

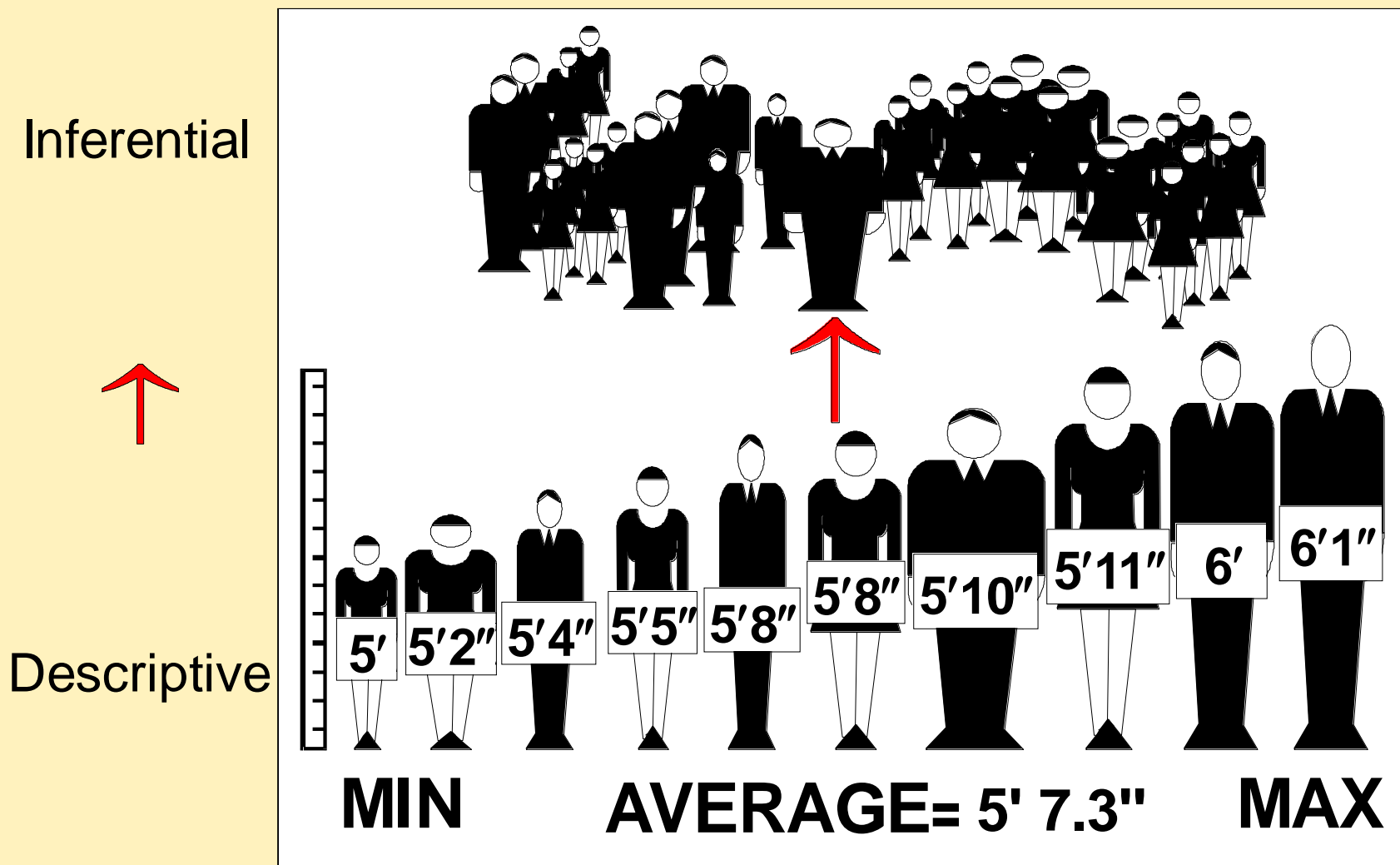
Introduction to Statistics – 08AUG2022

1: Fundamental Statistical Concepts

2: Examining Distributions

3: Describing Categorical Data

Two Broad Categories of Statistics

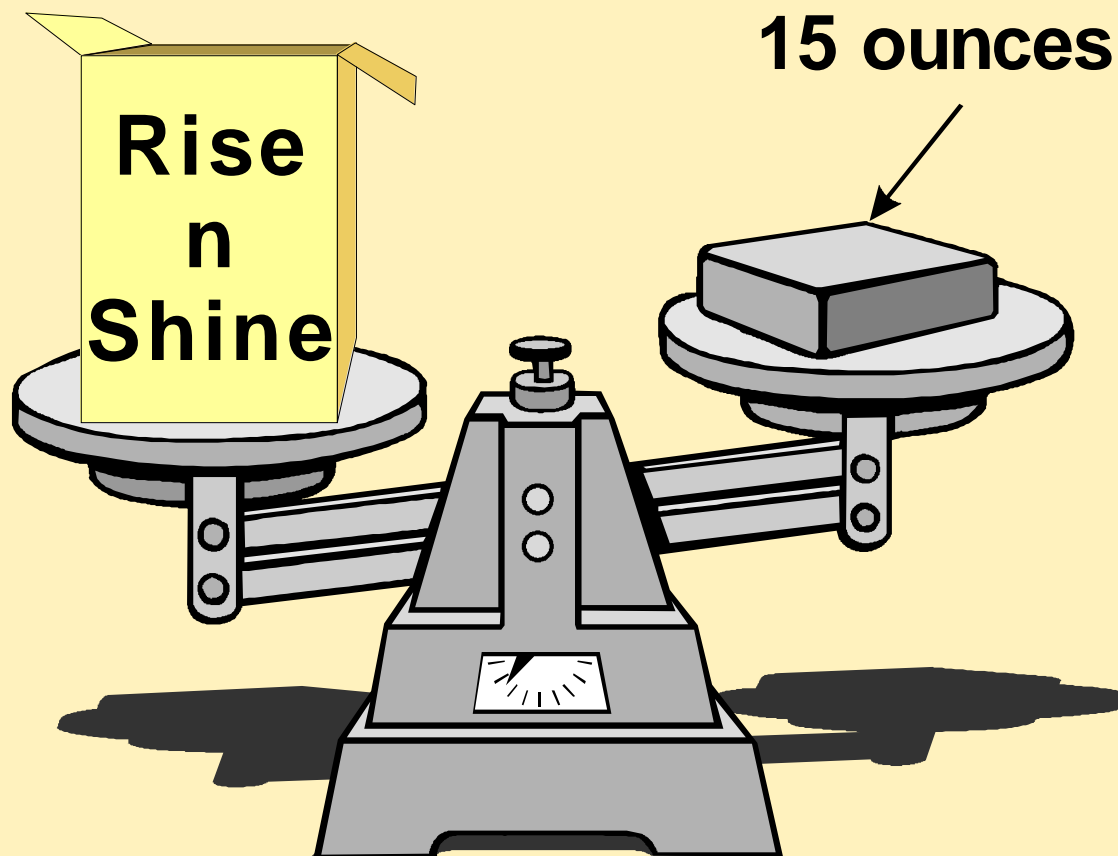


Defining the Problem

Before you begin any analysis, you should complete certain tasks.

1. Outline the purpose of the study.
2. Document the study questions.
3. Define the population of interest.
4. Determine the need for sampling.
5. Define the data collection protocol.

Cereal Example

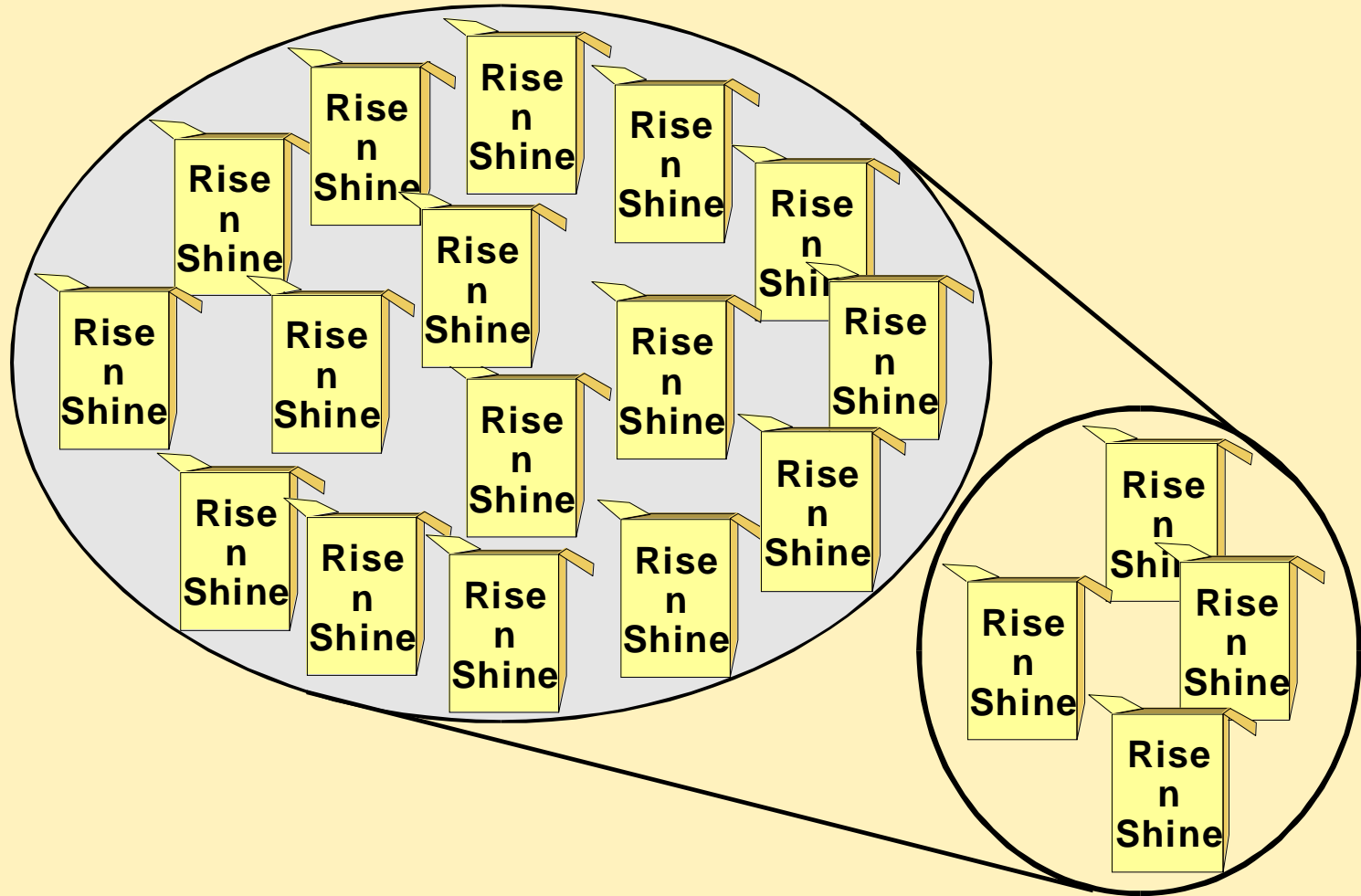


Defining the Problem

The purpose of the study is to determine whether Rise n Shine cereal boxes contain 15 ounces of cereal.

The study question is whether the average amount of cereal in Rise n Shine boxes is equal to 15 ounces.

Sample



Sampling

Sample the data by creating one or more data tables. The samples should be large enough to contain significant information and small enough to process effeciently.

Class Discussion

What are some potential problems that you can encounter when sampling from a larger population?

Parameters and Statistics

Statistics are used to approximate population parameters.

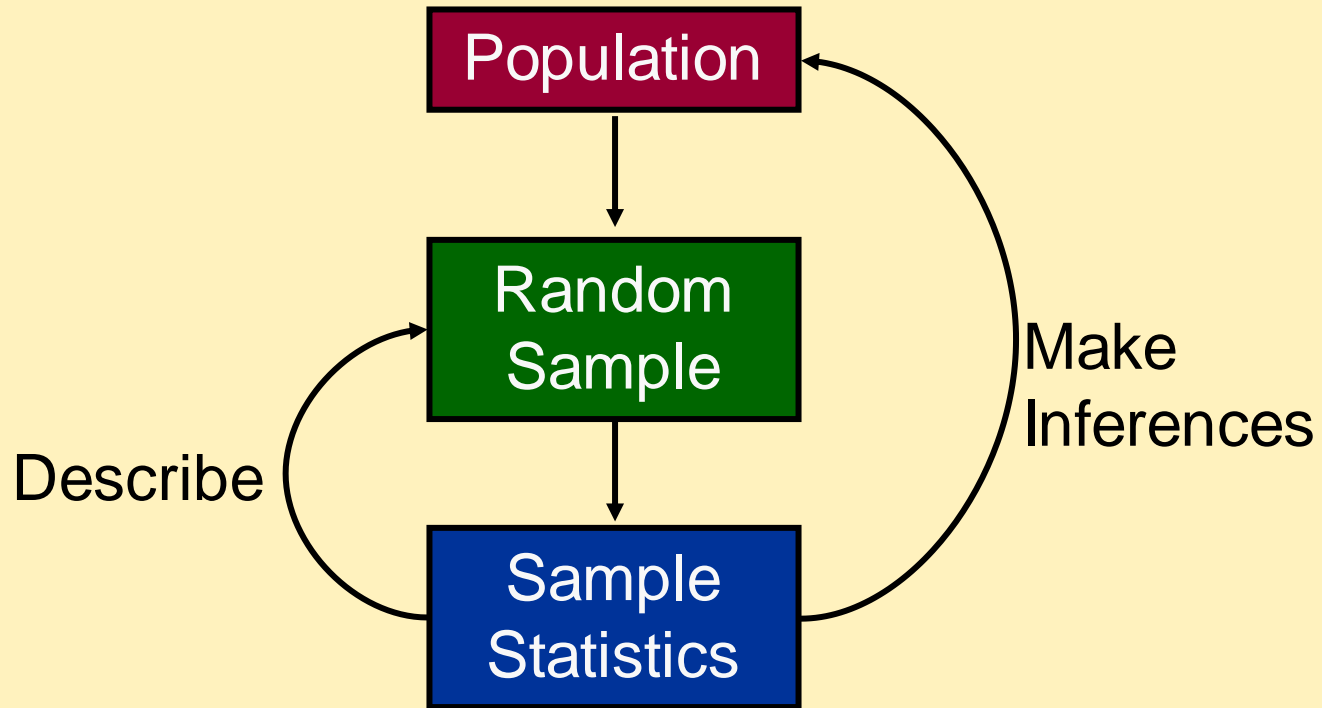
	Population Parameters	Sample Statistics
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard Deviation	σ	s

Describing Your Data

When you describe data, your goals are the following:

- Screen for unusual data values.
- Inspect the spread and shape of continuous variables.
- Characterize the central tendency.
- Draw preliminary conclusions about your data.

Process of Statistical Analysis



Section Summary

- Discussed some of the fundamental issues of statistics.
- Identified issues to be considered when you draw a sample.

Chapter 1: Introduction to Statistics

1: Fundamental Statistical Concepts

2: Examining Distributions

3: Describing Categorical Data

Objectives

- Examine distributions of data.
- Explain and interpret measures of location, dispersion, and shape.
- Use the Summary Statistics and Distribution Analysis tasks to produce descriptive statistics.
- Use the Distribution Analysis task to generate histograms, box-and-whisker plots, and normal probability plots.

Cereal Data Set

brand

weight

idnumber



•

•

•

•

•

•

•

•

•

•

•

•

Assumption for This Course

The sample drawn is *representative* of the population.

In other words, the sample characteristics should reflect the characteristics of the population as a whole.

Distributions

When you examine the distribution of values for the variable **weight**, you can determine the following:

- 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

Typical Values in a Distribution

- Mean: 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: The middle value^{*n*} (also known as the 50th percentile)
- Mode: The most common or frequent data value

Measures of central tendency

Introduction to Statistics – 09AUG2022

The background of the slide features a faded image of a person in a white shirt and tie, pointing towards the right. Overlaid on this image are various statistical and mathematical symbols, including the Greek letter alpha (α), the letter 'C', the hash symbol (#), the letter 'A', the letter 'P', the letter 'Z', the at-sign (@), and binary code (001).

1: Fundamental Statistical Concepts

2: Examining Distributions

3: Describing Categorical Data

Percentiles

98

95

92 75th Percentile=91

90

85

81 50th Percentile=80

79

70


63 25th Percentile=59

55

47


42

third quartile



Quartiles break up your data into quarters.

first quartile



The Spread of a Distribution: Dispersion

Measure	Definition
<i>range</i>	the difference between the maximum and minimum data values
<i>interquartile range</i>	the difference between the 25 th and 75 th percentiles
<i>variance</i>	a measure of dispersion of the data around the mean
<i>standard deviation</i>	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

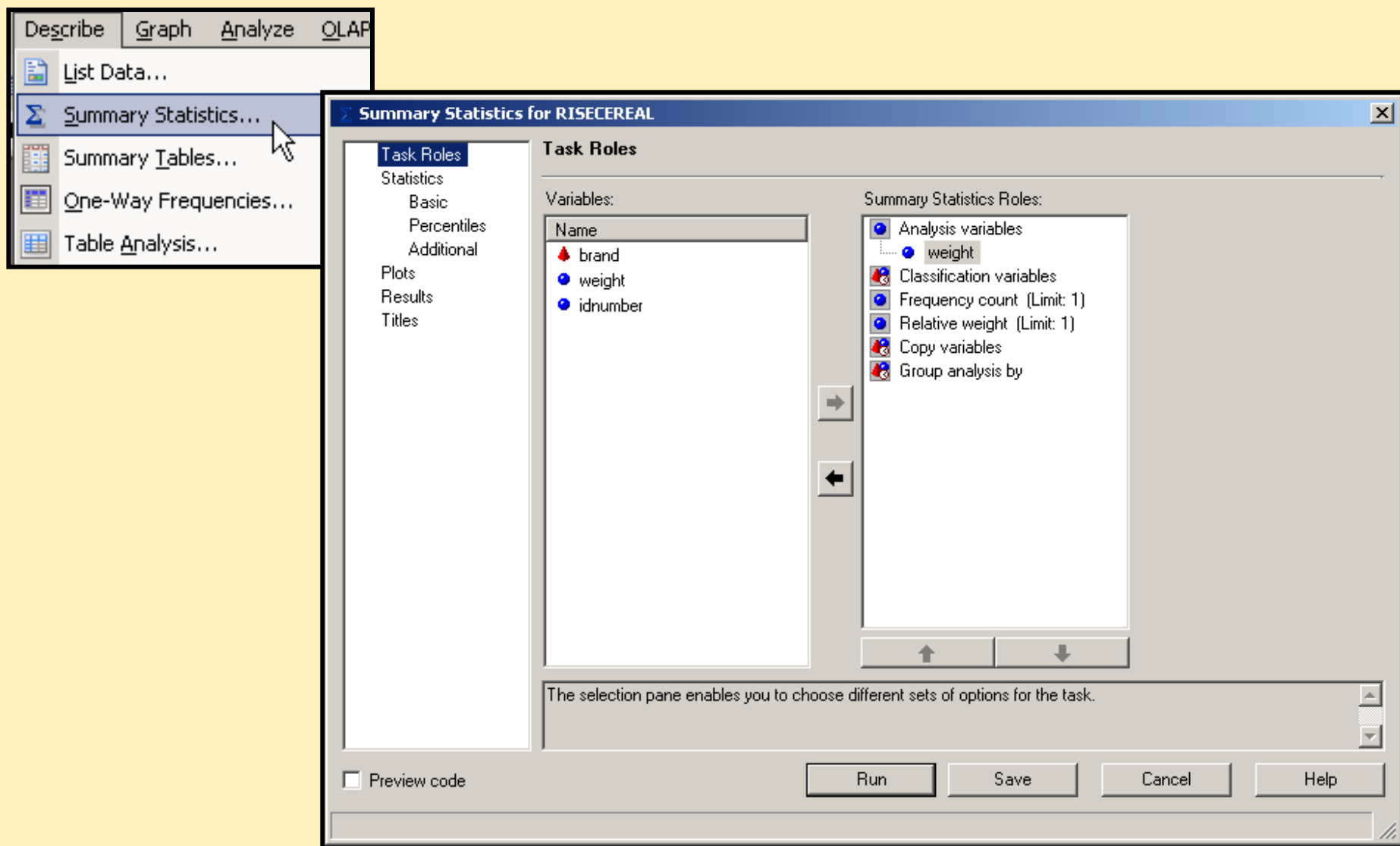
Nf : Coefficient of Variation

Coefficient of Variation

- ✓ ratio of the standard deviation to the mean
(standart dev to Expected Return)
- ✓ investment risk indicator for capital budgeting purposes.

$$\checkmark \quad \frac{\sigma}{\mu}$$

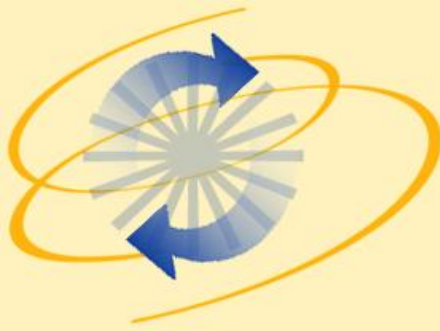
The Summary Statistics Task





Descriptive Statistics

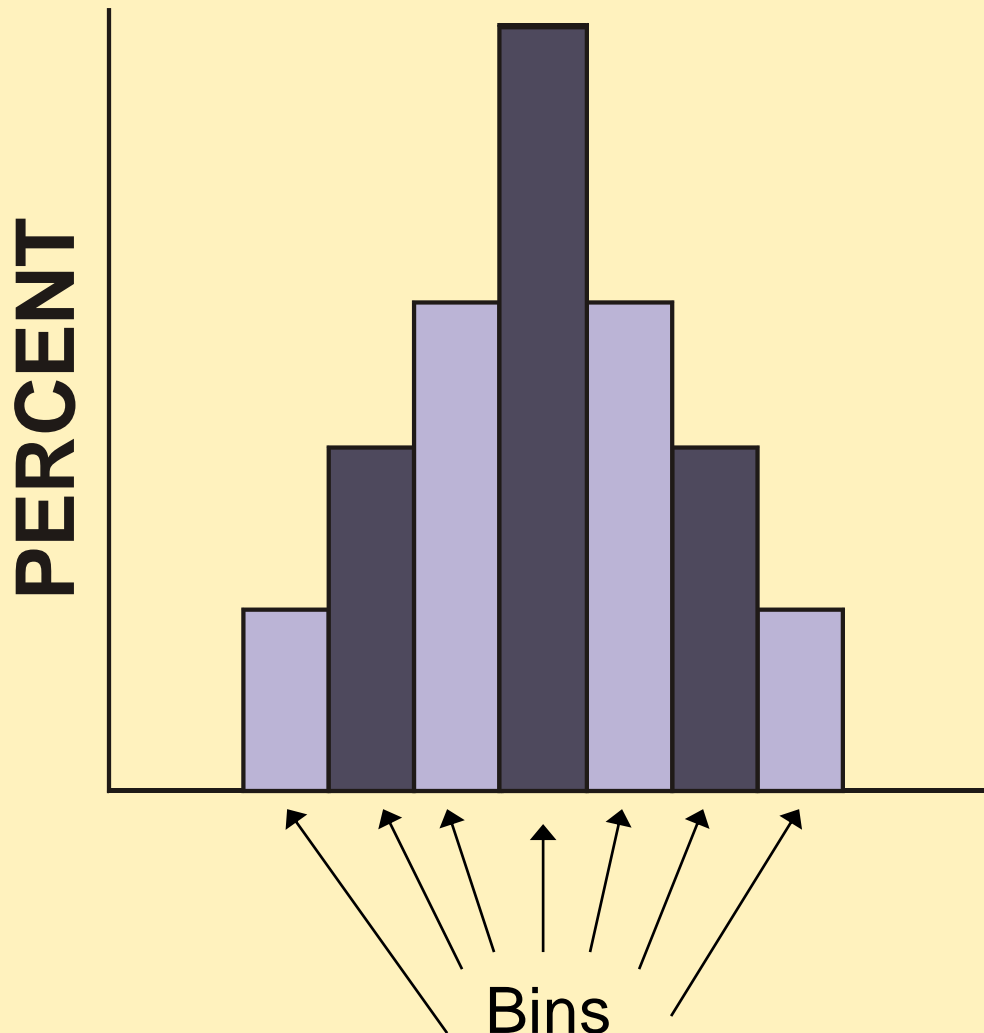
This demonstration illustrates creating the data sets for the course by running the SAS program in the class folder. Then use the Summary Statistics task to create descriptive statistics.



Exercises 1 and 2

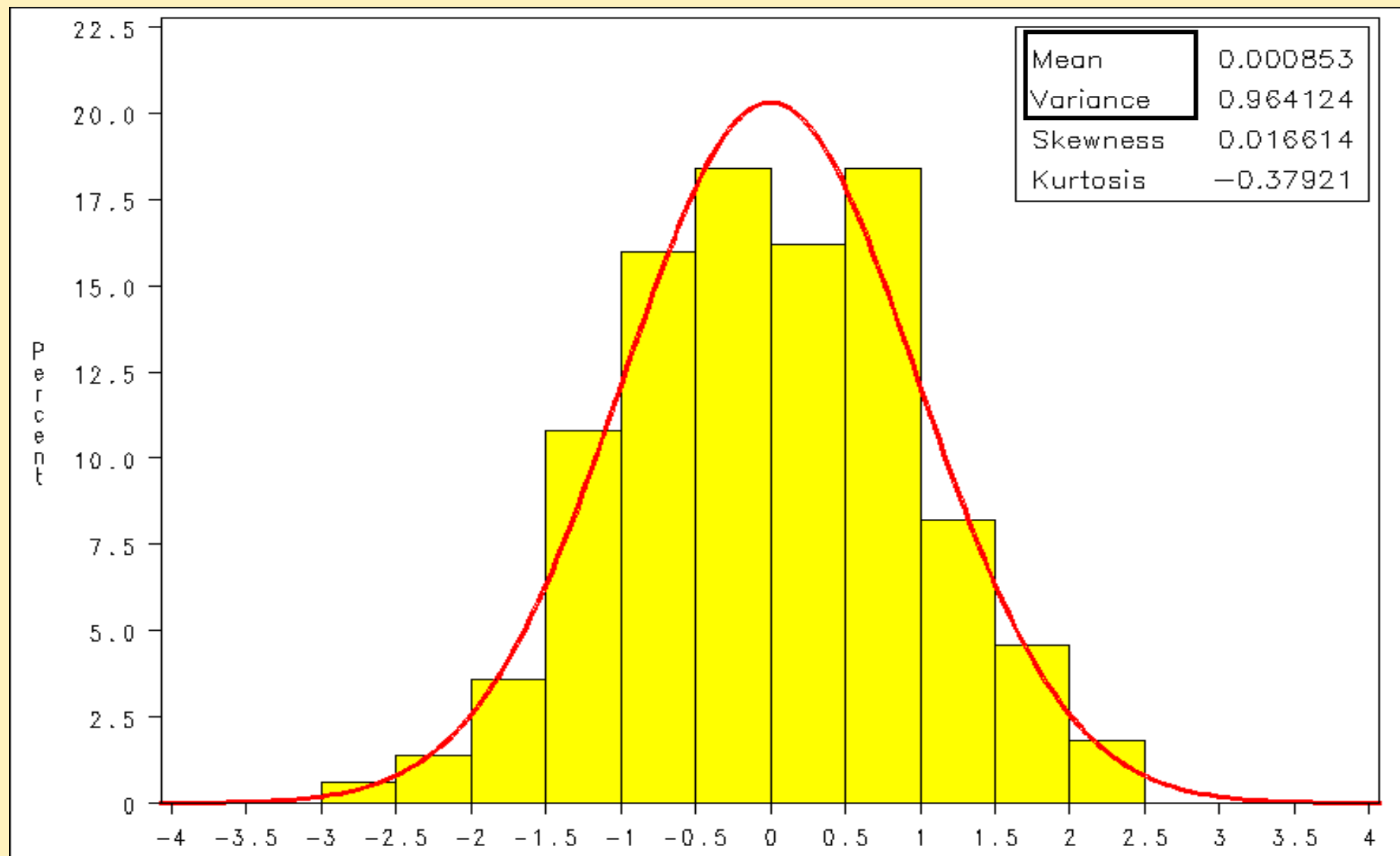
- Create a new project.
- Open and run the program EGBS.sas.
- After the program runs, view the **BOSTONNEWYORK** data set.
- Use the Summary Statistics task to create descriptive statistics for the variable **tottime**.

Picturing Distributions: Histogram



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

The Normal Distribution



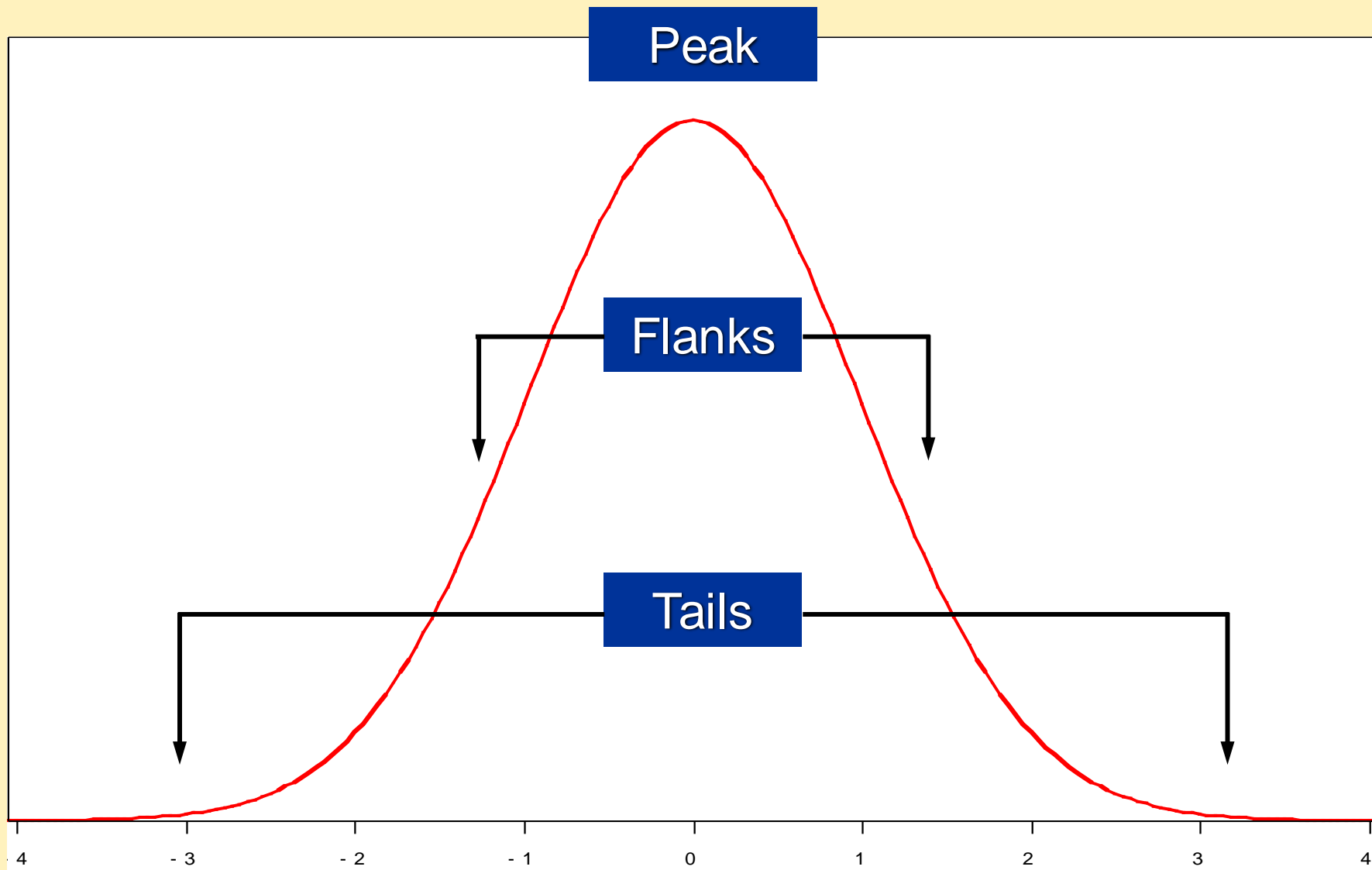
The Normal Distribution

The normal distribution

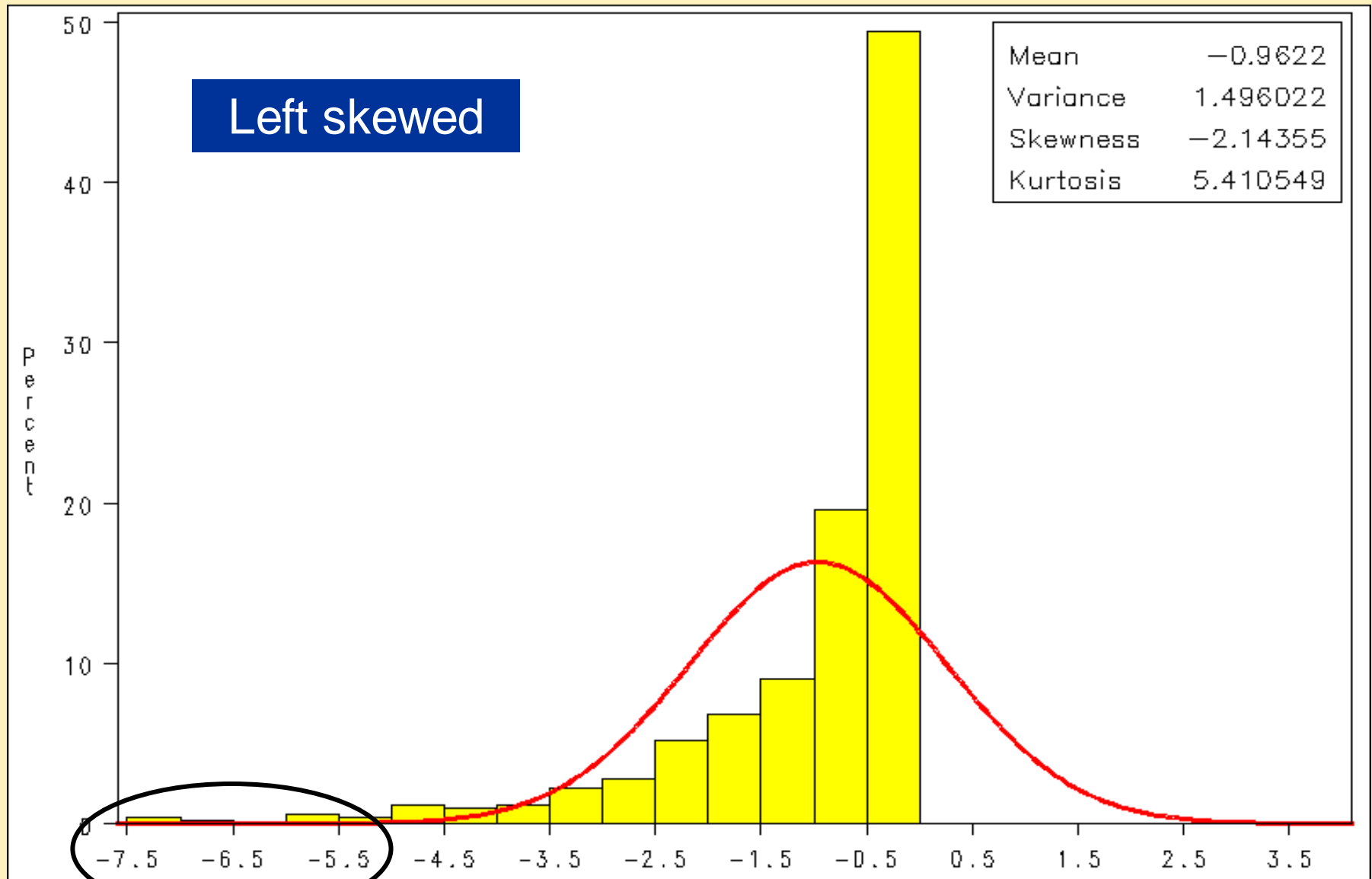
- is *symmetric*. If you draw a line down the center, you get the same shape on either side.
- is *fully characterized by the mean and standard deviation*. Given those two parameters, you know all there is to know about the distribution.
- is bell shaped.
- has mean \approx median \approx mode.

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.

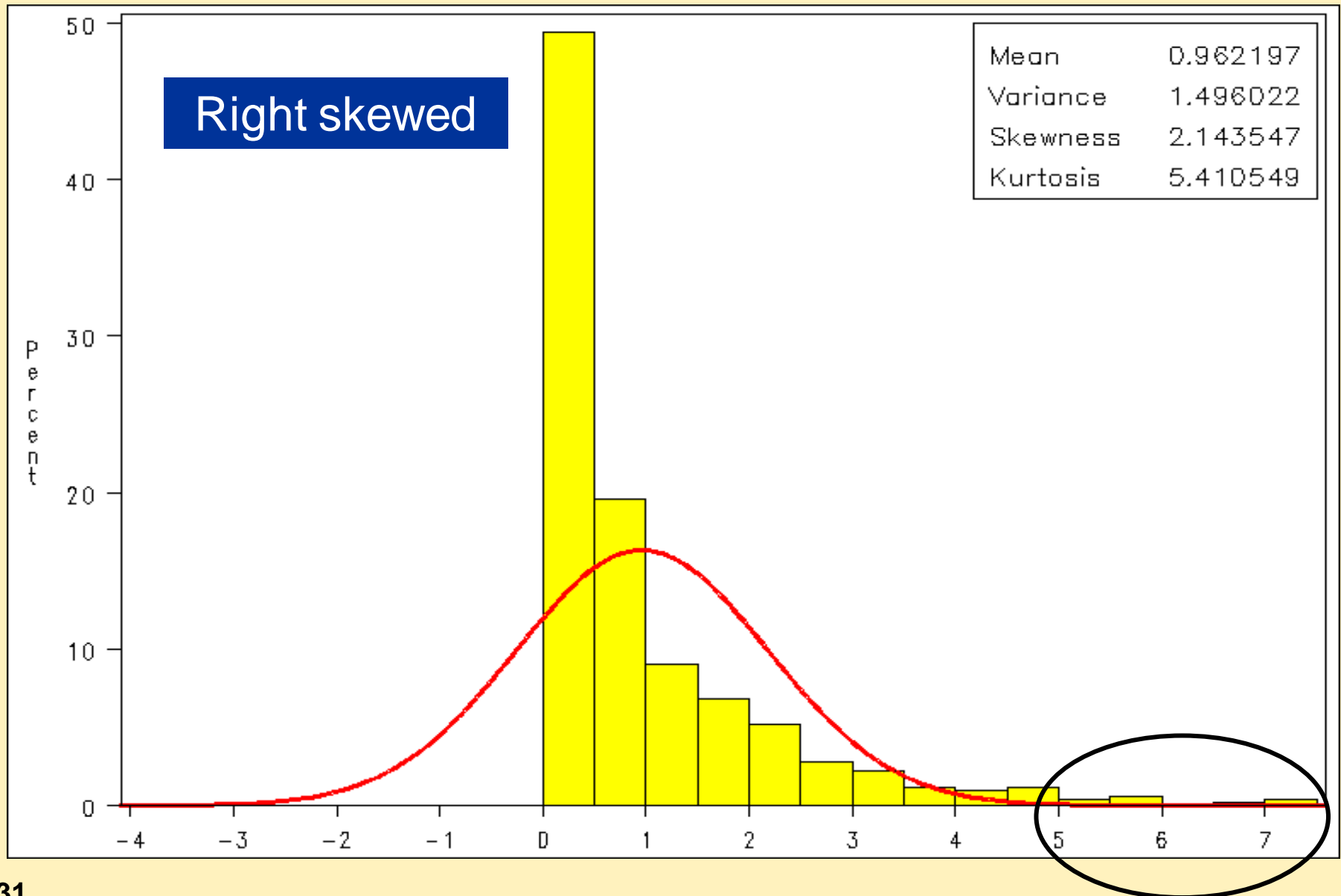
Characteristics of the Bell Curve



Measures of Shape: Skewness



Measures of Shape: Skewness

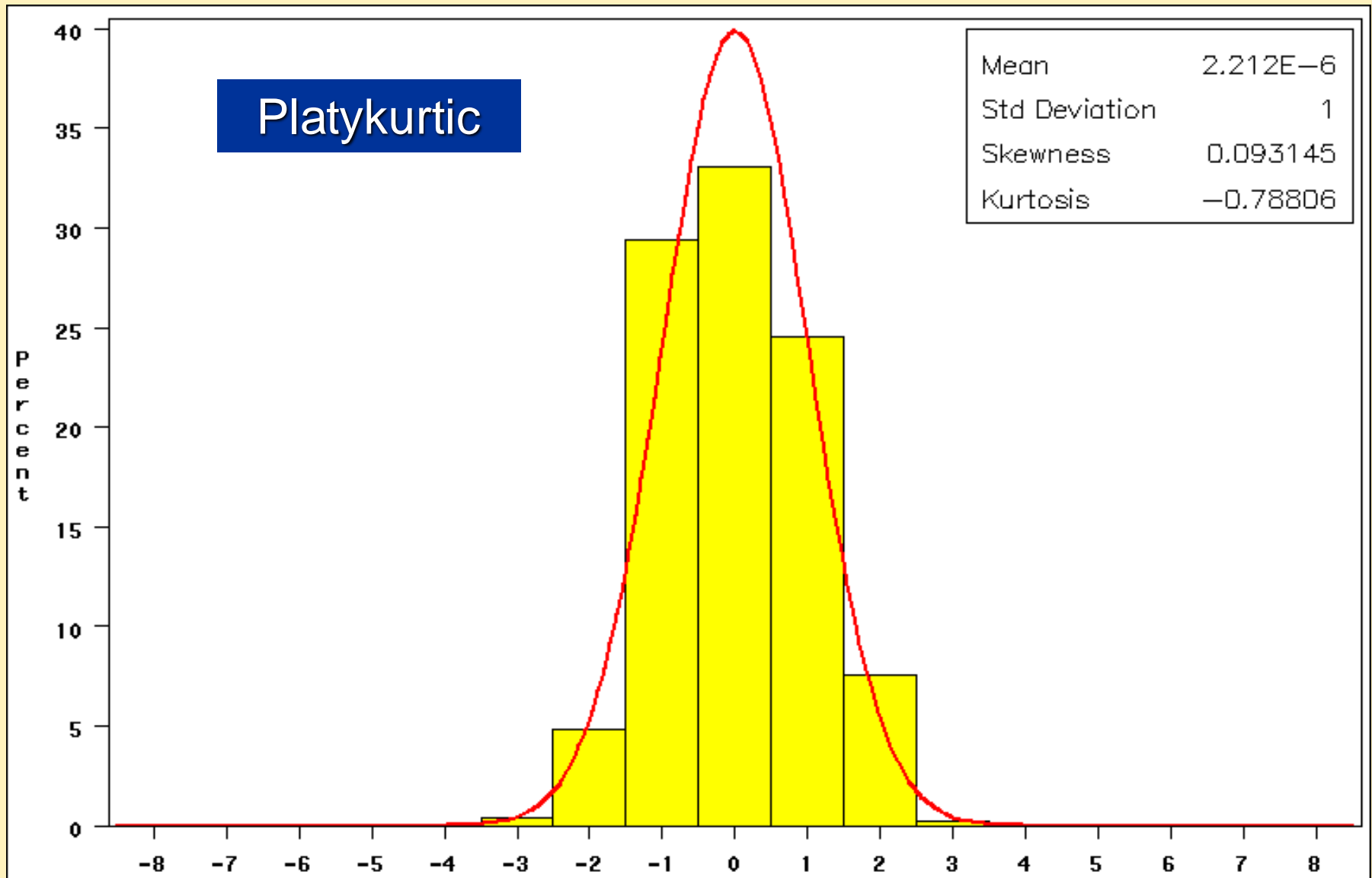


Skewness Risk

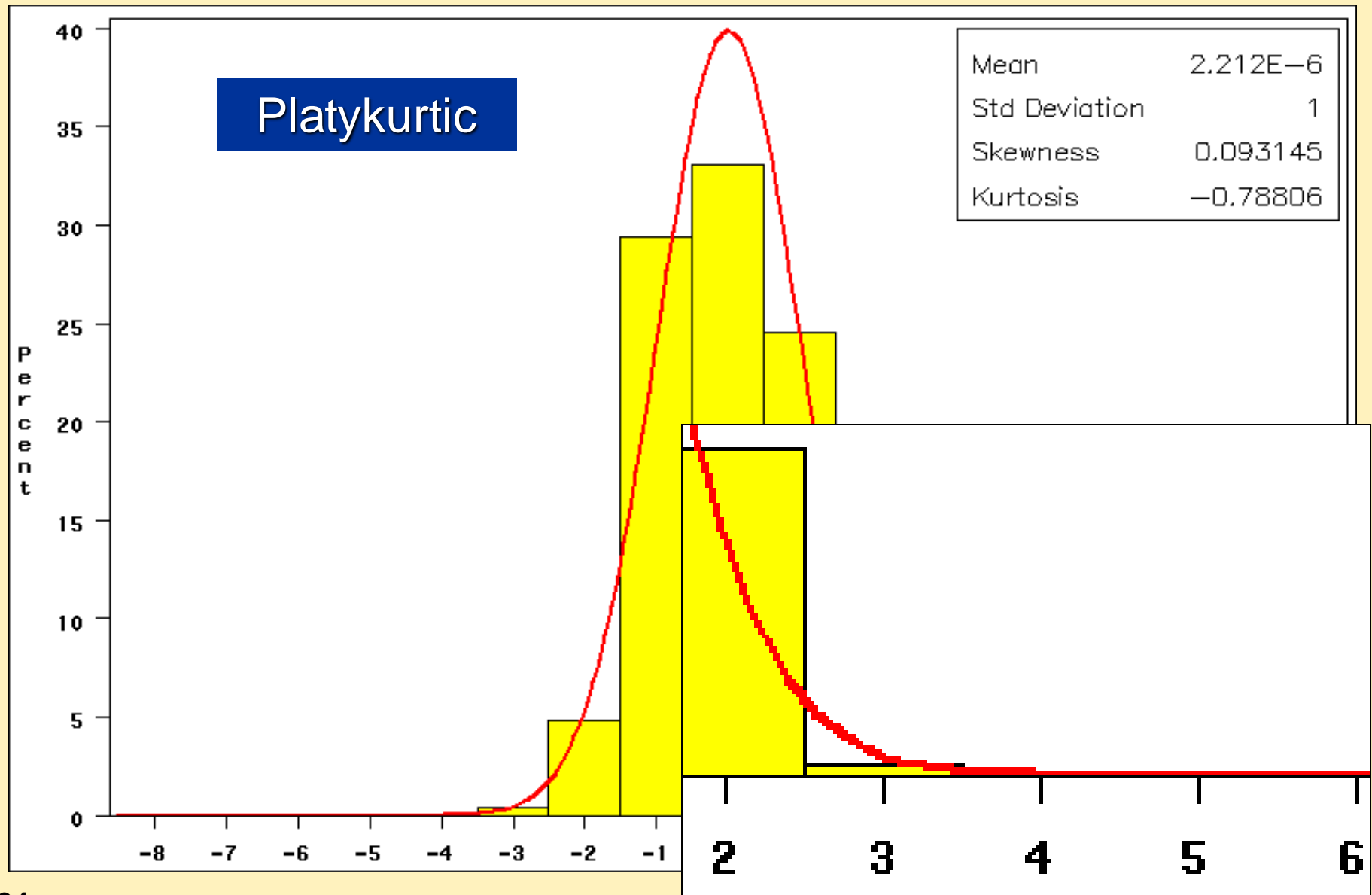
Skewness risk in financial modeling denotes that observations are not spread symmetrically around an average value.

Ignoring skewness risk, by assuming that variables are symmetrically distributed when they are not, will cause any model to understate the risk of variables with high skewness.

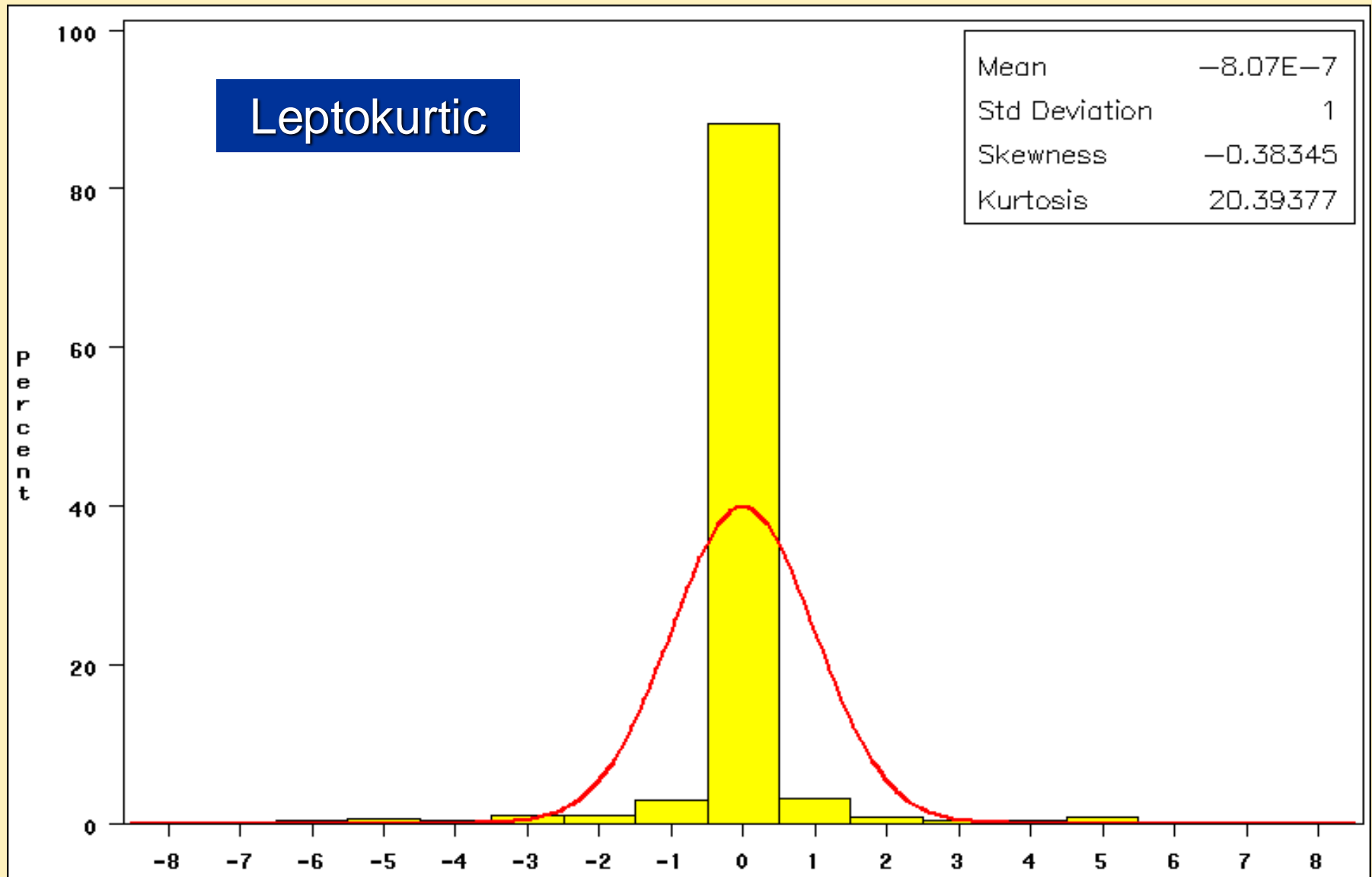
Measures of Shape: Kurtosis



Measures of Shape: Kurtosis



Measures of Shape: Kurtosis



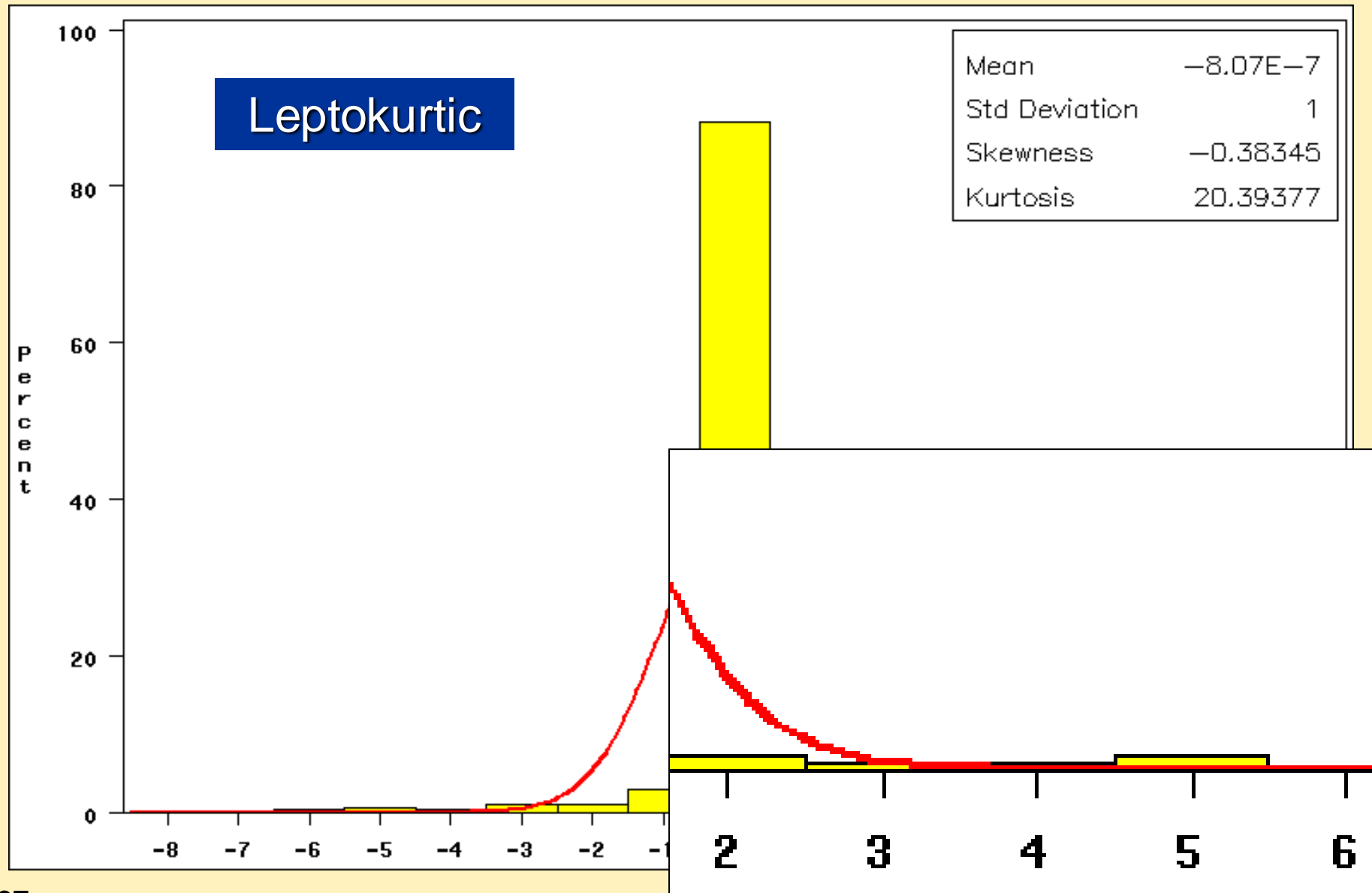
Kurtosis Risk

Observations are spread in a wider fashion than the normal distribution entails. In other words, fewer observations cluster near the average and more observations populate the extremes either far above or far below the average compared to the bell curve shape of the normal distribution.

Ignoring kurtosis risk will cause any model to understate the **risk** of variables with **high kurtosis**. For instance, **Long-Term Capital Management**, a hedge fund cofounded by **Myron Scholes**, ignored kurtosis risk to its detriment. After four successful years, this hedge fund had to be bailed out by major investment banks in the late 90s because it understated the kurtosis of many financial securities underlying the fund's own trading positions

Case: DOHOL returns Kurtosis Structure

Measures of Shape: Kurtosis



Poll: For a symmetric distribution, which of the following s...

*[PlaceWare Multiple Choice Poll. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*

Mean

Median

Either

The Distribution Analysis Task

The screenshot shows the SAS software interface with the 'Analyze' menu open and 'Distribution Analysis...' selected. The 'Distribution Analysis for RISECEREAL' dialog box is displayed, showing the 'Task Roles' tab. The 'Variables' list contains 'brand', 'weight', and 'idnumber'. The 'Distribution Analysis Roles' list includes 'Analysis variables', 'Group analysis by', 'Frequency count (Limit: 1)', 'Relative weight (Limit: 1)', and 'Classification variables (Limit: 2)'. The 'weight' variable is selected under 'Analysis variables'. The 'Preview code' checkbox is unchecked. The 'Run', 'Save', 'Cancel', and 'Help' buttons are at the bottom.

Analyze **QLAP** **Add-In** **Tools**

- Correlations...
- Distribution Analysis...**
- ANOVA
- Regression
- Multivariate
- Survival Analysis
- Capability Analysis
- Control Charts
- Pareto Chart...
- Time Series

Distribution Analysis for RISECEREAL

Task Roles

Task Roles

Variables:

Name
brand
weight
idnumber

Distribution Analysis Roles:

- ☒ Analysis variables
 - ☒ weight
- ☒ Group analysis by
- ☒ Frequency count (Limit: 1)
- ☒ Relative weight (Limit: 1)
- ☒ Classification variables (Limit: 2)

The selection pane enables you to choose different sets of options for the task.

☐ Preview code

Run **Save** **Cancel** **Help**



Examining Distributions

This demonstration illustrates how to create histograms using the Distribution Analysis task.

View/Application Share: AppShare: UNIVARIATE

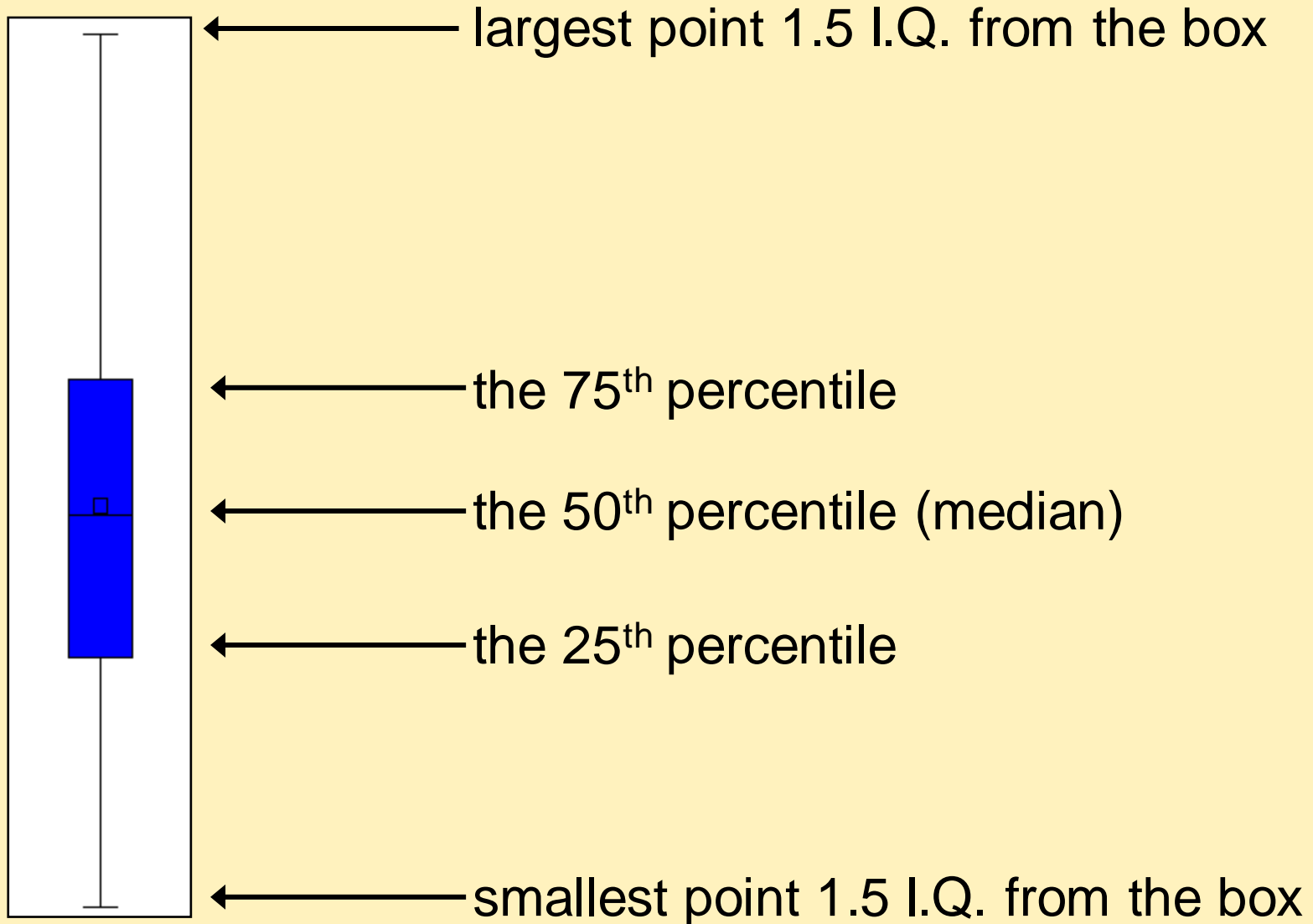
*[PlaceWare View/Application Share. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*

Graphical Displays of Distributions

You can produce three kinds of plots for examining the distribution of your data values:

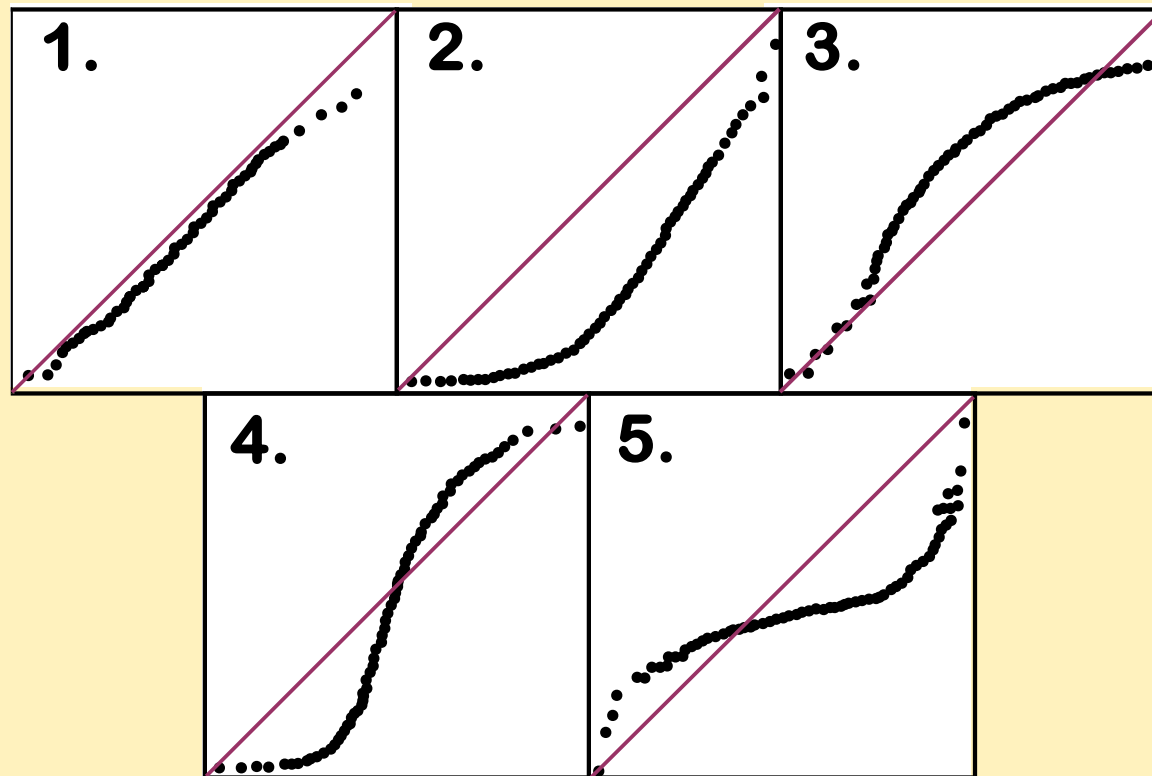
- histograms
- box plots
- normal probability plots

Box Plots



The mean is denoted by a square point.

Normal Probability Plots



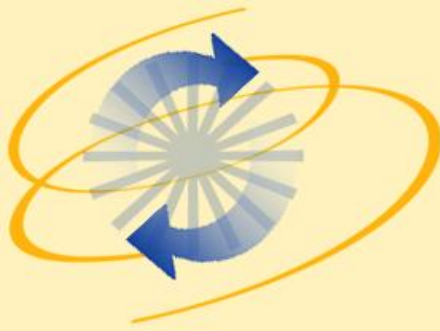


Examining Distributions

This demonstration illustrates how to create graphical displays of distributions using the Distribution Analysis task.

View/Application Share: AppShare: Box Plots and Probability P...

*[PlaceWare View/Application Share. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*



Exercises 3 and 4

- Create a new project and add the appropriate data sets for this session.
- Use the Distribution Analysis task to calculate statistics for the **BOSTONMARATHON** data set.
- Specify the appropriate options to compare each variable to a normal distribution.

Poll: According to your statistics, which of the following v...

*[PlaceWare Multiple Choice Poll. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*

age

tottime

firsthalf

secondhalf

None of the above

Section Summary

- Used the Summary Statistics and Distribution Analysis tasks to produce descriptive statistics.
- Interpreted measures of location, dispersion, and shape.
- Used the Distribution Analysis task to generate histograms, box plots, and normal probability plots.

Introduction to Statistics

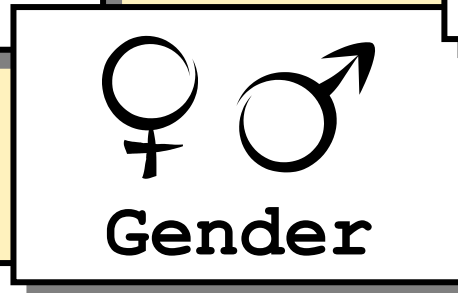
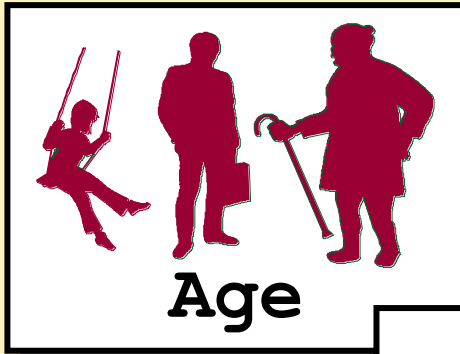


1: Fundamental Statistical Concepts

2: Examining Distributions

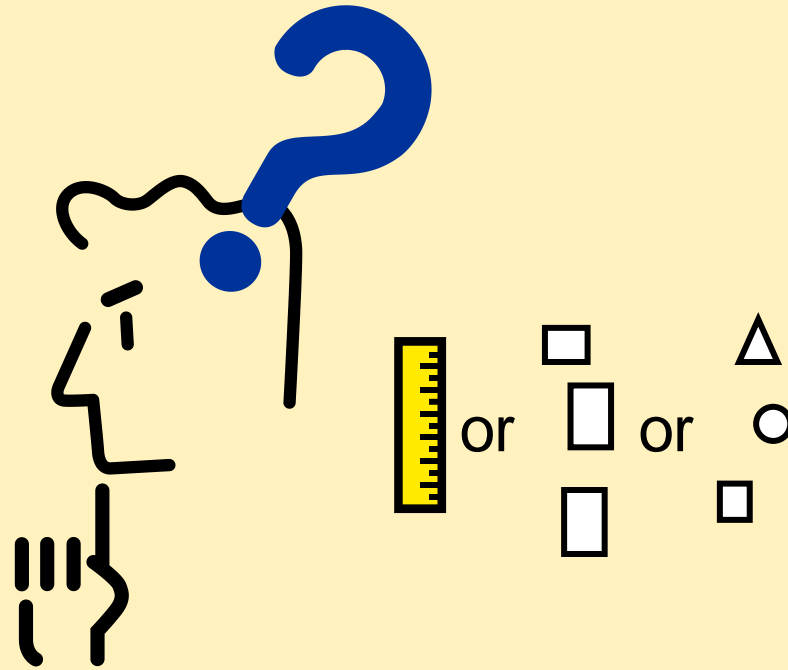
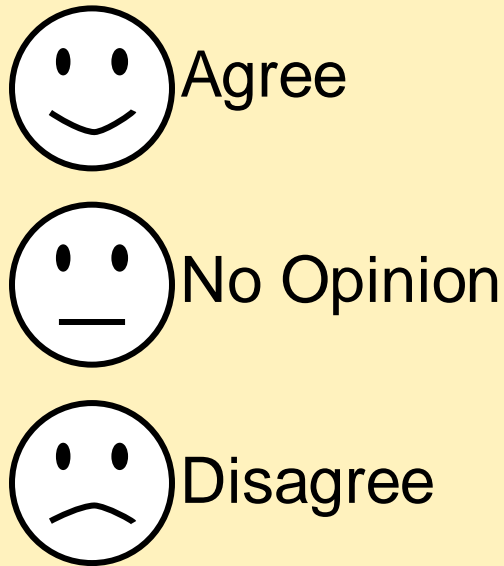
3: Describing Categorical Data

Sample Data Set



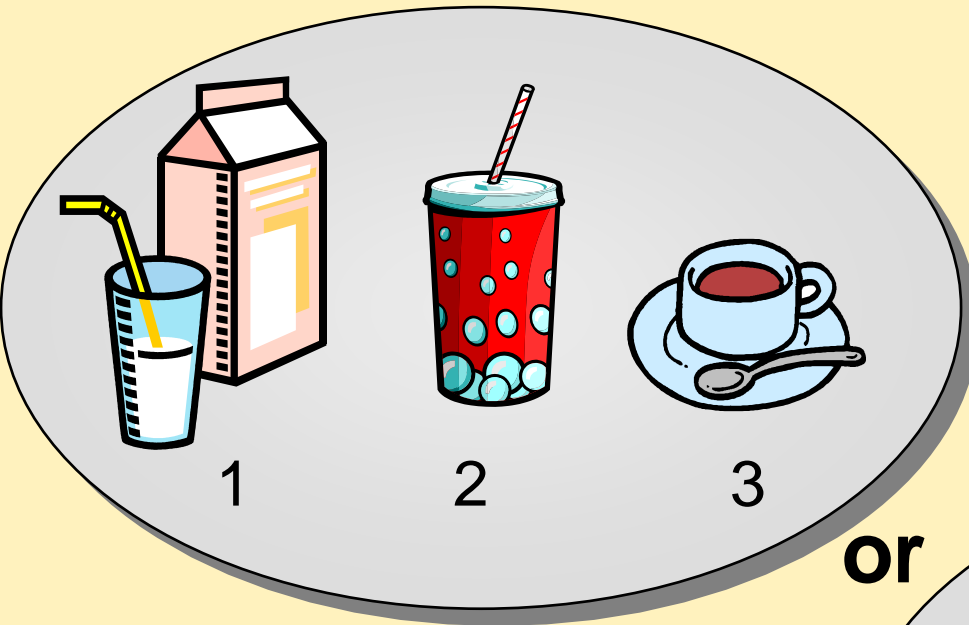
Identifying the Scale of Measurement

Variable



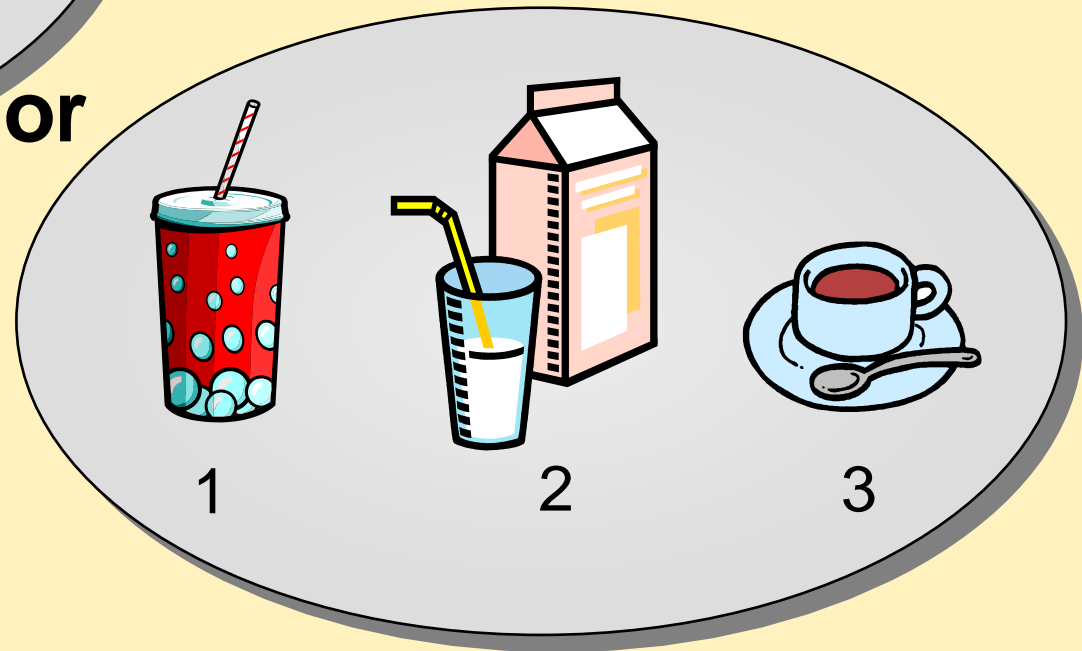
Before analyzing, identify the measurement scale for each variable.

Nominal Variables



Variable:
Kind of Beverage

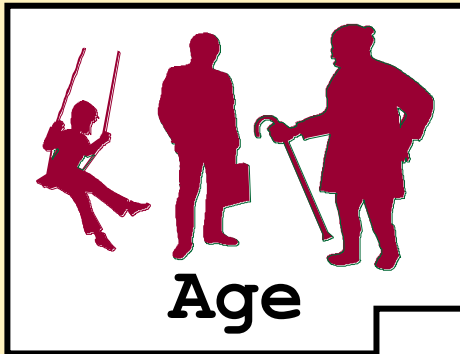
or



Order any way
you please!

Which Variables are Nominal?

Which of the variables in the example could be considered nominal?



Ordinal Variables

Variable: Size of Beverage



Small



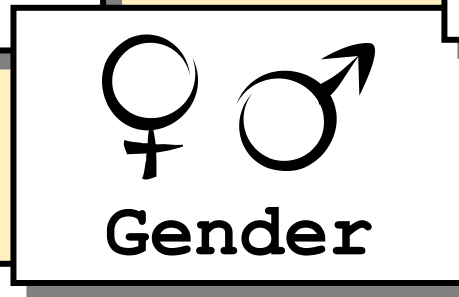
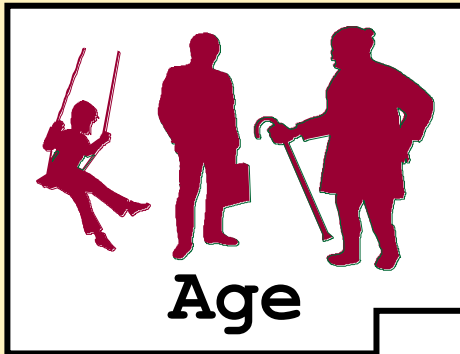
Medium



Large

Which Variables Are Ordinal?

Which of the variables in the example could be considered ordinal?



Which of the following variables in the data set is NOT categorical?

- gender
- age
- income
- purchase

Examining Categorical Variables

By examining the distribution of categorical variables, you can perform the following tasks:

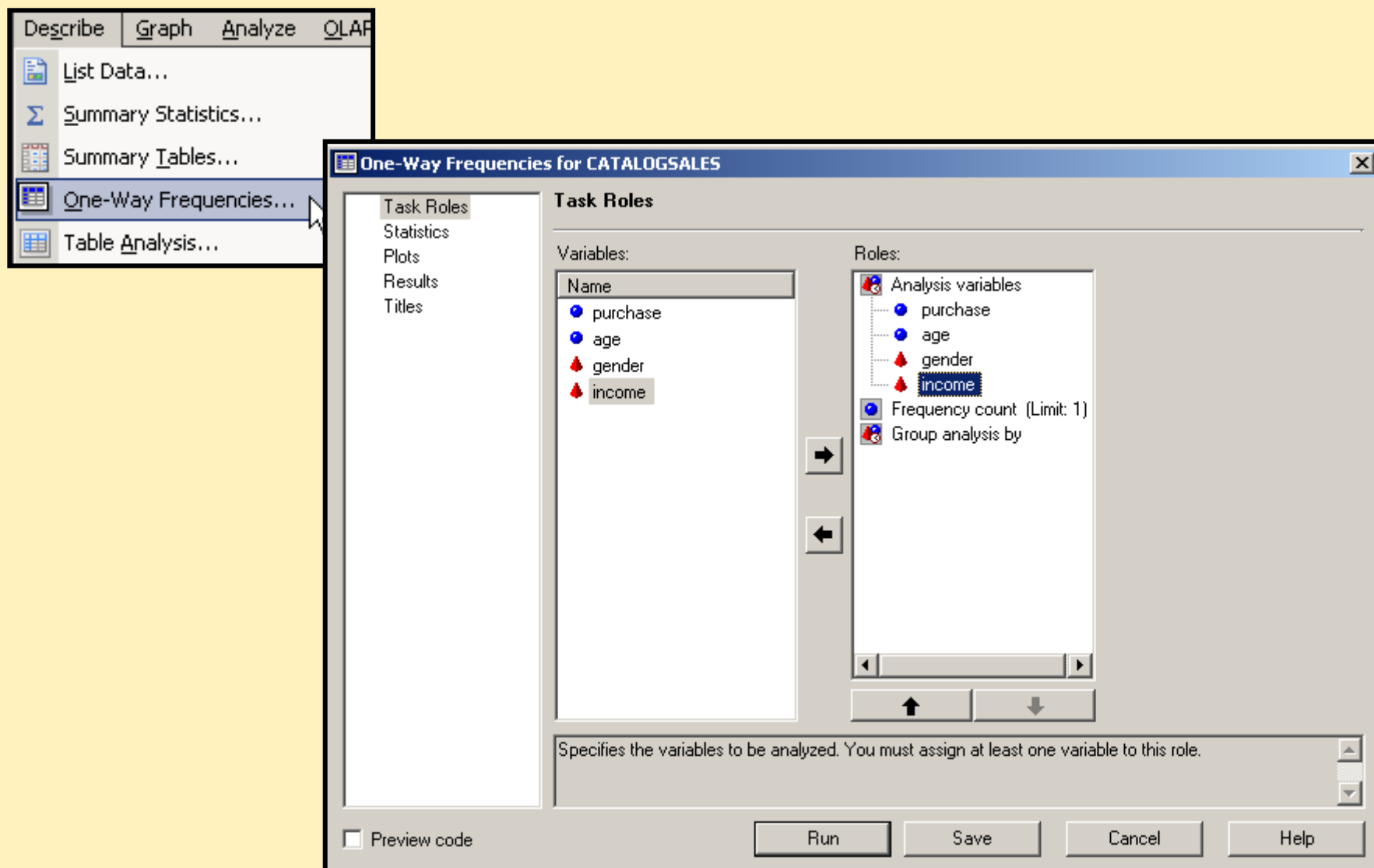
- Screen for unusual data values.
- Determine the frequency of data values.

Frequency Tables

A frequency table shows the number of observations that occur in certain categories or intervals. A one-way frequency table examines one variable.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

The One-Way Frequencies Task



Have you used the One-Way Frequencies task before?

- I never used it.
- I have used it somewhat.
- I use it often.



Examining Categorical Distributions

This demonstration illustrates using the One-Way Frequencies task to examine the **CATALOGSALES** data.

View/Application Share: AppShare: FREQ

*[PlaceWare View/Application Share. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*

Ordering Values

When you have an ordinal variable such as **income**, it is important to put the values in logical order for analysis purposes.

Present Order	Logical Order
---------------	---------------

High	Low
------	-----

Low	Medium
-----	--------

Medium	High
--------	------

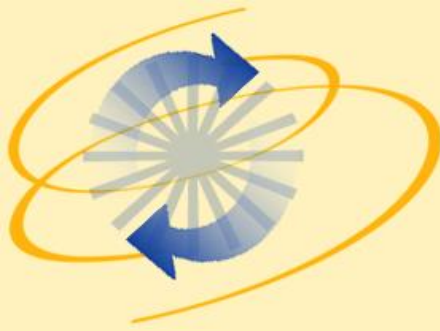


Ordering Values in the Frequency Table

This demonstration illustrates using a query to recode **income** as a variable named **inclevel** so that the sorted order corresponds to its logical order.

View/Application Share: AppShare:Re-Code Data

*[PlaceWare View/Application Share. Use **PlaceWare** > **Edit Slide Properties...** to edit.]*



Exercise 5

- View the **VEHICLESAFETY** data set.
- Determine the measurement scale for each variable.
- Use the One-Way Frequencies task to create descriptive statistics for the categorical variables.

Section Summary

- Explained the differences between categorical data and continuous data.
- Identified different scales of measurement for categorical variables.
- Presented methods for examining the distributions of categorical variables.

Chapter Summary

- Defined the difference between continuous and categorical variables.
- Described distributions for both continuous and categorical variables using statistics and graphics.
- Determined which SAS Enterprise Guide tasks and statistics were appropriate for each type of variable.

Appendix

✓ Probability Theory is a mathematical model for chance phenomena.