UTHM
Universiti Tun Hussein Onn Malaysia
1993-2023
AMANAH • PROFESIONAL • INOVATIF

BIT34503 Data Science

CHAPTER 5: EXPLORATARY DATA ANALYSIS (EDA)

5. EXPLORATARY DATA ANALYSIS (EDA)

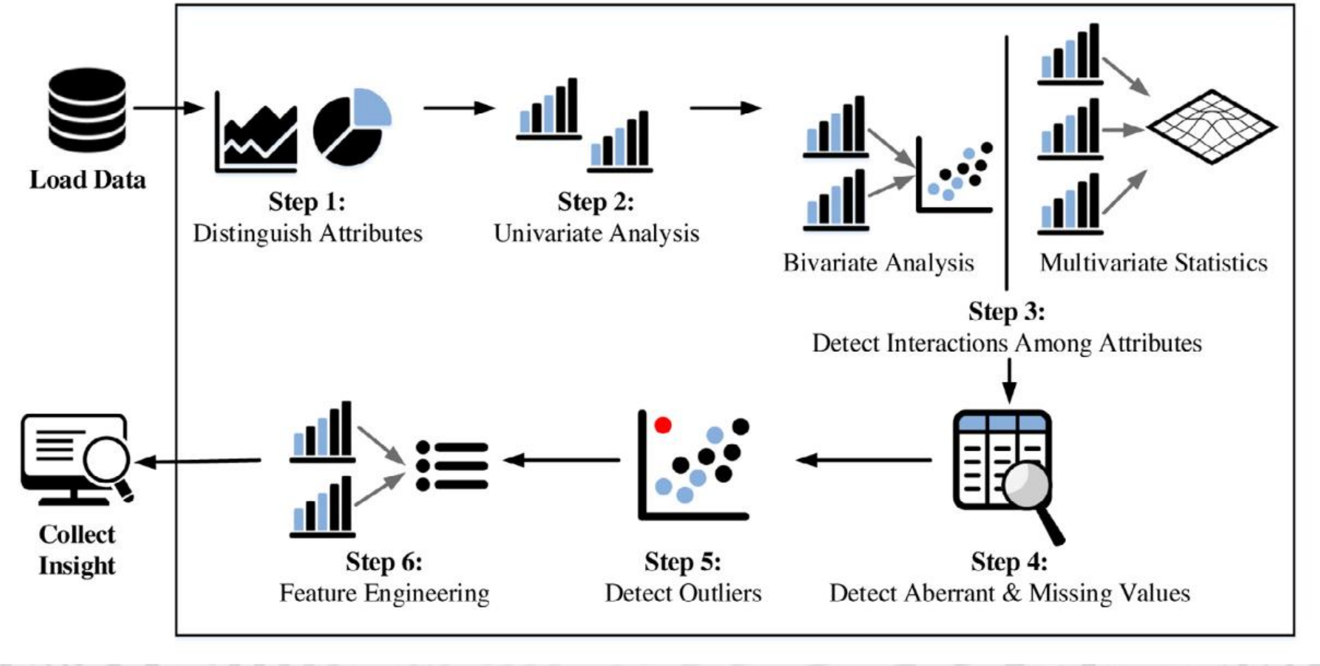
5.1 Goals of EDA

5.2 The role of graphics

5.3 Handling outliers

5.4 Dimension reduction

- Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.



5.1 Goals of EDA

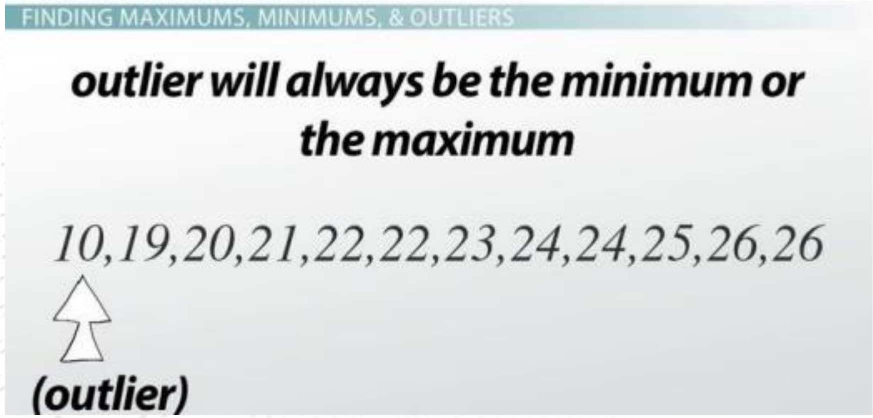
- to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as: a good-fitting, parsimonious model. a list of outliers.

5.2 The role of graphics

- Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations

5.3 Handling outliers

- Set up a filter in your testing tool. Even though this has a little cost, filtering out outliers is worth it. ...
- Remove or change outliers during post-test analysis. ...
- Change the value of outliers. ...
- Consider the underlying distribution. ...
- Consider the value of mild outliers.

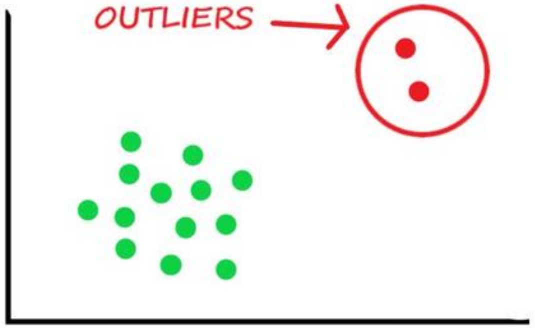


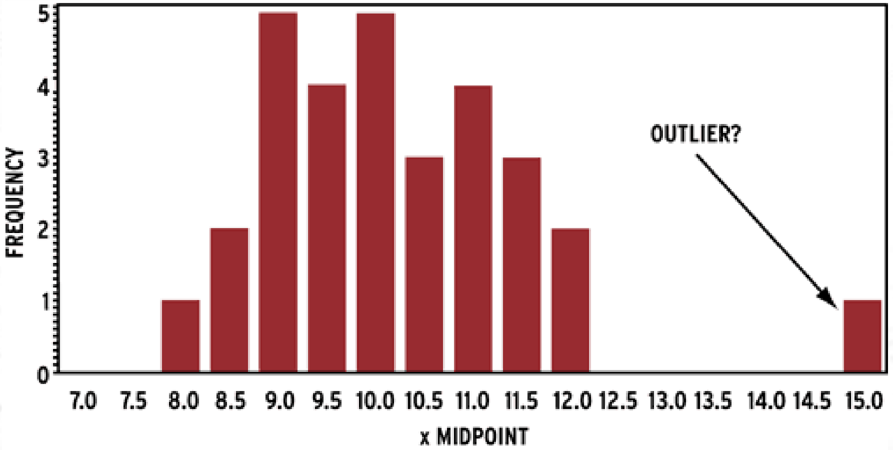
FINDING MAXIMUMS, MINIMUMS, & OUTLIERS

outlier will always be the minimum or the maximum

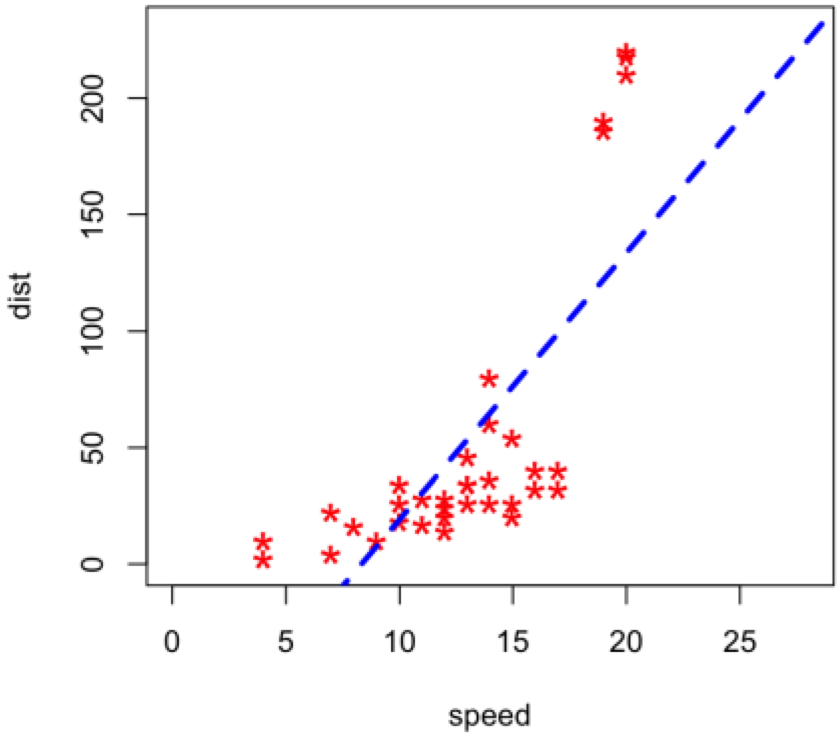
10,19,20,21,22,22,23,24,24,25,26,26

(outlier)

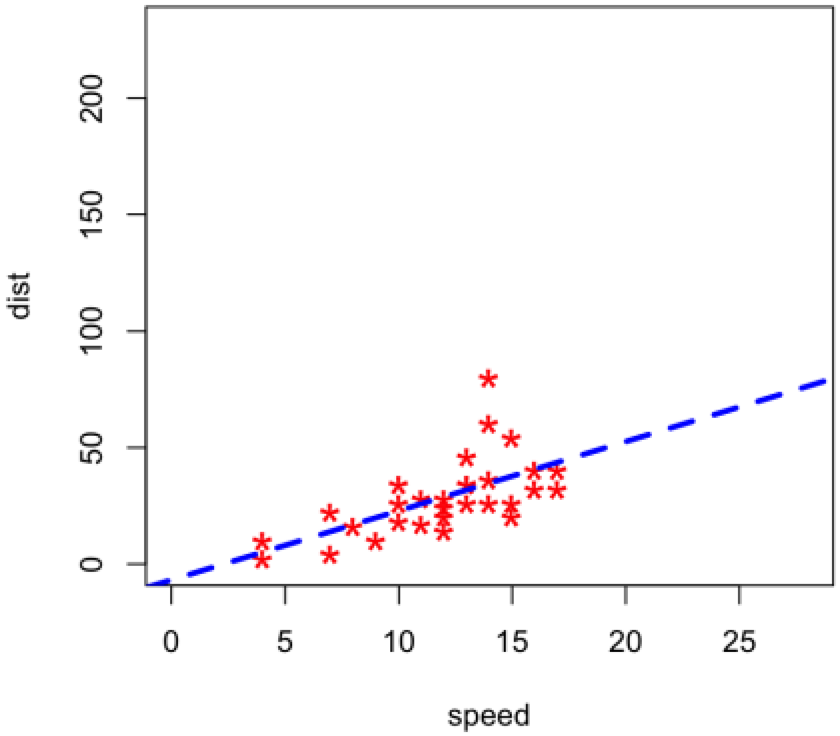




With Outliers

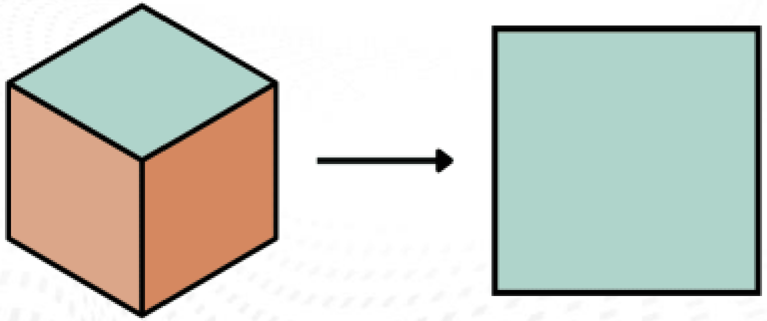


Outliers removed
A much better fit!



5.4 Dimension reduction

- Dimensionality reduction simply refers to the process of reducing the number of attributes in a dataset while keeping as much of the variation in the original dataset as possible. It is a data preprocessing step meaning that we perform dimensionality reduction before training the model.



Components of Dimensionality Reduction

- There are two components of dimensionality reduction:
  a. Feature selection
     - In this, we need to find a subset of the original set of variables. Also, need a subset which we use to model the problem.

  b. Feature Extraction
     - We use this, to reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Advantages of Dimensionality Reduction

- Avoiding overfitting.
- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.
- Improves model accuracy with less misleading data.
- Use less computing with lesser dimensions and with less data algorithms gets trained faster.

Lesson 11:
Introduction to Basic Statistics

11.1 Introduction to Statistics

11.2 Viewing Distributions

11.3 Hypothesis Testing

Objectives

Basics

- Define the basic concepts of statistics.
- Describe data with simple statistics.

Distributions

- Look at distributions of continuous variables.
- Describe the normal distribution.
- Use and interpret a histogram and a box plot.

Hypothesis testing

- Define some common terminology related to hypothesis testing.
- Explain *p*-value.

Lesson 11:
Introduction to Basic Statistics

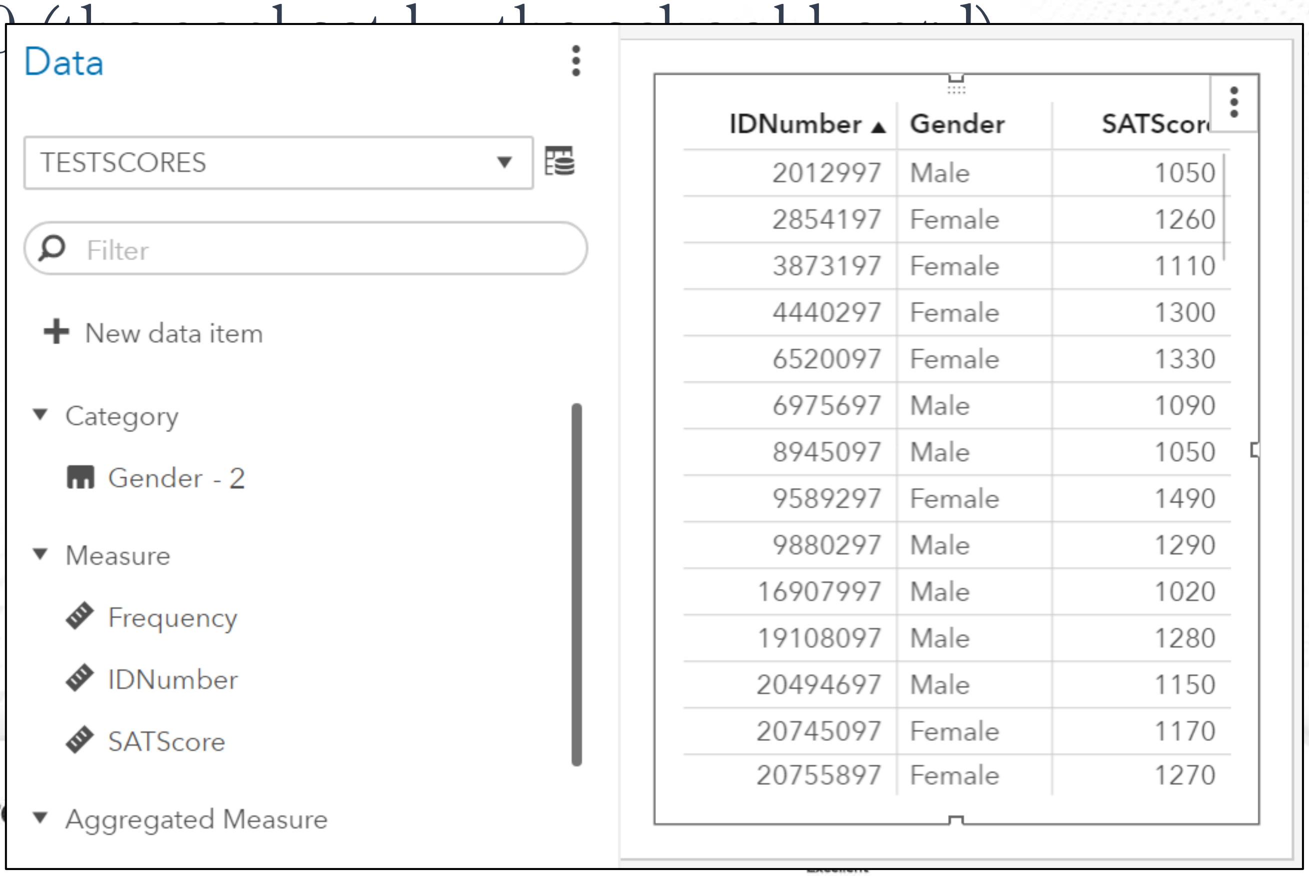
1.1 Introduction to Statistics

1.2 Viewing Distributions

1.3 Hypothesis Testing

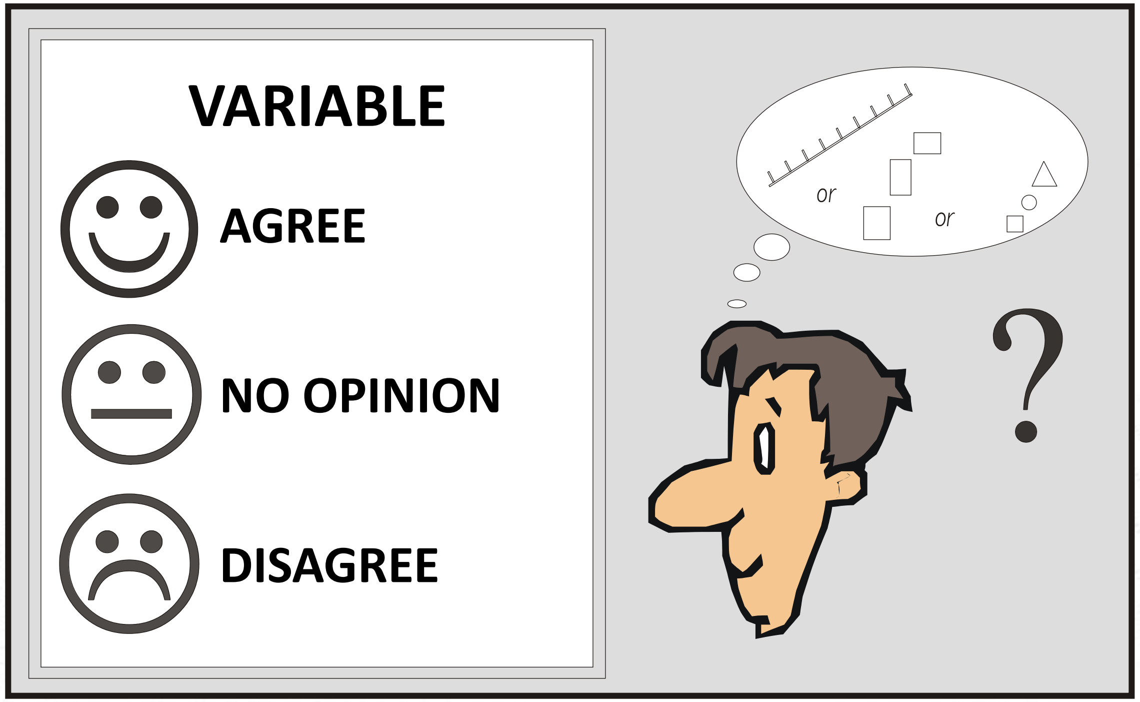
Defining the Problem: SATScore = 1200?

- The purpose of the study is to determine whether the average combined Math and Verbal scores on the Scholastic Aptitude Test (SAT) at Carver County magnet high schools is 1200



IDNumber ▲	Gender	SATScor
2012997	Male	1050
2854197	Female	1260
3873197	Female	1110
4440297	Female	1300
6520097	Female	1330
6975697	Male	1090
8945097	Male	1050
9589297	Female	1490
9880297	Male	1290
16907997	Male	1020
19108097	Male	1280
20494697	Male	1150
20745097	Female	1170
20755897	Female	1270

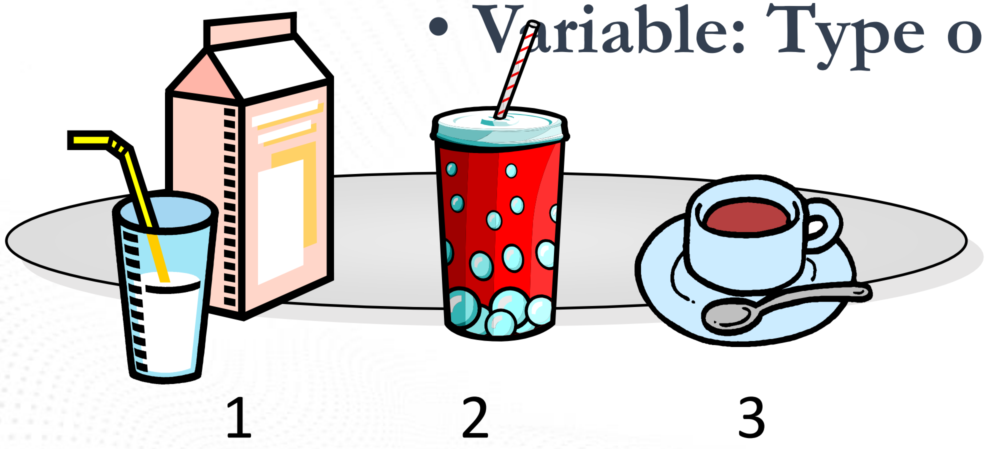
Identifying the Scale of Measurement



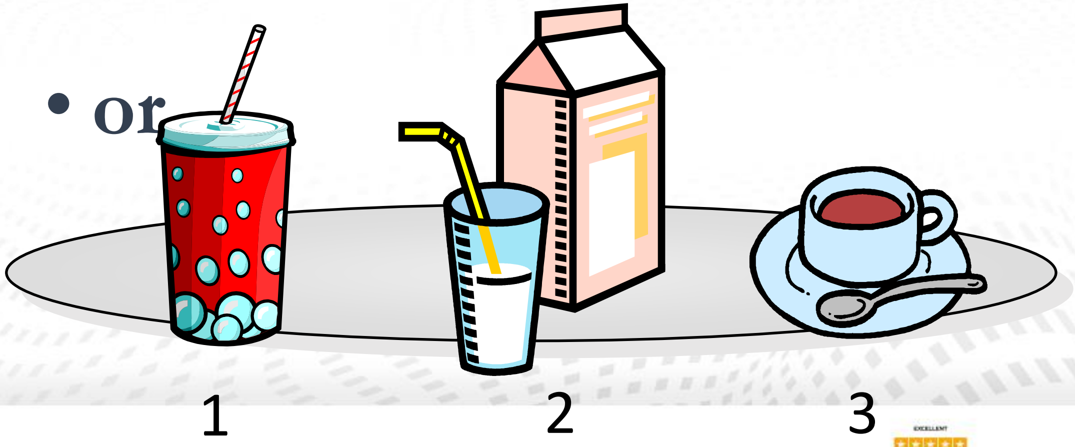
- Before analyzing the data, identify the measurement scale for each variable (continuous, nominal, or ordinal).

Nominal Variables

- Variable: Type of Beverage



- or



Ordinal Variables

Variable: Size of Beverage

Small

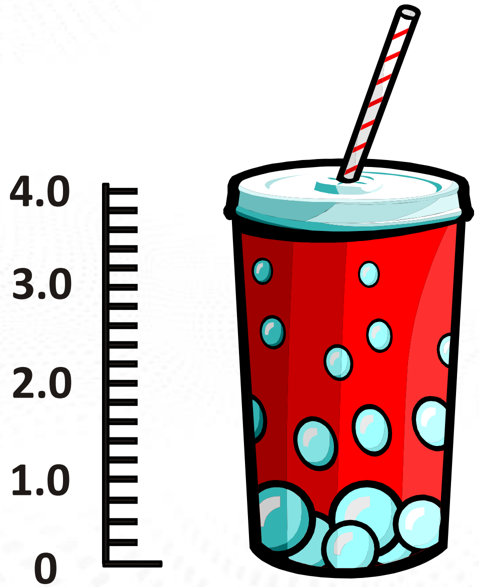
Medium

Large

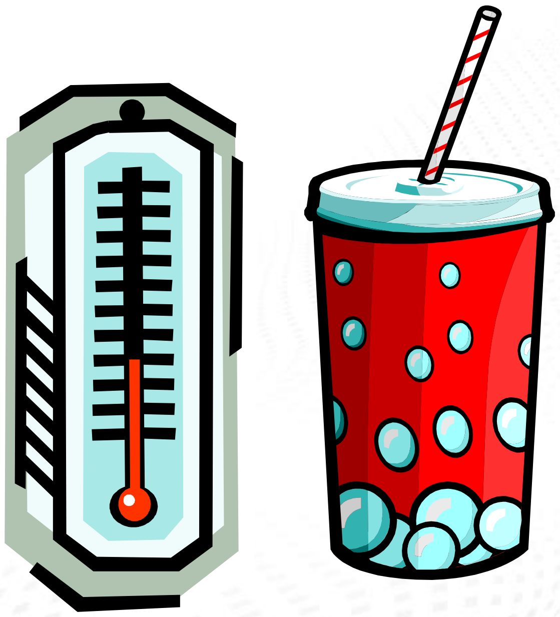
Continuous Variables

Variable:
Volume of Beverage

Variable:
Temperature of Beverage

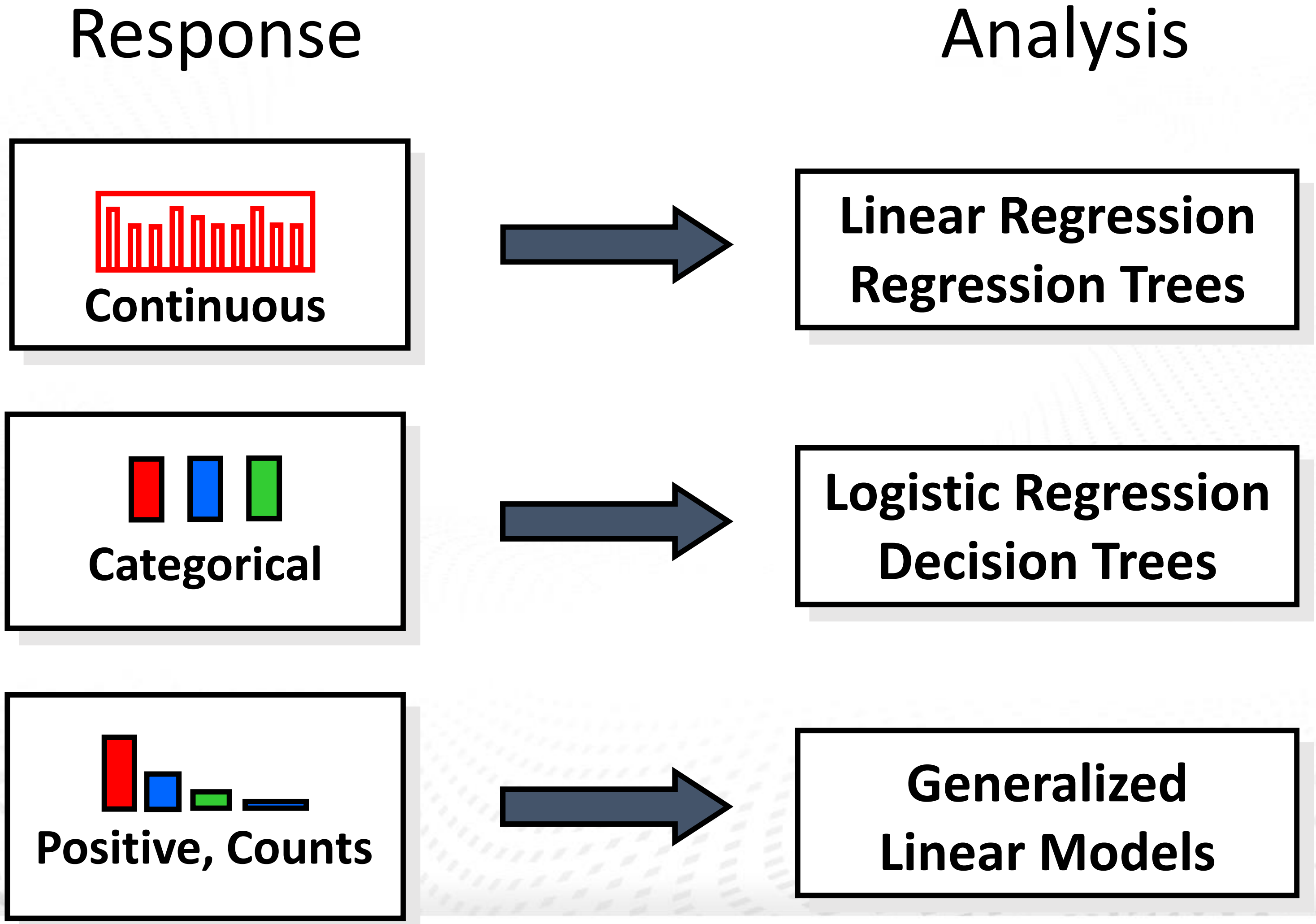


Ratio Level



Interval Level

Overview of Statistical Models



Response	Analysis
Continuous	Linear Regression Regression Trees
Categorical	Logistic Regression Decision Trees
Positive, Counts	Generalized Linear Models

Populations and Samples

Population – the entire collection of individual members of a group of interest

Sample – a subset of a population that is drawn to enable inferences to the population

- Assumption for this course: The sample drawn is representative of the population.

Parameters and Statistics

- Statistics are used to approximate population parameters.

	Population Parameters	Sample Statistics
Mean	μ	x
Variance	σ^2	s^2
Standard Deviation	σ	s

Descriptive Statistics

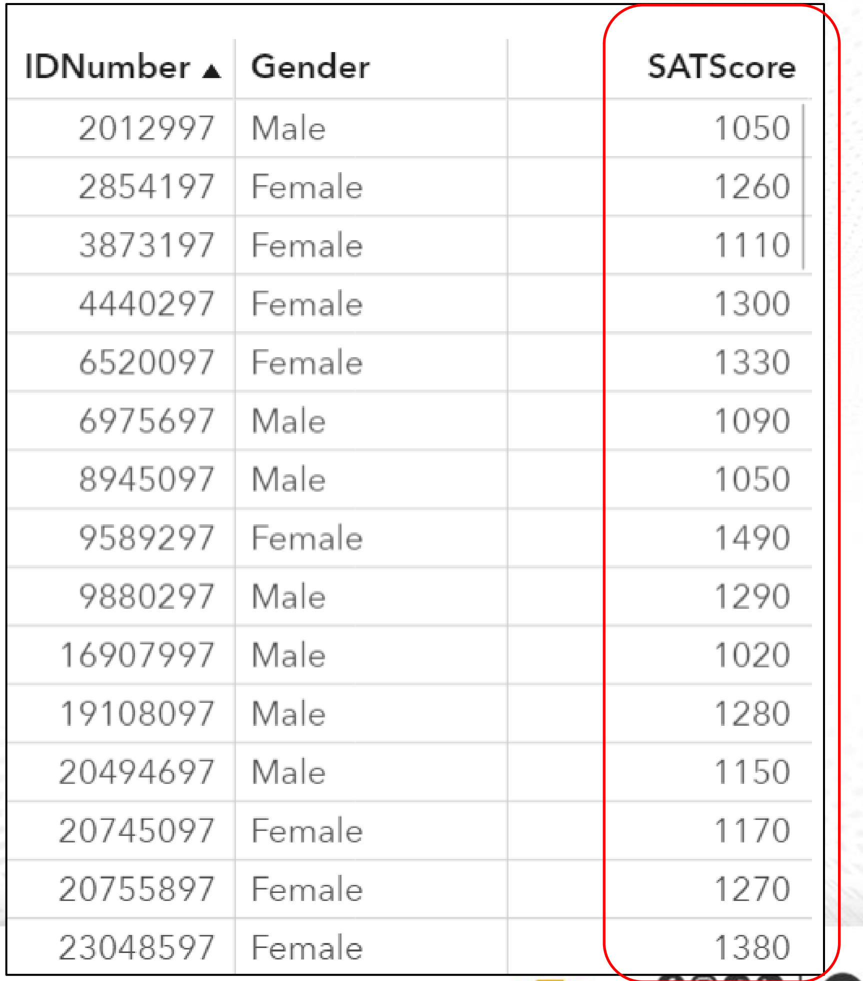
- The goal when you are describing data is to
  - screen for unusual sample data values
  - inspect the spread and shape of continuous variables
  - characterize the central tendency of the sample.

- Inferential Statistics

- The goal for statistical inference is to
  - estimate or predict unknown parameter values from a population, using a sample
  - make probabilistic statements about population attributes.

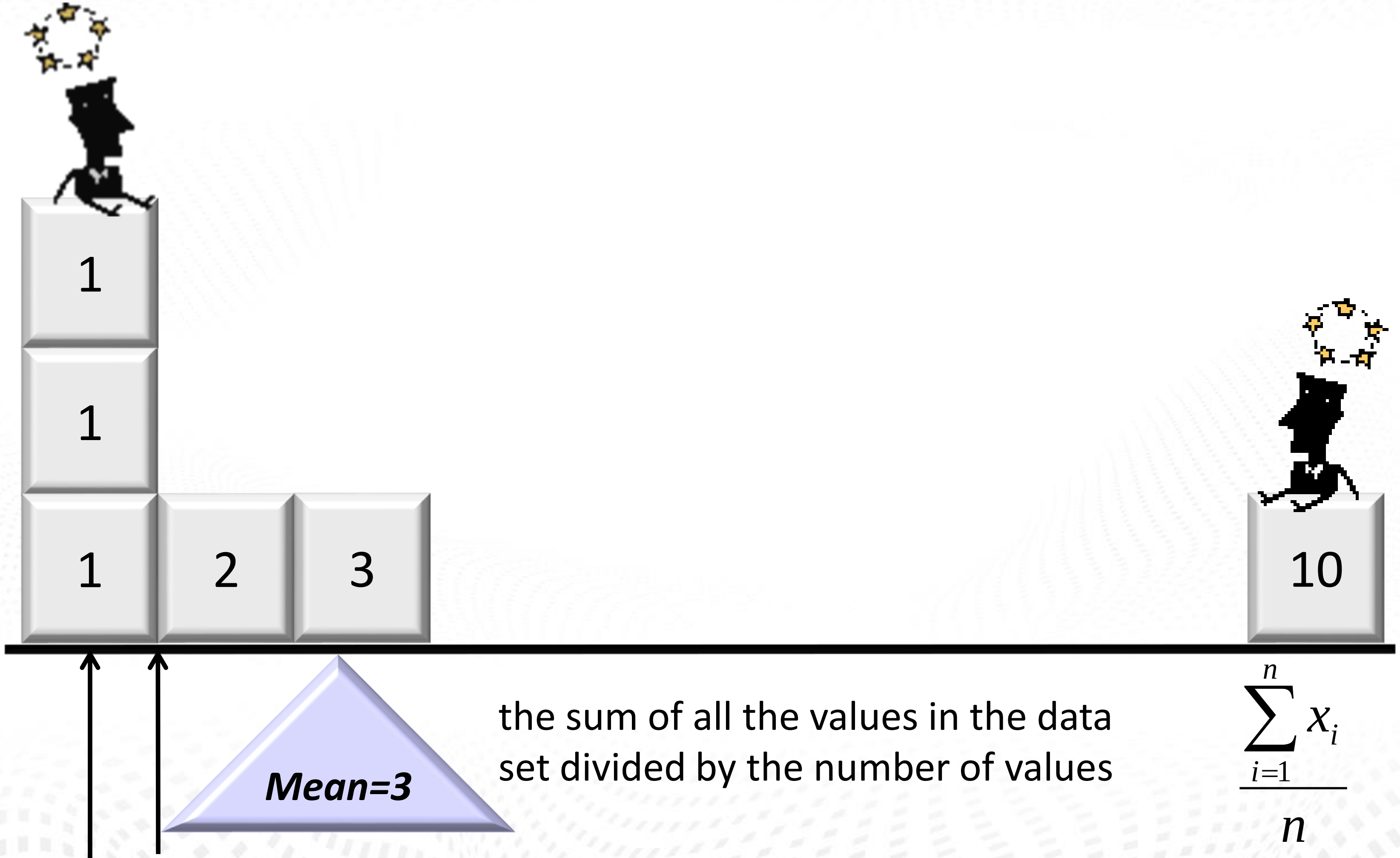
Example: Distribution of SATScore

- When you examine the distribution of values for the variable **SATScore**, you can determine the following characteristics:
  - the range of possible data values
  - the frequency of data values
  - whether the data values accumulate in the middle of the distribution or at one end



IDNumber ▲	Gender	SATScore
2012997	Male	1050
2854197	Female	1260
3873197	Female	1110
4440297	Female	1300
6520097	Female	1330
6975697	Male	1090
8945097	Male	1050
9589297	Female	1490
9880297	Male	1290
16907997	Male	1020
19108097	Male	1280
20494697	Male	1150
20745097	Female	1170
20755897	Female	1270
23048597	Female	1380

Central Tendency: Mean, Median, and Mode



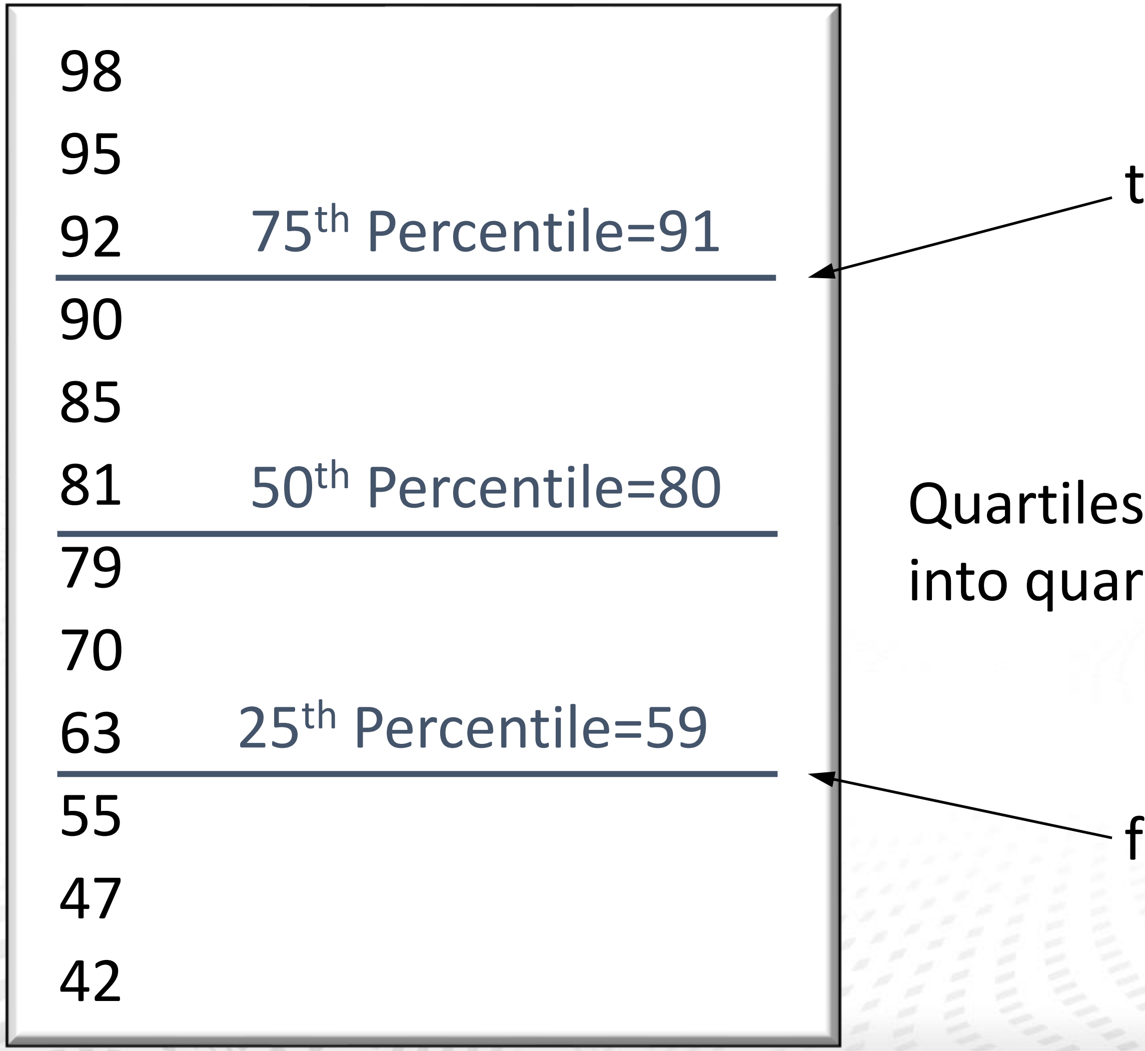
Mean=3 the sum of all the values in the data set divided by the number of values

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

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

Mode=1 the most common or frequent data value

Percentiles



98
95
92 75th Percentile=91
90
85
81 50th Percentile=80
79
70
63 25th Percentile=59
55
47
42

third quartile

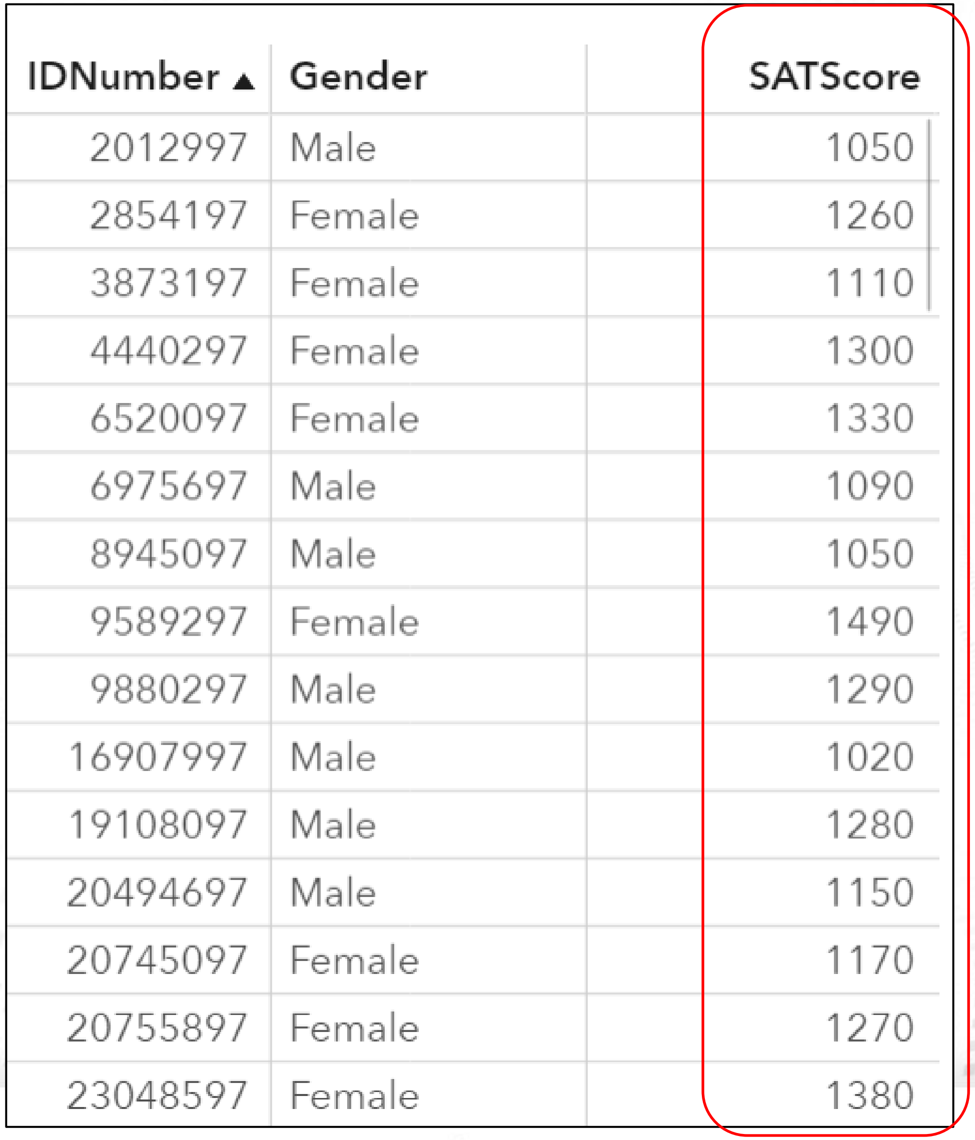
first quartile

Quartiles divide your data into quarters.

The Spread of a Distribution: Dispersion

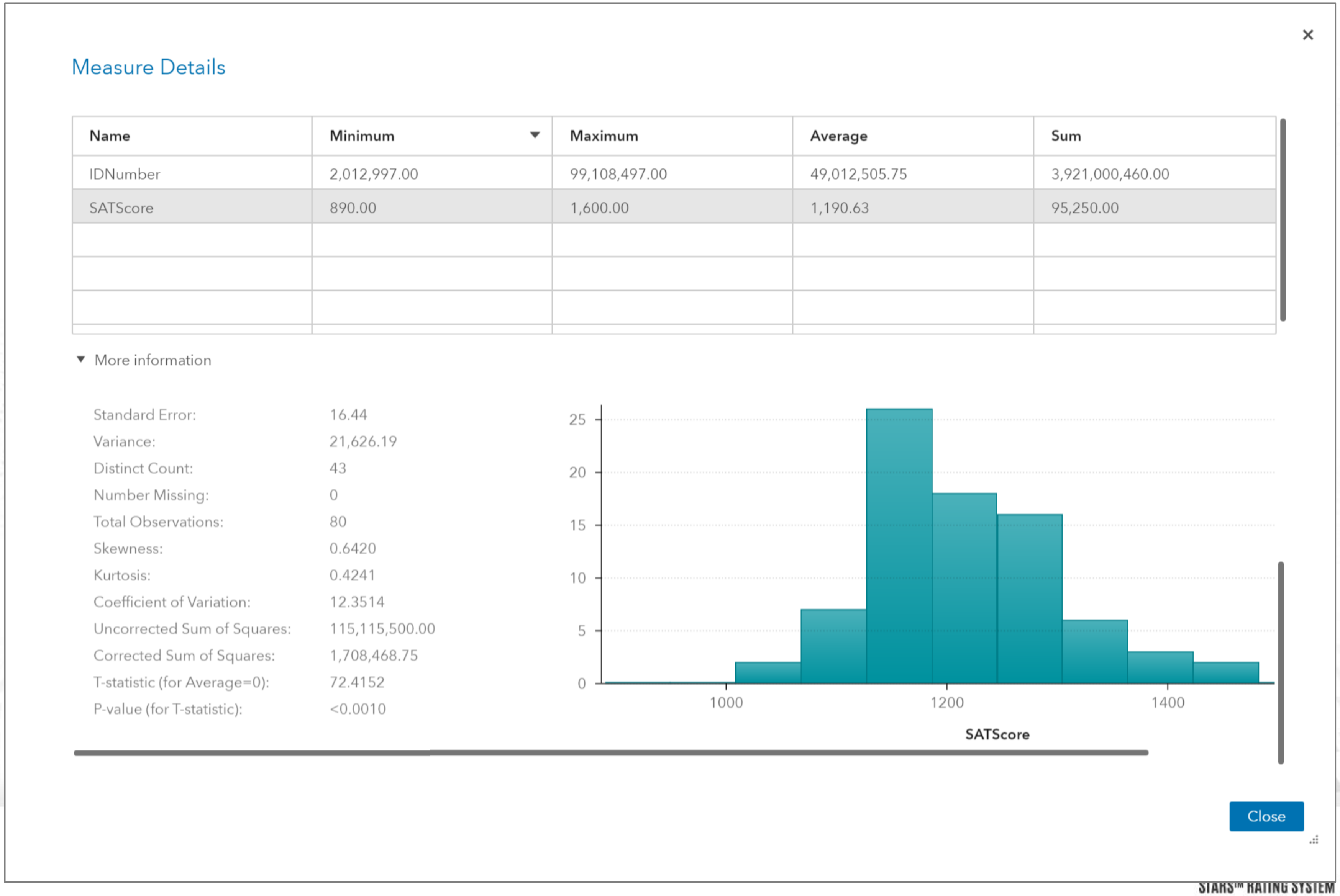
Measure	Definition
Range	The difference between the maximum and minimum data values
Interquartile Range	The difference between the 25th and 75th percentiles
Variance	A measure of dispersion of the data around the mean
Standard Deviation	A measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

Example: Describe SATScore



IDNumber ▲	Gender	SATScore
2012997	Male	1050
2854197	Female	1260
3873197	Female	1110
4440297	Female	1300
6520097	Female	1330
6975697	Male	1090
8945097	Male	1050
9589297	Female	1490
9880297	Male	1290
16907997	Male	1020
19108097	Male	1280
20494697	Male	1150
20745097	Female	1170
20755897	Female	1270
23048597	Female	1380

Example: Describe SATScore



Measure Details

Name	Minimum	Maximum	Average	Sum
IDNumber	2,012,997.00	99,108,497.00	49,012,505.75	3,921,000,460.00
SATScore	890.00	1,600.00	1,190.63	95,250.00

▼ More information

Standard Error:	16.44
Variance:	21,626.19
Distinct Count:	43
Number Missing:	0
Total Observations:	80
Skewness:	0.6420
Kurtosis:	0.4241
Coefficient of Variation:	12.3514
Uncorrected Sum of Squares:	115,115,500.00
Corrected Sum of Squares:	1,708,468.75
T-statistic (for Average=0):	72.4152
P-value (for T-statistic):	<0.0010

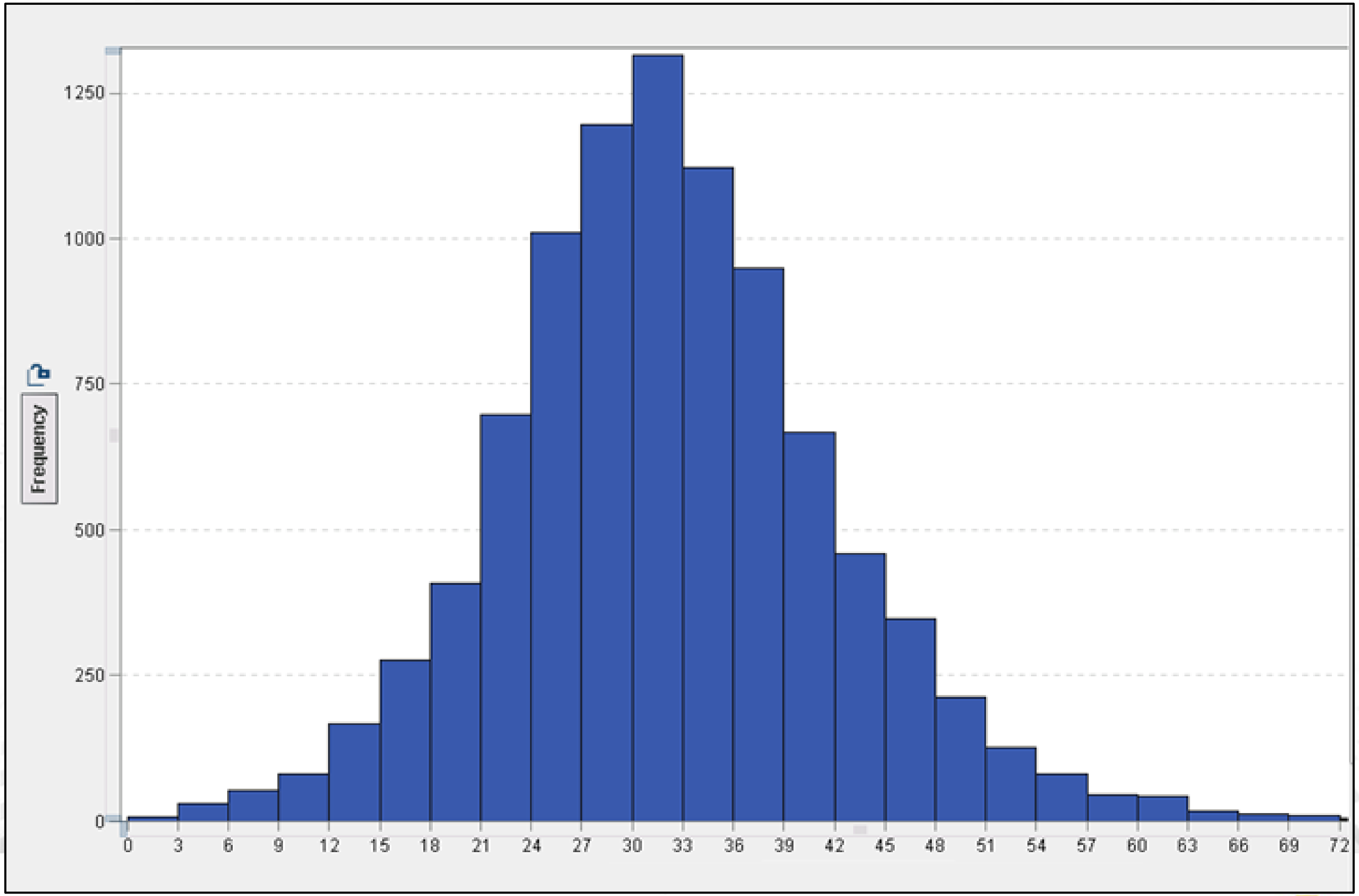
Lesson 11:
Introduction to Basic Statistics

1.1 Introduction to Statistics

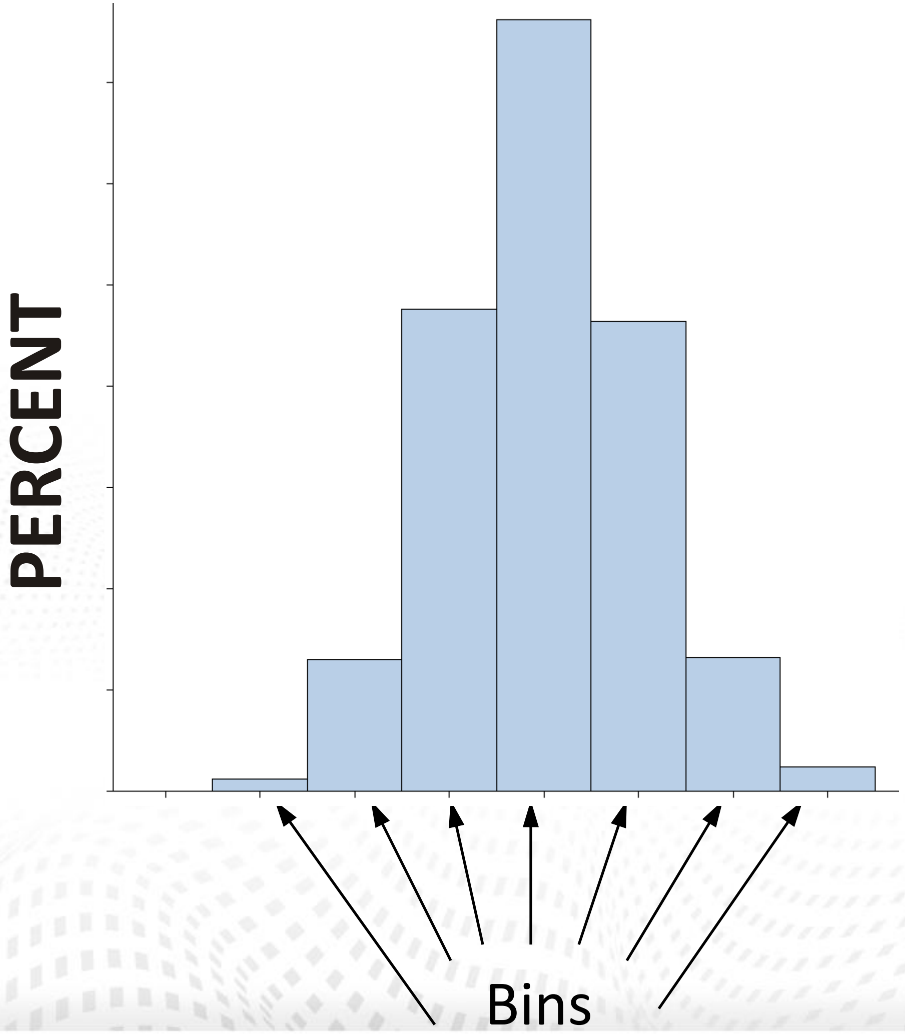
1.2 Viewing Distributions

1.3 Hypothesis Testing

What Distribution Is This?

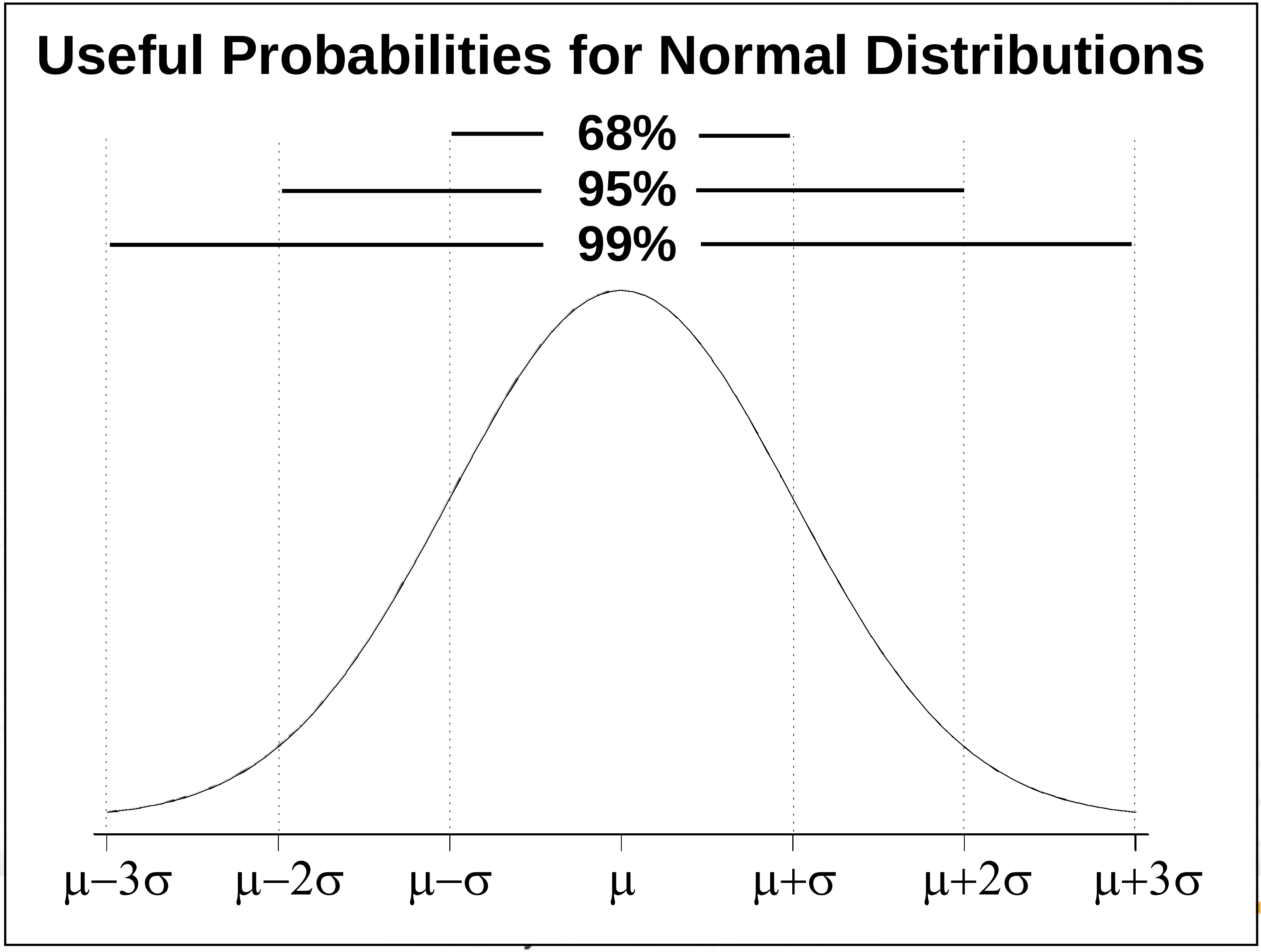


Picturing Distributions: Histogram



- 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 Distributions



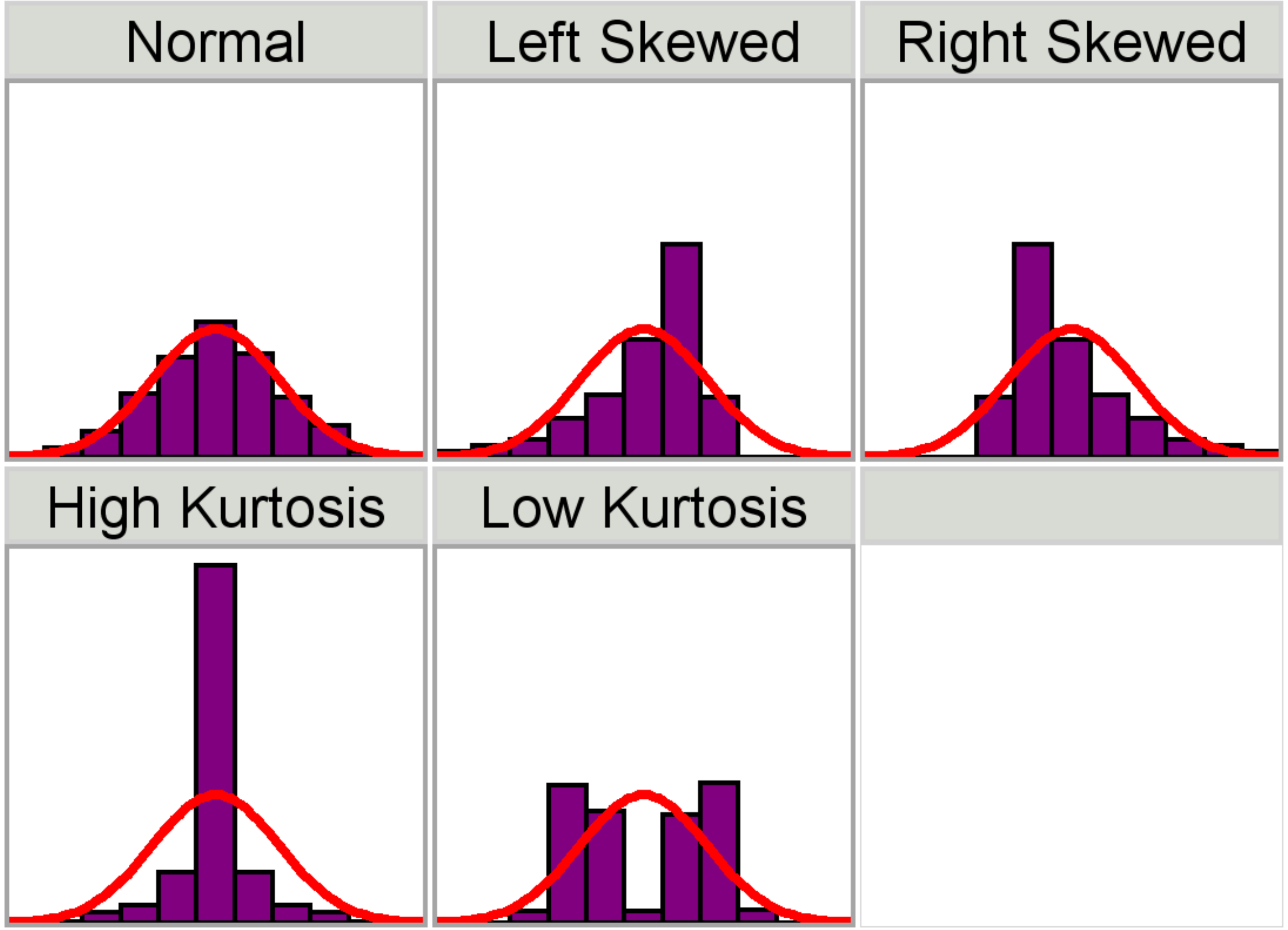
Normal Distributions

A *normal distribution*

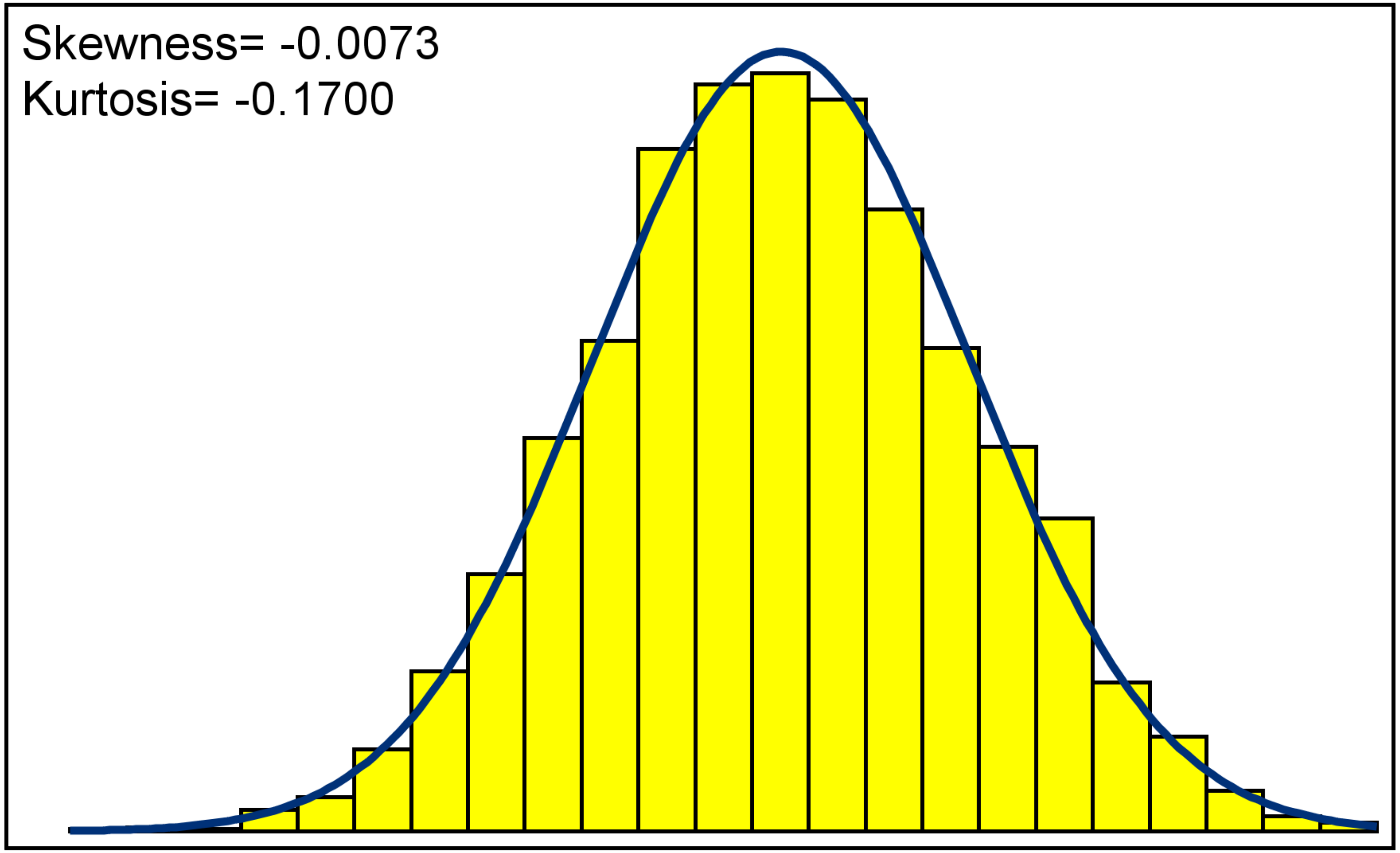
- is ***symmetric***. If you draw a line down the center, you get the same shape on either side.
- is ***fully characterized*** by the mean and standard deviation. Given the values of those two parameters, you know all that there is to know about the distribution.
- is bell shaped.
- has mean = median = mode.

- The **red** line on each of the following graphs represents the shape of the normal distribution with the mean and variance estimated from the sample data.

Data Distributions Compared to Normal

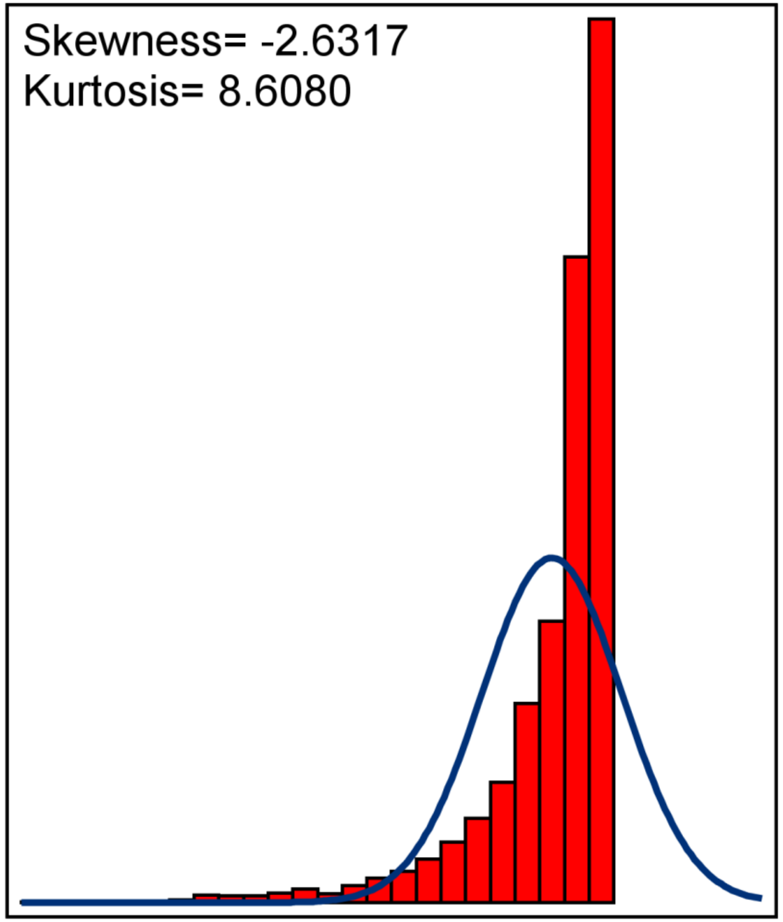


Normal Distribution

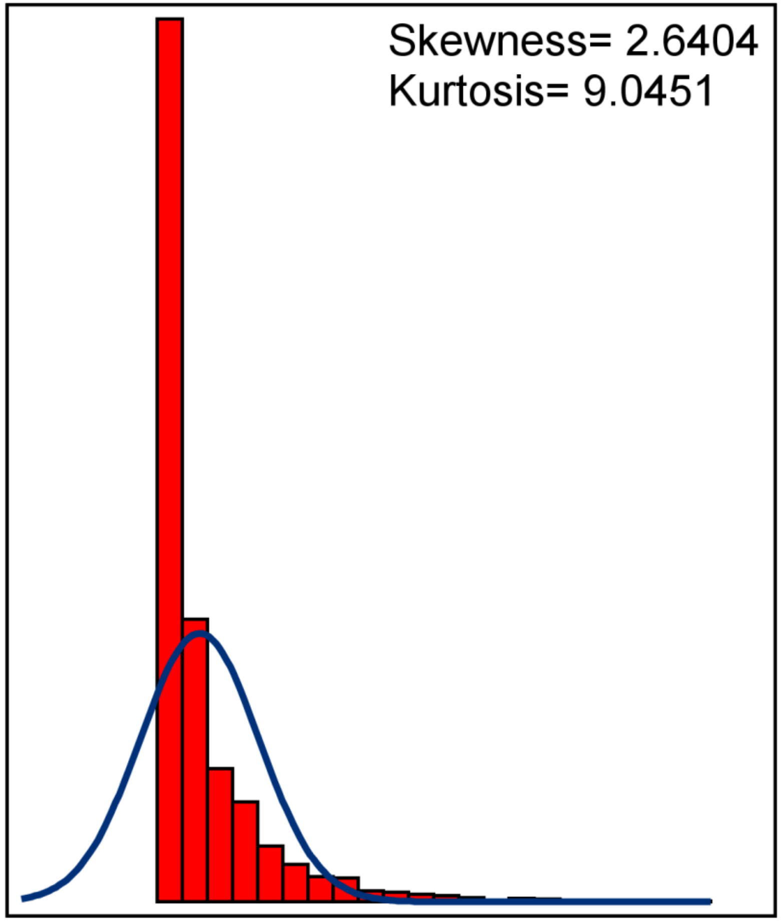


A Normal Distribution

Skewness

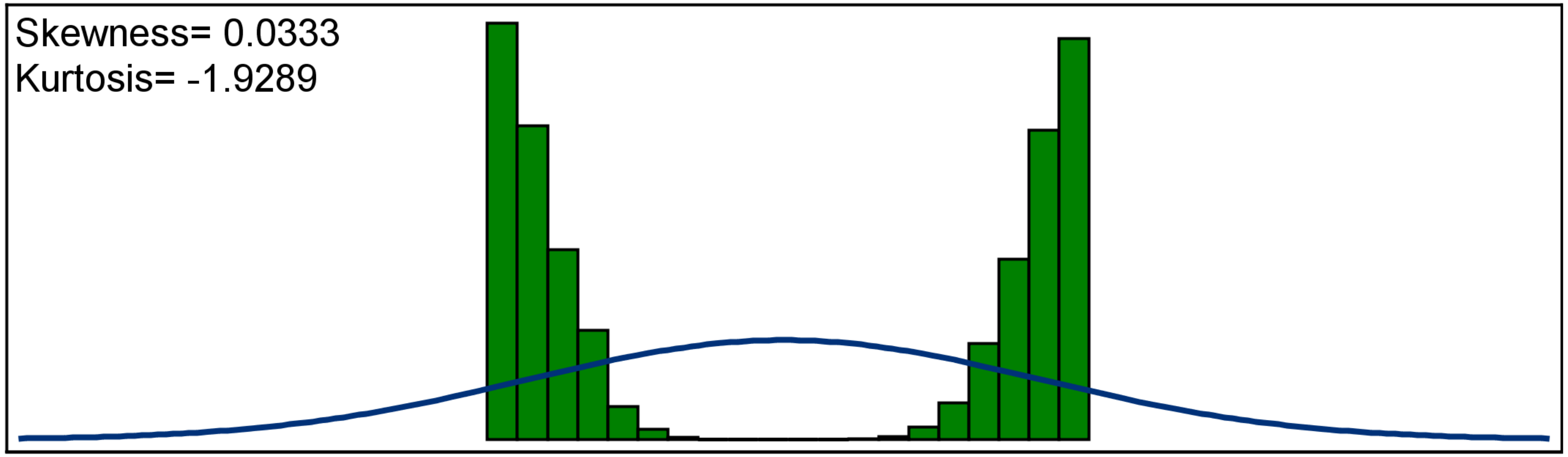


A Left Skewed Distribution

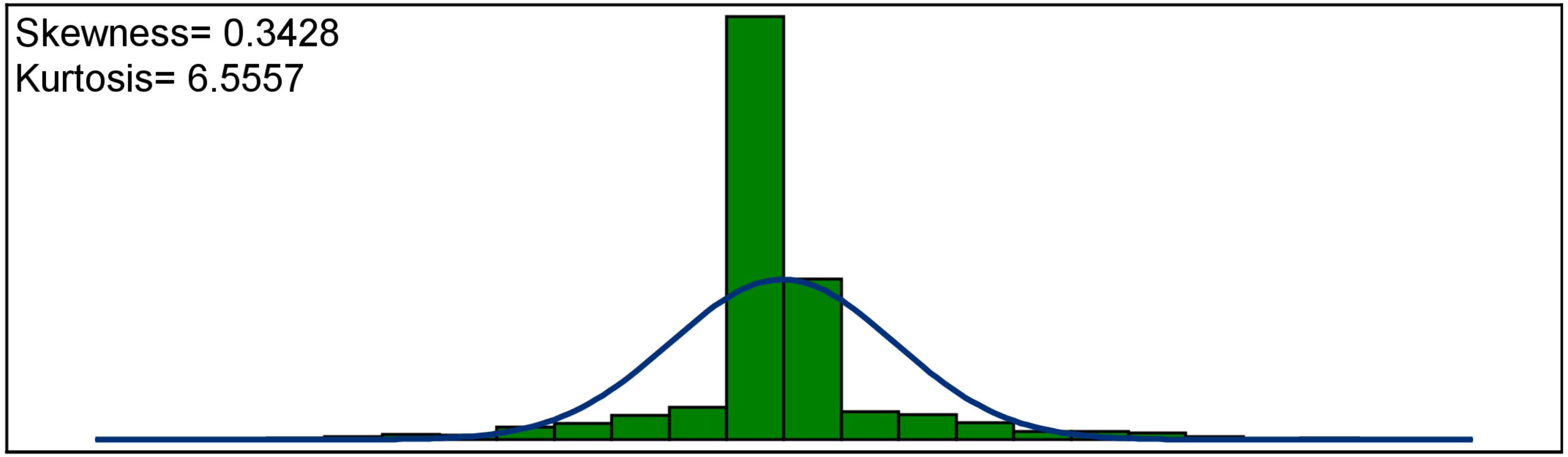


A Right Skewed Distribution

Kurtosis



A Platykurtotic Distribution

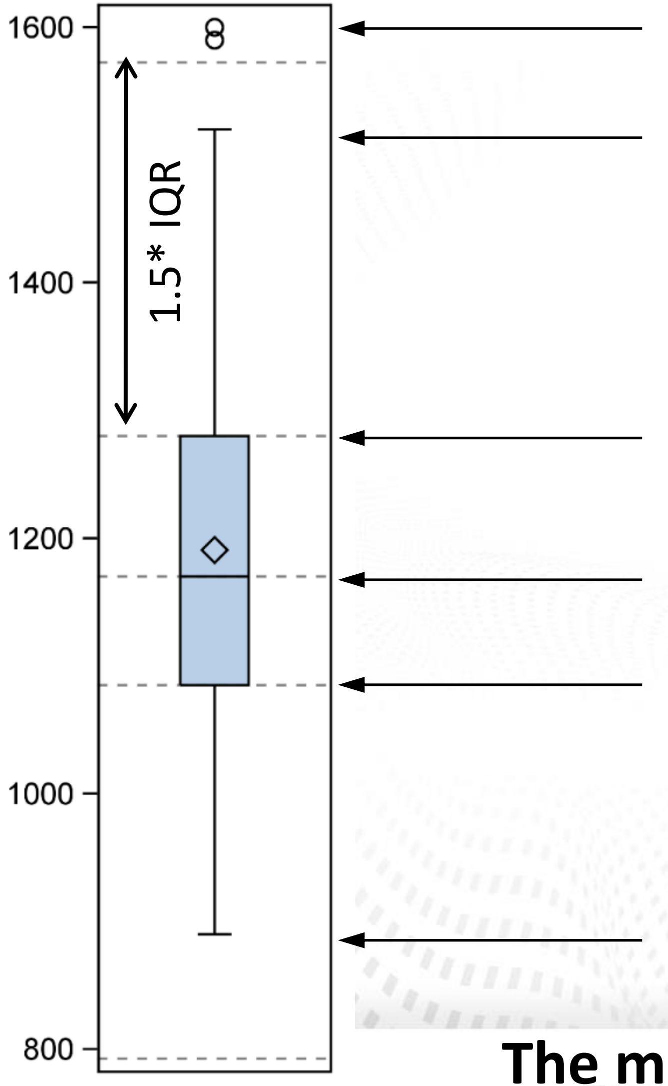


A Leptokurtotic Distribution

Graphical Displays of Distributions

- You can produce the following two types of plots for examining the distribution of your data values:
    - histograms
    - box plots

Box Plots



outliers > 1.5 IQR from the box

largest point <= 1.5 IQR from the box

the 75th percentile

the 50th percentile (median)

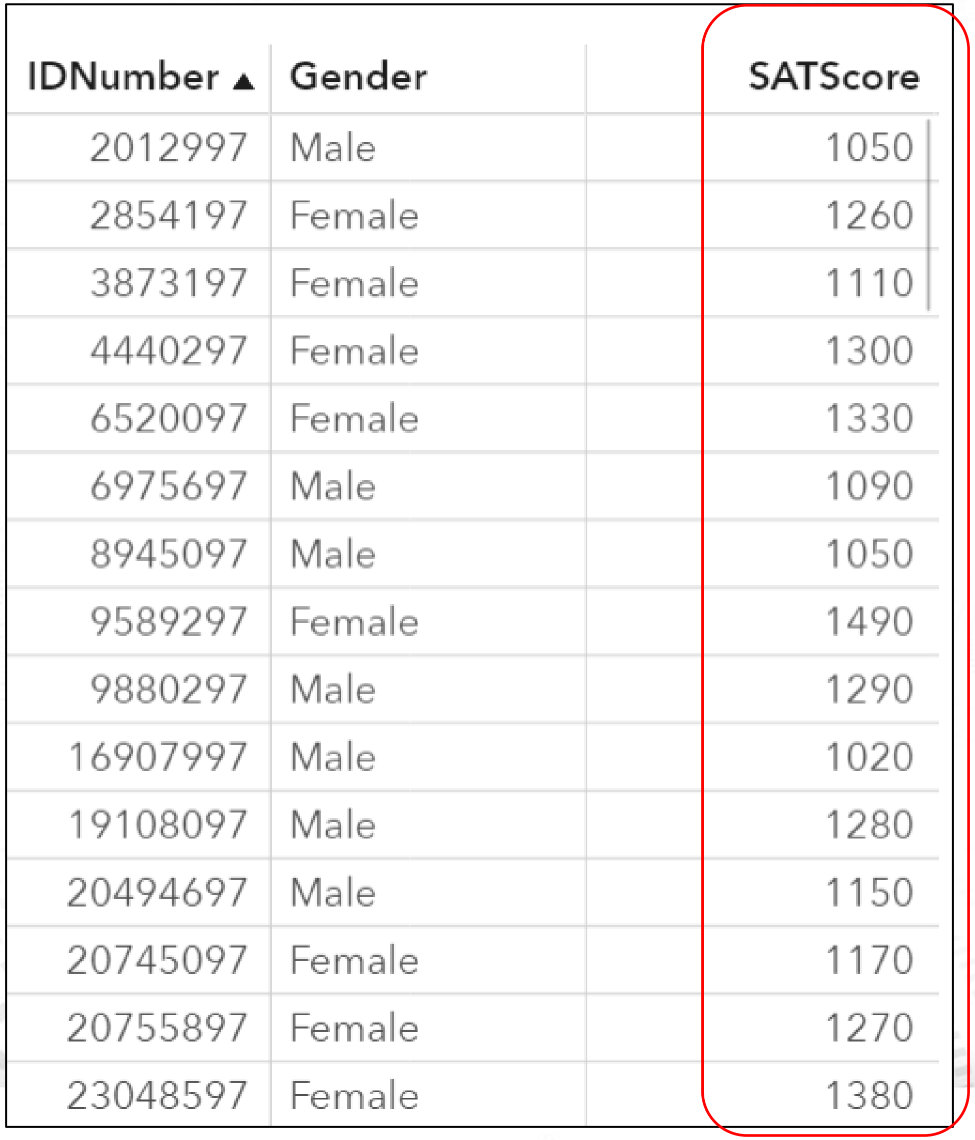
the 25th percentile

smallest point <= 1.5 IQR from the box

Information about the variability of your data and the extreme values

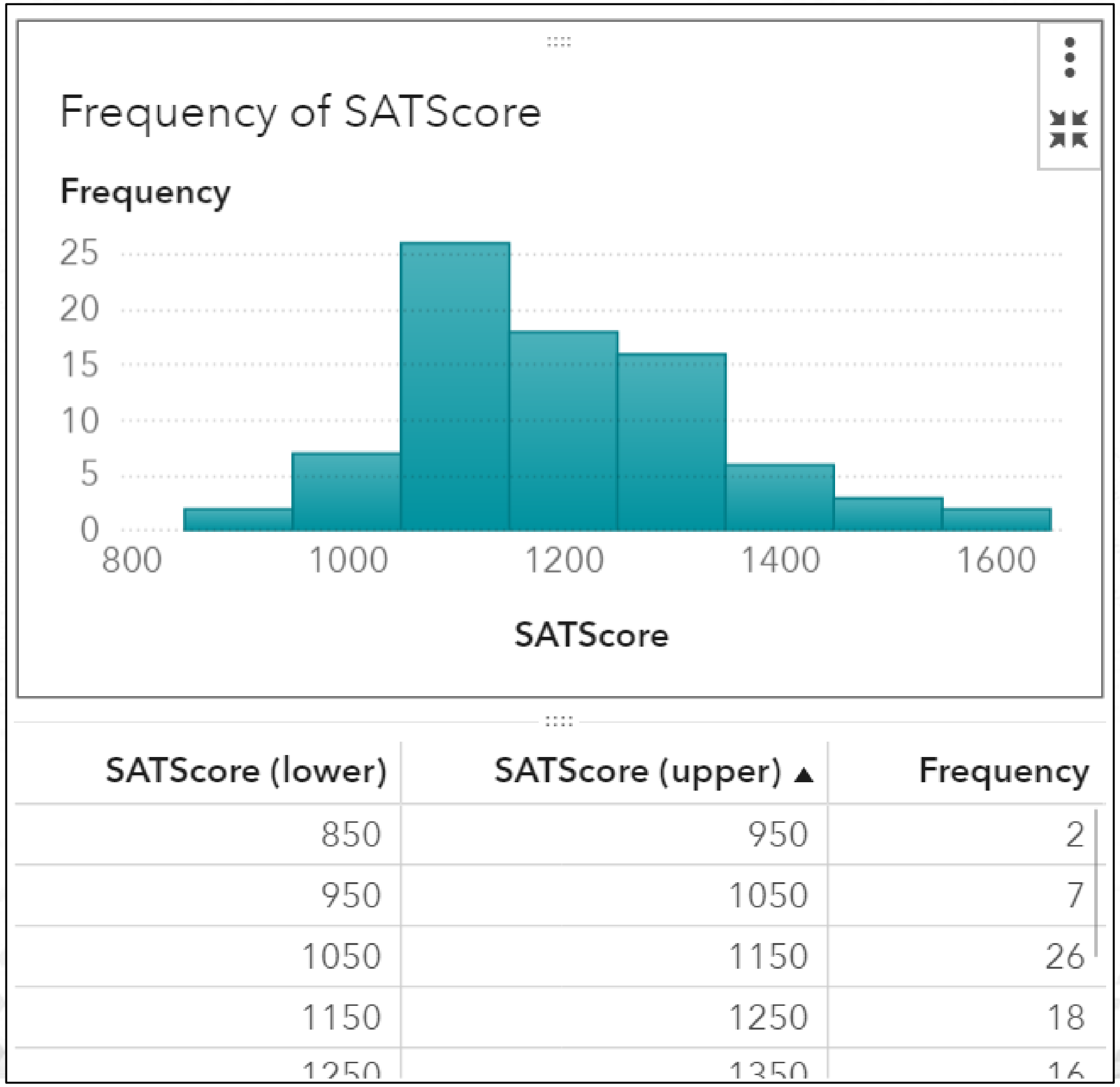
The mean is denoted by a ◊.

Example: Describe SATScore



IDNumber ▲	Gender	SATScore
2012997	Male	1050
2854197	Female	1260
3873197	Female	1110
4440297	Female	1300
6520097	Female	1330
6975697	Male	1090
8945097	Male	1050
9589297	Female	1490
9880297	Male	1290
16907997	Male	1020
19108097	Male	1280
20494697	Male	1150
20745097	Female	1170
20755897	Female	1270
23048597	Female	1380

Example: Describe SATScore



SATScore (lower)	SATScore (upper) ▲	Frequency
850	950	2
950	1050	7
1050	1150	26
1150	1250	18
1250	1350	16

Example: Describe SATScore

outliers →

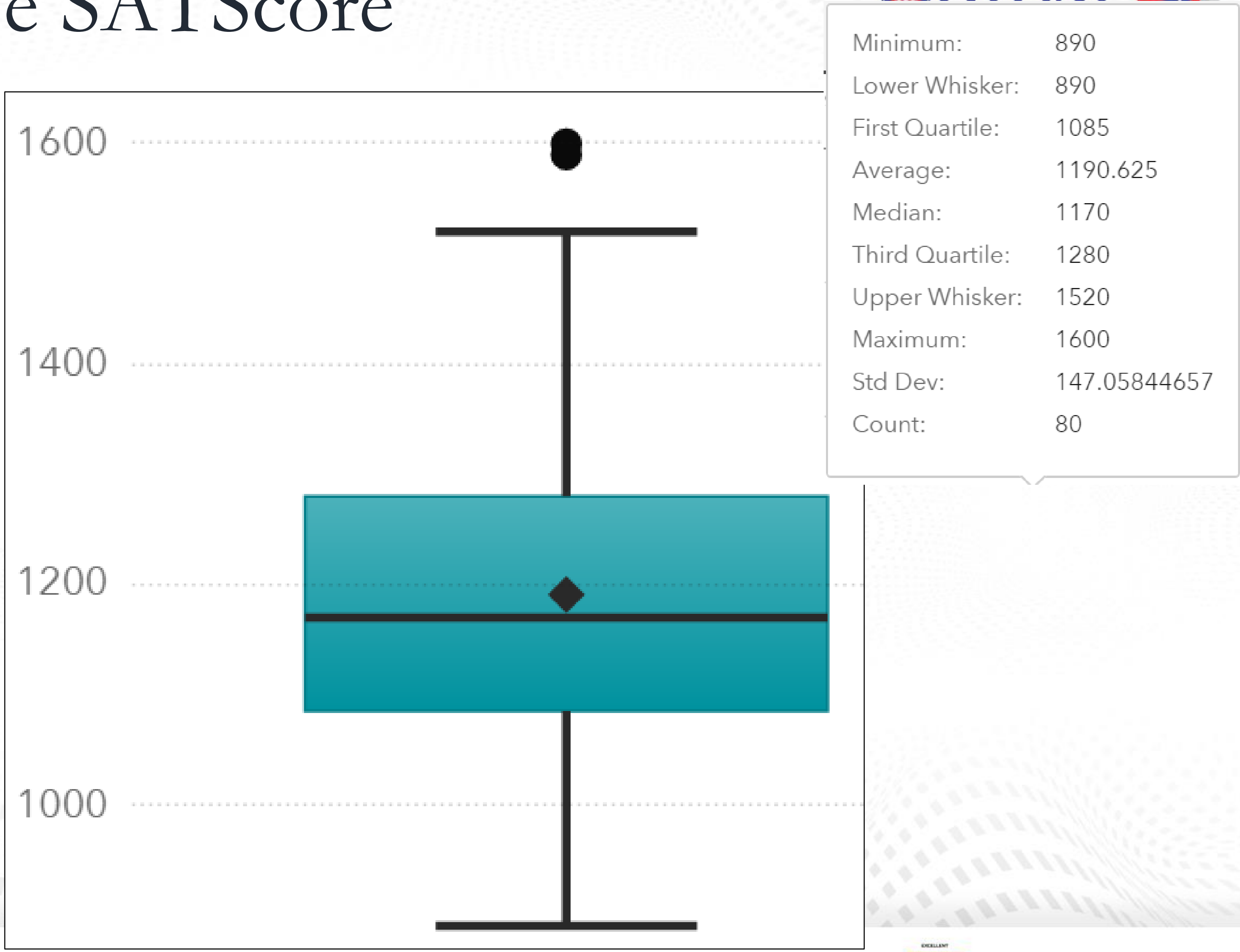
Q3 + 1.5* IQR →

Q3: 75% percentile →

median →

Q1: 25% percentile →

minimum →



Minimum:	890
Lower Whisker:	890
First Quartile:	1085
Average:	1190.625
Median:	1170
Third Quartile:	1280
Upper Whisker:	1520
Maximum:	1600
Std Dev:	147.05844657
Count:	80

Examining Distributions

This demonstration illustrates the use of SAS Visual Analytics for calculating statistics and for creating histograms and box plots.

Practice

This practice reinforces the concepts discussed previously.

Lesson 11:
Introduction to Basic Statistics

1.1 Introduction to Statistics

1.2 Viewing Distributions

1.3 Hypothesis Testing

Coin Example

Coin Example

Is this a fair coin?

H_0: probability head = 50%

H_1: probability head ≠ 50%

In other words:

H_0: Null hypothesis: Coin is fair.

H_1: Alternative: Coin is ***not*** fair.

- **Approach:**

✓Flip the coin a number of times.

✓Formulate the test statistic (for example, number of times *heads*).

✓If the test statistic is too large or too small, then H_0 is rejected.

- **What is too large or too small?**

Coin Example

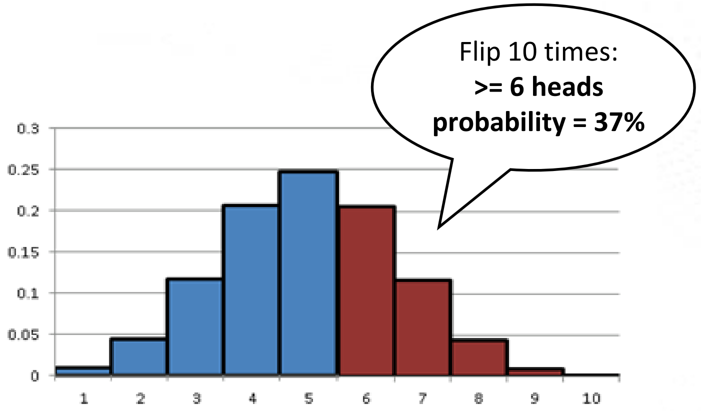
Suppose the coin is fair (H_0):
What is the probability of a particular outcome?

If the probability of the outcome
(or *test statistic*) is small (say < 5%),
then reject H_0 and accept H_1:
The coin is not fair.

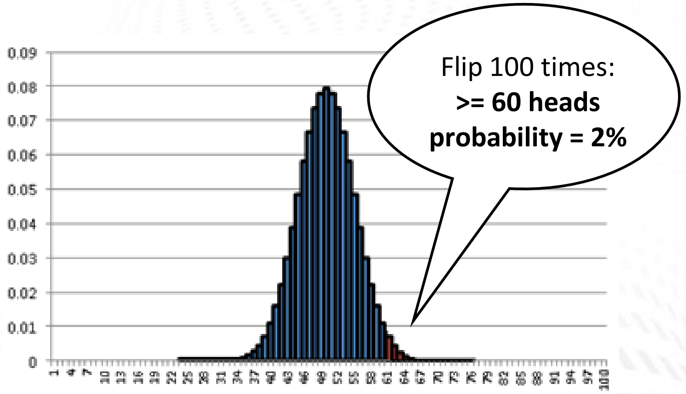
- Flip 10 times, 6 heads has probability 37%.
  Do not reject H_0.
- Flip 100 times, 60 heads has probability 2%.
  Do reject H_0.

The probability of an outcome of
the test statistic is called the ***p-value***.

You can make a ***mistake***:
Reject H_0 when it is true, or do not reject H_0 when it is not true.

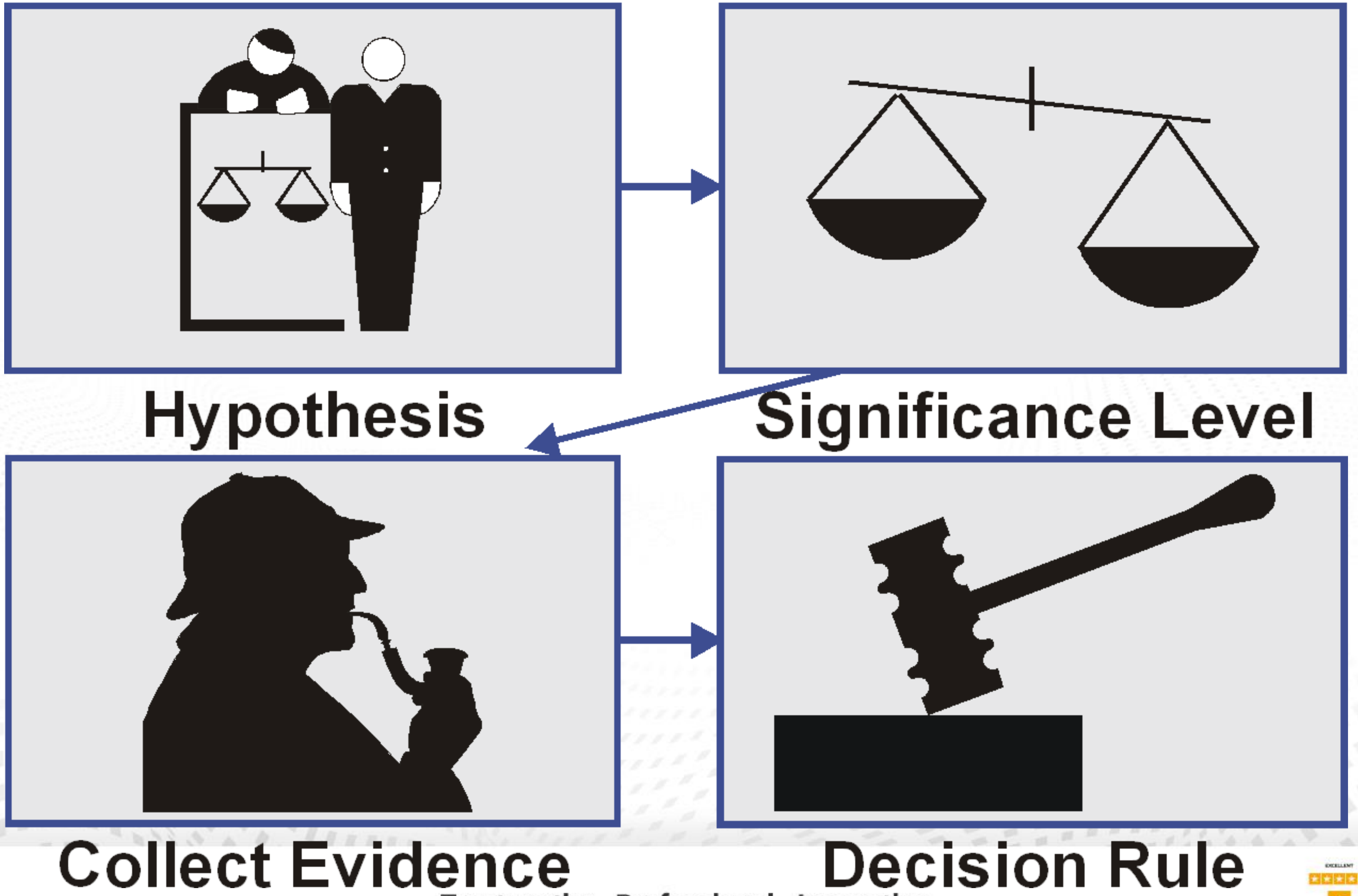


Flip 10 times:
>= 6 heads
probability = 37%



Flip 100 times:
>= 60 heads
probability = 2%

Judicial Analogy



11.01 Question

If you have a fair coin and flip it 100 times, is it possible for it to land on heads 100 times?

○ Yes

○ No

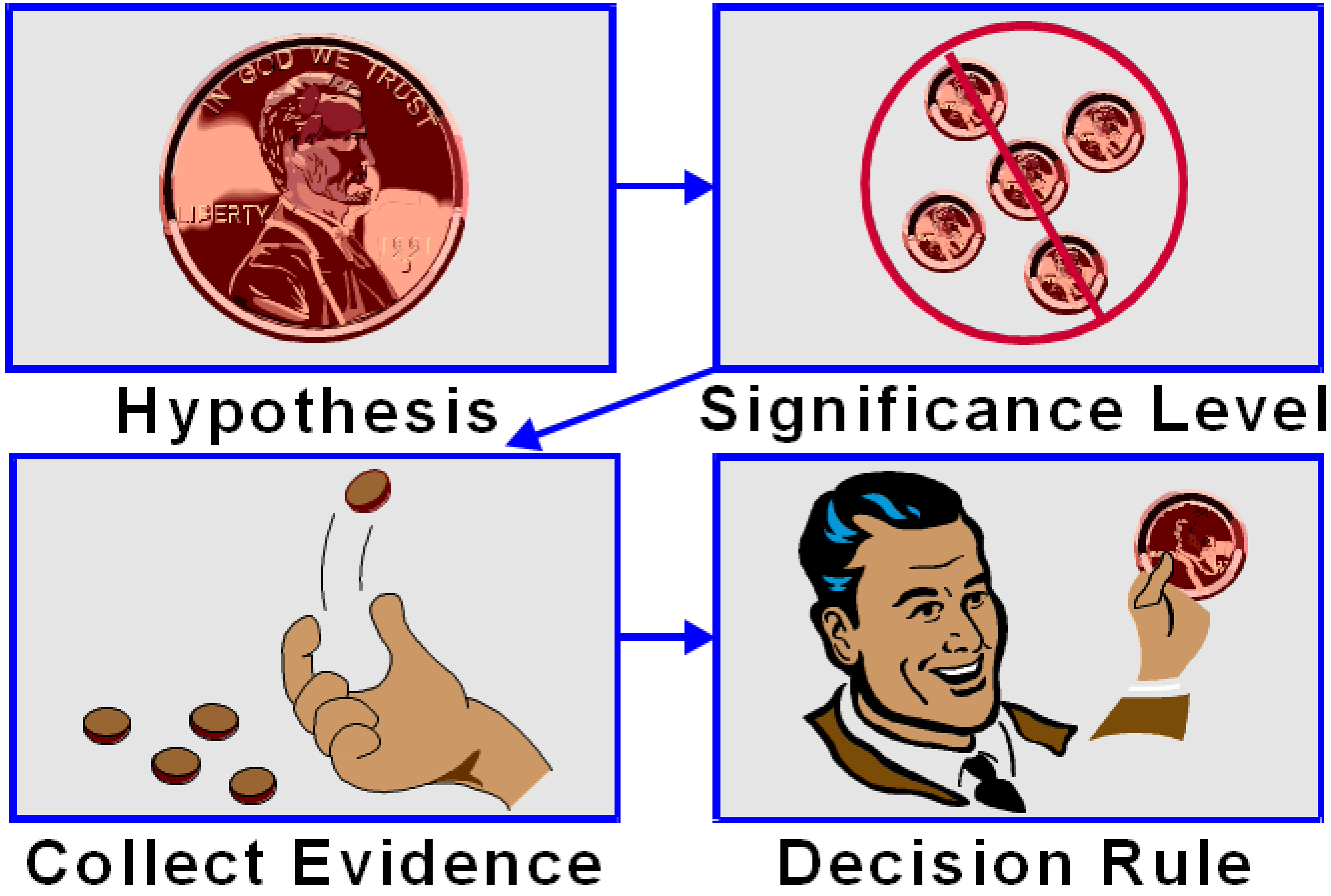
11.01 Question – Correct Answer

If you have a fair coin and flip it 100 times, is it possible for it to land on heads 100 times?

[✓ ○ Yes]

○ No

Coin Analogy



Types of Errors

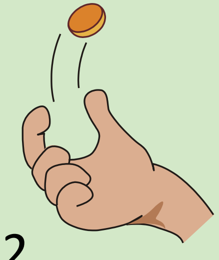
You used a decision rule to make a decision, but was the decision correct?

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

Coin Experiment: Effect Size Influence

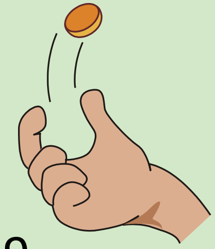
Flip a coin 100 times and decide whether it is fair.

55 Heads
45 Tails



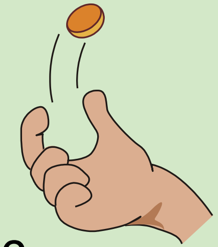
p-value=.3682

40 Heads
60 Tails



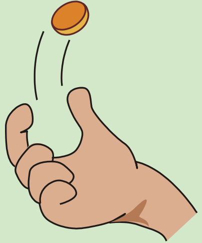
p-value=.0569

37 Heads
63 Tails



p-value=.0120

15 Heads
85 Tails

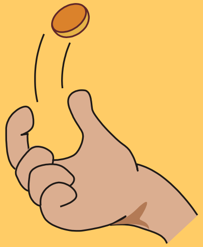


p-value<.0001

Coin Experiment: Sample Size Influence

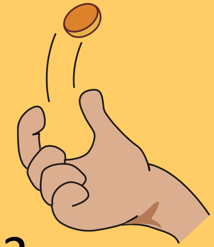
Flip a coin and get 40% heads, and decide whether it is fair.

4 Heads
6 Tails



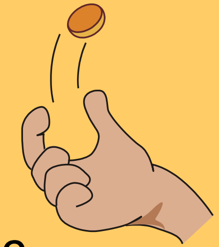
p-value=.7539

16 Heads
24 Tails



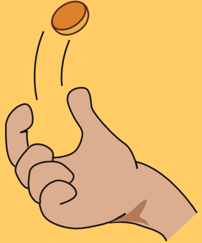
p-value=.2682

40 Heads
60 Tails



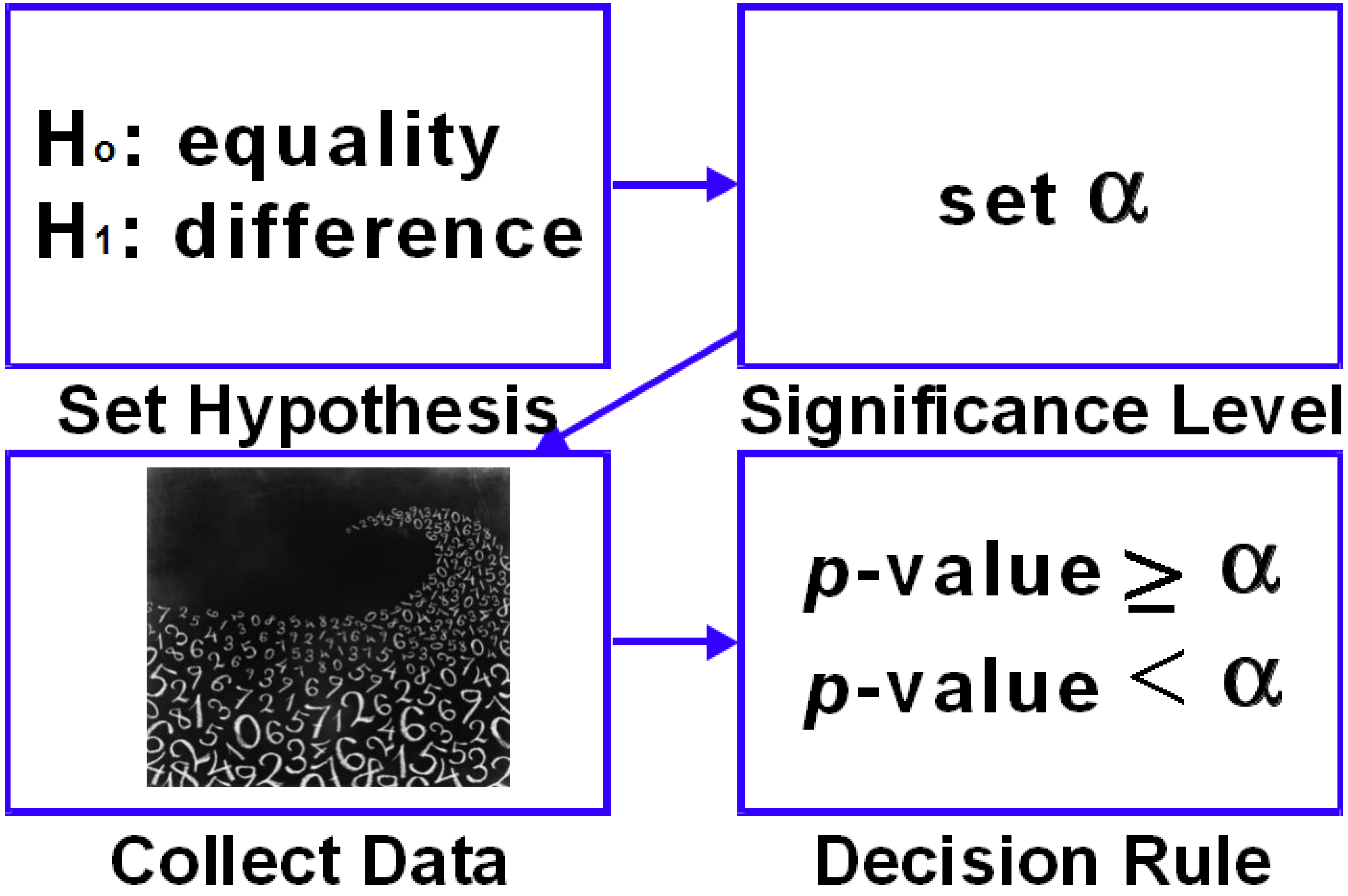
p-value=.0569

160 Heads
240 Tails



p-value<.0001

Statistical Hypothesis Test



H_0: equality
H_1: difference

Set Hypothesis

set α

Significance Level

Collect Data

p-value $\geq \alpha$
p-value $< \alpha$

Decision Rule

Comparing α and the *p*-Value

- In general, you
  - reject the null hypothesis if *p*-value < α
  - fail to reject the null hypothesis if *p*-value ≥ α.

Defining the Problem: SATScore = 1200?

- The purpose of the study is to determine whether the average combined Math and Verbal scores on the Scholastic Aptitude Test (SAT) at Carver County magnet high schools is 1200 (the goal set by the school board).

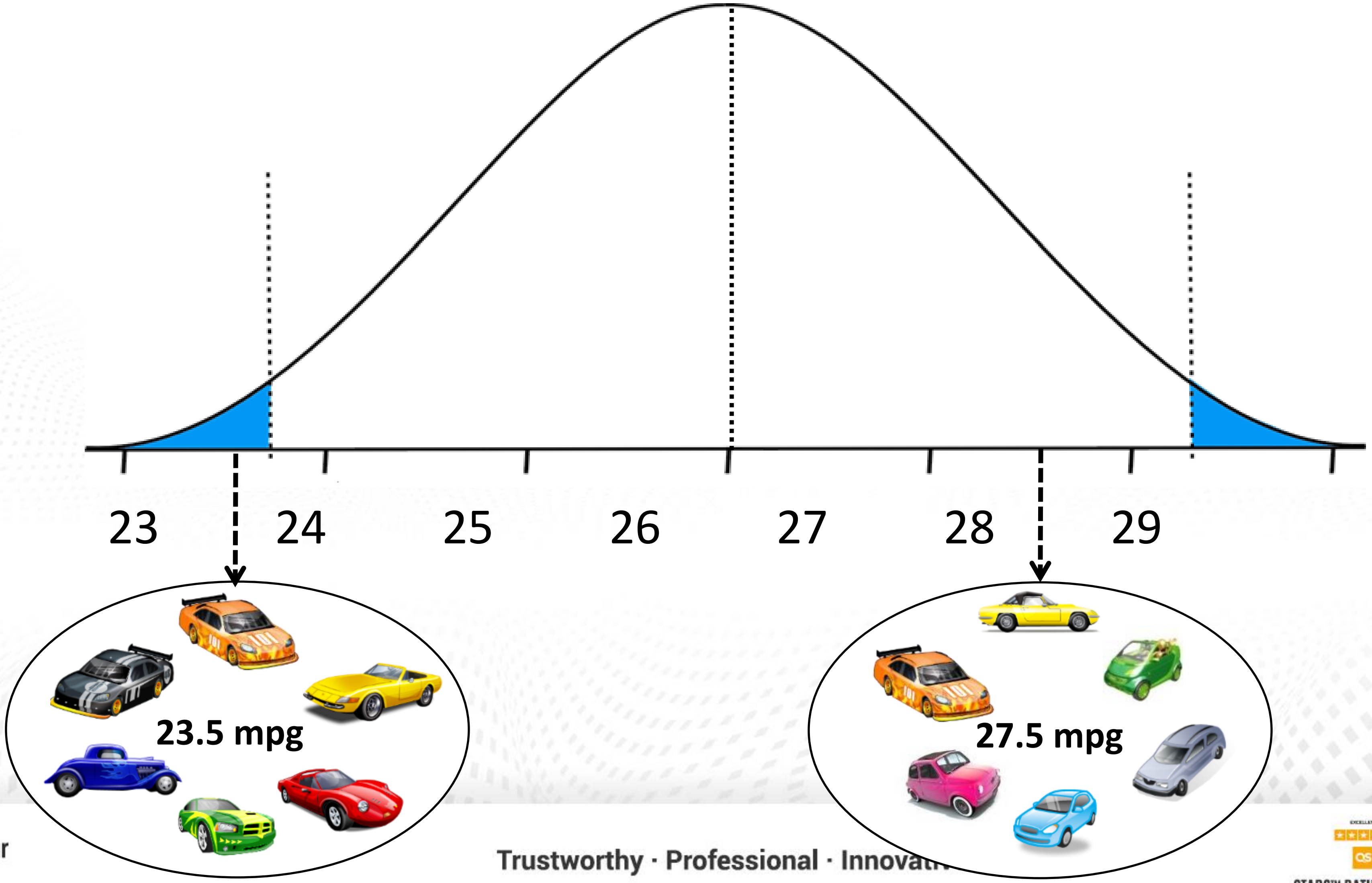
- H_0: population mean of **SATScore** = 1200
- H_1: population mean of **SATScore** ≠ 1200

Point Estimates

$\overline{\chi}$ estimates μ

S estimates σ

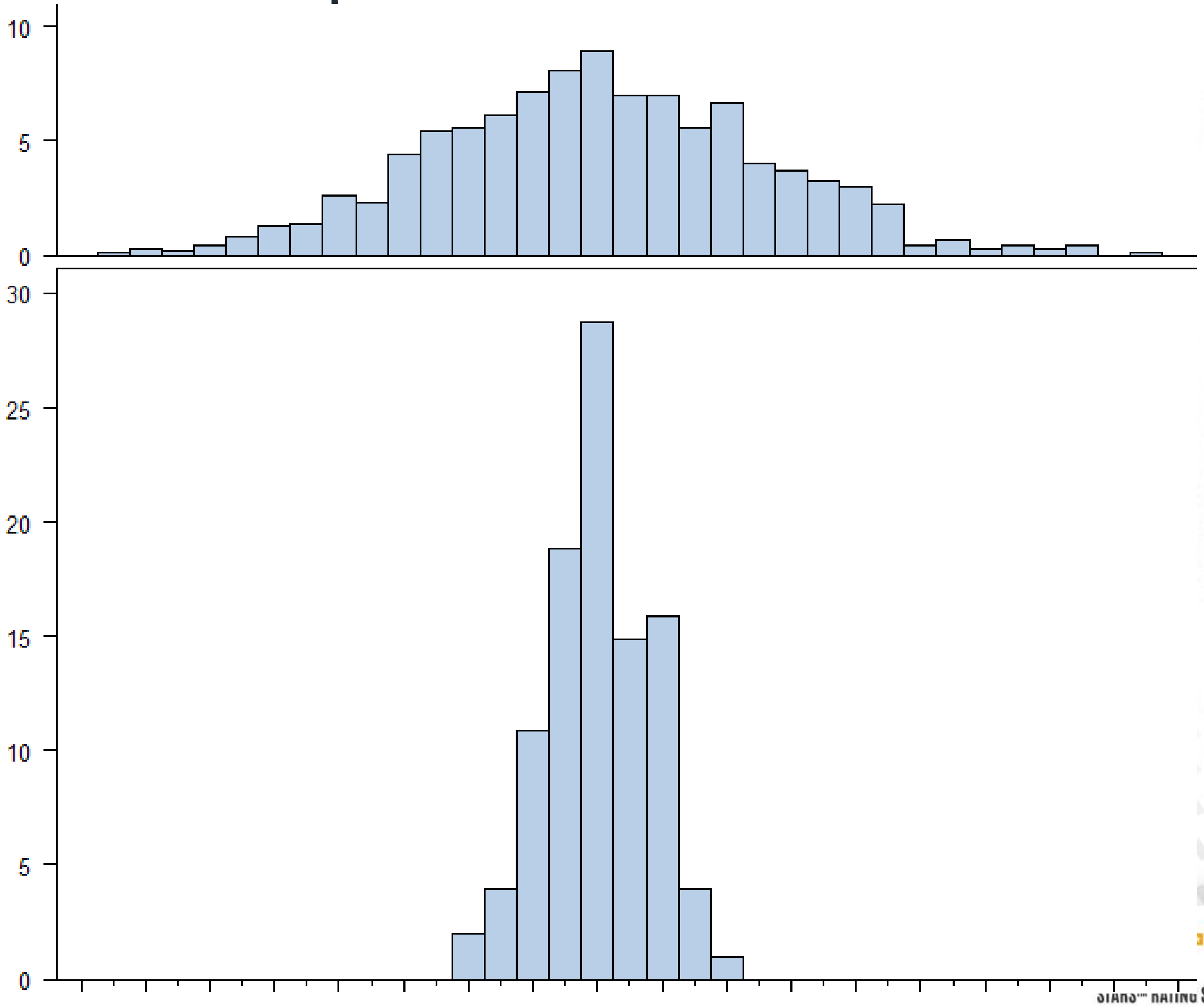
Variability among Samples



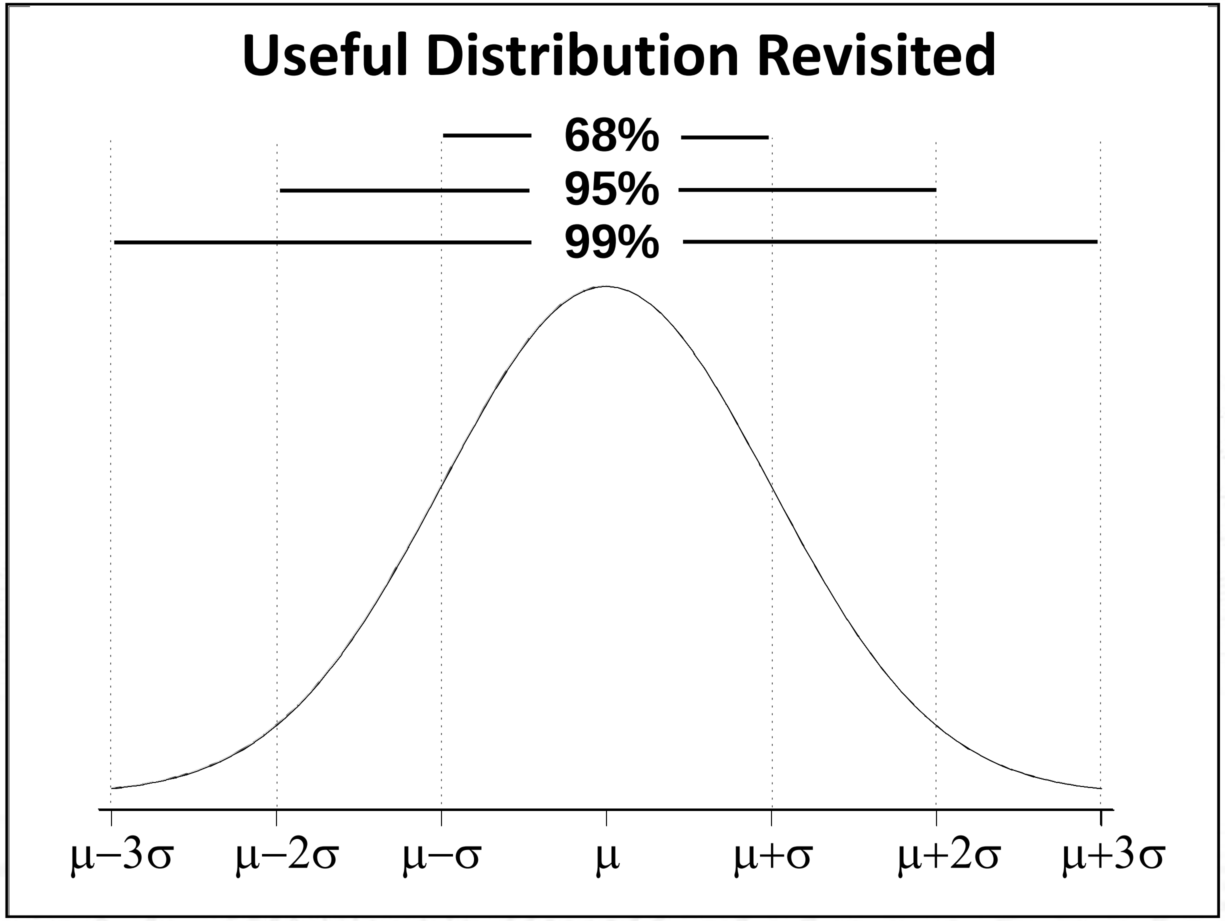
Distribution of Sample Means

SAT score

Means of SAT score (n=10)



Normal Distribution for the Mean



The types of confidence intervals in this course assume that the sample means are normally distributed.

Standard Error of the Mean

- A statistic that measures the variability of your estimate is the *standard error of the mean*.

- It differs from the sample standard deviation because
  - the sample standard deviation is a measure of the variability of ***data***
  - the standard error of the mean is a measure of the variability of ***sample means***.

    - Standard error of the mean = $\dfrac{s}{\sqrt{n}} = s_{\bar{X}}$

Performing a Hypothesis Test

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

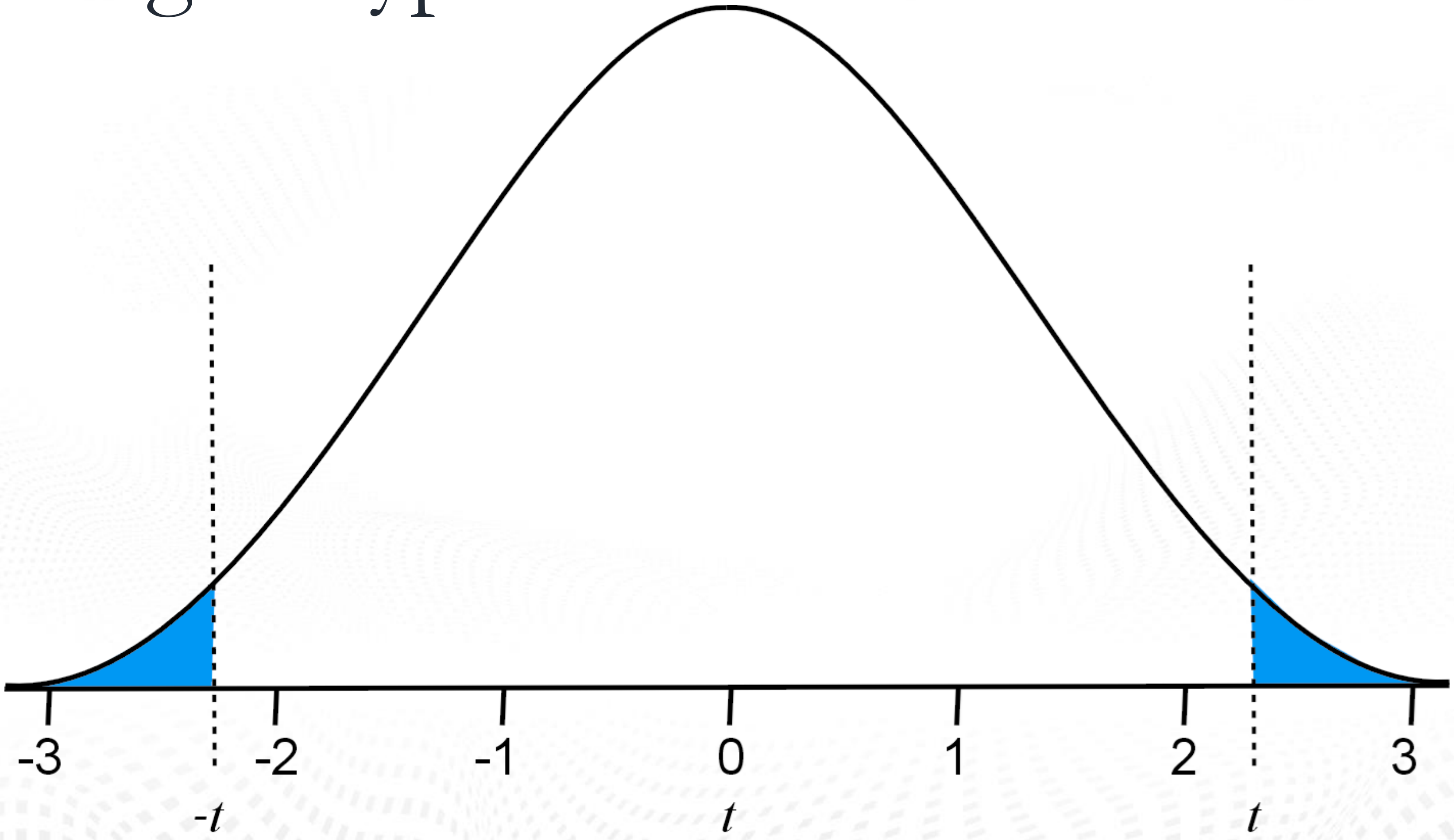
$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

For the test scores example:

$$t = \frac{(1190.625 - 1200)}{16.4416} = -0.5702$$

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.

Performing a Hypothesis Test



The *t* statistic can be positive or negative.

Hypothesis Testing

This demonstration illustrates using the *t* statistic from the Measure Details to test the hypothesis that the mean of the SAT Math and Verbal scores equals 1200.

Practice

This practice reinforces the concepts discussed previously.

UTHM Universiti Tun Hussein Onn Malaysia 30 1993-2023

AMANAH • PROFESIONAL • INOVATIF

Thank you