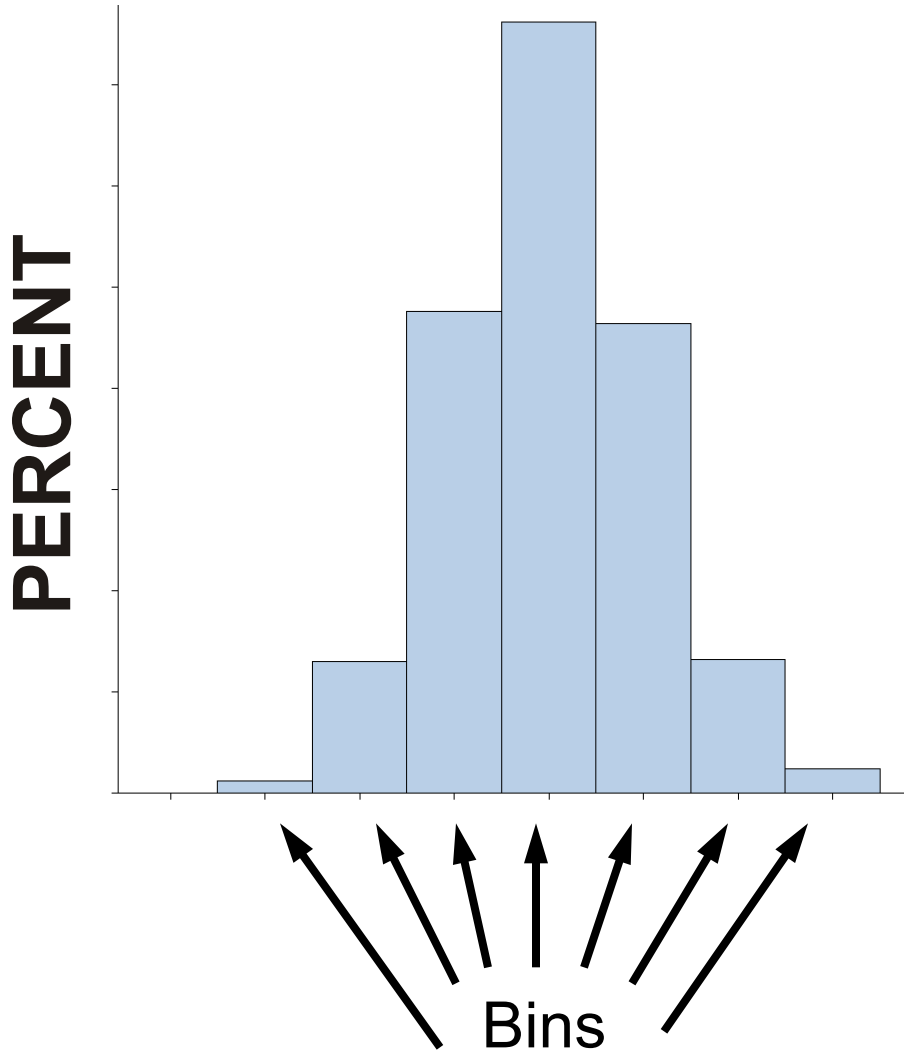
```
proc means data=bootcamp.ameshousing3 printalltypes;
     class House_Style;
     var SalePrice;
     title 'Descriptive Statistics by House Style';
run;

proc means data=bootcamp.ameshousing3
           maxdec=2
           n mean median std q1 q3 qrange;
     var SalePrice;
     title 'Selected Descriptive Statistics';
run;
```

Describing Your Data

- 4 Keys Things to Look for:
  1. Center / Location
  2. Spread (Variation)
  3. Shape
  4. Anomalous Observations

Picturing Distributions: Histogram

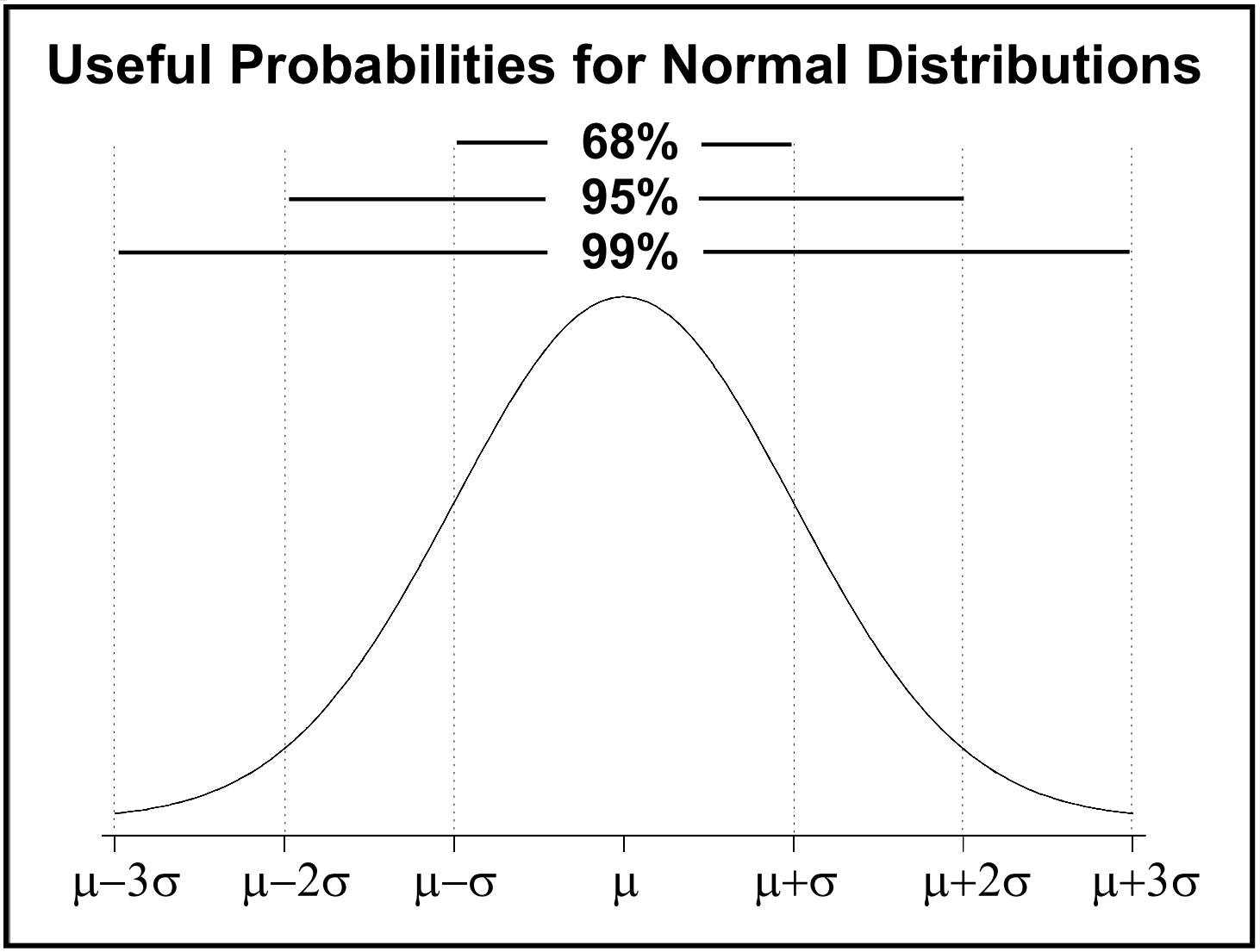


- Each bar in the histogram represents a group of values (a *bin*).
- The height of the bar represents the frequency or percent of values in the bin.
- SAS determines the width and number of bins automatically, or you can specify them.

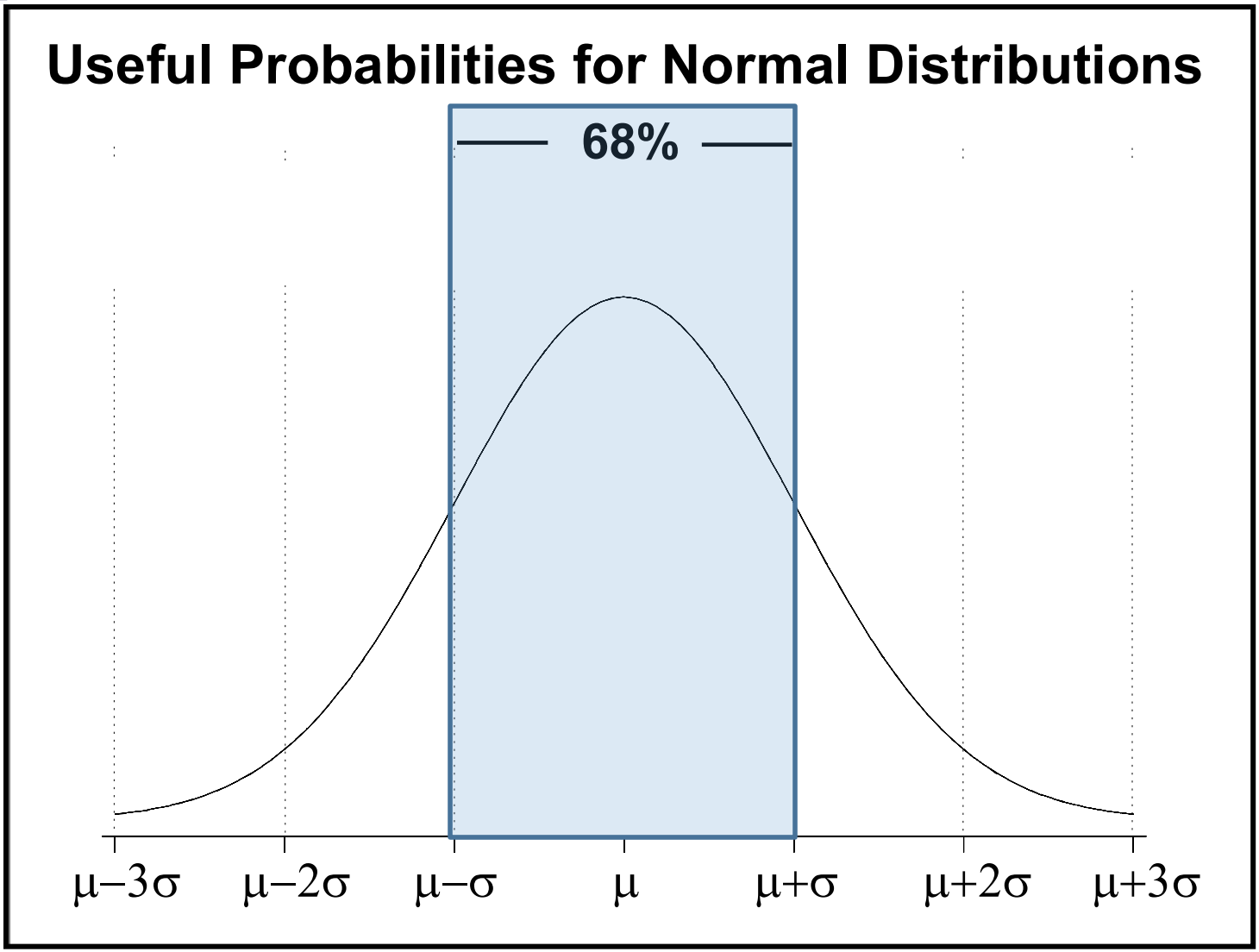
Normal Distribution Characteristics

- Symmetric
- Fully Defined by the mean and standard deviation.
- Bell Shaped / Unimodal
- Mean = Median = Mode
- Asymptotic to the x-axis (bounds are $-\infty$ to $+\infty$)

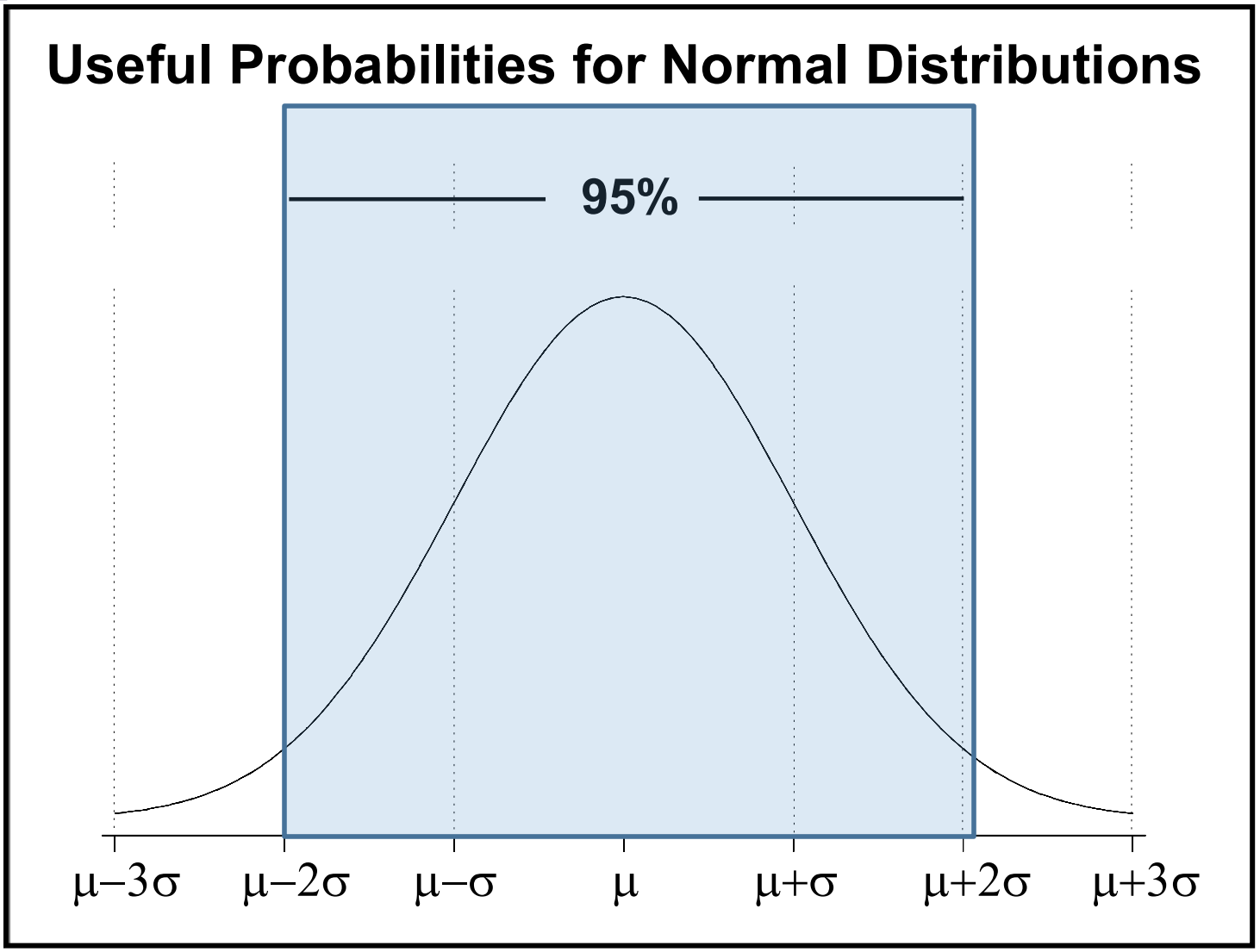
Empirical Rule



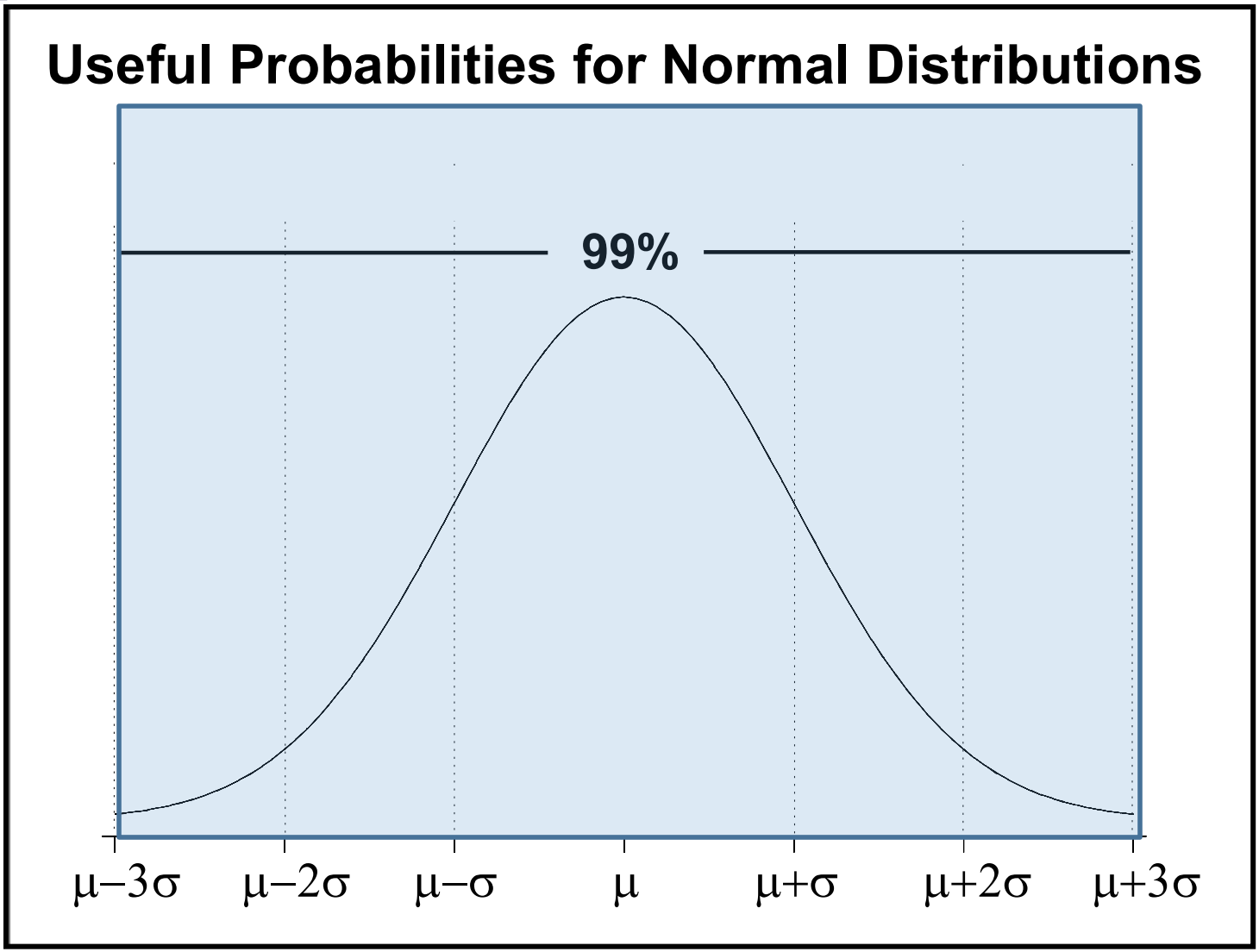
Empirical Rule



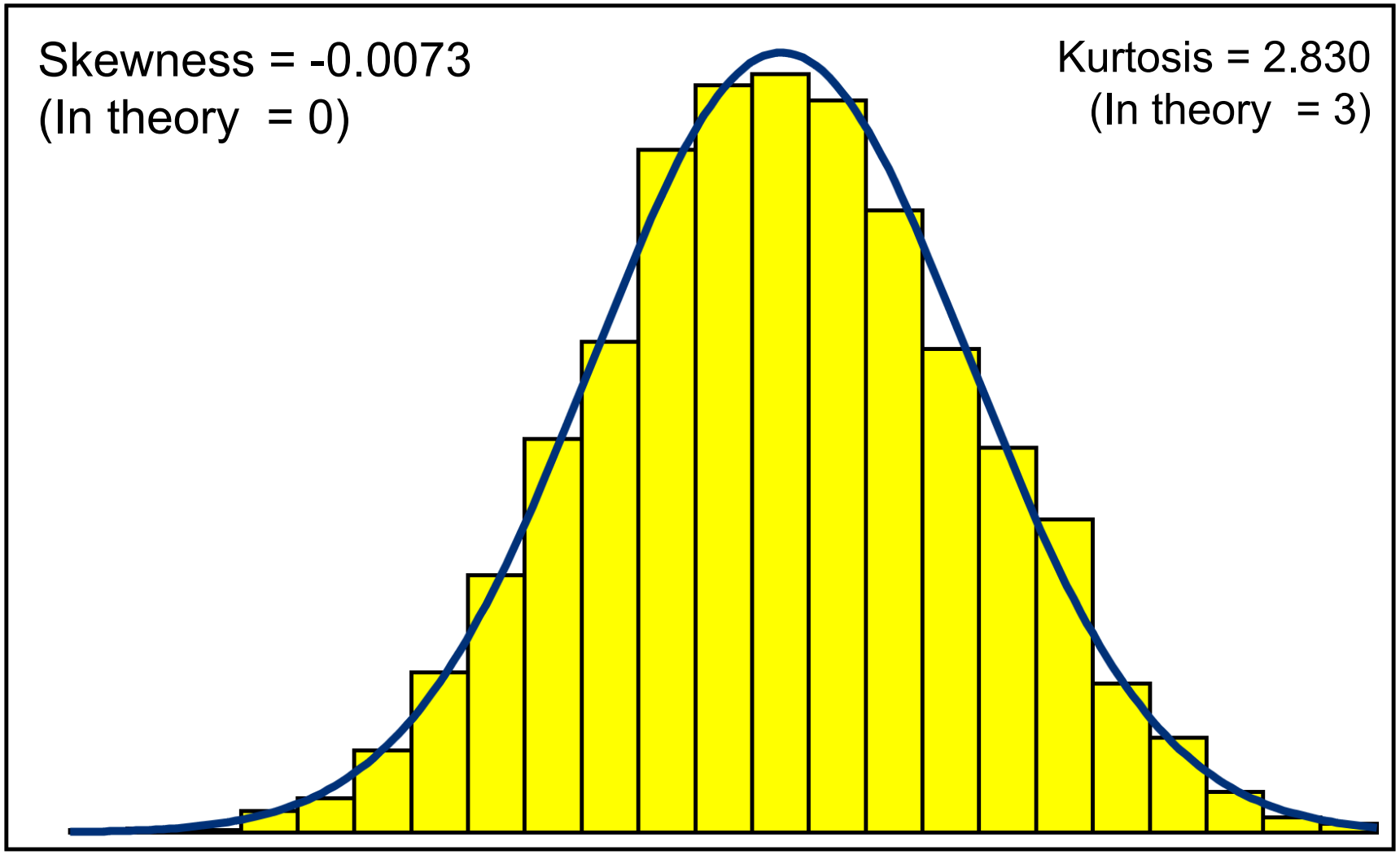
Empirical Rule



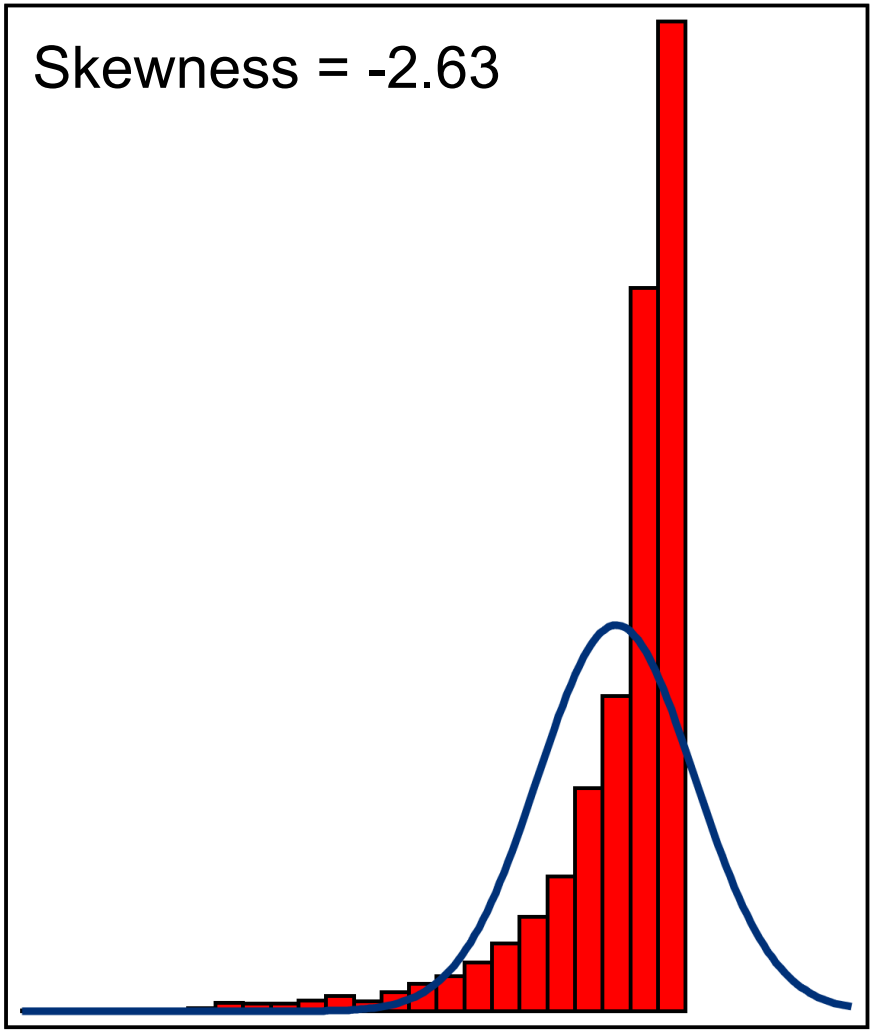
Empirical Rule



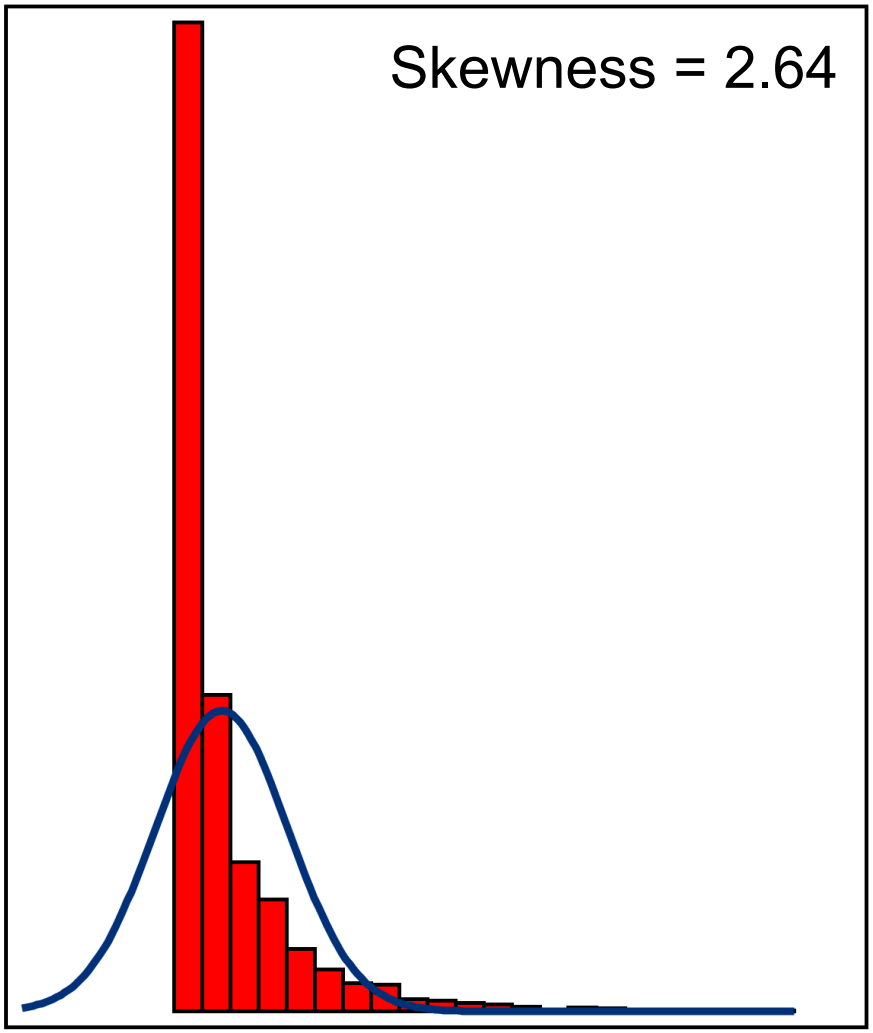
A Normal Distribution



Skewness



A Left Skewed Distribution



A Right Skewed Distribution

Kurtosis

- Number that describes the “tailedness” of a distribution.
- The normal distribution has a kurtosis of 3.
- We often focus on ***excess* kurtosis** which compares a distribution to the Normal distribution by simply subtracting 3.

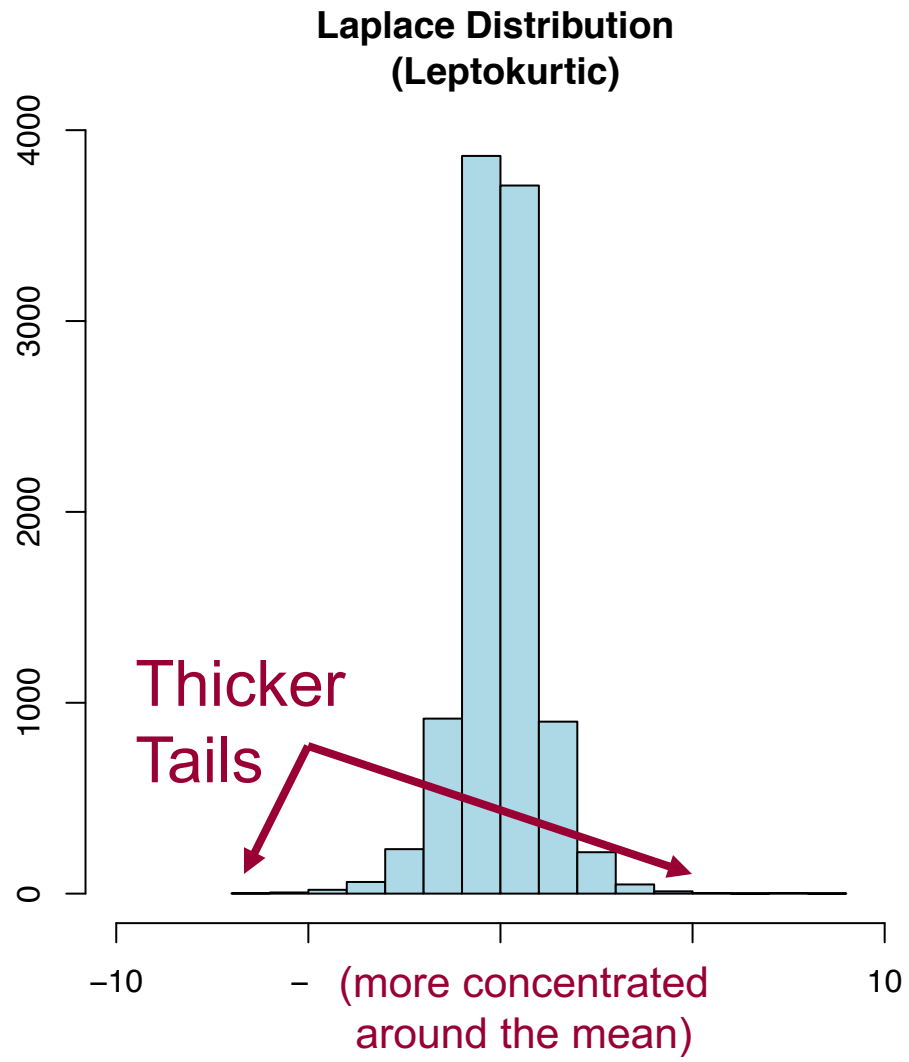
Platykurtic/Leptokurtic

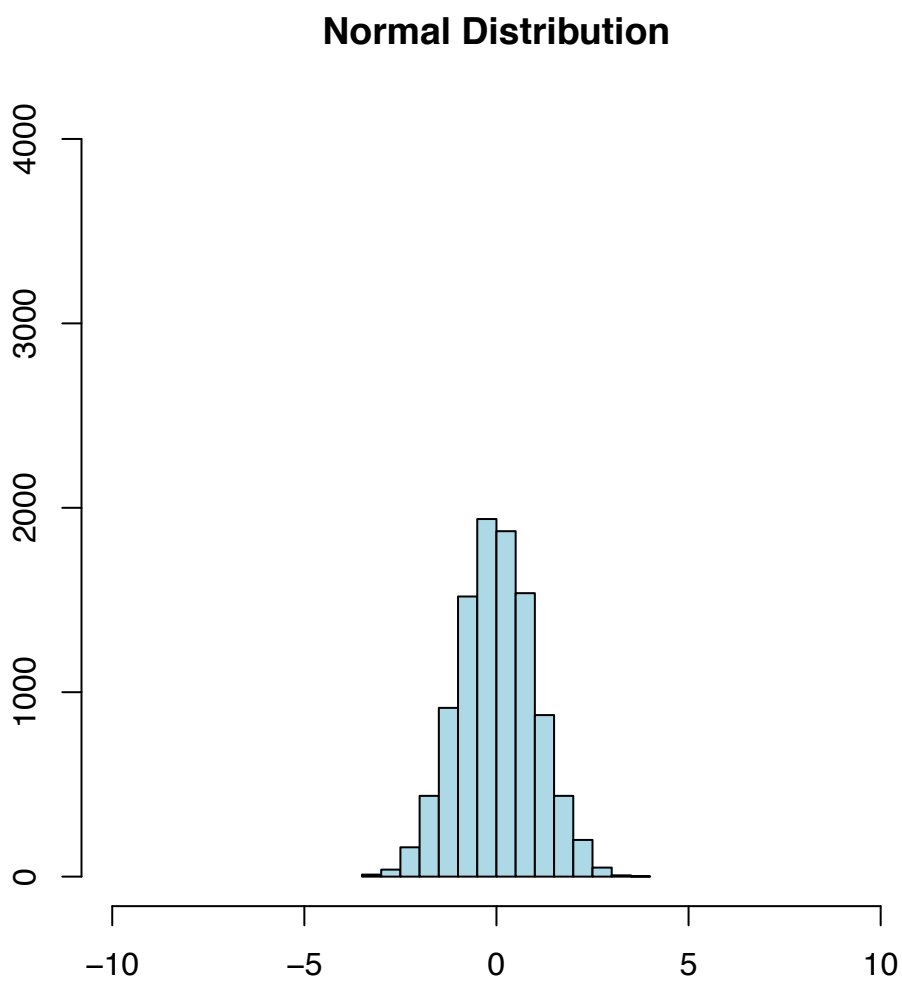
How the tails of your distribution compare to a normal distribution *with the same mean and variance.*

- **Platykurtic** (or platykurtotic): thin-tailed. Smaller amount of data in the tails

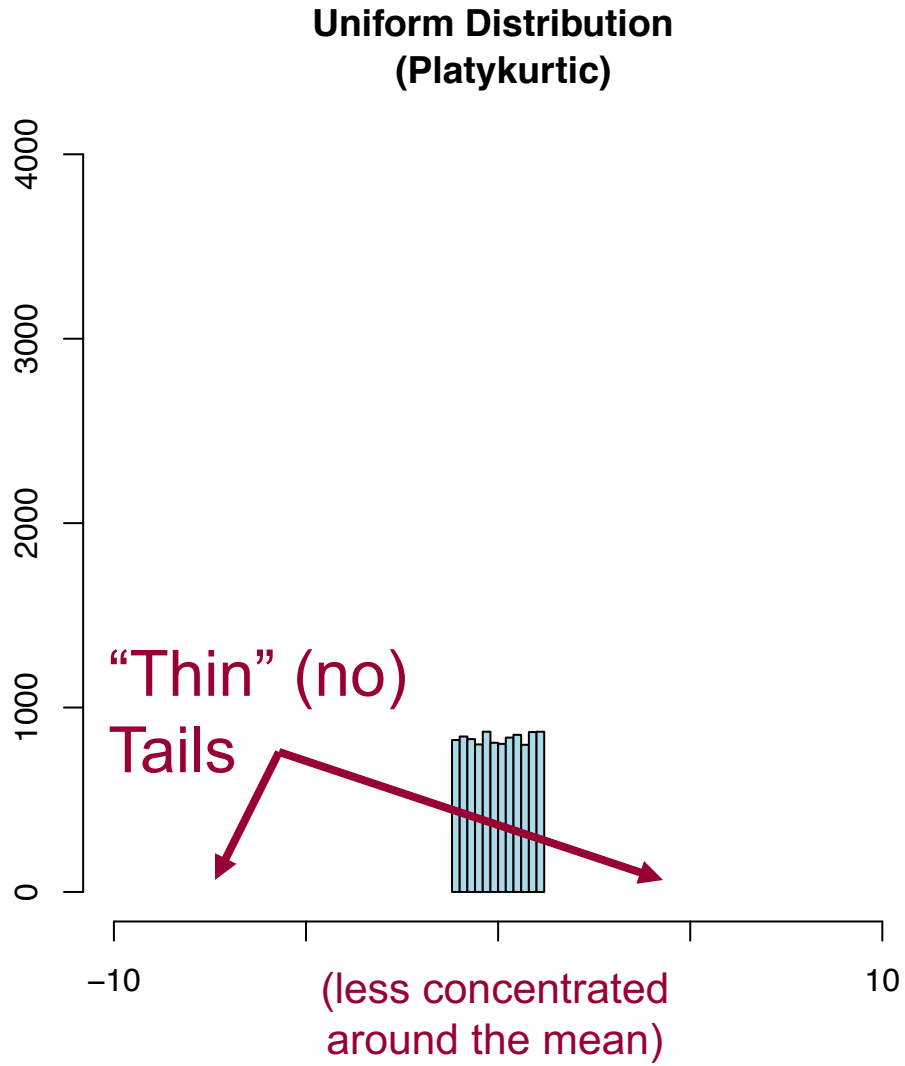
- **Leptokurtic** (or leptokurtotic): heavy-tailed. Larger amount of data in the tails

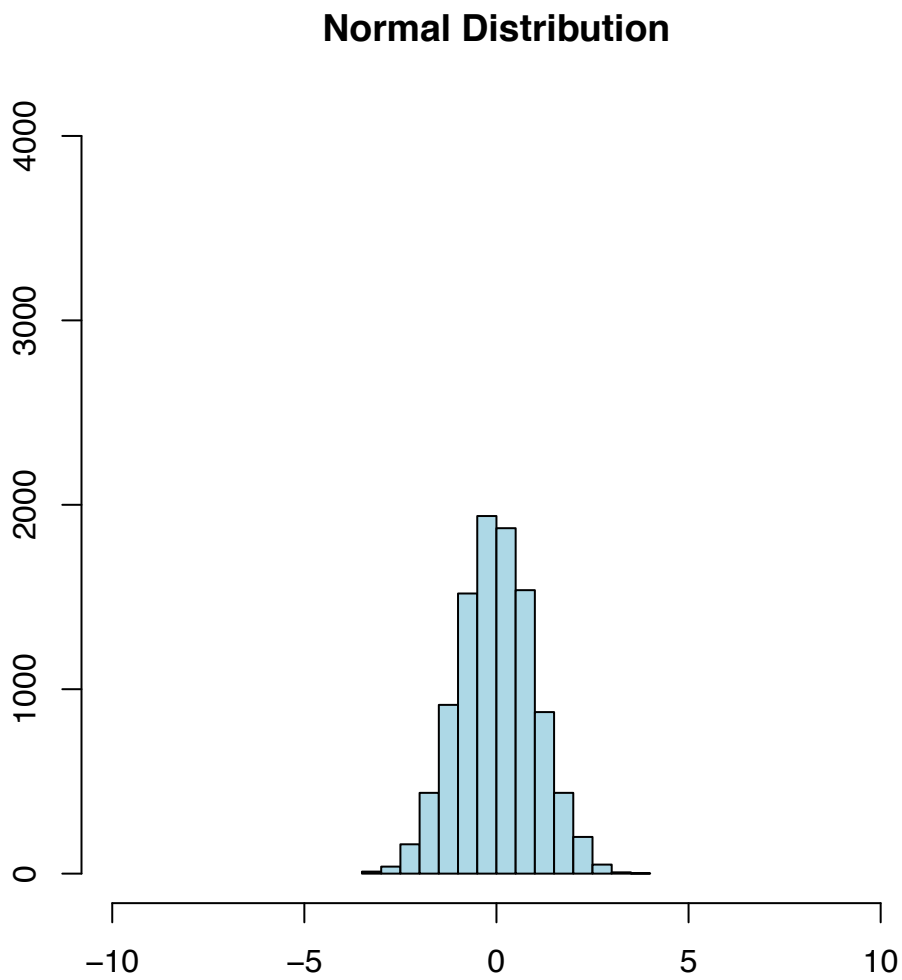
Leptokurtic Example

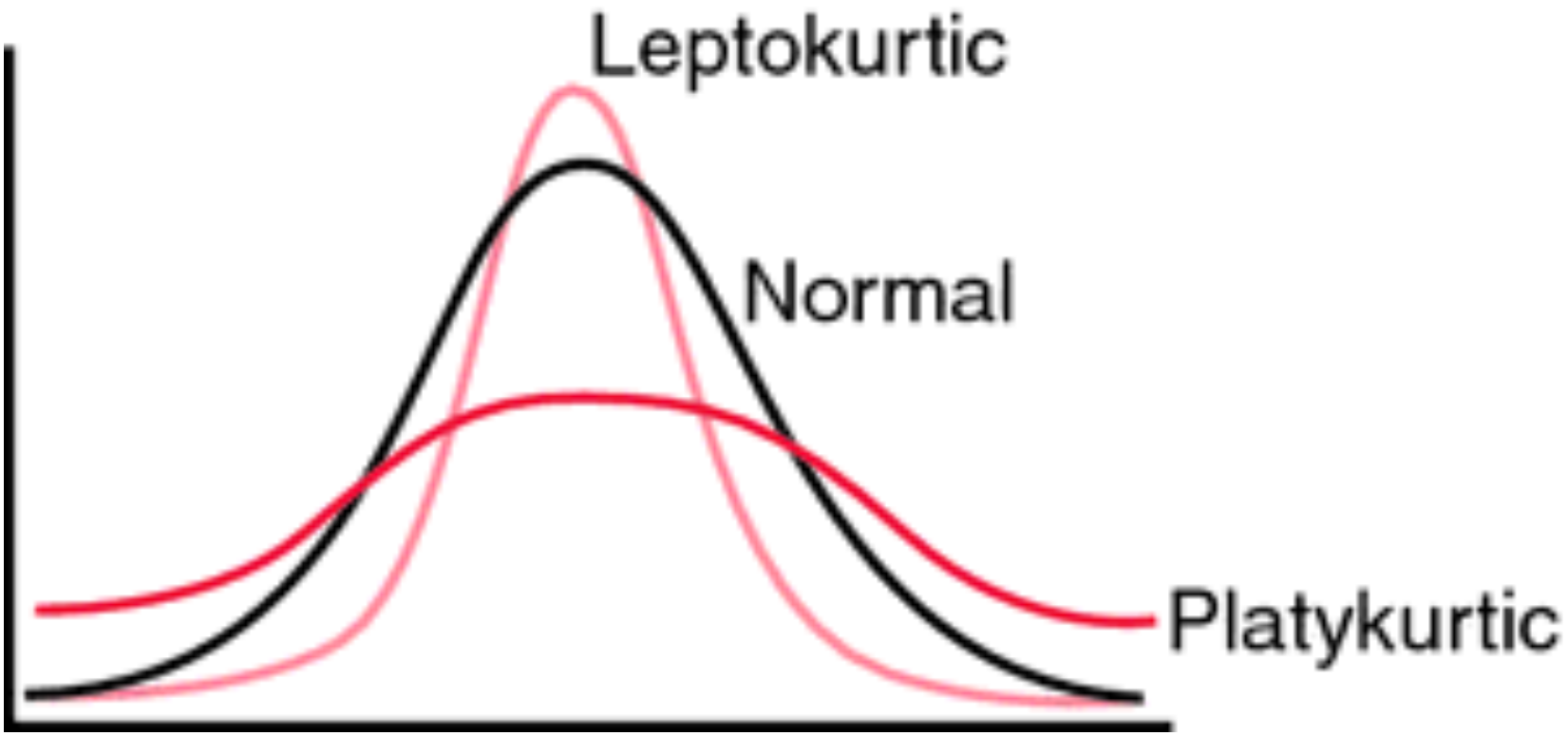




Platykurtic Example



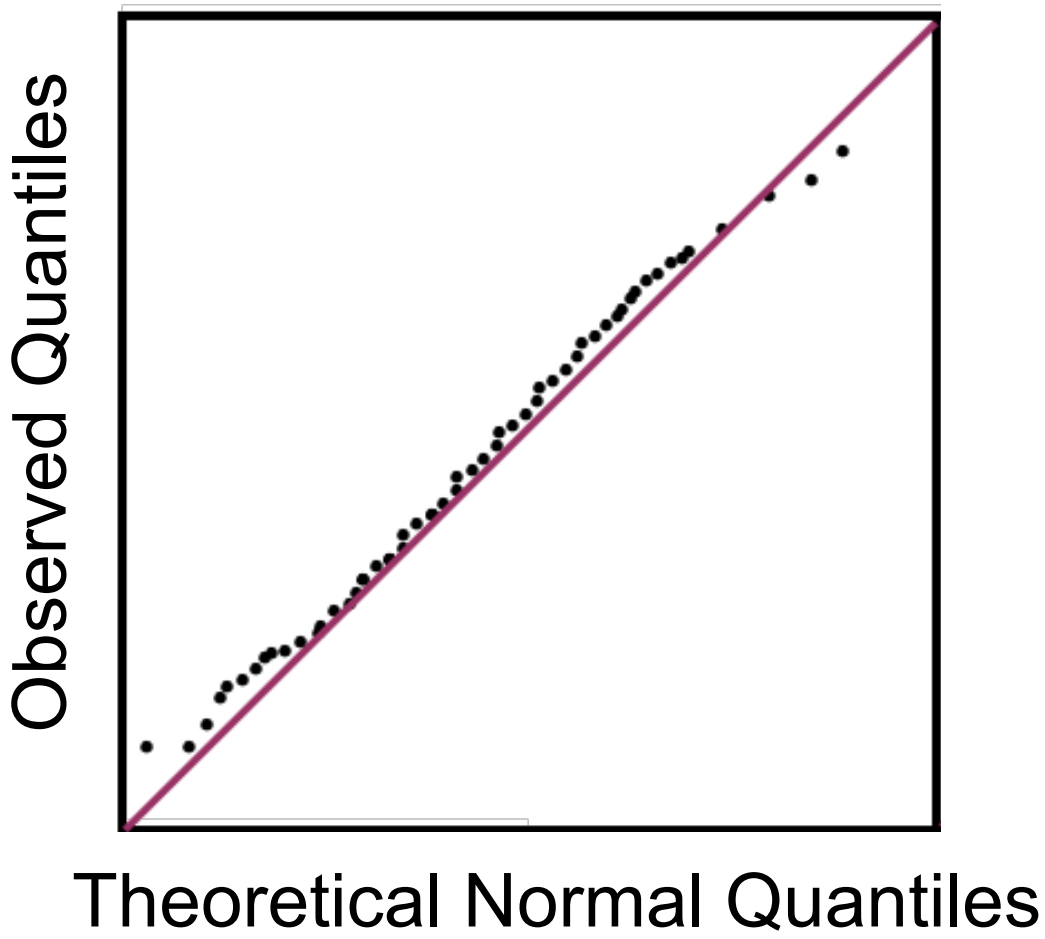



Leptokurtic
Normal
Platykurtic

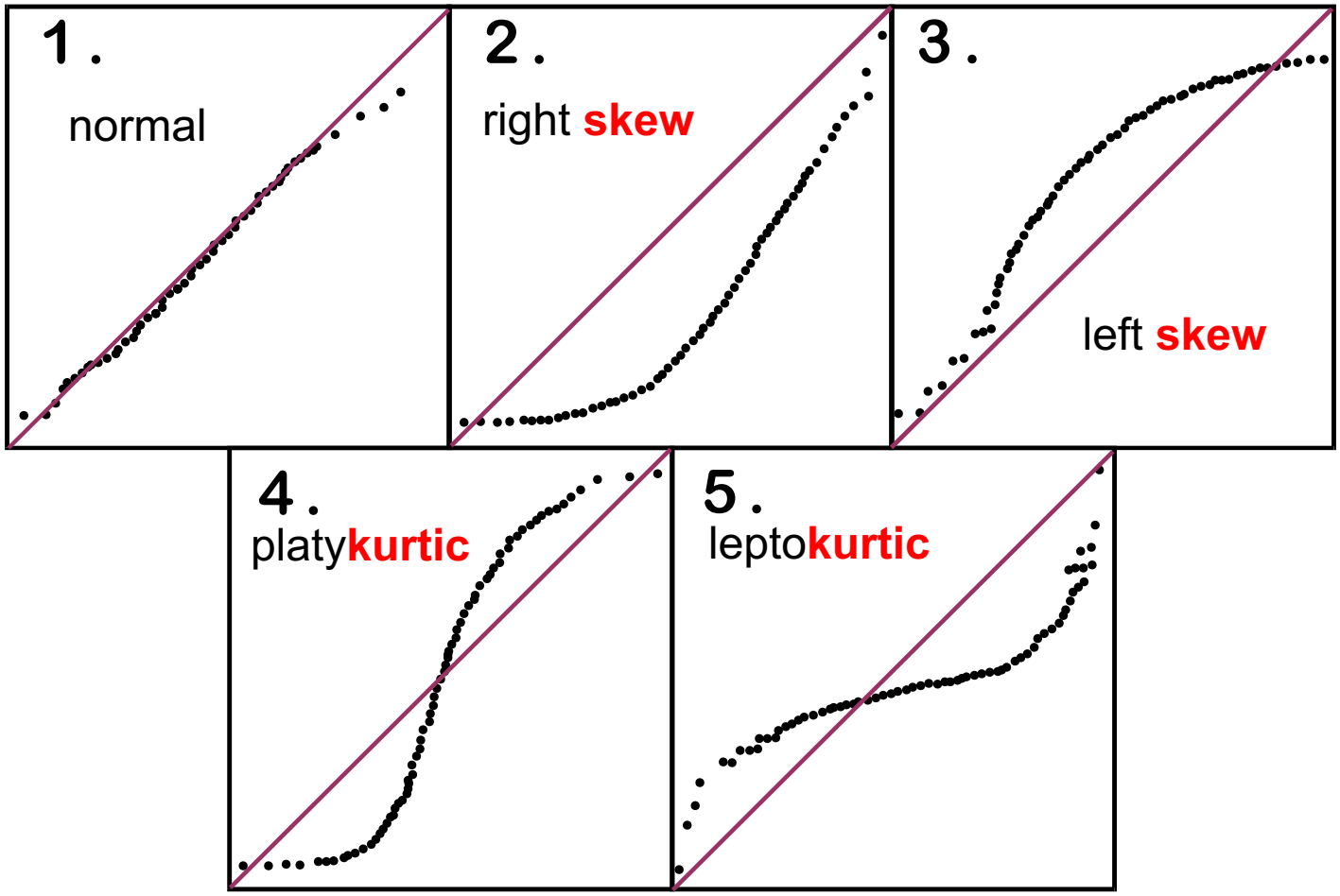
Graphical Displays of Distributions

- You can produce three types of plots for examining the distribution of your data values:
  - Histograms
  - Normal Probability Plots (QQ-plots)
  - Box Plots

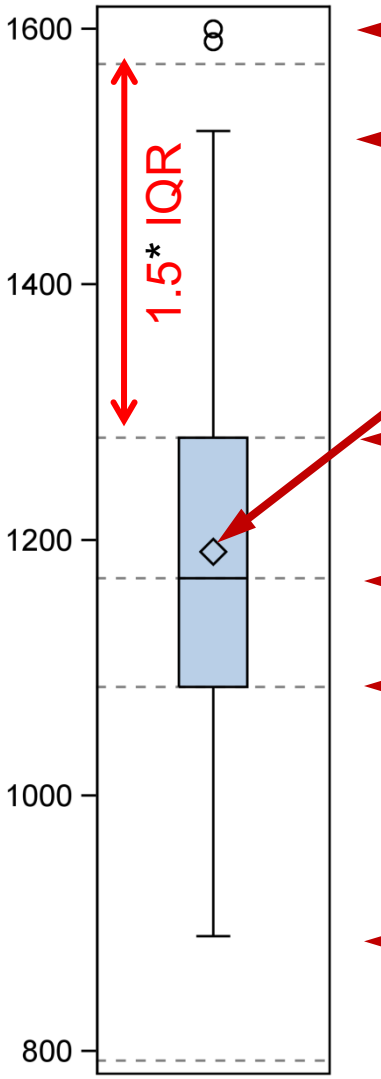
Normal Probability Plots (QQ Plots)



Normal Probability Plots



Box Plots



outliers > 1.5 IQR from the box

largest point <= 1.5 IQR from the box

The mean is denoted by a ◊.

the 75th percentile

the 50th percentile (median)

the 25th percentile

smallest point <= 1.5 IQR from the box

Statistics, Histograms, QQ-plots

```
proc univariate data=bootcamp.ameshousing3;
     var SalePrice;
     histogram SalePrice / normal(mu=est sigma=est) kernel;
     inset skewness kurtosis / position=ne;
     probplot SalePrice / normal(mu=est sigma=est);
     inset skewness kurtosis;
     title 'Descriptive Statistics of Sales Price';
run;
```

Statistical Graphics Procedures in SAS

- **PROC SGSCATTER creates single-cell and multi-cell scatter plots and scatter plot matrices with optional fits and ellipses.**
- **PROC SGPLOT creates single-cell plots with a variety of plot and chart types.**
- PROC SGPANEL creates single-page or multi-page panels of plots and charts conditional on classification variables.
- PROC SGRENDER provides a way to create plots from graph templates that you modified or wrote yourself.

The SGPLOT Procedure

```
PROC SGPLOT <option(s)>;
      DOT category-variable </option(s)>;
      HBAR category-variable < /option(s) >;
      HBOX response-variable </option(s)>;
      HISTOGRAM response-variable < /option(s)>;
      NEEDLE X= variable Y= numeric-variable </option(s)>;
      REG X= numeric-variable Y= numeric-variable
                </option(s)>;
      SCATTER X= variable Y= variable </option(s)>;
      VBAR category-variable < /option(s)>;
      VBOX response-variable </option(s)>;
RUN;
```

Vertical Box-Plot with SGPLOT

```
proc sgplot data=bootcamp.ameshousing3;
     vbox SalePrice / ; *datalabel=Overall_Qual;
     refline 135000 / axis=y label;
     title "Box Plots of Sales Prices";
run;
```

LAB 1

Don’t forget to take the lab check on Moodle!

CONFIDENCE INTERVALS

Point Estimates

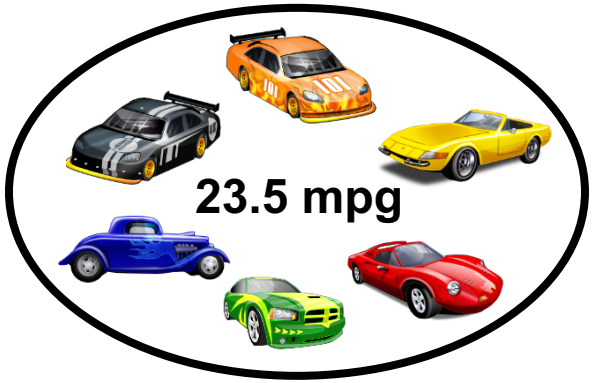
$\overline{x}$ estimates μ

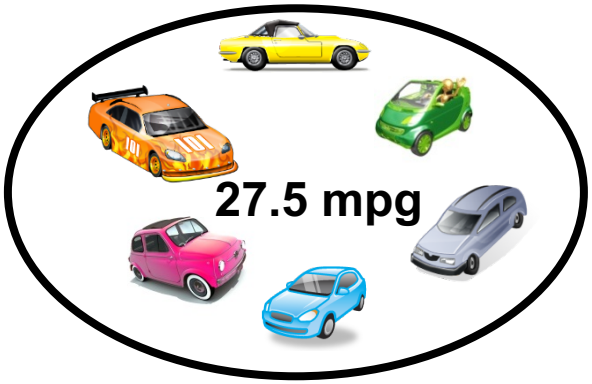
s estimates σ

Variability among Samples

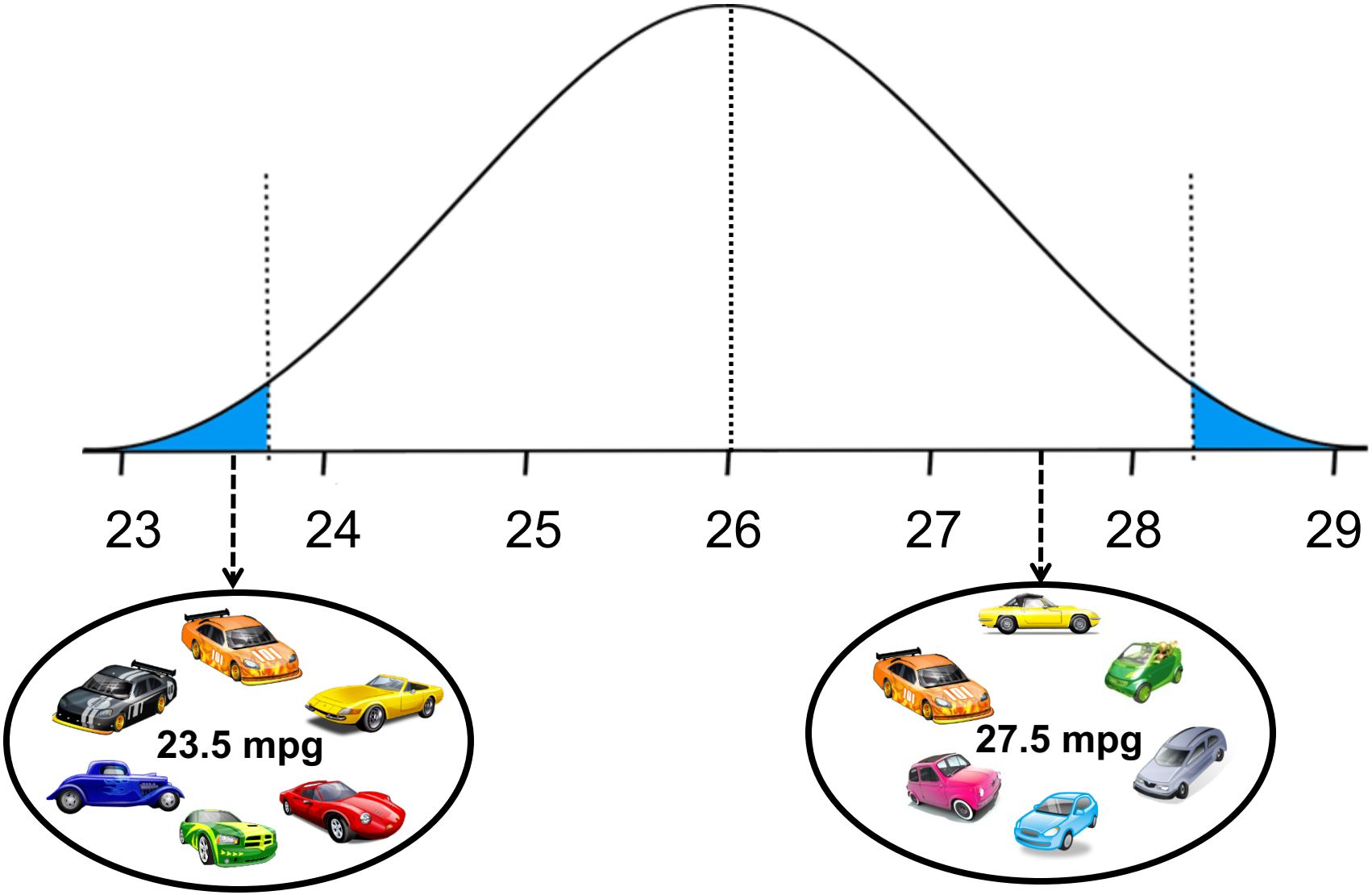
Average gas mileage of cars made after 2000?

Too many cars/options to measure. So we sample.



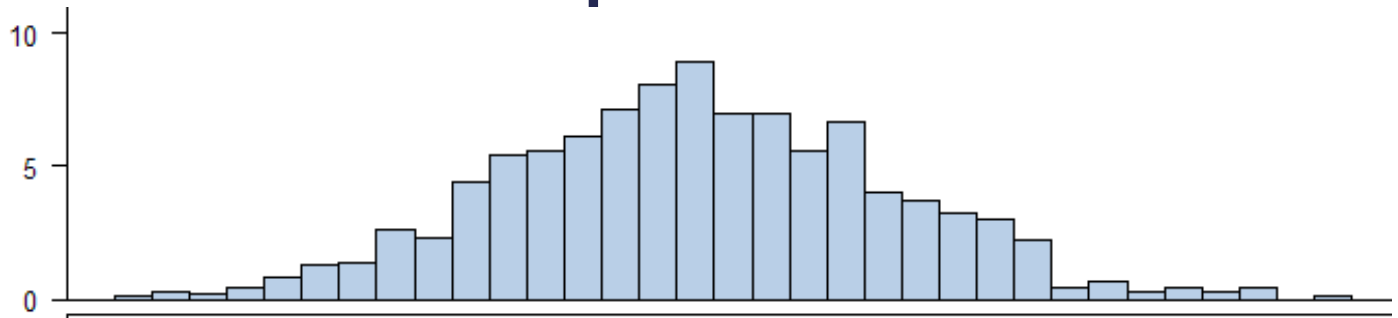


Distribution of Sample Means

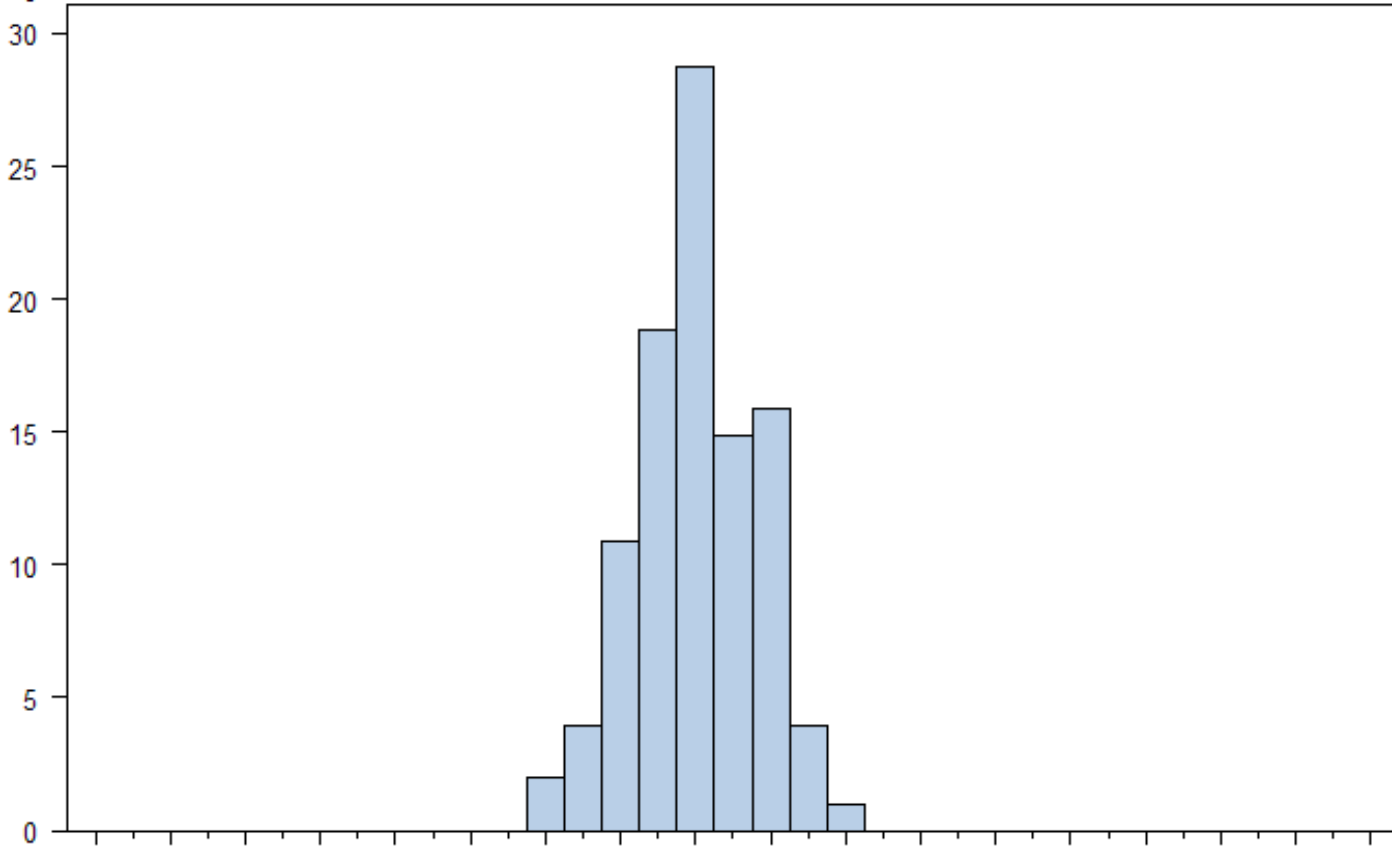


Distribution of Sample Means

Distribution of mpg

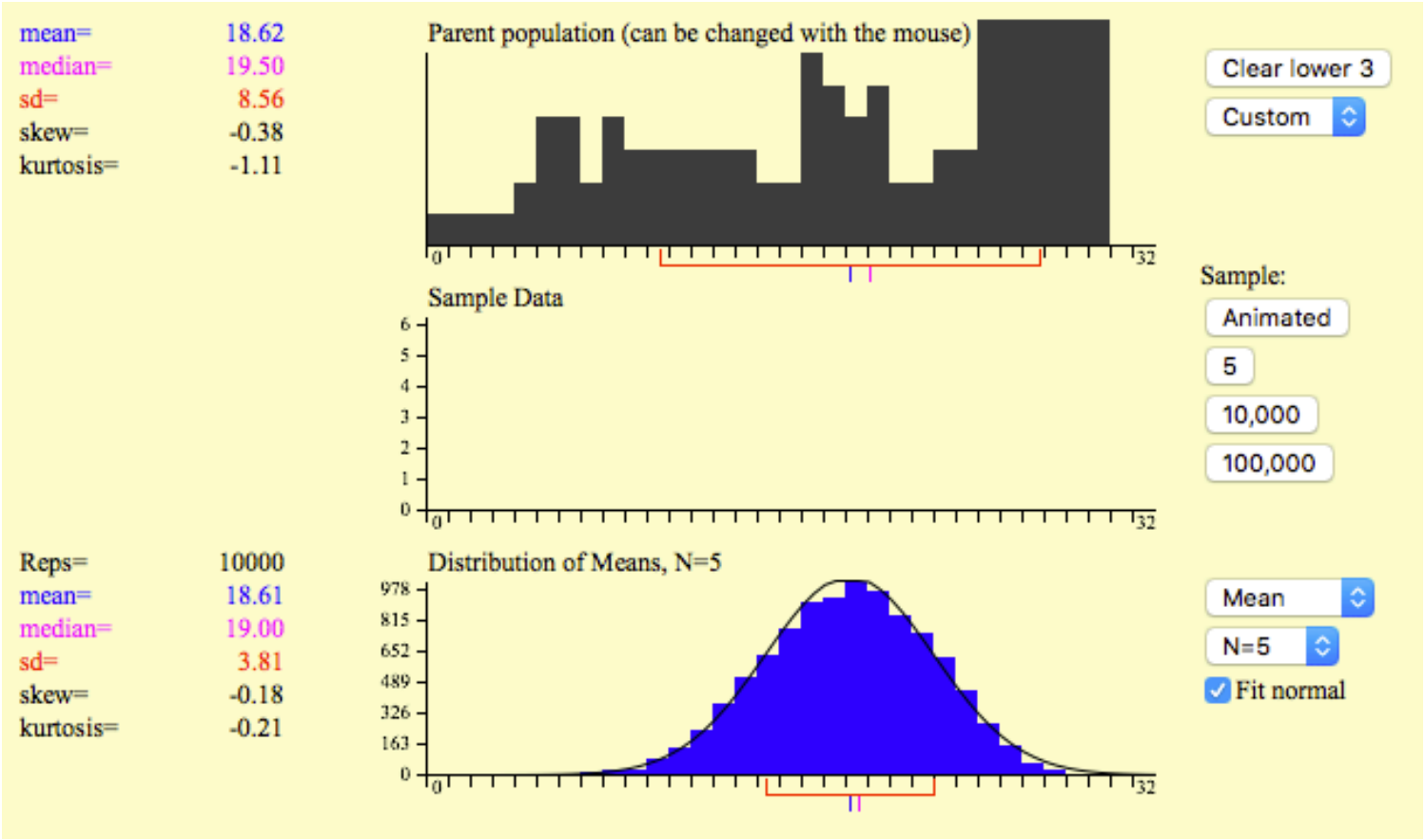


Distribution of sample means (n=10)



Interactive Demo

http://onlinestatbook.com/stat_sim/sampling_dist/



Normal Distribution for the Mean

Can assume the sample mean is normally distributed
if *EITHER:*

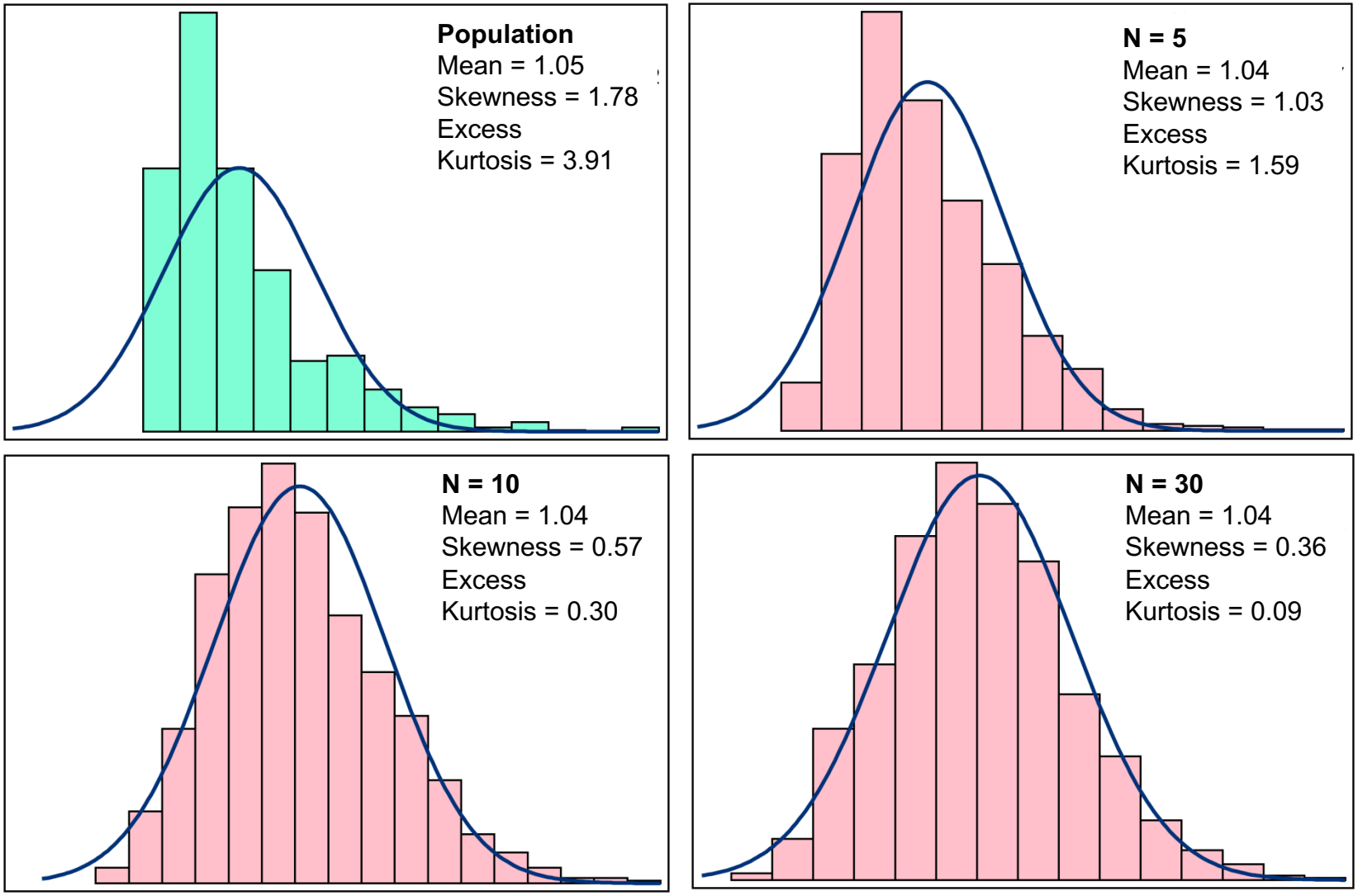
1. The population is normally distributed

   OR

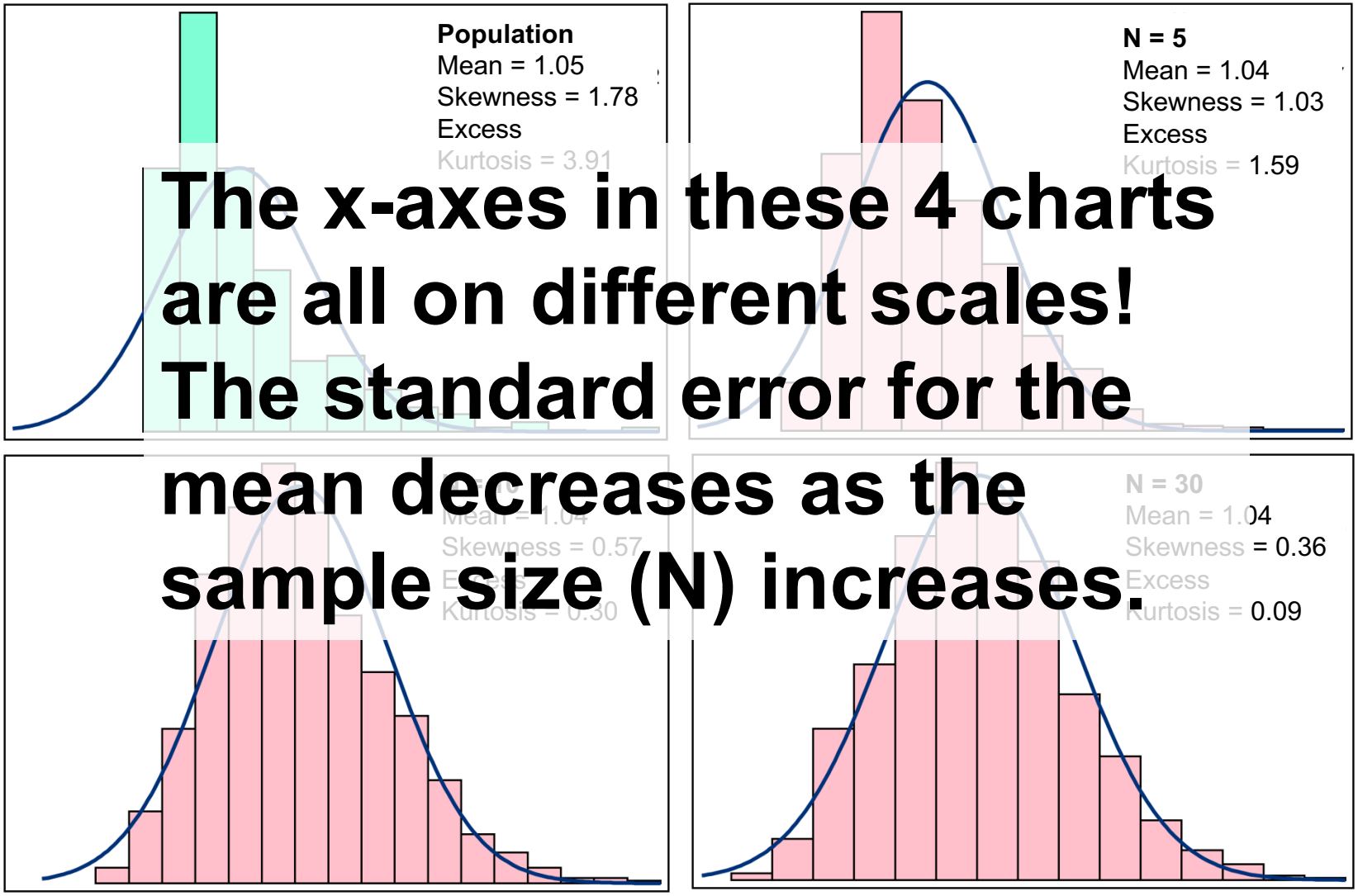
2. The sample size is large (>50 observations).

The Central Limit Theorem states that the distribution of sample means is approximately normal, regardless of the population distribution's shape, if the sample size is "large enough".

Central Limit Theorem, Illustrated

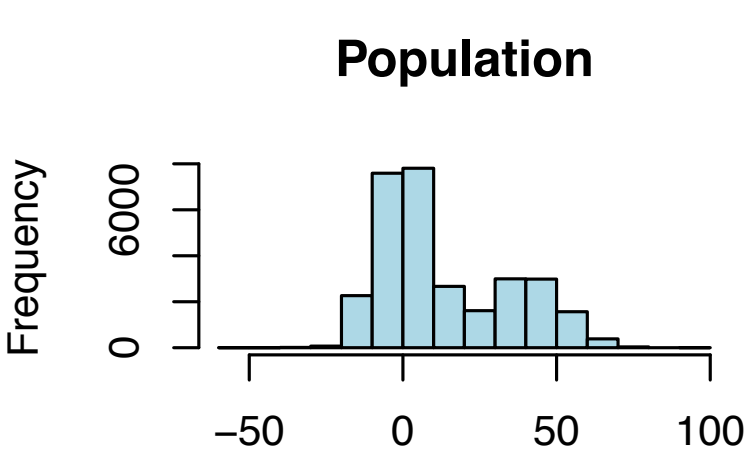


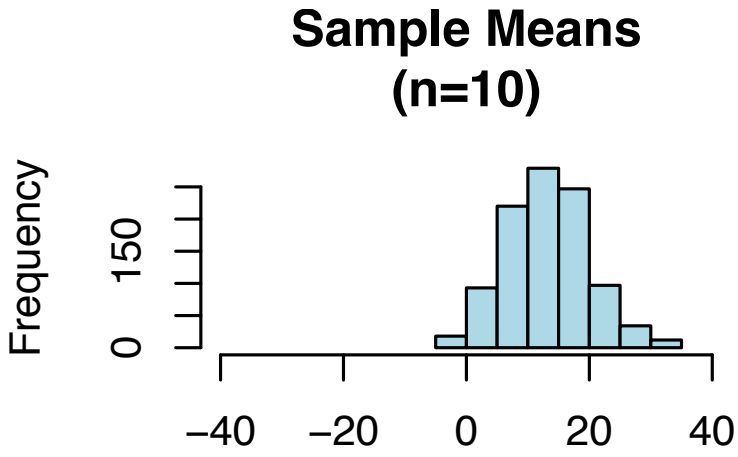
Central Limit Theorem, Illustrated

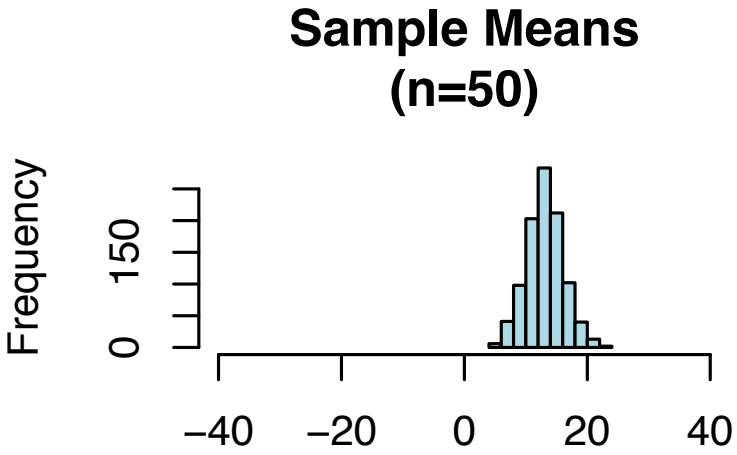


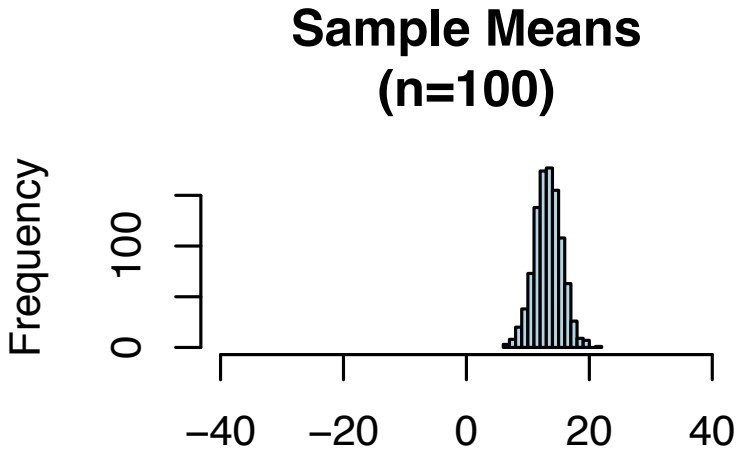
The x-axes in these 4 charts are all on different scales! The standard error for the mean decreases as the sample size (N) increases.

Central Limit Theorem, Re-Illustrated









Standard Error of the Mean

- *Standard Error* measures the variability of your estimate.

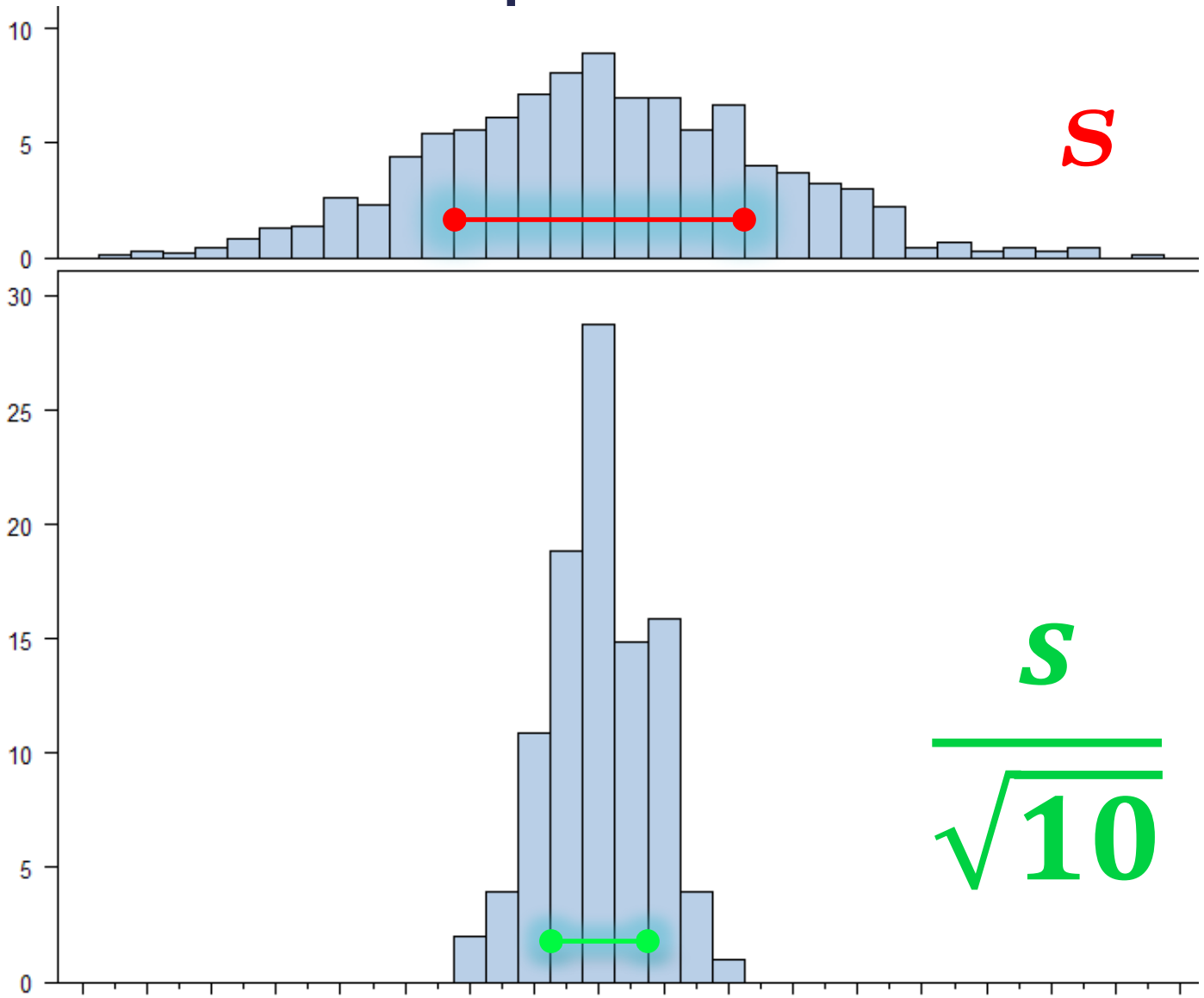
- It differs from the sample standard deviation:
  - *Sample standard deviation* is a measure of the variability of **data**
  - *Standard error of the mean* is a measure of the estimated variability of **sample means**

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

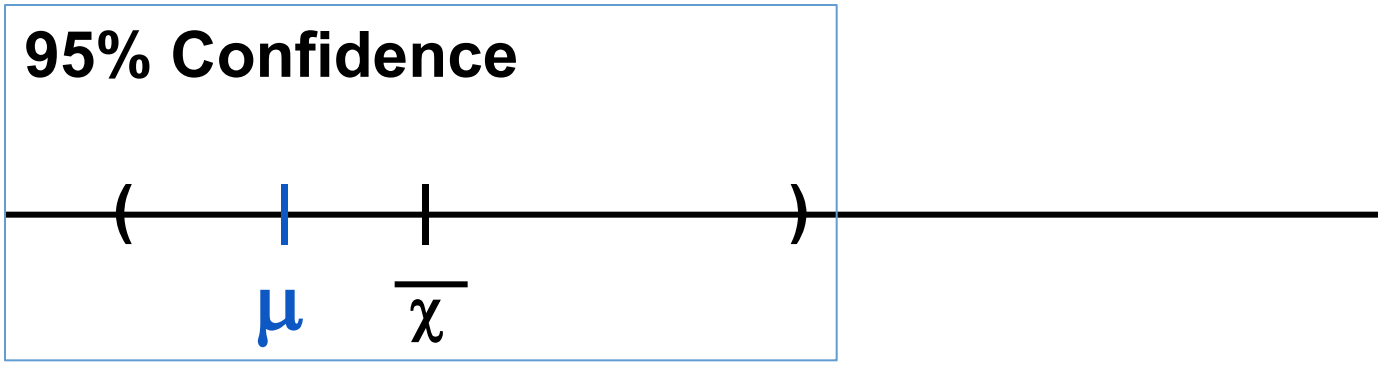
Distribution of Sample Means

Distribution of mpg

Distribution of sample means (n=10)



Confidence Intervals



- A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
  - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

Confidence Interval for the Mean

"margin of error"

$$\bar{x} \pm t \cdot s_{\bar{x}} \quad \text{or} \quad (\bar{x} - t \cdot s_{\bar{x}},\ \bar{x} + t \cdot s_{\bar{x}})$$

where

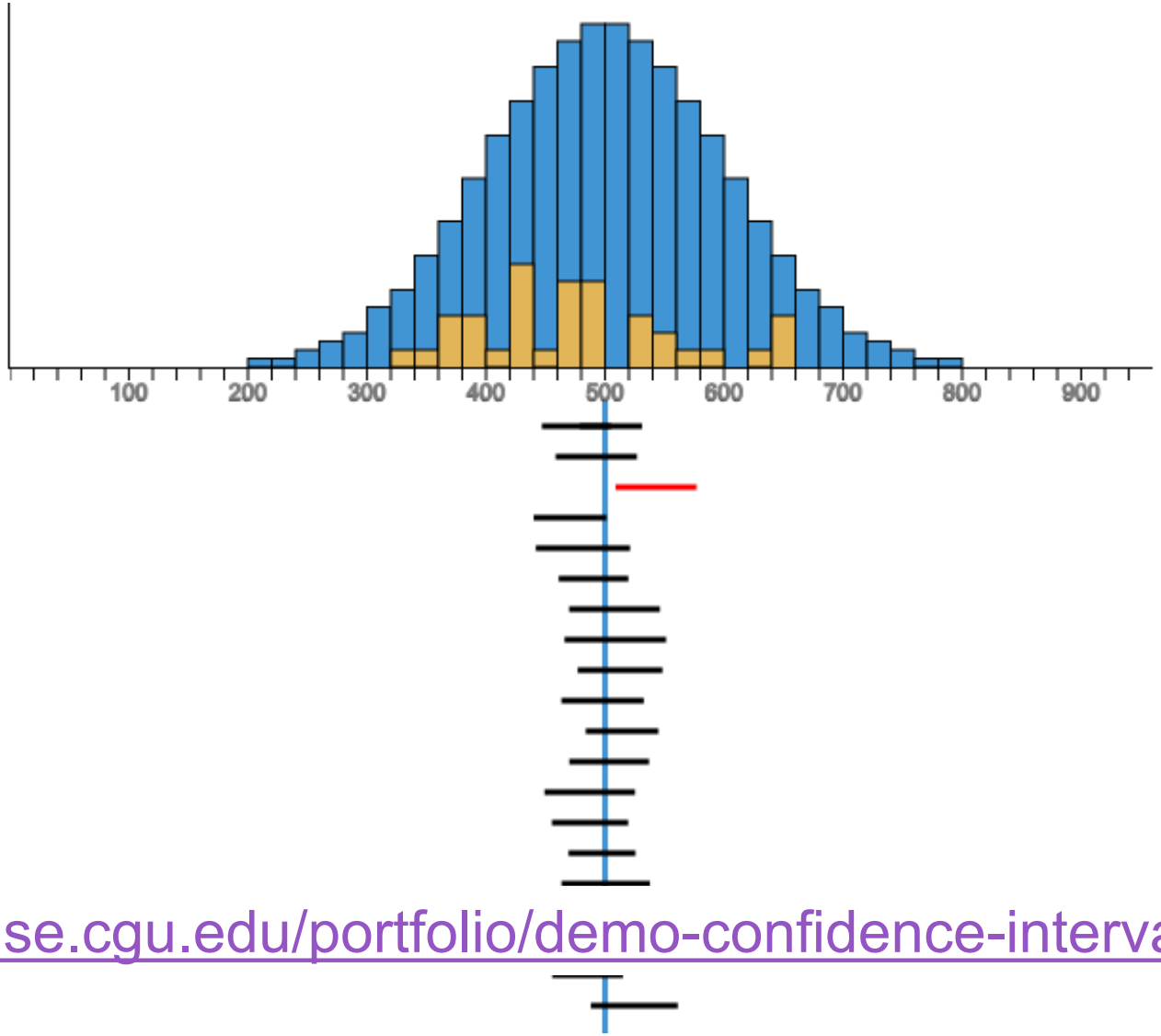
$\bar{x}$ is the sample mean.

t is the *t* value corresponding to the confidence level and *n*-1 degrees of freedom, where *n* is the sample size.

$s_{\bar{x}}$ is the standard error of the mean.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Confidence Intervals Sometimes Miss



http://wise.cgu.edu/portfolio/demo-confidence-interval-creation/

Confidence Interval for the Mean

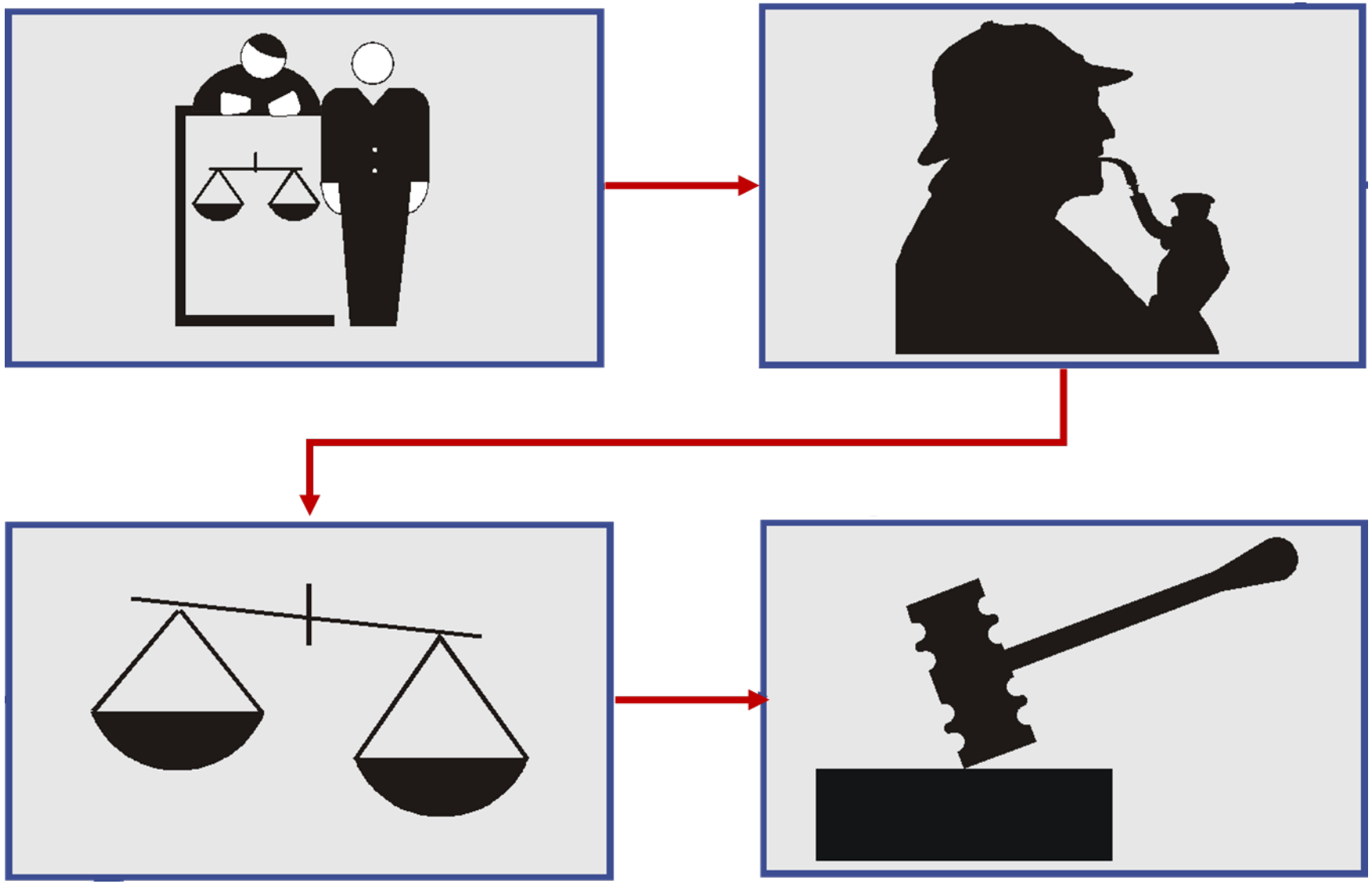
```
proc means data=bootcamp.ameshousing3 maxdec=2
           n mean std stderr clm;
    var SalePrice;
    title '95% Confidence Interval for Sales Price';
run;
```

Poll

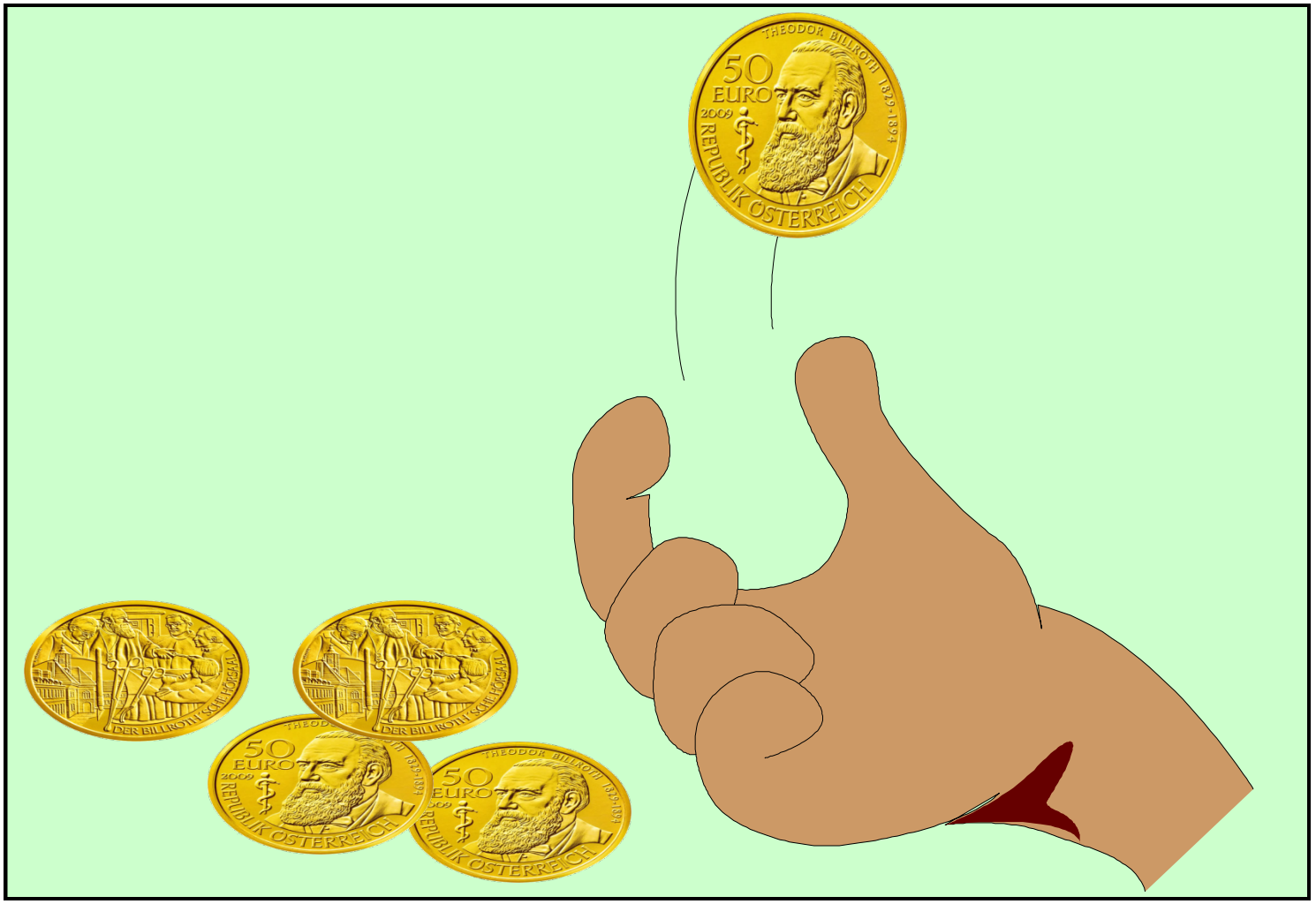
Quiz

HYPOTHESIS TESTING

Judicial Analogy



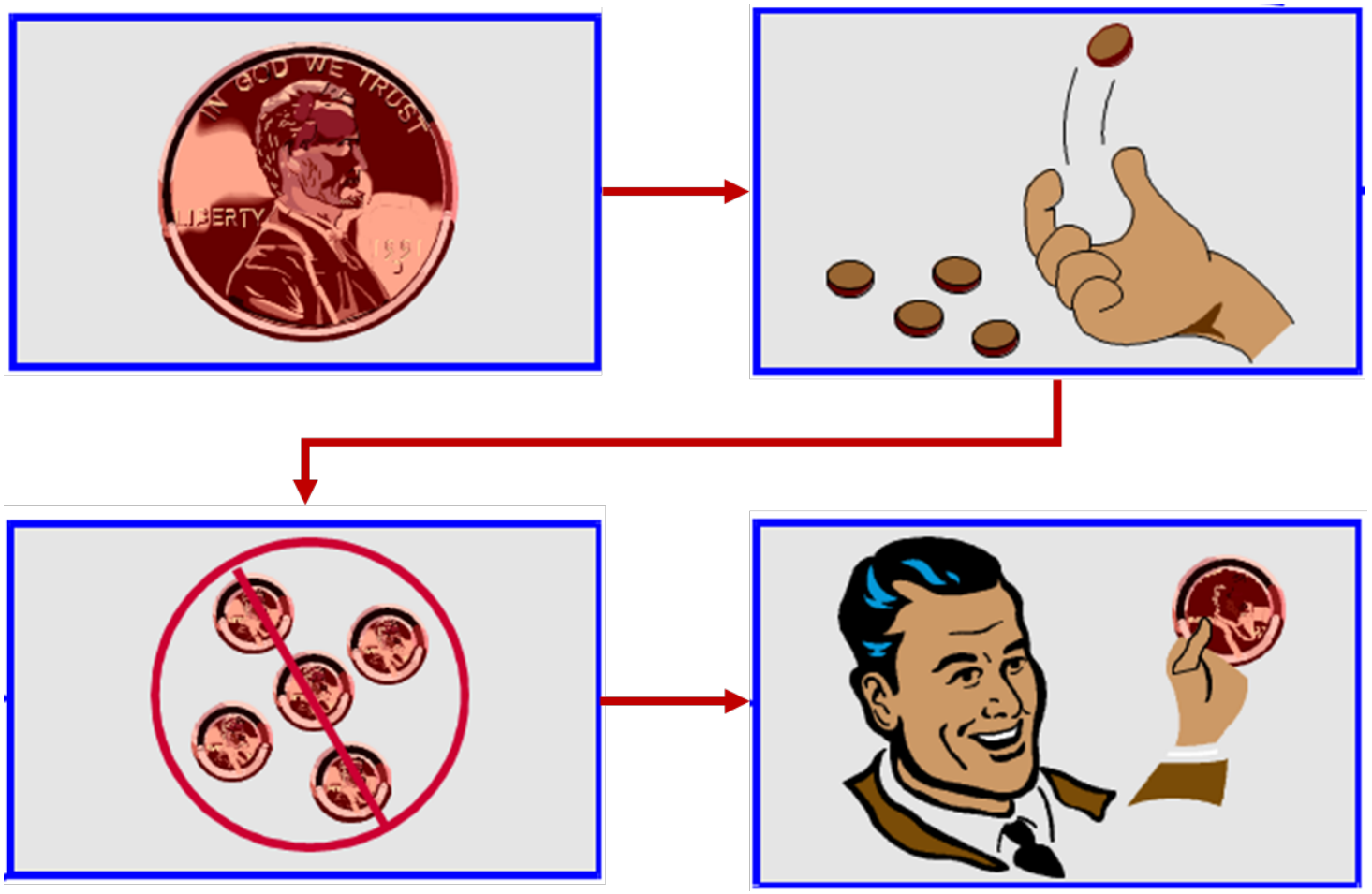
Coin Example



Poll

Quiz

Coin Analogy



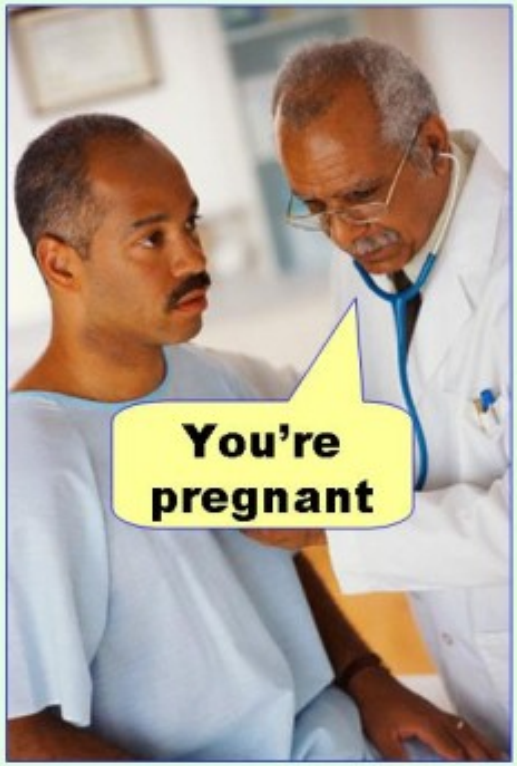
Types of Errors

- You used a decision rule to make a decision, but was the decision correct?

ACTUAL / DECISION	H_0 Is True	H_0 Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

Null – Not Pregnant

Type I error
(false positive)

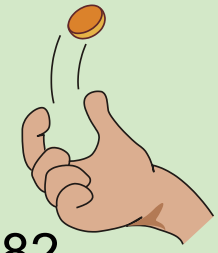


Type II error
(false negative)

Coin Experiment – Effect Size Influence

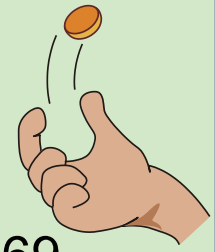
- Flip a coin 100 times and decide whether it is fair.

55 Heads
45 Tails



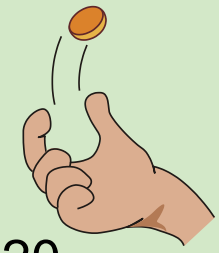
p-value=.3682

40 Heads
60 Tails



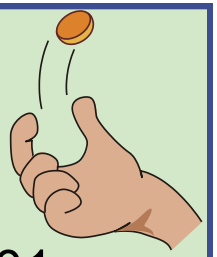
p-value=.0569

37 Heads
63 Tails



p-value=.0120

15 Heads
85 Tails

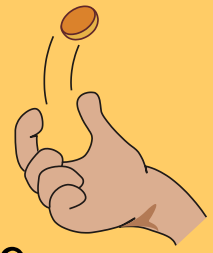


p-value<.0001

Coin Experiment – Sample Size Influence

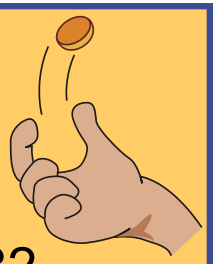
- Flip a coin and get 40% heads and decide whether it is fair.

4 Heads
6 Tails



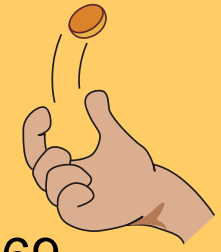
p-value=.7539

16 Heads
24 Tails



p-value=.2682

40 Heads
60 Tails



p-value=.0569

160 Heads
240 Tails

p-value<.0001

Comparing α and the *p*-Value

- In general, you do one of the following:
  - REJECT the null hypothesis if *p*-value $\leq \alpha$
  - FAIL TO REJECT the null hypothesis if *p*-value $> \alpha$.

Poll

Quiz

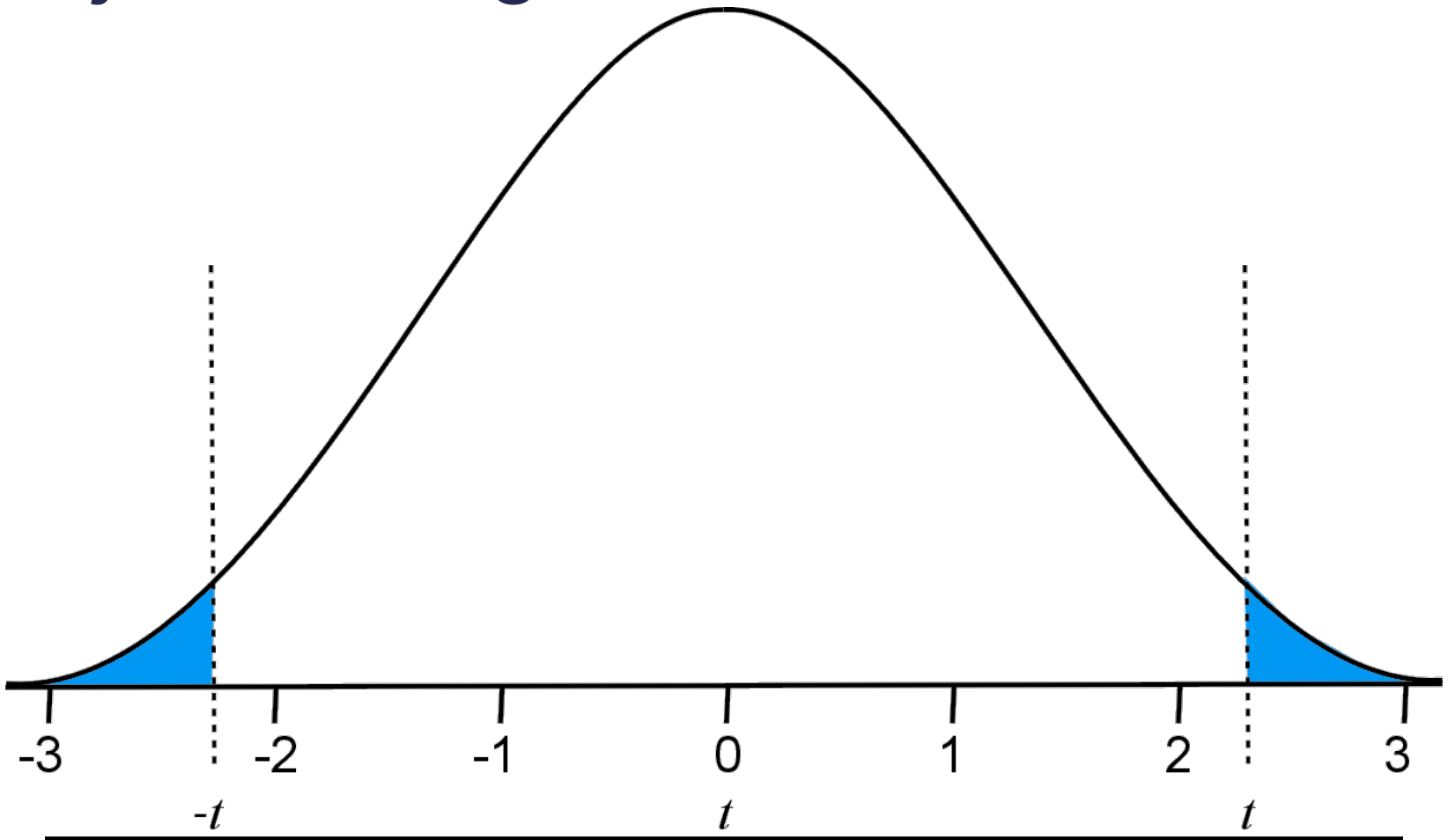
Performing a Hypothesis Test

- To test the null hypothesis H_0: $\mu = \mu_0$, SAS software calculates the *Student's t* statistic value:

$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

- The null hypothesis is rejected when the calculated value is more extreme (either positive or negative) than would be expected by chance ***if H_0 were true***.

Rejection Region for Two-Sided Test



The *t* statistic can be positive or negative.

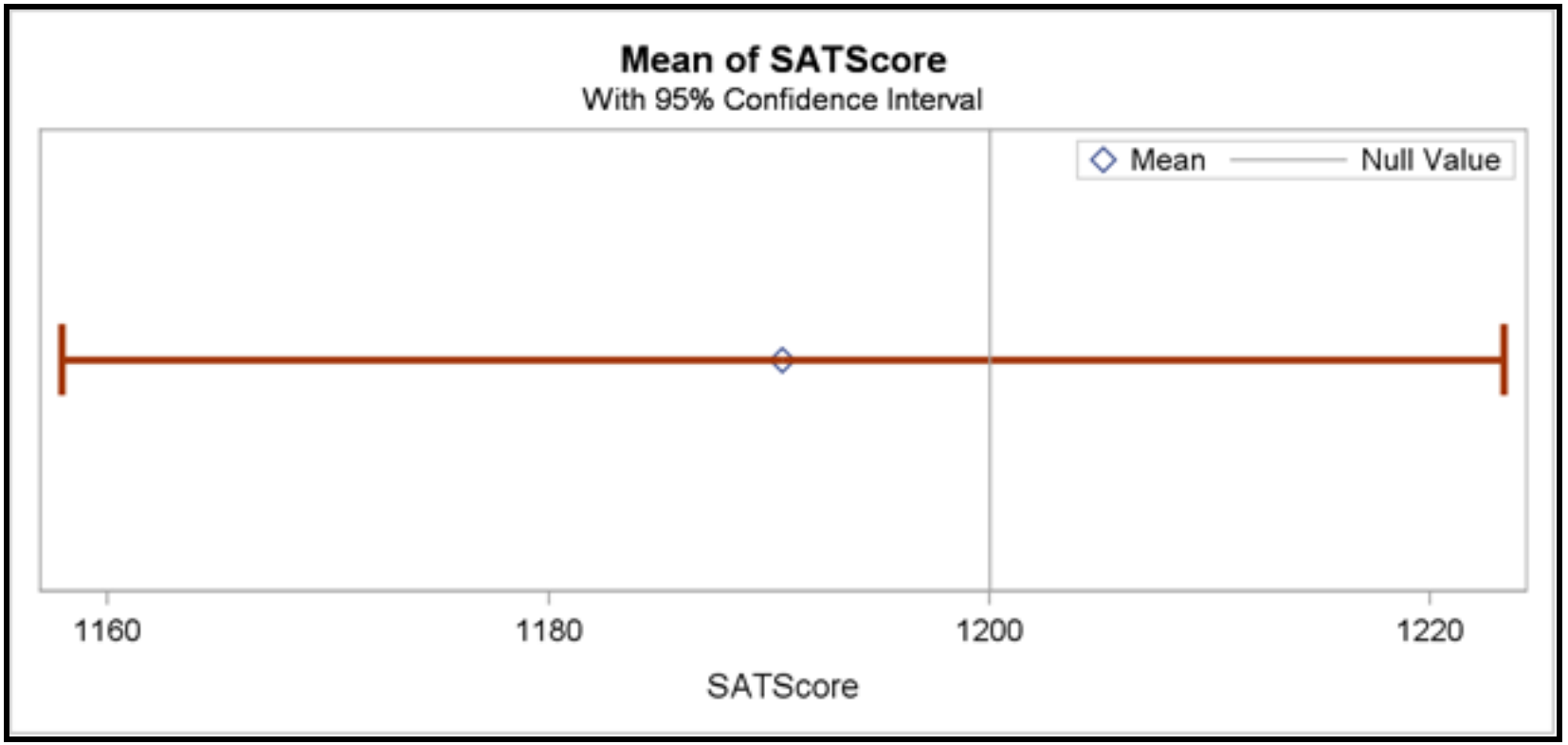
Hypothesis Test

```
proc univariate data=bootcamp.ameshousing3 mu0=135000;
    var SalePrice;
    title 'Testing Whether the Mean of Sales Price = $135K';
run;
```

Hypothesis Test (Another Way)

```
proc ttest data=bootcamp.ameshousing3
           plots(shownull)=interval
           H0=135000;
    var SalePrice;
    title "One-Sample t-Test Testing Mean"
          " SalePrice=$135K";
run;
```

Confidence Interval Plots



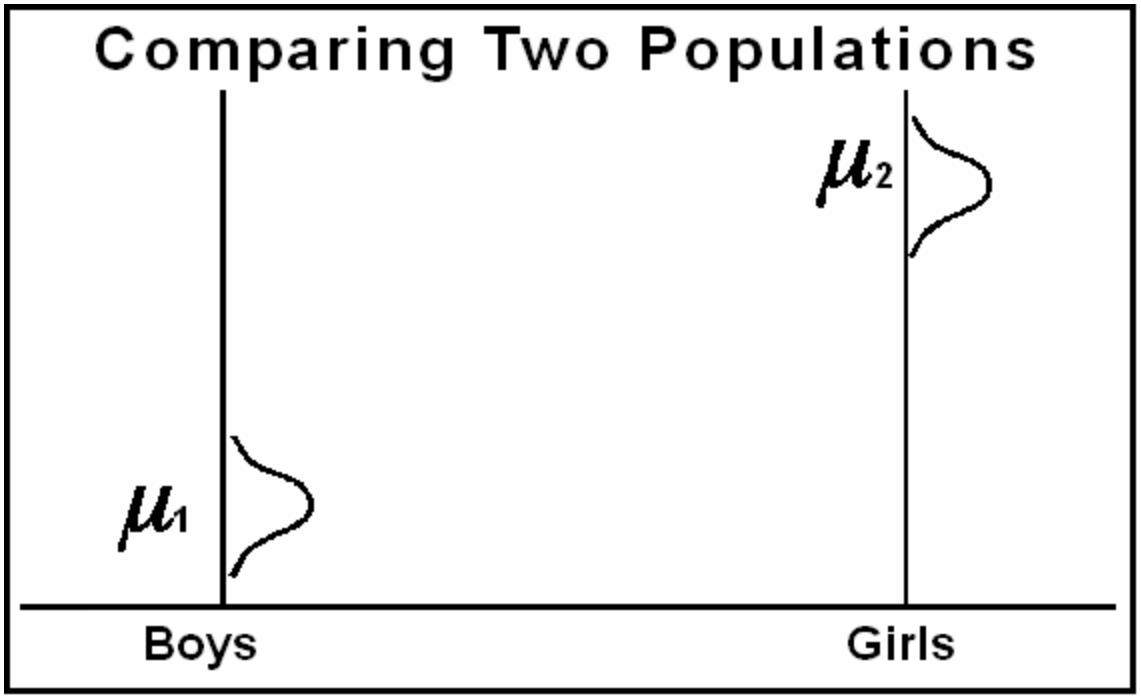
Poll

Quiz

TWO-SAMPLE *t* TESTS

Assumptions

- Independent observations
- Normally distributed data for each group
- Equal variances for each group



F Test for Equality of Variances

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{\max(s_1^2,\ s_2^2)}{\min(s_1^2,\ s_2^2)}$$

Equal Variance *t* Test and *p*-Values

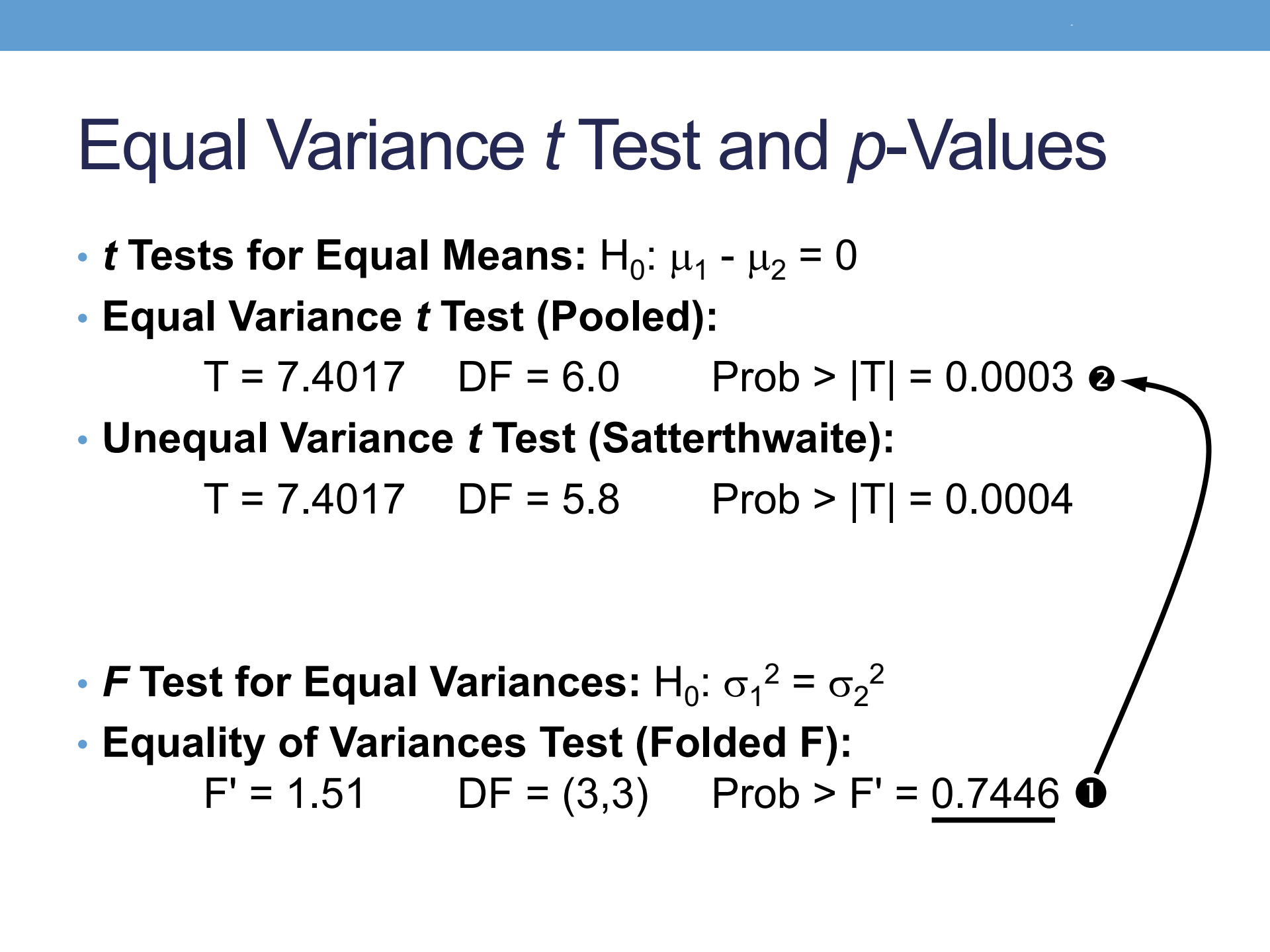
- ***t* Tests for Equal Means:** H_0: $\mu_1 - \mu_2 = 0$
- **Equal Variance *t* Test (Pooled):**

  T = 7.4017 DF = 6.0 Prob > |T| = 0.0003 ❷
- **Unequal Variance *t* Test (Satterthwaite):**

  T = 7.4017 DF = 5.8 Prob > |T| = 0.0004

- ***F* Test for Equal Variances:** H_0: $\sigma_1^2 = \sigma_2^2$
- **Equality of Variances Test (Folded F):**

  F' = 1.51 DF = (3,3) Prob > F' = 0.7446 ❶



Unequal Variance *t* Test and *p*-Values

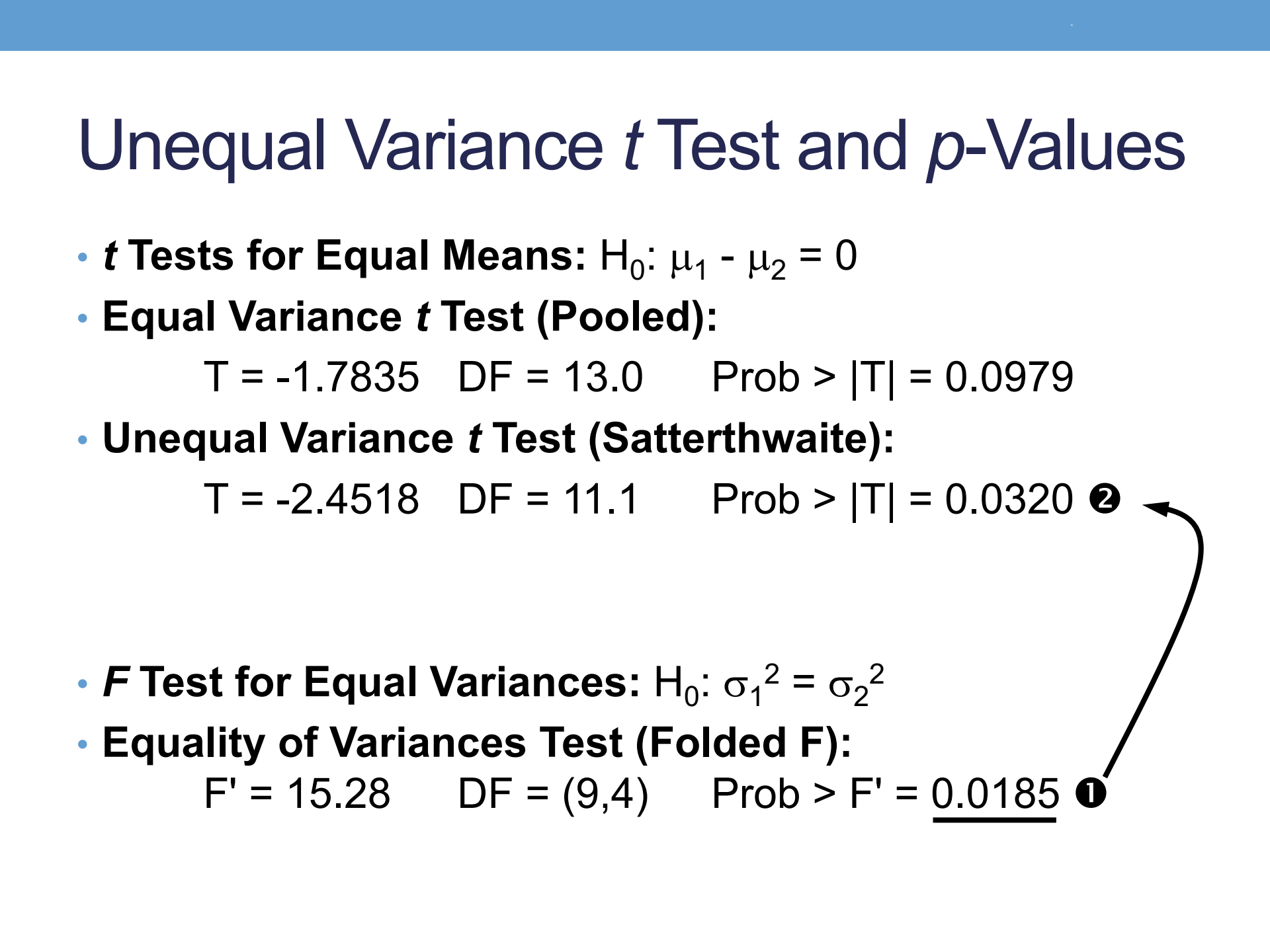
- ***t* Tests for Equal Means:** H_0: $\mu_1 - \mu_2 = 0$
- **Equal Variance *t* Test (Pooled):**

  T = -1.7835 DF = 13.0 Prob > |T| = 0.0979
- **Unequal Variance *t* Test (Satterthwaite):**

  T = -2.4518 DF = 11.1 Prob > |T| = 0.0320 ❷

- ***F* Test for Equal Variances:** H_0: $\sigma_1^2 = \sigma_2^2$
- **Equality of Variances Test (Folded F):**

  F' = 15.28 DF = (9,4) Prob > F' = 0.0185 ❶



Two-Sample *t*-Test

(Text Books)

```
proc ttest data=bootcamp.ameshousing3 plots(shownull)=interval;
     class central_air;
     var saleprice;
     title "Two-Sample t-Test Comparing Price of
               Houses With and Without Central Air";
run;
```

Two-Sample *t*-Test

(Reality)

```
proc ttest data=bootcamp.ameshousing3 plots(shownull)=interval;
     class central_air;
     var year_built;
     title "Two-Sample t-Test Comparing Age of
               Houses With and Without Central Air";
run;
```

LAB 2

Don’t forget to take the lab check on Moodle!