Statistics I: Introduction to ANOVA, Regression, and Logistic Regression

Course Notes

Statistics I: Introduction to ANOVA, Regression, and Logistic Regression Course Notes was developed by Peggy Cavalieri, Judith Jayawickrama, Bob Lucas, Mike Patetta, Kevin Scott, and Sue Walsh. Revisions were made by Paul Marovich, Mike Patetta, and Sue Walsh. Additional contributions were made by Chris Bond, Lynne Bresler, Brent Cohen, Duane Hayes, Wayne Thompson, and Doug Wielenga. This Live Web course was developed by Melinda Thielbar, with additional contributions by Paul Marovich and Catherine Truxillo. Editing and production support was provided by the Curriculum Development and Support Department.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

Statistics I: Introduction to ANOVA, Regression, and Logistic Regression Course Notes

Copyright © 2004 by SAS Institute Inc., Cary, NC 27513, USA. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

Book code 59901, course code LWSTAT1, prepared date 27Apr04.

Table of Contents

Course Description

This five-session Live Web course is designed for SAS software users who perform statistical analyses using SAS/STAT software. The course is a prerequisite to many of the courses in the statistical analysis curriculum. The course covers a range of statistical topics and the use of SAS software to carry out statistical analyses. Topics include statistical inference, analysis of variance, multiple regression, categorical data analysis, and logistic regression. You learn to construct graphs to explore and summarize data, construct confidence intervals for means and test simple hypotheses, and apply multiple comparison techniques.

During class, you practice with hands-on exercises in your own SAS session. Between sessions, you complete self-paced assignments to reinforce the concepts covered.

To learn more…

SAS Education

A full curriculum of general and statistical instructor-based training is available at any of the Institute's training facilities. Institute instructors can also provide on-site training.

For information on other courses in the curriculum, contact the SAS Education Division at 1-919-531-7321, or send e-mail to training@sas.com. You can also find this information on the Web at support.sas.com/training as well as in the Training Course Catalog.

SAS Publishing

For a list of other SAS books that relate to the topics covered in this Course Notes, USA customers can contact our SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the Publications Catalog on the Web at support.sas.com/pubs for a complete list of books and a convenient order form.

Prerequisites

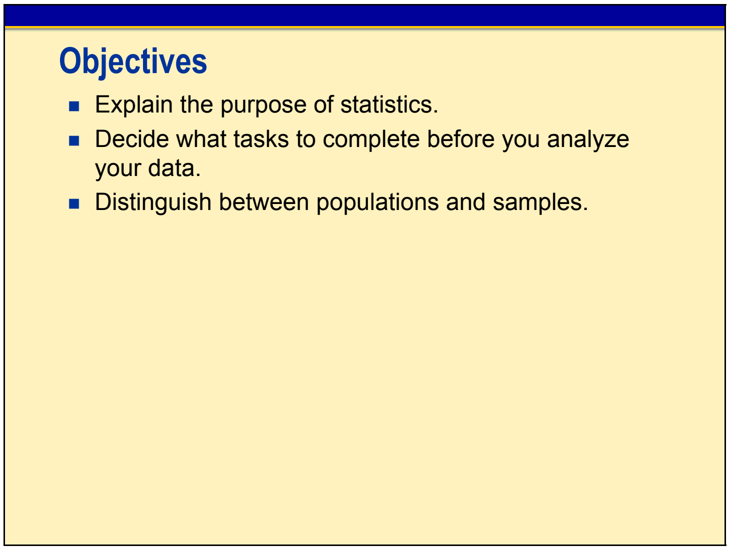
Before attending this course, you should

- have completed an undergraduate course in statistics covering *p*-values, hypothesis testing, analysis of variance, and regression
- be able to execute SAS programs and create SAS data sets.

You can gain the SAS experience by completing the *SAS® Programming I: Essentials* course.

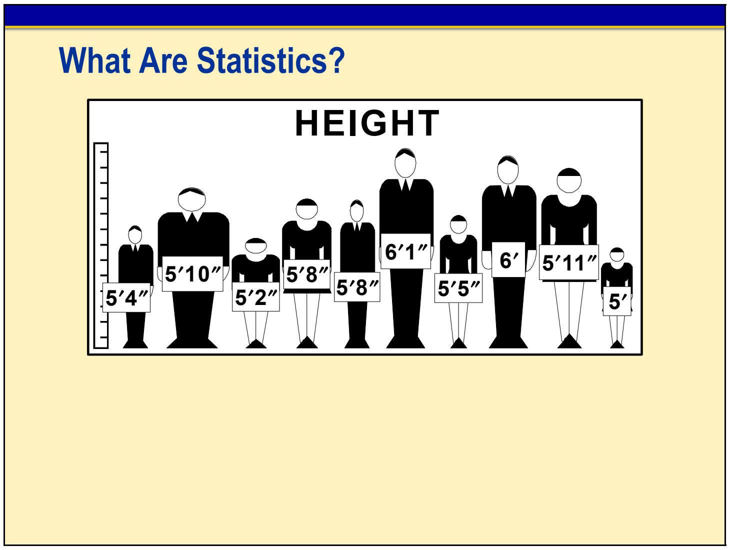
Module 1 Descriptive Statistics for Continuous and Categorical Data

1.1 Fundamental Statistical Concepts



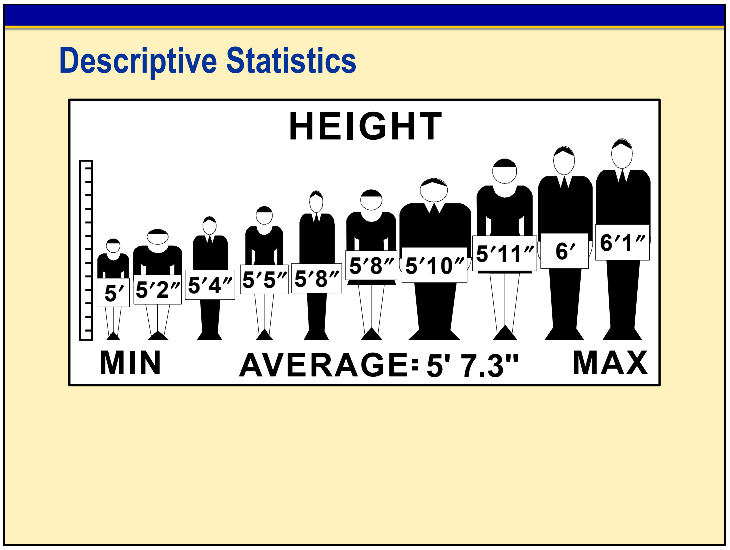
Objectives

- Explain the purpose of statistics.
- Decide what tasks to complete before you analyze your data.
- Distinguish between populations and samples.



What Are Statistics?

One purpose of statistics is to make sense of your data. Statistics provide information about your data so that you can answer questions and make informed decisions.

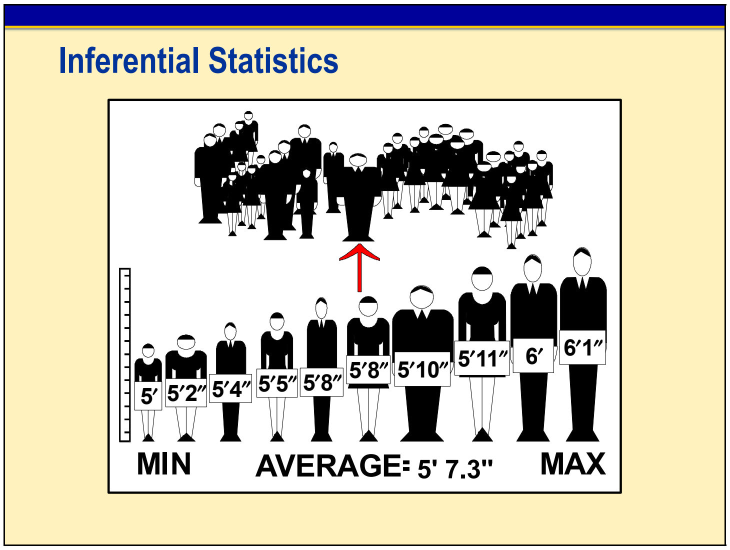


Descriptive Statistics

The discipline of statistics has these two broad categories:

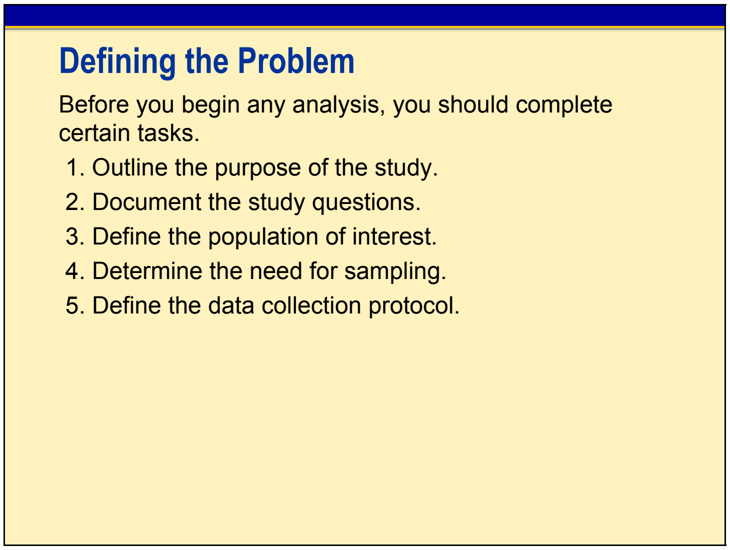
- descriptive statistics
- inferential statistics.

Descriptive statistics are used to organize, summarize, and focus on the main characteristics of your data, making it more usable.



Inferential Statistics

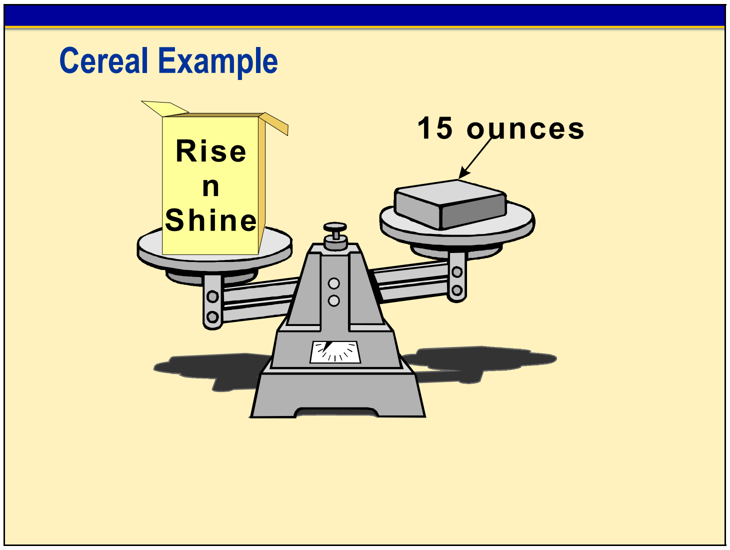
Inferential statistics make generalizations or inferences from your data to a larger set of data, based on probability theory.



Defining the Problem

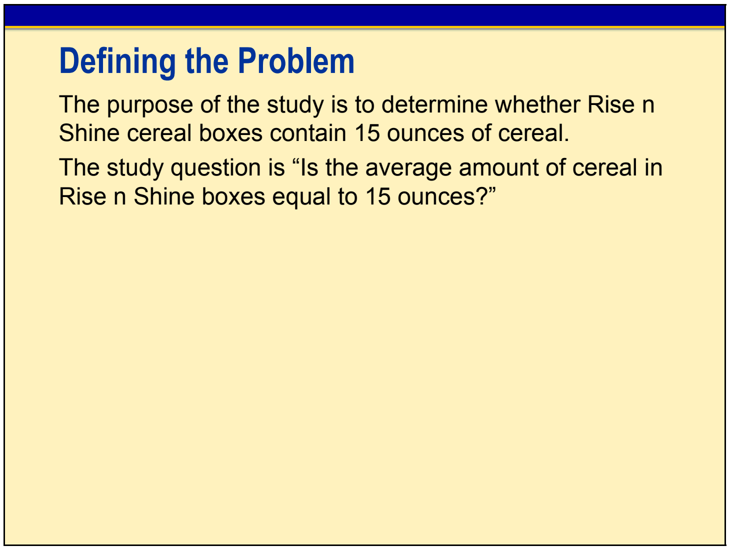
Before you begin any analysis, you should complete certain tasks.

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

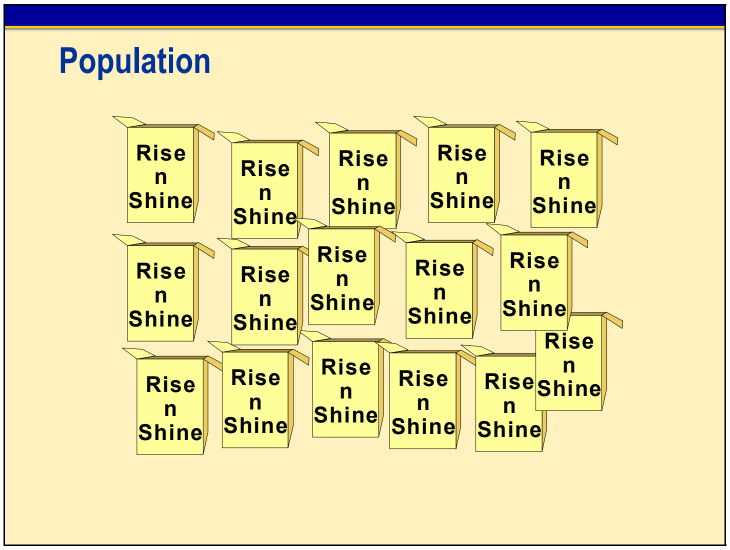
Example: A consumer advocacy group is questioning whether a brand of cereal named Rise n Shine contains the advertised amount of cereal per box. The box states that it contains 15 ounces of cereal. There are approximately one million boxes of Rise n Shine cereal in grocery stores.



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 “Is the average amount of cereal in Rise n Shine boxes equal to 15 ounces?”



Population

A *population* is the set of all measurement values of interest.

In the cereal example, the population is the number of ounces of cereal in each Rise n Shine cereal box, not the actual cereal boxes.

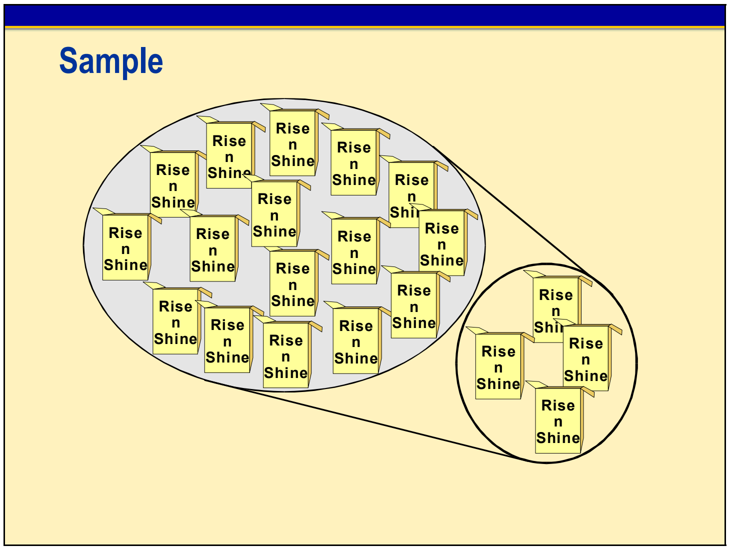
Populations can be categorized as either concrete or theoretical:

- A population is *concrete* if you can identify every subject in the population. For example, at any one point in time (for example, as of June 30, 1999), you can identify each person on the company payroll. These people constitute a concrete population.
- A population is *theoretical* if it is constantly changing. For example, because Rise n Shine cereal continues to be produced and packaged, the population changes almost continuously.

Because there are approximately one million cereal boxes in the grocery stores, you would need to record approximately one million measurements to examine the entire population.

Is it feasible to examine the entire population?

No, the population consists of approximately one million measurements. This would require too much time and too many resources to conduct the study and analyze the results.

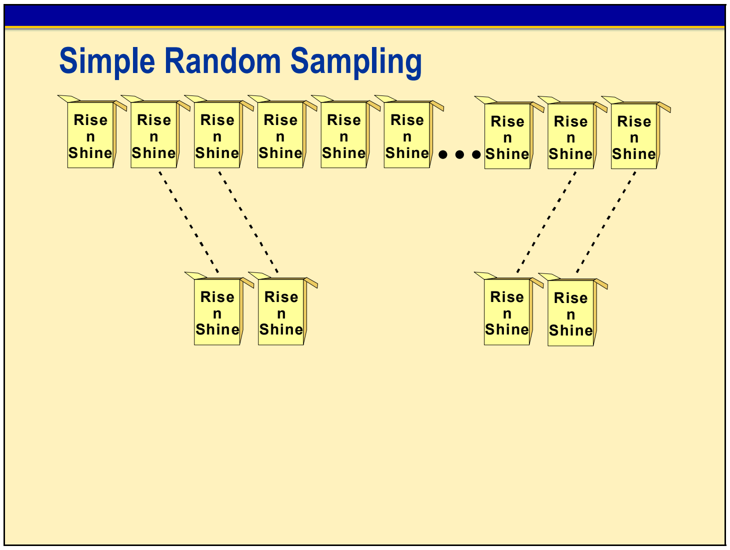


Sample

A *sample* is a subset of the population. The sample should be random to help ensure that it is representative of the population.

A *representative* sample has characteristics that are similar to the population’s characteristics.

For the cereal example, this means that the average weight of cereal in a representative sample of Rise n Shine boxes should be close to the average weight of all Rise n Shine boxes.



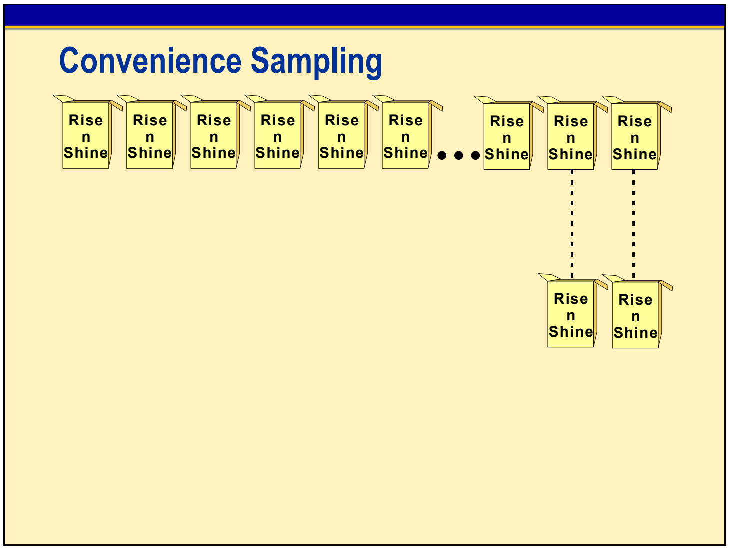
Simple Random Sampling

One sampling method that helps ensure a representative sample is *simple random sampling*.

In a simple random sample, every member of the population has an equal chance of being included.

In the cereal example, the number of ounces of cereal in each box has an equal chance of being selected from the population.

You can perform random sampling with and without replacement using the SURVEYSELECT procedure. See Appendix B, “Sampling Macros” for more information.



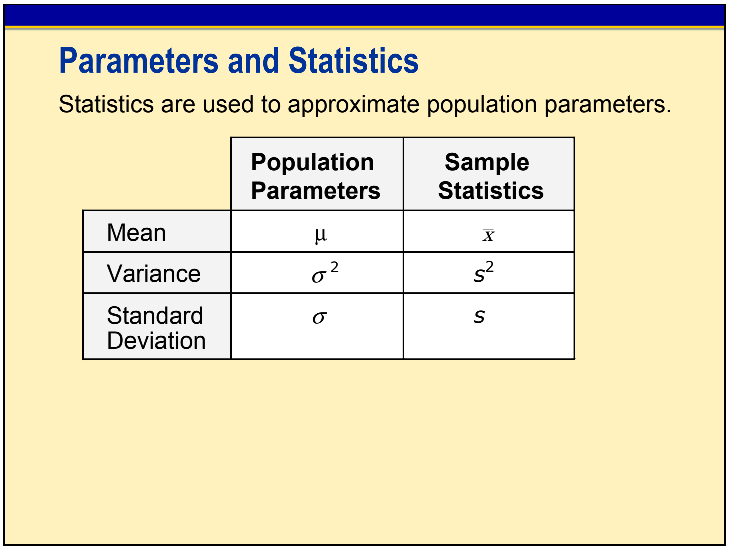
Convenience Sampling

Why not select cereal boxes from one grocery store near your home?

When you select values in a population that are easily available to you, you are using *convenience sampling*.

Convenience sampling might lead to biased samples. A *biased* sample is not representative of the population from which it is drawn.

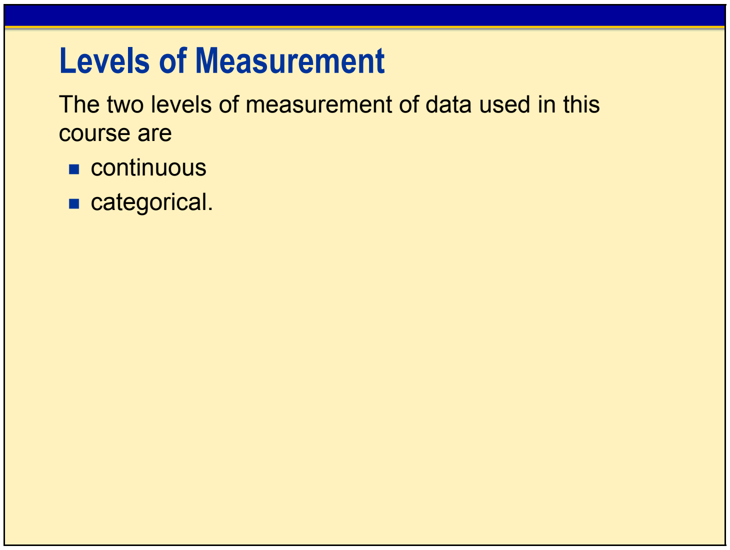
In the cereal example, the average weight of a biased sample might not be close to the true average of the population. This could cause the consumer advocacy group to draw erroneous conclusions about the cereal Rise n Shine.



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



Levels of Measurement

The two levels of measurement of data used in this course are

- continuous
- categorical.

In order to use the appropriate method of data summarization and data analysis, it is important to recognize the level of measurement of your data.

On a *continuous* scale,

- the variable has an unlimited number of possible values within a given range
- the values are numeric only.

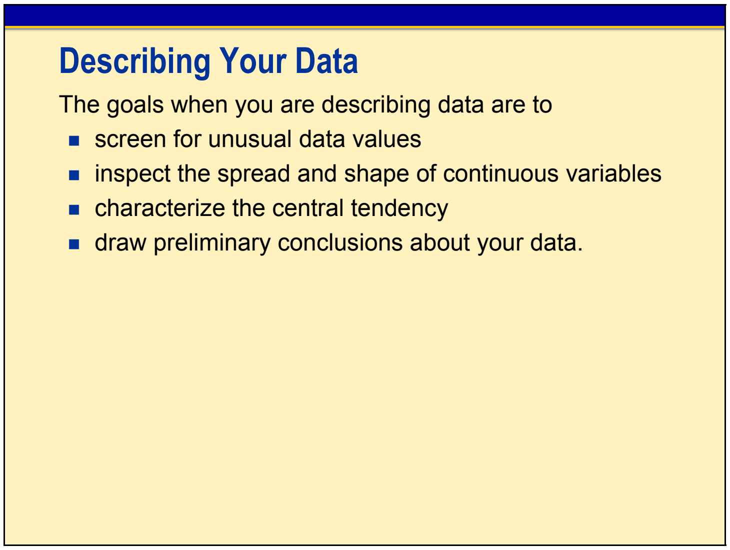
The variable for ounces of cereal is measured on a continuous scale.

On a *categorical* scale,

- the variable usually has a small number of distinct values within a given range
- the values can be character or numeric.

A variable such as brand of cereal is measured on a categorical scale.

Categorical data is also referred to as discrete data.



Describing Your Data

The goals when you are describing data are to

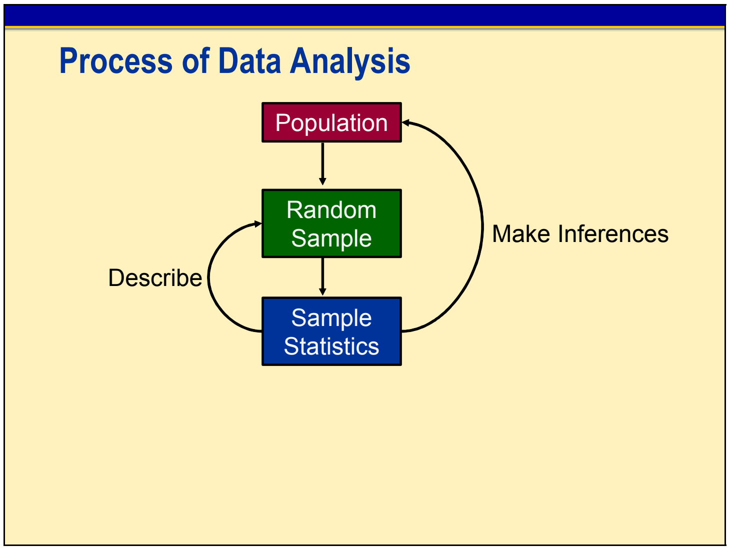
- screen for unusual data values
- inspect the spread and shape of continuous variables
- characterize the central tendency
- draw preliminary conclusions about your data.

After you select a random sample of the data, you can start describing the data. Although you want to draw conclusions about your population, you first want to explore and describe your data before you use inferential statistics.

Why?

- Data must be as error-free as possible.
- Unique aspects, such as data values that cluster or show some unusual shape, could be missed.
- An extreme value of a variable could be missed and cause gross errors in the interpretation of the statistics.

Some popular scientists have suggested that all great scientific discoveries have been due to outliers. An outlying observation indicates an event that is unexpected and does not follow existing theories. In resolving the anomaly, new theories are born.



Process of Data Analysis

Population

Random Sample

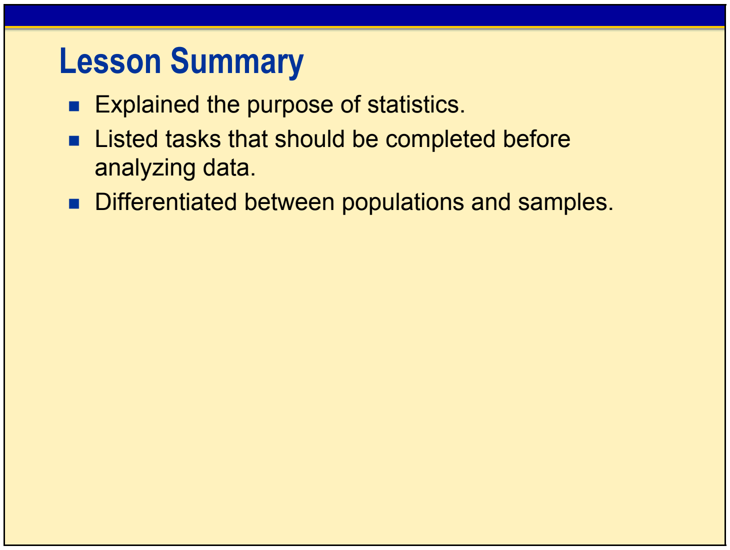
Sample Statistics

Describe

Make Inferences

These processes are involved in a statistical analysis:

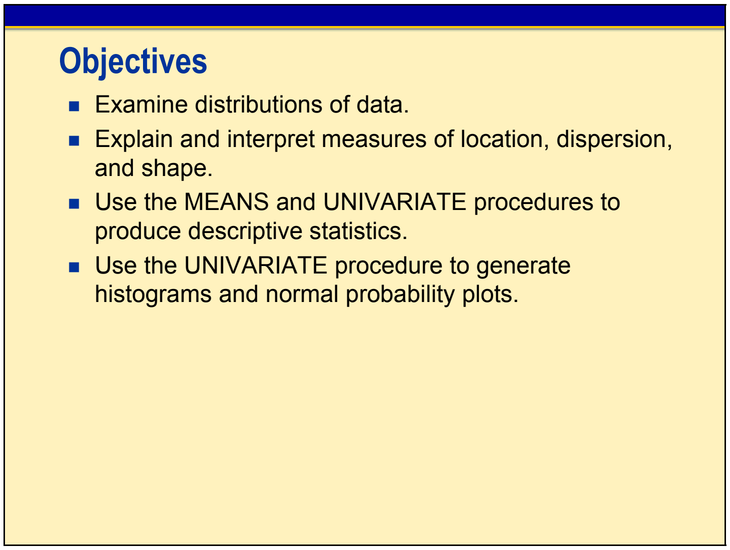
1. Identify the population of interest.
2. Draw a random sample.
3. Compute sample statistics to describe the sample.
4. Use sample information to make inferences about the population.



Lesson Summary

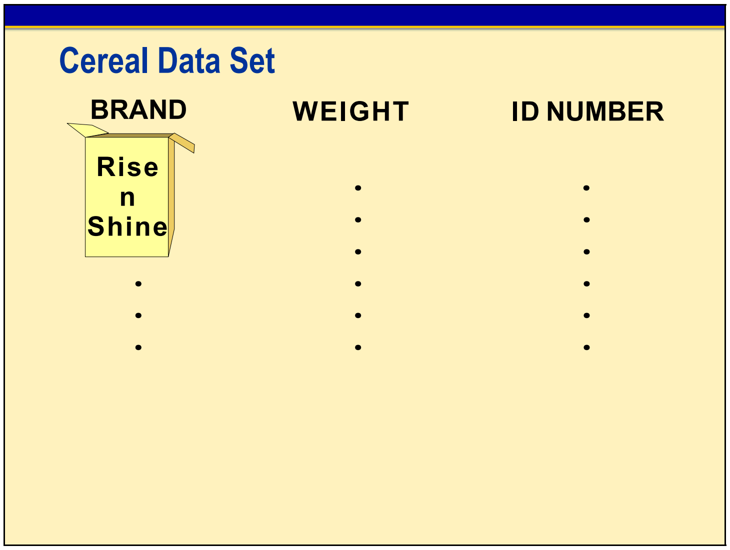
- Explained the purpose of statistics.
- Listed tasks that should be completed before analyzing data.
- Differentiated between populations and samples.

1.2 Examining Distributions



Objectives

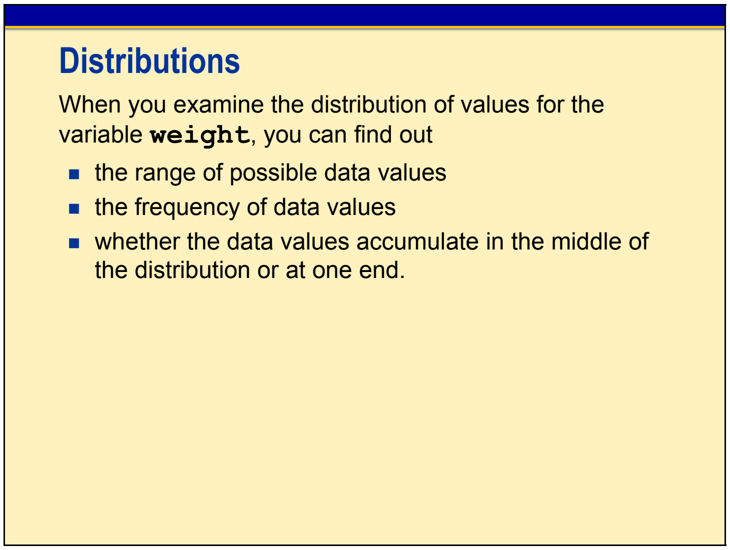
- Examine distributions of data.
- Explain and interpret measures of location, dispersion, and shape.
- Use the MEANS and UNIVARIATE procedures to produce descriptive statistics.
- Use the UNIVARIATE procedure to generate histograms and normal probability plots.



Cereal Data Set

BRAND WEIGHT ID NUMBER

Example: A consumer advocacy group wants to determine whether Rise n Shine cereal boxes contain 15 ounces of cereal. A random sample of 40 boxes was selected. The identification number of each box (**`idnumber`**) and the amount of cereal in ounces (**`weight`**) were recorded. The data is stored in the **`sasuser.b_rise`** data set.



Distributions

When you examine the distribution of values for the variable **`weight`**, you can find out

- 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.

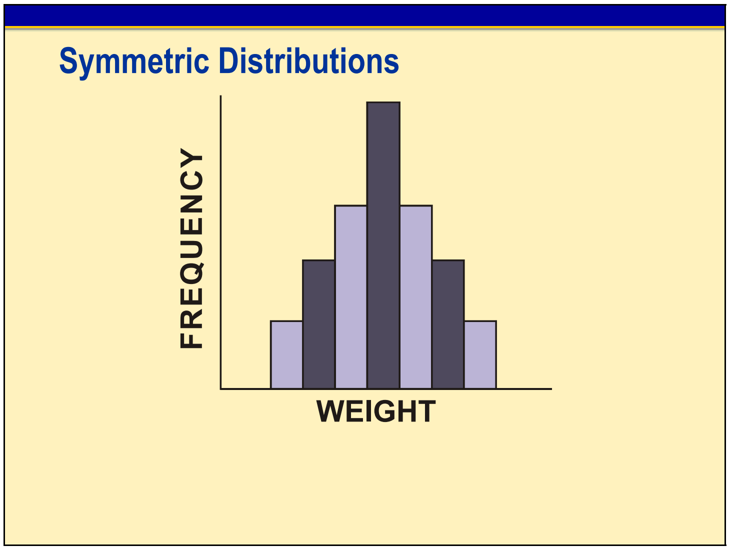
A *distribution* is a collection of data values arranged in order, along with the relative frequency. For any kind of problem, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.

For the cereal example, these questions can be addressed using graphical techniques.

- Are the values of **`weight`** symmetrically distributed?
- Are any values of **`weight`** unusual?

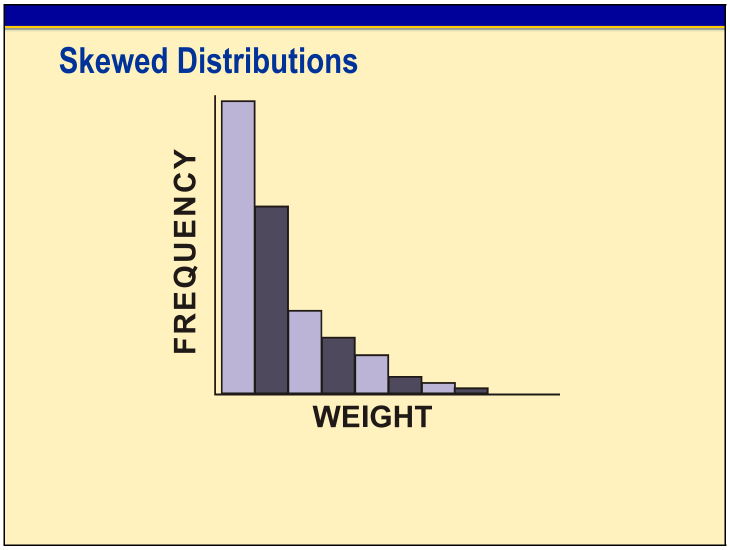
You can answer these questions using descriptive statistics.

- What is the best estimate of the average for **`weight`** for the population?
- What is the best estimate of the average spread or dispersion of the values of **`weight`** for the population?



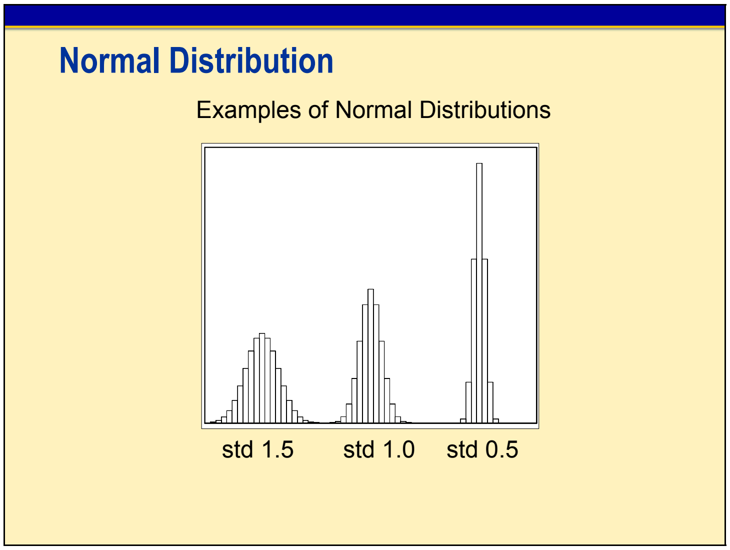
Symmetric Distributions

In a *symmetric distribution*, the right side of the distribution is a mirror image of the left side, and the mean is equal to the median.



Skewed Distributions

In a *skewed distribution*, many data values accumulate at one end of the distribution, and the mean is **not** equal to the median.



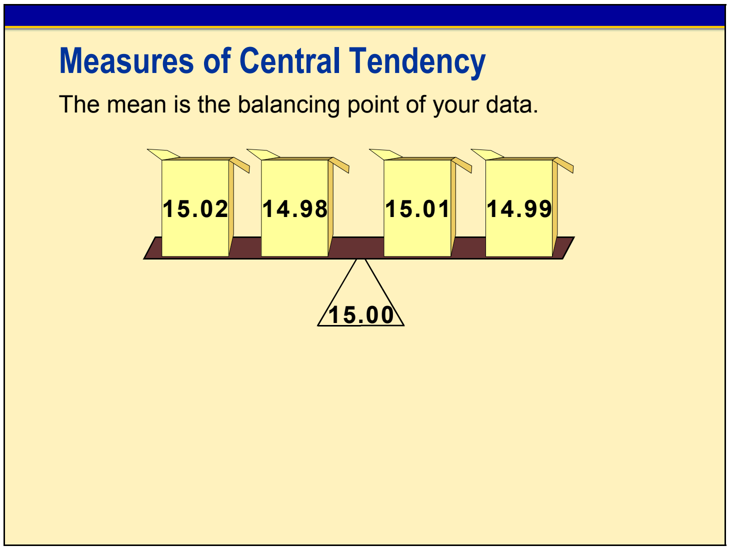
Normal Distribution

Examples of Normal Distributions

The *normal distribution*

- is bell-shaped and symmetric
- is completely characterized by its mean and standard deviation
- has mean=median=mode.

✎ The important properties of the normal distribution are part of the inferential statistics discussion.



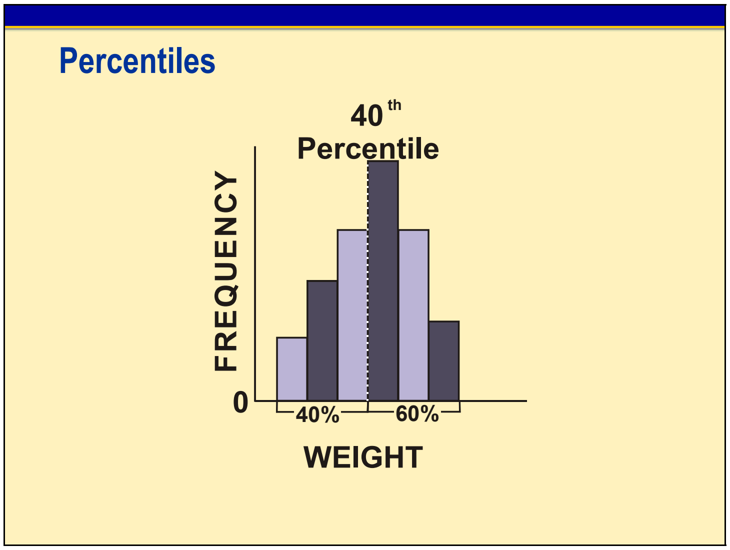
Measures of Central Tendency

The mean is the balancing point of your data.

Descriptive statistics that locate the center of your data are called *measures of central tendency*. The most common measure of central tendency is the sample mean.

A property of the sample mean is that the sum of the differences of each data value from the mean is always 0. That is, $\sum\left(Y_i - \bar{Y}\right) = 0$.

The mean is the physical balancing point of your data.



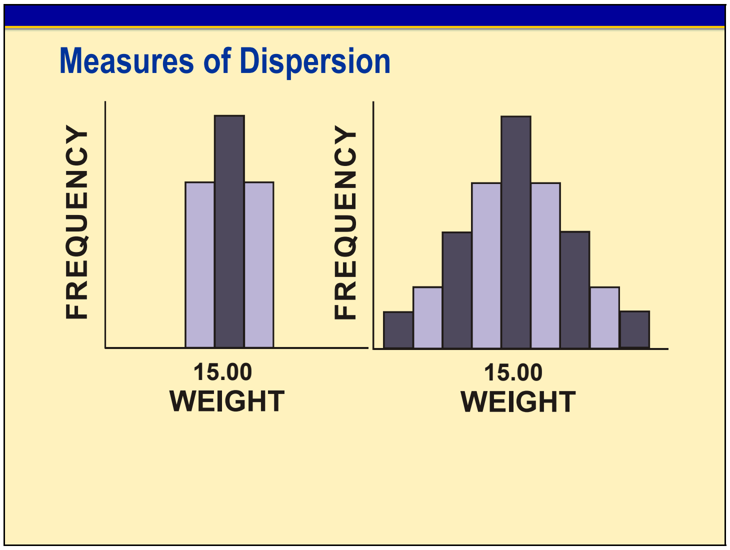
Percentiles

Percentiles locate a position in your data larger than a given proportion of data values.

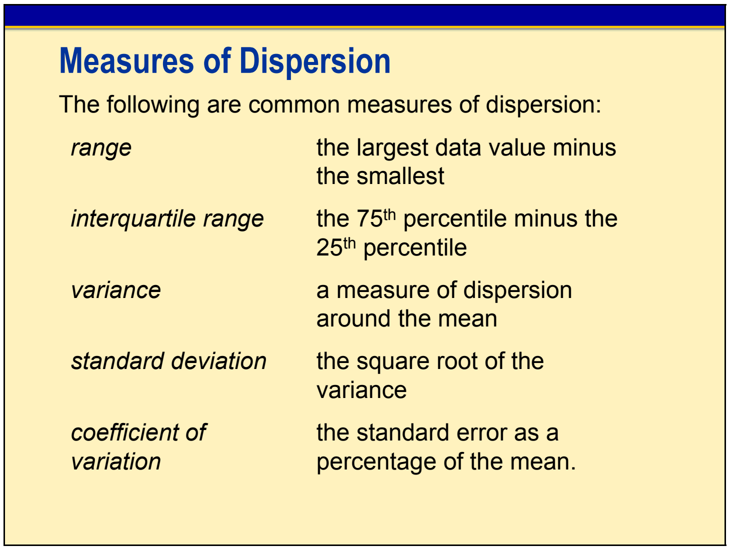
For example, in the graph above, 40% of the data values fall below or are equal to the 40th percentile, whereas 60% of the data values fall above the 40th percentile.

Commonly reported percentile values are

- the 25th percentile, also called the *first quartile*
- the 50th percentile, also called the *median*
- the 75th percentile, also called the *third quartile*.



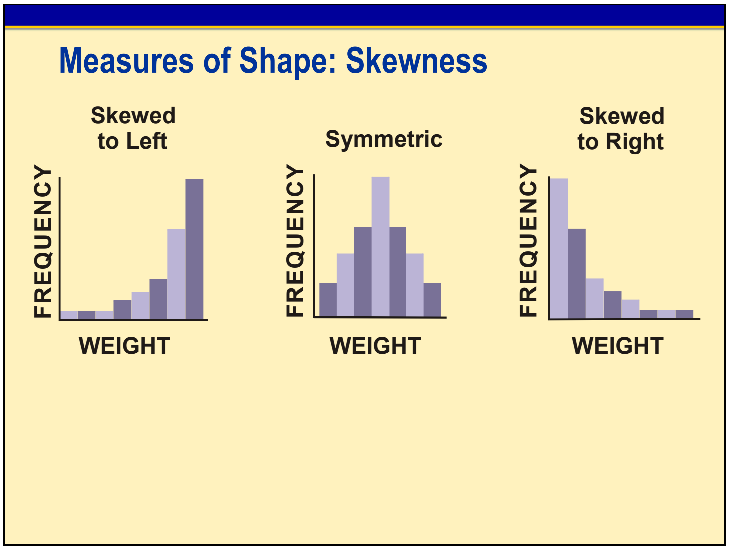
Measures of Dispersion



Measures of Dispersion

The following are common measures of dispersion:

range	the largest data value minus the smallest
interquartile range	the 75th percentile minus the 25th percentile
variance	a measure of dispersion around the mean
standard deviation	the square root of the variance
coefficient of variation	the standard error as a percentage of the mean.



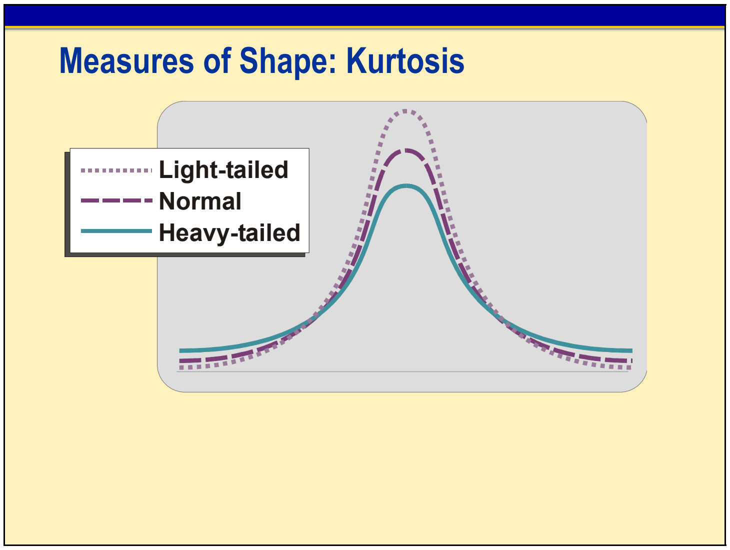
Measures of Shape: Skewness

Measures of shape describe the shape of your distribution. Two common measures of shape are the skewness and kurtosis statistics.

The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to 0.

If your distribution is more spread out on the

- left side, the statistic is negative, and the mean is less than the median
- right side, the statistic is positive, and the mean is greater than the median.

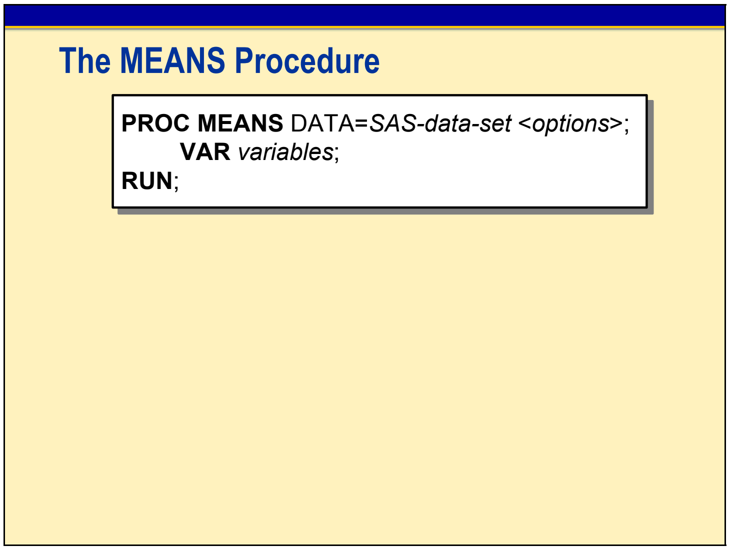


Measures of Shape: Kurtosis

The *kurtosis* statistic measures the tendency of your data to be distributed toward the tails, or ends, of the distribution. A distribution that is approximately normal has a kurtosis statistic close to 0.

If your distribution has

- heavy tails compared to the normal distribution, the statistic is positive
- light tails compared to the normal distribution, the statistic is negative.



The MEANS Procedure

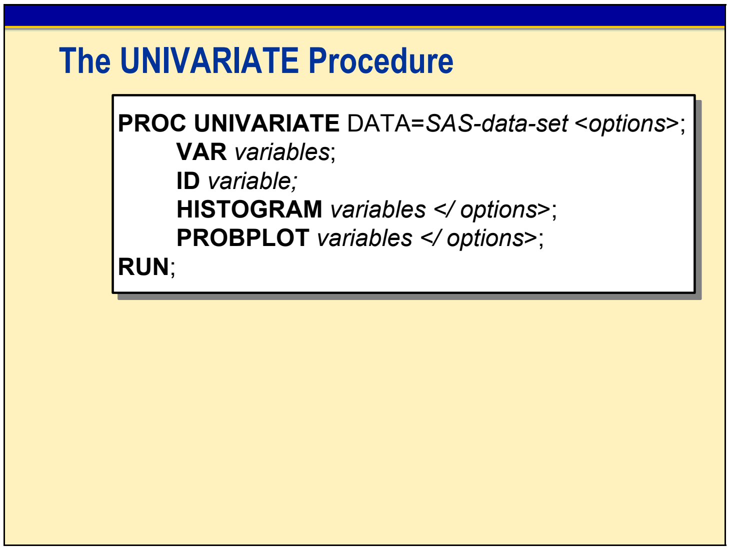
```
PROC MEANS DATA=SAS-data-set <options>;
       VAR variables;
RUN;
```

The MEANS procedure is a Base SAS procedure for generating descriptive statistics of your data.

Selected MEANS procedure statement:

VAR specifies numeric variables for which you want to calculate descriptive statistics. If no VAR statement appears, all numeric variables in the data set are analyzed.

For assistance with the correct syntax and options for a SAS procedure you can type **`help`** followed by the name of the procedure in the command box. This opens the Help window for that procedure. When you are in the appropriate Help window, select **syntax** to see all options available for that procedure.



The UNIVARIATE Procedure

```
PROC UNIVARIATE DATA=SAS-data-set <options>;
        VAR variables;
        ID variable;
        HISTOGRAM variables </ options>;
        PROBPLOT variables </ options>;
RUN;
```

The UNIVARIATE procedure not only computes descriptive statistics; it also provides greater detail on the distributions of the variables.

Selected UNIVARIATE procedure statements:

VAR specifies numeric variables to analyze. If no VAR statement appears, all numeric variables in the data set are analyzed.

ID specifies a variable used to label the five lowest and five highest values in the output.

HISTOGRAM creates high-resolution histograms.

PROBPLOT creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution.

Descriptive Statistics

m1demo01.sas, m1demo02.sas, m1demo03.sas

Example: Use the PRINT procedure to list the first 10 observations in the data set **`sasuser.b_rise`**. Then use PROC MEANS and PROC UNIVARIATE to generate descriptive statistics for **`weight`**.

```
options nodate nonumber;
proc print data=sasuser.b_rise (obs=10);
   title 'Listing of the Cereal Data Set';
run;
```

```
                  Listing of the Cereal Data Set

              Obs       brand         weight    idnumber

                1    Rise n Shine    15.0136    33081197
                2    Rise n Shine    14.9982    37070397
                3    Rise n Shine    14.9930    60714297
                4    Rise n Shine    15.0812     9589297
                5    Rise n Shine    15.0418    85859397
                6    Rise n Shine    15.0639    99108497
                7    Rise n Shine    15.0613    70847197
                8    Rise n Shine    15.0255    53750297
                9    Rise n Shine    15.0176     3873197
               10    Rise n Shine    15.0122    43493297
```

```
proc means data=sasuser.b_rise maxdec=4;
   var weight;
   title 'Descriptive Statistics Using PROC MEANS';
run;
```

Selected PROC MEANS statement option:

MAXDEC= specifies the maximum number of decimal places to use when printing numeric values.

```
                  Descriptive Statistics Using PROC MEANS

                            The MEANS Procedure

                        Analysis Variable : weight

     N            Mean         Std Dev         Minimum         Maximum
    ──────────────────────────────────────────────────────────────────
    40         15.0360          0.0265         14.9831         15.0980
    ──────────────────────────────────────────────────────────────────
```

By default, PROC MEANS prints the number of nonmissing observations, the mean, the standard deviation, the minimum value, and the maximum value.

```
proc univariate data=sasuser.b_rise;
   var weight;
   id idnumber;
   title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```

PROC UNIVARIATE Output

```
                 Descriptive Statistics Using PROC UNIVARIATE

                               The UNIVARIATE Procedure
                                  Variable:  weight

                                       Moments

            N                            40    Sum Weights                 40
            Mean                   15.03596    Sum Observations      601.4384
            Std Deviation        0.02654963    Variance            0.00070488
            Skewness             0.39889232    Kurtosis            -0.1975717
            Uncorrected SS       9043.23122    Corrected SS        0.02749044
            Coeff Variation      0.17657424    Std Error Mean      0.00419787


                             Basic Statistical Measures

                    Location                    Variability

                Mean     15.03596     Std Deviation           0.02655
                Median   15.03480     Variance              0.0007049
                Mode     15.01220     Range                   0.11490
                                      Interquartile Range     0.03650

        NOTE: The mode displayed is the smallest of 2 modes with a count of 2.


                             Tests for Location: Mu0=0

                   Test           -Statistic-    -----p Value------

                   Student's t    t  3581.811    Pr > |t|    <.0001
                   Sign           M        20    Pr >= |M|   <.0001
                   Signed Rank    S       410    Pr >= |S|   <.0001
```

PROC UNIVARIATE Output (continued)

```
                              Quantiles (Definition 5)

                                Quantile      Estimate

                                100% Max       15.0980
                                99%            15.0980
                                95%            15.0863
                                90%            15.0726
                                75% Q3         15.0525
                                50% Median     15.0348
                                25% Q1         15.0160
                                10%            15.0095
                                5%             14.9956
                                1%             14.9831
                                0% Min         14.9831

                                  Extreme Observations

          ------------Lowest-----------        -----------Highest-----------

             Value    idnumber      Obs           Value    idnumber      Obs

           14.9831    30834797       37         15.0639    99108497        6
           14.9930    60714297        3         15.0812     9589297        4
           14.9982    37070397        2         15.0858    73461797       21
           15.0093    46028397       14         15.0868    40177297       27
           15.0096    59149297       40         15.0980    23573597       35
```

The output indicates that

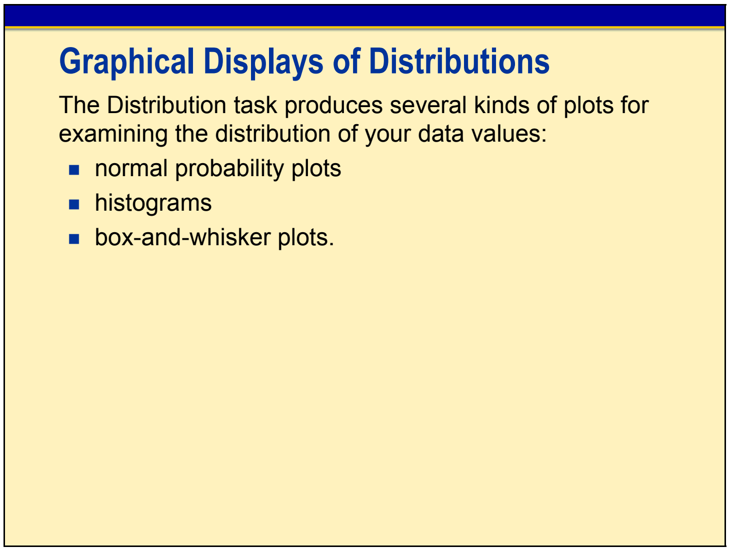
- the mean, or center point, of the data is 15.03596 ounces. This is approximately equal to the median (15.0348), which indicates the distribution is fairly symmetric.
- the standard deviation is 0.02655, which means that the average variability around the mean is approximately 0.027 ounces.
- the distribution is slightly skewed to the right.
- the distribution has lighter tails than the normal distribution.
- the range of the data is 0.1149, the difference between 14.9831 and 15.098.
- the interquartile range focuses on the variation of the middle 50% of the data and is 0.0365.
- the cereal box with the largest amount of cereal has an identification number of 23573597, which is observation number 35 in the data set.

The *mode* is the most frequent data value. The note in the output listing indicates that the mode displayed is the smallest of two modes with a count of two. If there are no replicated values in your data, the mode does not exist and, therefore, is reported as missing.

If you would like a table of the modes and their respective frequencies, add the MODES option in the PROC UNIVARIATE statement.

In the Quantiles table, Definition 5 indicates that PROC UNIVARIATE is using the default definition for calculating percentile values. You can use the PCTLDEF= option in the PROC UNIVARIATE statement to specify one of five methods. These methods are listed in Appendix C, “Percentile Definitions.”

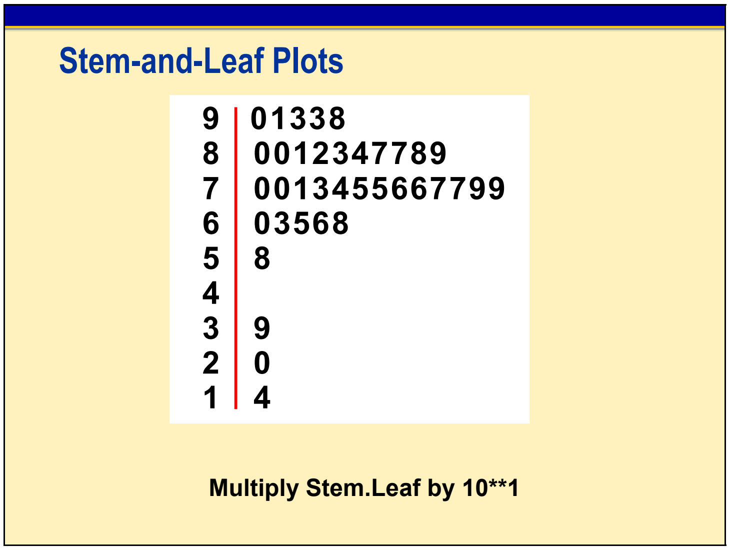
Exercise: Refer to your course workbook.



Graphical Displays of Distributions

The Distribution task produces several kinds of plots for examining the distribution of your data values:

- normal probability plots
- histograms
- box-and-whisker plots.



Stem-and-Leaf Plots

```
9 | 01338
8 | 0012347789
7 | 0013455667799
6 | 03568
5 | 8
4 |
3 | 9
2 | 0
1 | 4
```

Multiply Stem.Leaf by 101**

A *stem-and-leaf plot* is a histogram that provides specific information about the numeric values in your data.

Consider this data, which represents test scores on a statistics exam:

```
14 20 39 58 60 63 65 66 68 70 70 71 73 74 75 75 76 76 77
77 79 79 80 80 81 82 83 84 87 87 88 89 90 91 93 93 98
```

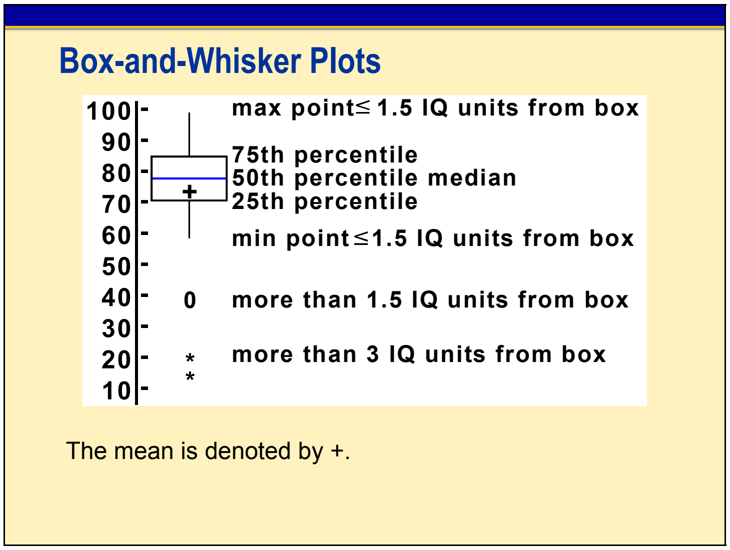
The legend Multiply Stem.Leaf by 10**1 indicates how to convert values from the stem-and-leaf plot to actual data values. In this case, multiply the stems by 10. If more than 48 observations fall within a single interval, PROC UNIVARIATE produces a horizontal bar chart.

For this example, the stems of the plot correspond to the tens digits, and the leaves correspond to the ones digits. Thus, you can see that 98 occurred once, 93 occurred twice, and so on.

The stem-and-leaf plot shows the

- raw data
- shape of the distribution.

In this example, the distribution is heavily skewed to the lower test scores.

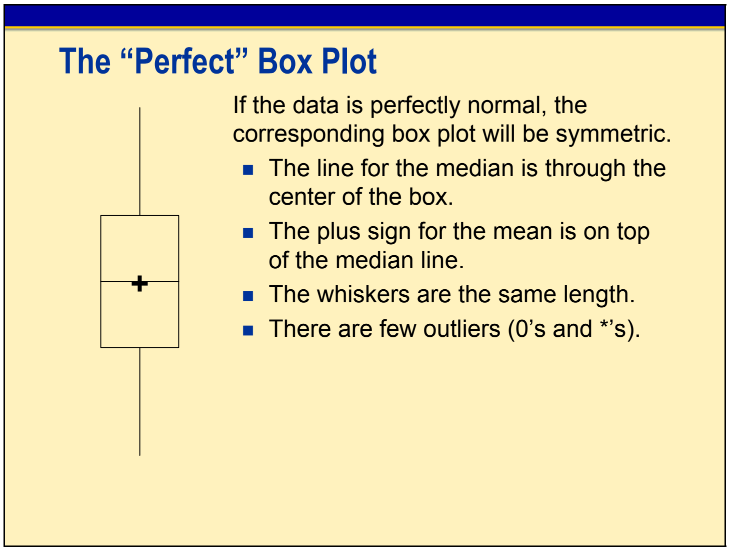


Box-and-Whisker Plots

The mean is denoted by +.

Box-and-whisker plots provide information about the variability of data and the extreme data values. The box represents the median (middle value) of your data, and you get a rough impression of the symmetry of your distribution by comparing the mean and median. The whiskers extend from the box as far as the data extend, to a distance of, at most, 1.5 interquartile units.

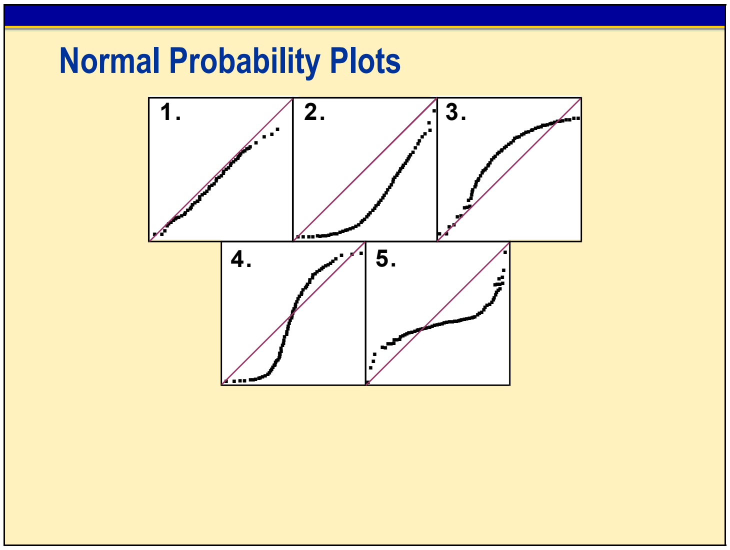
The above plot is of the test scores from a statistics exam. The plot shows the data is skewed and has a few extreme values.



The "Perfect" Box Plot

If the data is perfectly normal, the corresponding box plot will be symmetric.

- The line for the median is through the center of the box.
- The plus sign for the mean is on top of the median line.
- The whiskers are the same length.
- There are few outliers (0's and *'s).



Normal Probability Plots

A *normal probability plot* is a visual method for determining whether your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis is the expected percentiles from a standard normal distribution. In other words, the plot is an overlay plot of your observed data versus your expected data if your data came from a normal distribution.

The above diagrams illustrate some possible normal probability plots for data from a

1. normal distribution (the observed data follows the reference line)
2. skewed-to-the-right distribution
3. skewed-to-the-left distribution
4. light-tailed distribution
5. heavy-tailed distribution.

Examining Distributions

m1demo04.sas

Example: Use the PLOT option in PROC UNIVARIATE to produce plots for the variable **weight** in the **sasuser.b_rise** data set. Also, generate a histogram for **weight** and a graphically enhanced normal probability plot.

```
options ps=50 ls=64;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=sasuser.b_rise plot;
   var weight;
   id idnumber;
   histogram weight;
   probplot weight / normal (mu=est sigma=est
                             color=blue w=1);
   title;
run;
```

✎ You cancel all previously defined titles by submitting a TITLE statement.

Selected PROC UNIVARIATE statement option:

PLOT produces a stem-and-leaf plot, a box-and-whisker plot, and a normal probability plot.

Selected UNIVARIATE procedure statements:

HISTOGRAM creates high-resolution histograms.

PROBPLOT creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution.

Selected PROBPLOT statement option:

NORMAL superimposes a reference line on the normal probability plot, using the estimates of mu and sigma from the data. In this example, the reference line will be blue with a width of 1.

Below are the stem-and-leaf plot and the box-and-whisker plot. The plots show that the data is fairly symmetric with no extreme data values.

Partial PROC UNIVARIATE Output

```
Stem Leaf                     #             Boxplot
1509 8                        1                |
1508 167                      3                |
1507                                           |
1506 1234                     4                |
1505 0058                     4             +-----+
1504 122446                   6             |     |
1503 0279                     4             *--+--*
1502 00367                    5             |     |
1501 002246689                9             +-----+
1500 9                        1                |
1499 38                       2                |
1498 3                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-2
```

To convert values from the stem-and-leaf plot to the actual data values, you must multiply the stems by 10**-2, or 0.01.

The normal probability plot is shown below. The plus signs represent where the data values would fall if they came from a normal distribution. The asterisks represent the observed data values. Because the asterisks follow a fairly straight line and cover up many plus signs, you can conclude that there does not appear to be any severe departure from the normal distribution.

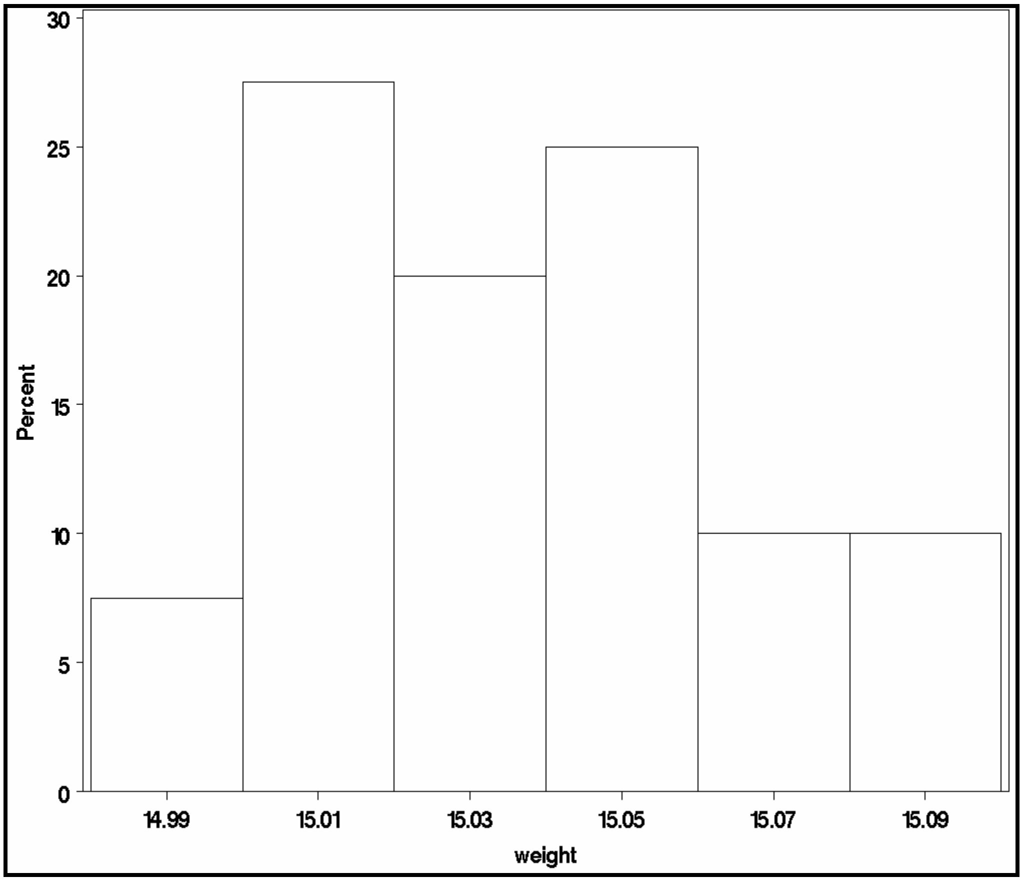
Partial PROC UNIVARIATE Output

```
                     Normal Probability Plot
15.095+                                              +*++
      |                                      * * *+++
      |                                      ++++
      |                                  ****
      |                               ***+
      |                           *****
      |                       ++**
      |                   +++***
      |              **+*****
      |            **++
      |        *+*+
14.985+   *++++
       +----+----+----+----+----+----+----+----+----+----+
           -2        -1         0        +1        +2
```

A histogram is a distribution with a unique feature. Instead of the frequency of the values being plotted on the vertical axis, the **percent** of the values is recorded. Therefore, the summation of the percentages of the bins is 100. The histogram of the variable **`weight`** is shown below. The horizontal axis values represent the midpoints of the bins. The vertical axis is the percent of the values in the specific bin.

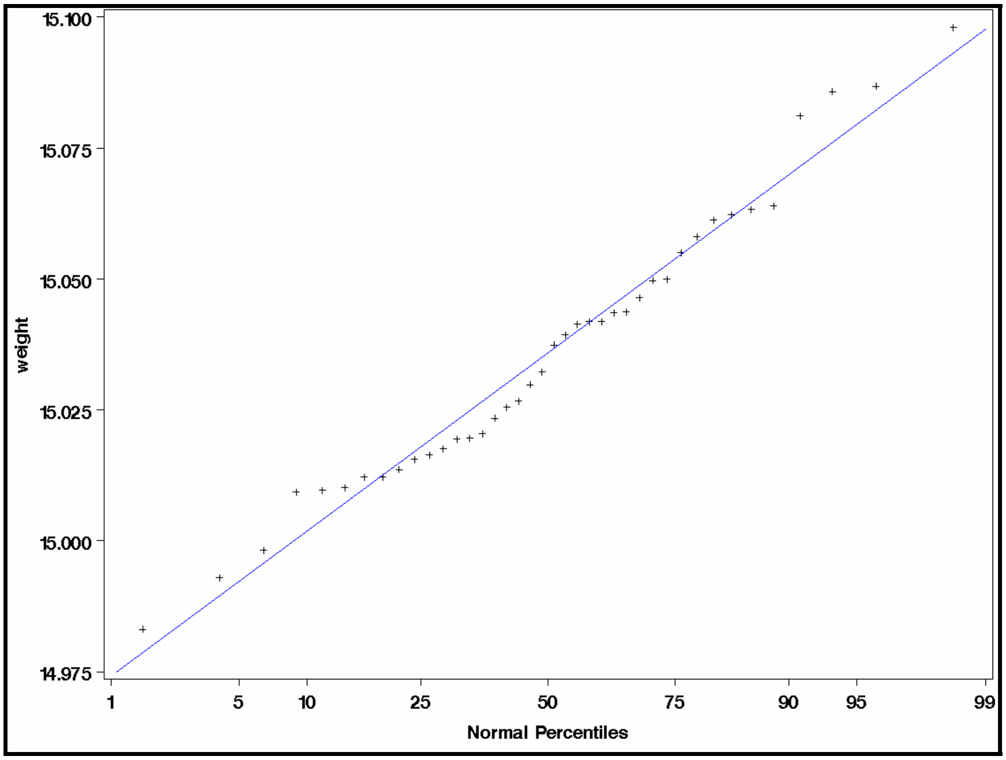
For example, the bin identified with the midpoint of 15.01 has approximately 27% of the values; in addition, you can state that 27% of the values fall between the bin end points of 15.00 and 15.02. In a similar way, you can state that approximately 7% of the values fall between 14.98 and 15.00.

Partial PROC UNIVARIATE Graph Output

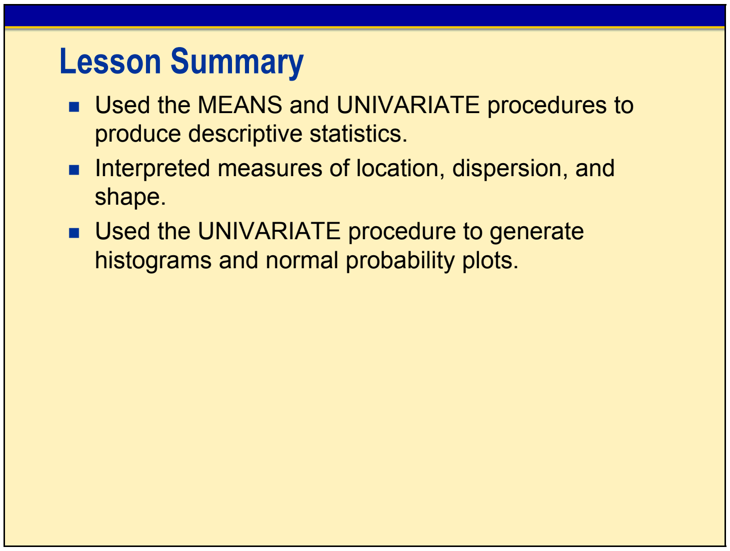


The graphically enhanced normal probability plot is shown below, using the PROBPLOT statement. The 45-degree line represents where the data values would fall if they came from a normal distribution. The plus signs represent the observed data values. Because the plus signs follow the 45-degree line in the graph below, you can conclude that there does not appear to be any severe departure from the normal distribution.

Partial PROC UNIVARIATE Graph Output



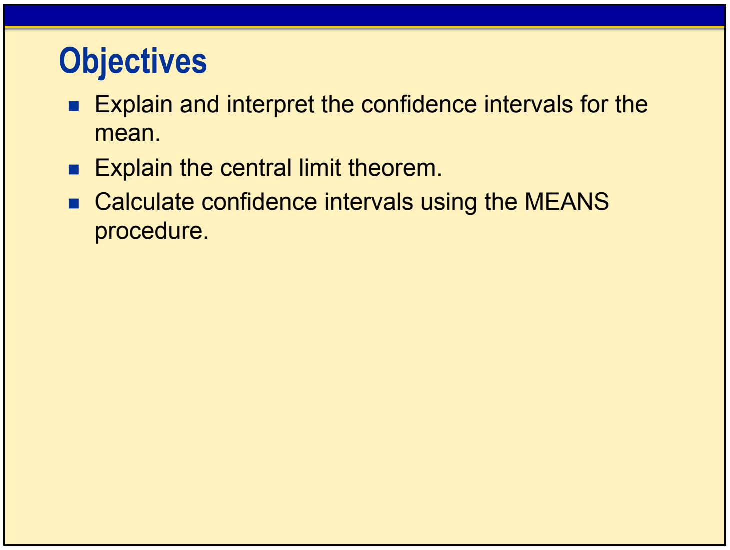
Exercise: Refer to your course workbook.



Lesson Summary

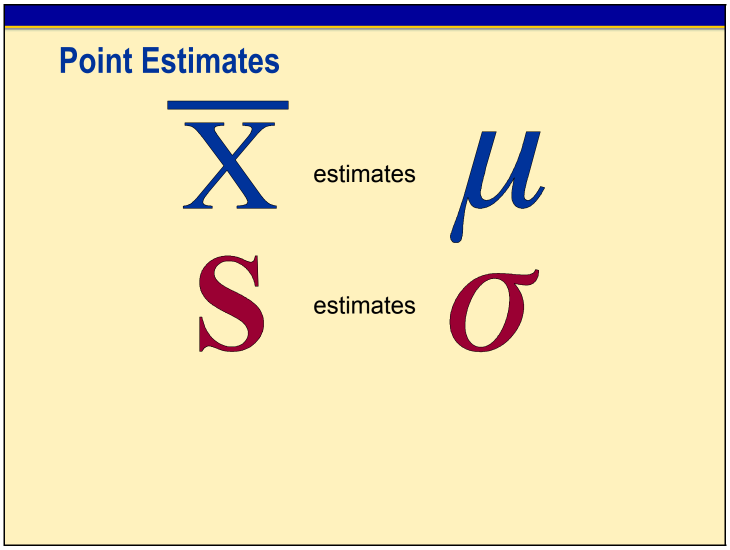
- Used the MEANS and UNIVARIATE procedures to produce descriptive statistics.
- Interpreted measures of location, dispersion, and shape.
- Used the UNIVARIATE procedure to generate histograms and normal probability plots.

1.3 Confidence Intervals



Objectives

- Explain and interpret the confidence intervals for the mean.
- Explain the central limit theorem.
- Calculate confidence intervals using the MEANS procedure.



Point Estimates

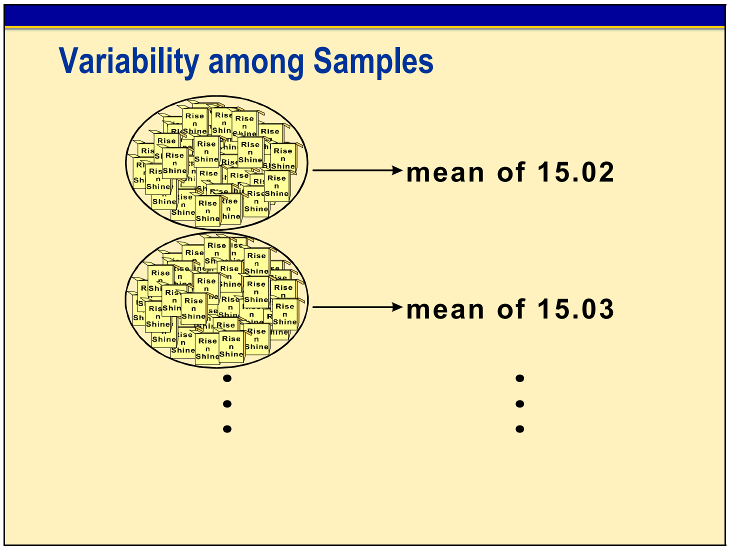
$\overline{X}$ estimates μ

S estimates σ

A *point estimate* is a sample statistic used to estimate a population parameter.

- An estimate of the average **`weight`** is 15.036, and an estimate of the standard deviation is 0.027.
- Because you only have an estimate of the unknown population mean, you need to know the variability of your estimate.

A point estimate does not take into account the accuracy of the calculated statistic.



Variability among Samples

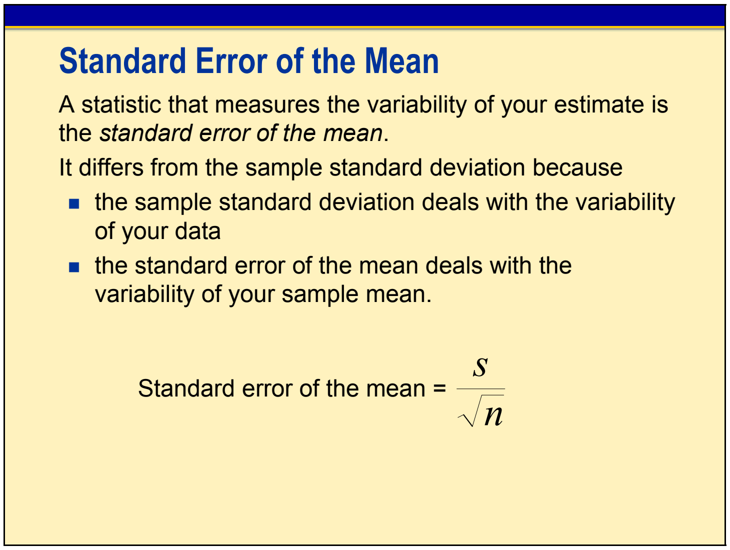
mean of 15.02

mean of 15.03

Why are you not absolutely certain that the mean weight for Rise n Shine cereals is 15.036?

The answer is because the sample mean is only an estimate of the population mean. If you collected another sample of cereal boxes, you would have another estimate of the mean.

Therefore, different samples yield different estimates of the mean for the same population. How close these sample means are to one another determines the variability of the estimate of the population mean.



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 deals with the variability of your data
- the standard error of the mean deals with the variability of your sample mean.

Standard error of the mean = $\frac{s}{\sqrt{n}}$

The standard error of the mean is computed as

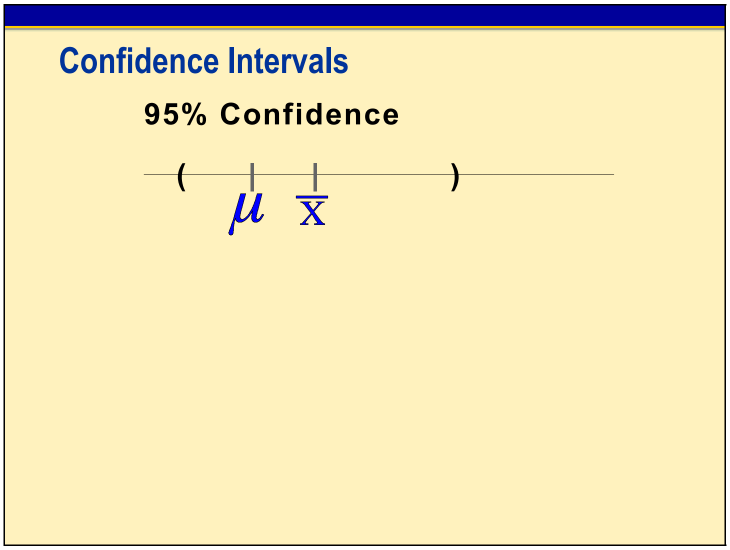
$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

s is the sample standard deviation

n is the sample size.

The standard error of the mean for the variable **`weight`** is (0.02654963 / SQRT(40)), or approximately 0.004. This is a measure of how much error you can expect when you use the sample mean to predict the population mean. Therefore, the smaller the standard error is, the more accurate your sample estimate is.



Confidence Intervals

95% Confidence

A *confidence interval*

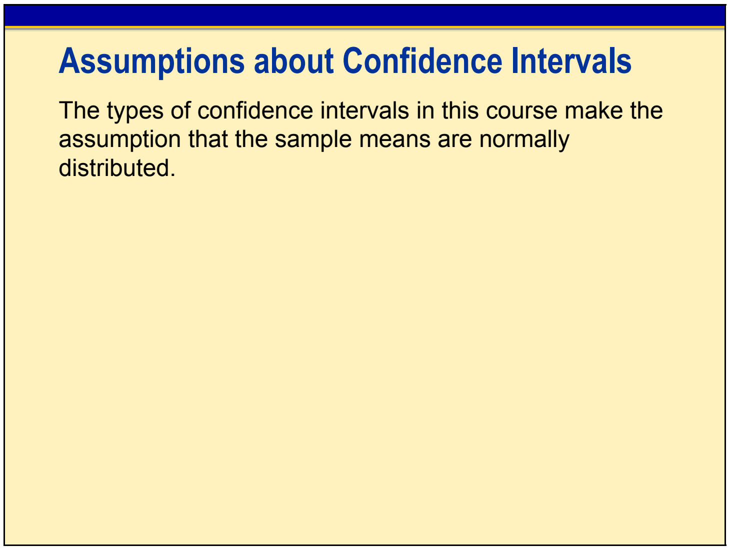
- is a range of values that you believe to contain the population parameter of interest
- places an upper and lower bound around a sample statistic.

To construct a confidence interval, a significance level must be chosen.

A 95% confidence interval is commonly used to assess the variability of the sample mean. In the cereal example, you interpret a 95% confidence interval by stating that you are 95% confident the interval contains the mean number of ounces of cereal for your population.

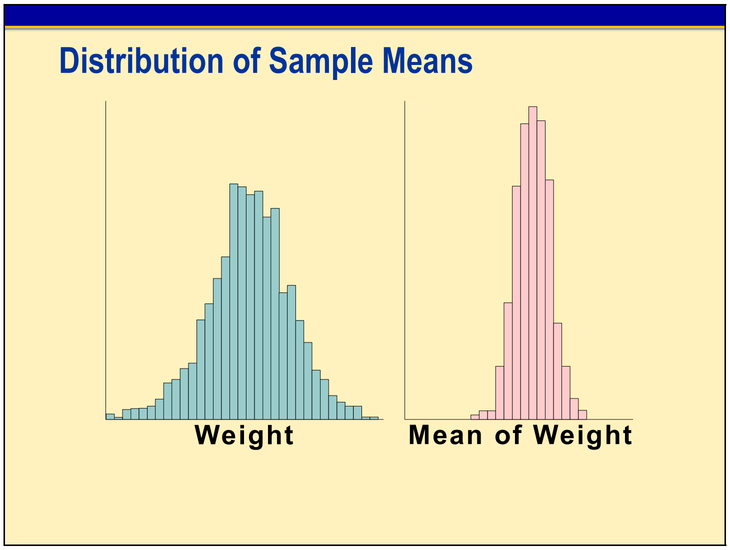
Do you want to be as confident as possible?

Yes, but if you increase the confidence level, the width of your interval increases. As the width of the interval increases, it becomes less useful.



Assumptions about Confidence Intervals

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



Distribution of Sample Means

What is a distribution of sample means?

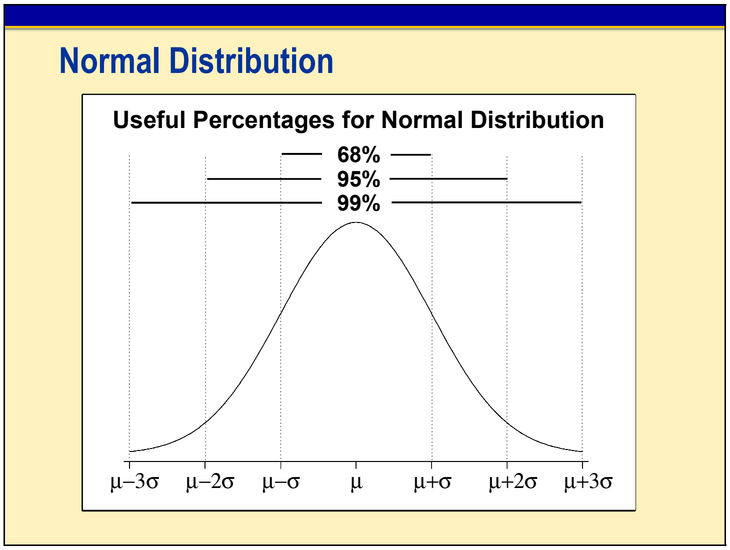
In the cereal example, it is the distribution of all possible sample means of ounces of cereal.

If you collect another sample of weights of cereal boxes, you would have another sample mean. In fact, if you collect 100 more samples, you would have 100 different sample means.

To illustrate the distinction between the distribution of the data values and the distribution of the sample means, suppose 500 samples of cereal weights of a sample size of 10 were collected.

- The first chart could represent all 5000 observations in the data.
- The second chart could be a plot of the means from each of 500 samples of size 10.

The distribution of sample means is not as wide. In other words, the distribution of sample means has a smaller variance.



Normal Distribution

Useful Percentages for Normal Distribution

Why does the distribution of sample means have to be normally distributed?

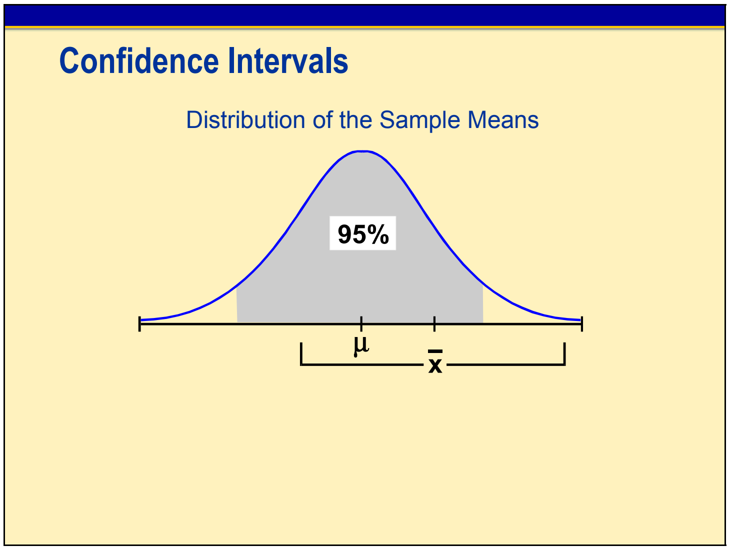
The normal distribution describes percentages. For example, approximately

- 68% of the data falls within one standard deviation of the mean
- 95% of the data falls within two standard deviations of the mean
- 99.7% of the data falls within three standard deviations of the mean.

If the distribution of sample means is normal, you can use the percentages described by the normal distribution when constructing a confidence interval. The percentage corresponds to the confidence level.

Therefore, if you construct a 95% confidence interval, you have a 95% chance of constructing a confidence interval that contains the population mean.

If the distribution of sample means is not normal, you have no idea what percentage corresponds to a 95% confidence interval (unless the distribution of sample means is another known distribution).



Confidence Intervals

Distribution of the Sample Means

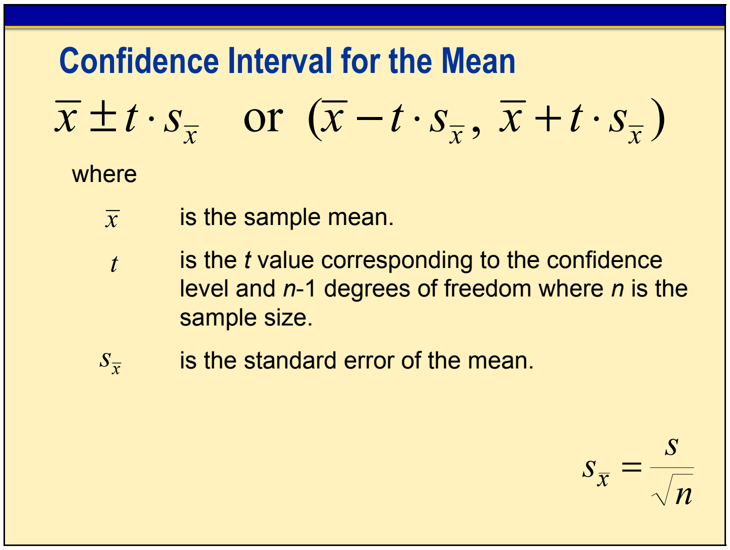
The graph above is the distribution of sample means. The shaded region represents 95% of the area in the distribution.

When constructing a 95% confidence interval, the length of the interval

- covers 95% of the area under the distribution of sample means when it is centered over μ, the population mean
- corresponds to a 95% probability of capturing the population mean when the interval is constructed.

Therefore, if the sample mean falls in the shaded region in the distribution of sample means, the interval constructed will contain the population mean.

Notice that μ is captured in this interval.



Confidence Interval for the Mean

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

where

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

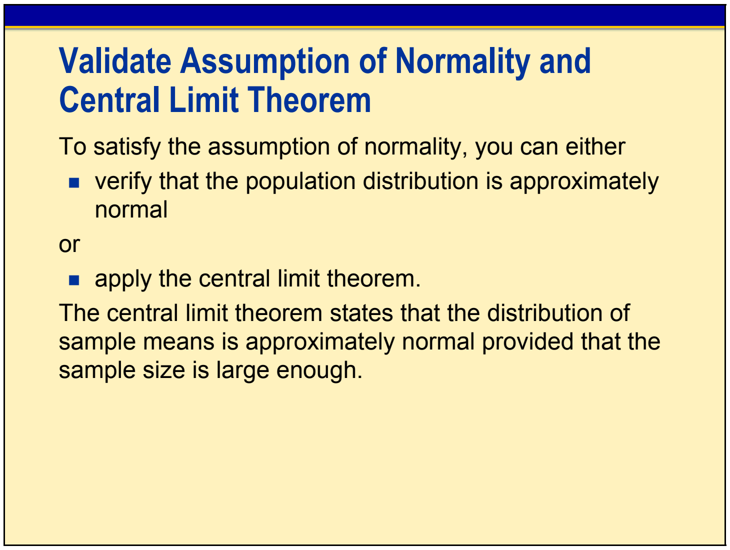
t is the *t* value corresponding to the confidence level and *n*-1 degrees of freedom where *n* is the sample size.

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

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

Inspect the formula for a confidence interval. Each part of the formula except for the sample mean affects the size of the confidence interval. Observe that the confidence limits will be wider if the

- standard error of the mean increases because either the sample standard deviation increases or the sample size decreases
- *t* value increases because higher confidence is required.



Validate Assumption of Normality and Central Limit Theorem

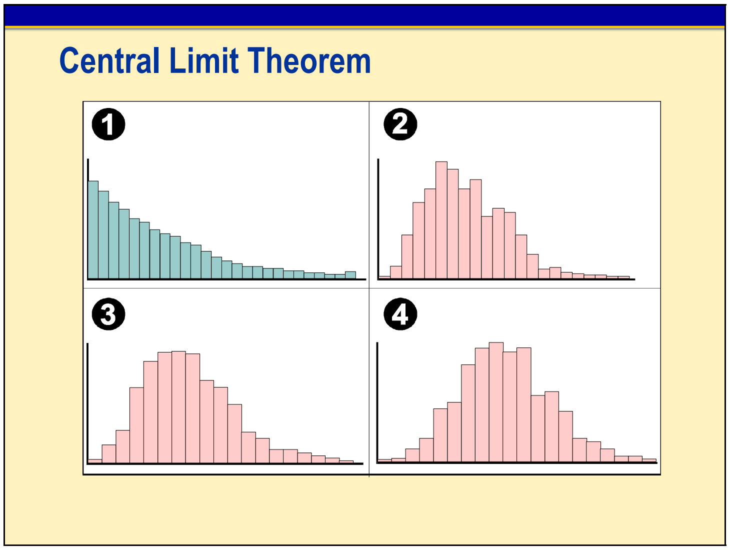
To satisfy the assumption of normality, you can either

- verify that the population distribution is approximately normal

or

- apply the central limit theorem.

The central limit theorem states that the distribution of sample means is approximately normal provided that the sample size is large enough.



Central Limit Theorem

The above graphs illustrate the tendency of a distribution of sample means to approach normality as the sample size increases.

The first chart is a histogram of data values drawn from an exponential distribution. The remaining charts are histograms of the sample means for samples of differing sizes drawn from the same exponential distribution.

1. Data from exponential distribution
2. 1000 samples of size 5
3. 1000 samples of size 10
4. 1000 samples of size 30

✐ For the sample size of 30, the distribution is approximately bell-shaped and symmetric, even though the sample data is highly skewed.

Confidence Intervals

m1demo05.sas

Example: Use the MEANS procedure to generate a 95% confidence interval for the mean of **weight** in the **sasuser.b_rise** data set.

```
proc means data=sasuser.b_rise n mean sterr clm;
   var weight;
   title '95% Confidence Interval for WEIGHT';
   title2 'Means Must be Normally Distributed';
run;
```

Selected PROC MEANS statement options:

N prints the number of nonmissing values.

MEAN prints the mean.

CLM calculates confidence limits for the mean.

The output is shown below.

```
                 95% Confidence Interval for WEIGHT
                  Means Must be Normally Distributed

                         The MEANS Procedure

                     Analysis Variable : weight

                                  Lower 95%       Upper 95%
      N             Mean      CL for Mean     CL for Mean
     ________________________________________________________

     40      15.0359600       15.0274690       15.0444510
     ________________________________________________________
```

In the cereal example, you are 95% confident that the population mean ounces for the Rise n Shine cereal boxes is contained in the interval 15.0275 and 15.0445. Because the interval between the upper and lower limits is small from a practical point of view, you can conclude that the sample mean is a fairly accurate estimate of the population mean.

How do you increase the accuracy of your estimate using the same confidence level?

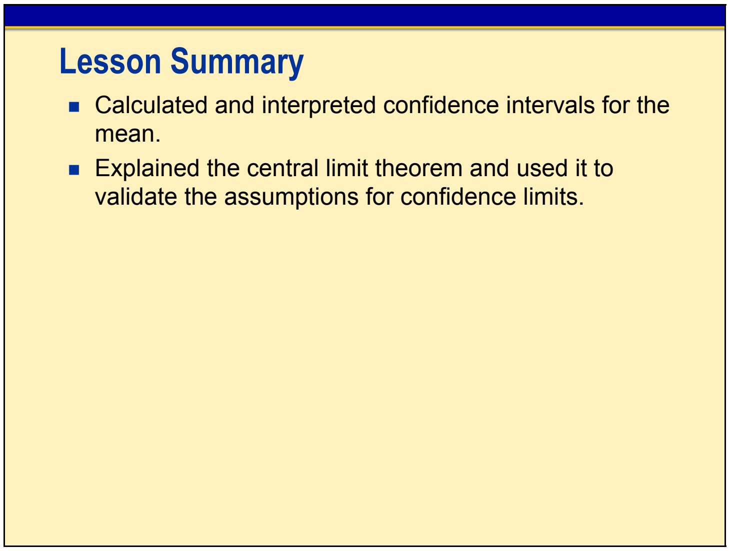
If you increase your sample size, you reduce the standard error of the sample mean and therefore reduce the width of your confidence interval. Thus, your estimate will be more accurate.

Do 95% of all cereal weights for all Rise n Shine boxes fall between 15.0275 and 15.0445?

No, confidence intervals deal with the variability of your sample mean.

✎ You can use the ALPHA= option in the PROC MEANS statement to construct confidence intervals with a different confidence level.

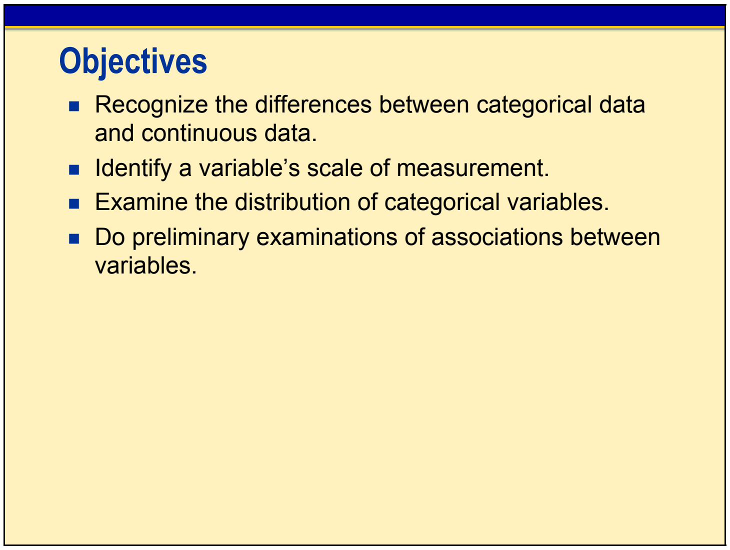
Exercise: Refer to your course workbook.



Lesson Summary

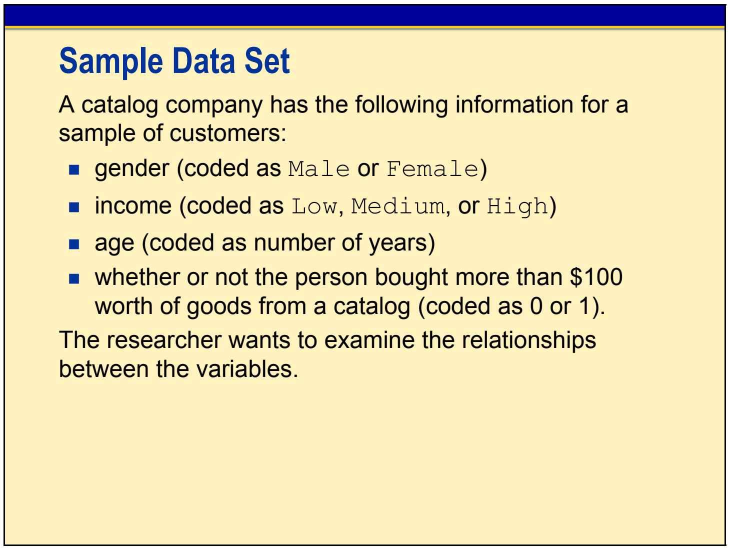
- Calculated and interpreted confidence intervals for the mean.
- Explained the central limit theorem and used it to validate the assumptions for confidence limits.

1.4 Descriptive Statistics with Categorical Data



Objectives

- Recognize the differences between categorical data and continuous data.
- Identify a variable's scale of measurement.
- Examine the distribution of categorical variables.
- Do preliminary examinations of associations between variables.



Sample Data Set

A catalog company has the following information for a sample of customers:

- gender (coded as `Male` or `Female`)
- income (coded as `Low`, `Medium`, or `High`)
- age (coded as number of years)
- whether or not the person bought more than $100 worth of goods from a catalog (coded as 0 or 1).

The researcher wants to examine the relationships between the variables.

Example: A company that sells its products via a catalog wants to identify those customers to whom advertising efforts should be directed. It has been decided that customers who spend 100 dollars or more are the target group. Based on the orders received over the last six months, the company wants to characterize this group of customers. The data is stored in the **`sasuser.b_sales`** data set.

The variables in the data set are

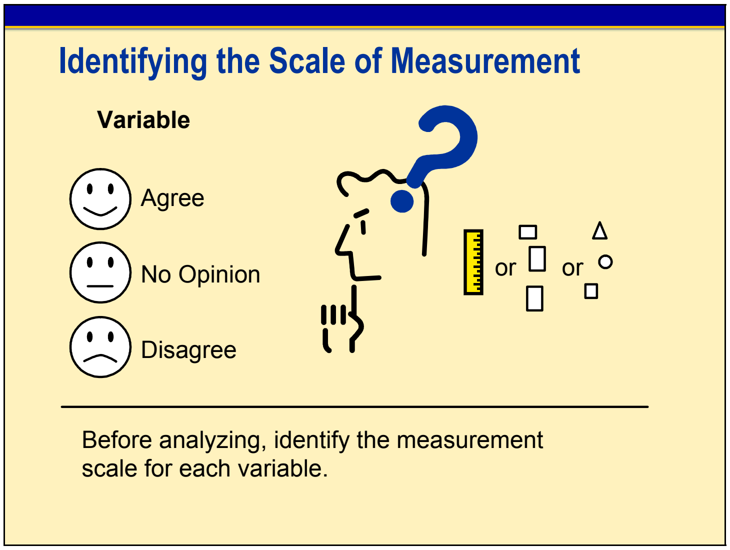
`purchase` purchase price (1=$100 or more, 0=under $100)

`age` age of customers in years

`gender` gender of customer (`Male`, `Female`)

`income` annual income (`Low`, `Medium`, `High`).

This is a hypothetical data set.



Identifying the Scale of Measurement

Variable

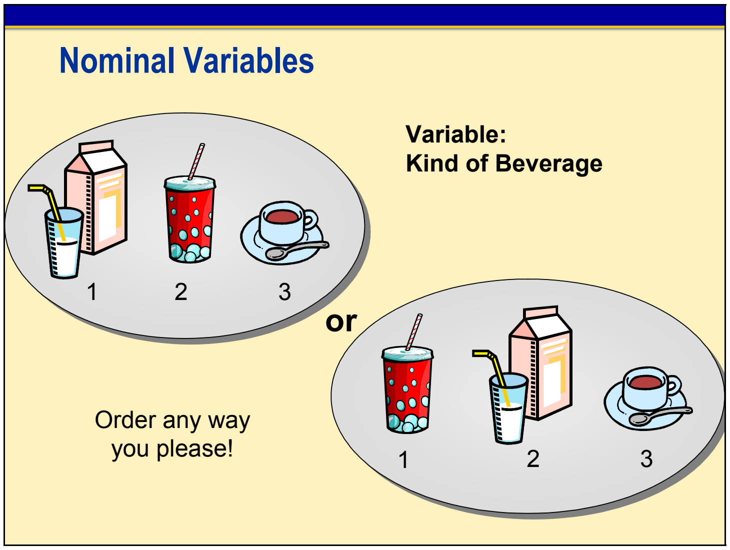
Agree

No Opinion

Disagree

Before analyzing, identify the measurement scale for each variable.

There are a variety of statistical methods for analyzing categorical data. To choose the appropriate method, you must determine the scale of measurement for your response variable.



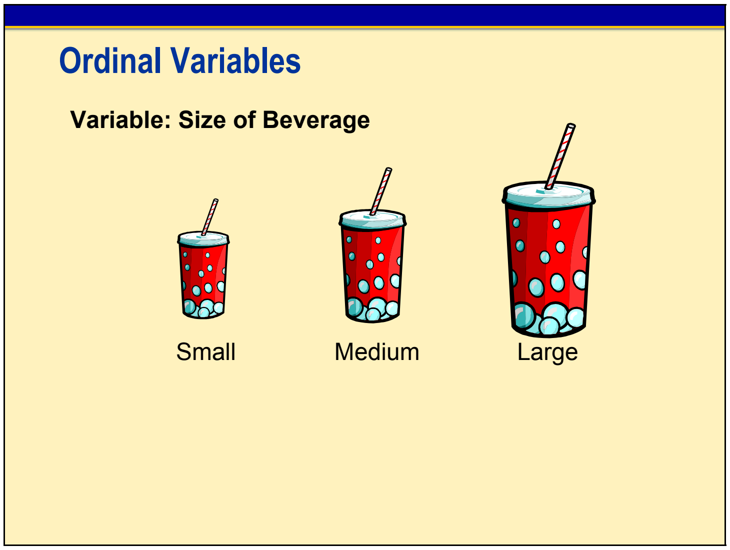
Nominal Variables

Variable:
Kind of Beverage

Order any way
you please!

Nominal variables have values with no logical ordering.

In the **`sasuser.b_sales`** data set, **`gender`** is a nominal variable.



Ordinal Variables

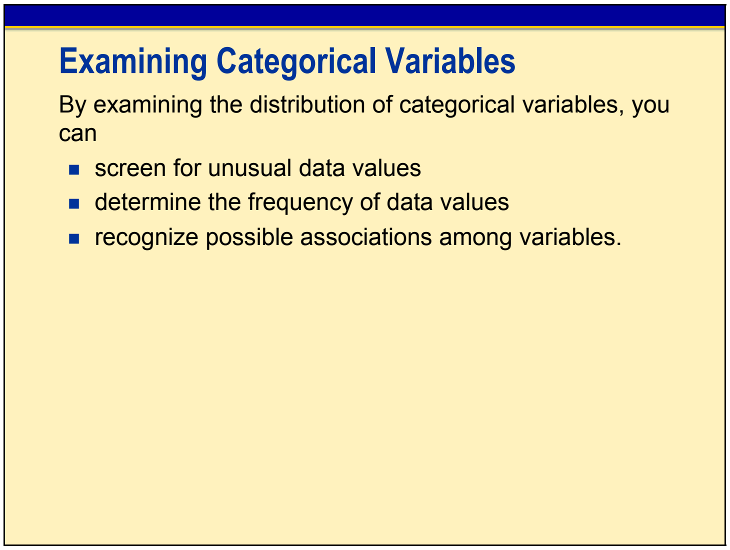
Variable: Size of Beverage

Small Medium Large

Ordinal variables have values with a logical order. However, the relative distances between the values are not clear.

In the **`sasuser.b_sales`** data set, **`income`** is an ordinal variable.

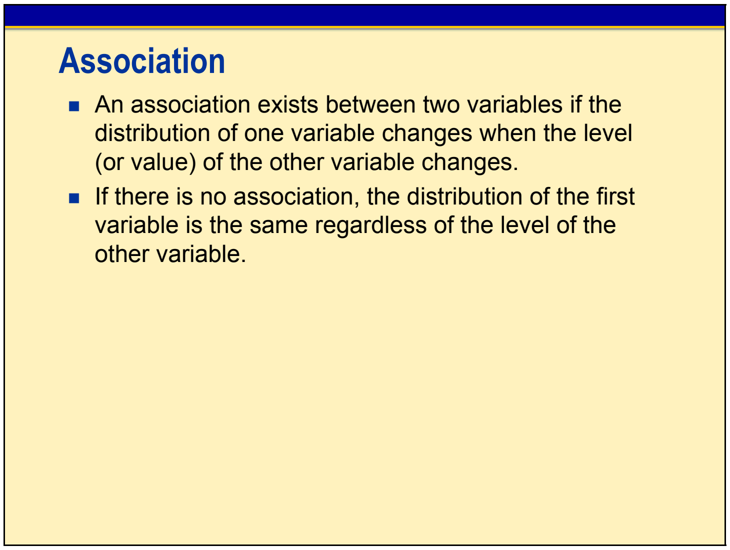
After you choose the appropriate scale of measurement, you can describe the relationship between categorical variables with the use of frequency tables.



Examining Categorical Variables

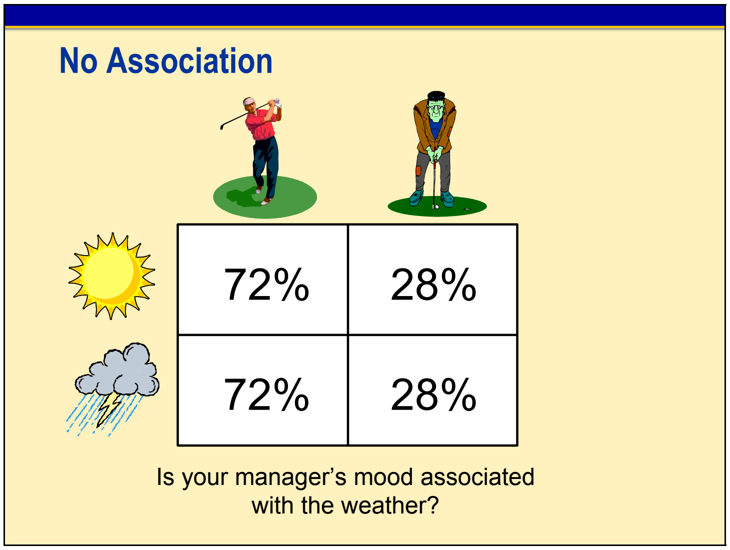
By examining the distribution of categorical variables, you can

- screen for unusual data values
- determine the frequency of data values
- recognize possible associations among variables.



Association

- An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

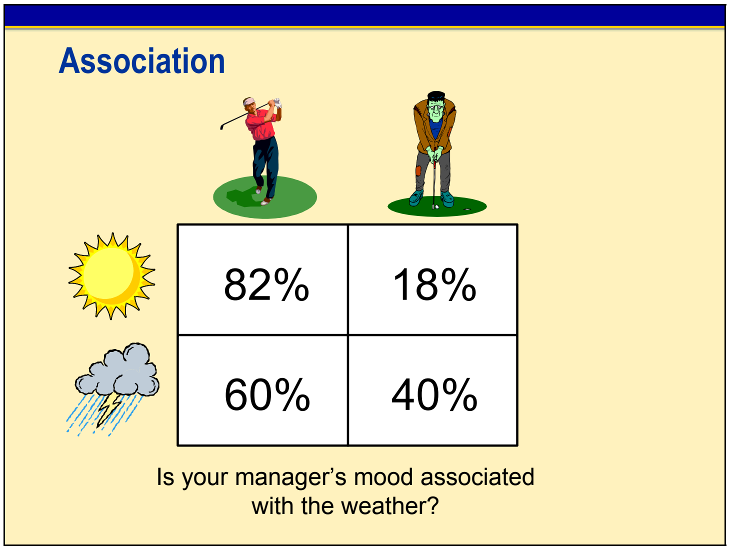


No Association

☀	72%	28%
⛈	72%	28%

Is your manager's mood associated with the weather?

There appears to be no association here because the row percentages are the **same** in each column.

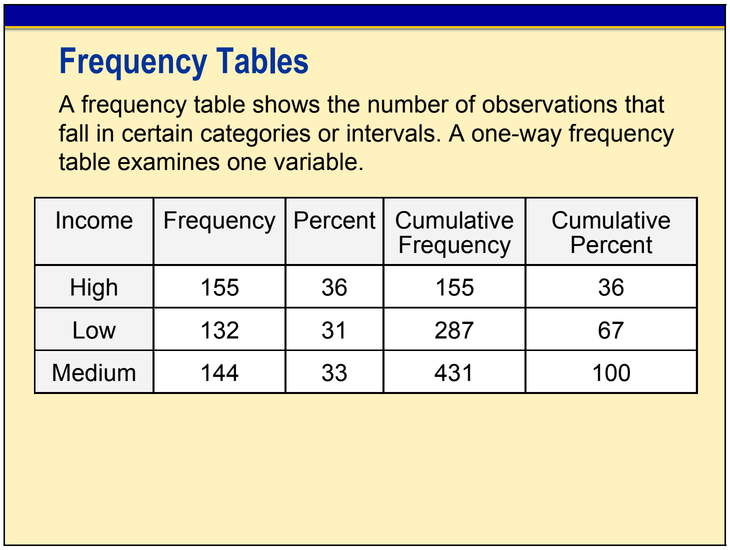


Association

☀	82%	18%
🌧	60%	40%

Is your manager's mood associated with the weather?

There appears to be an association here because the row percentages are **different** in each column.



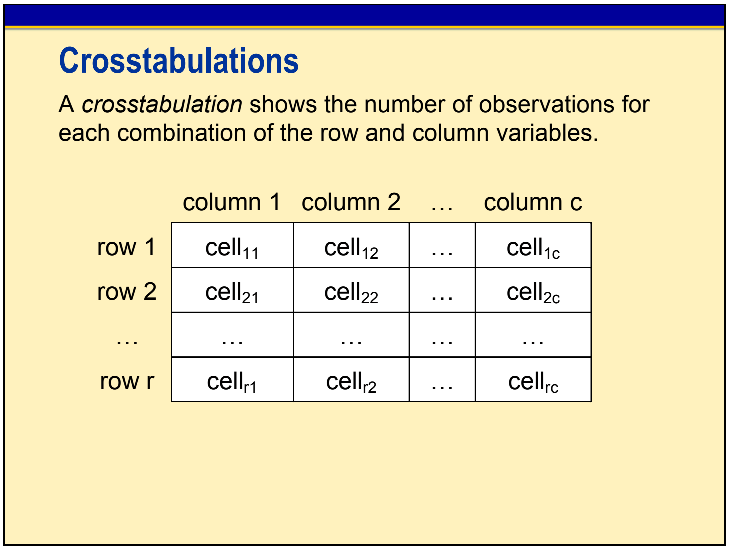
Frequency Tables

A frequency table shows the number of observations that fall 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

Typically, there are four types of frequency measures included in a frequency table:

frequency	is the number of times the value appears in the data set.
percent	is 100 times the relative frequency. This represents the percent of the data that has this value.
cumulative frequency	accumulates the frequency of each of the values by adding the second frequency to the first and so on.
cumulative percent	accumulates the percent each value represents by adding the second percent to the first and so on.



Crosstabulations

A *crosstabulation* shows the number of observations for each combination of the row and column variables.

	column 1	column 2	…	column c
row 1	cell$_{11}$	cell$_{12}$	…	cell$_{1c}$
row 2	cell$_{21}$	cell$_{22}$	…	cell$_{2c}$
…	…	…	…	…
row r	cell$_{r1}$	cell$_{r2}$	…	cell$_{rc}$

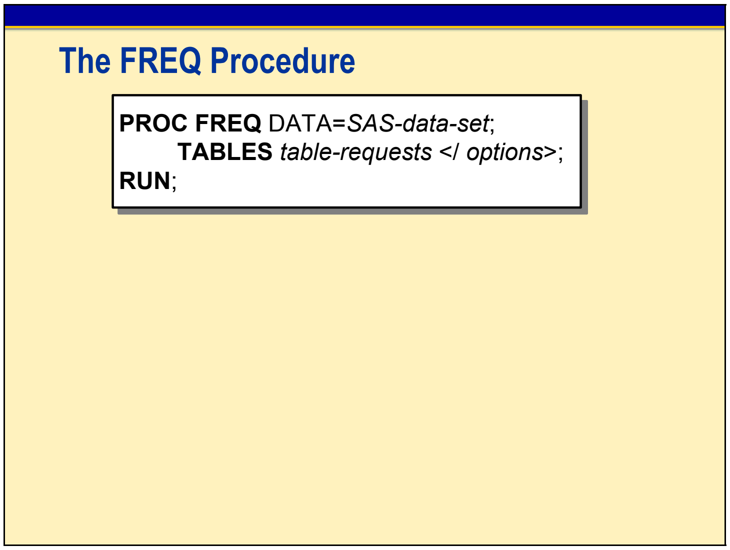
By default, a crosstabulation has four measures in each cell:

frequency — number of observations falling in a category formed by the row variable and the column variable

percent — number of observations in each cell as a percentage of the total number of observations

row pct — number of observations in each cell as a percentage of the total number of observations in that row

col pct — number of observations in each cell as a percentage of the total number of observations in that column.



The FREQ Procedure

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

Selected FREQ procedure statement:

TABLES requests tables and specifies options for producing tests. The general form of a table request is *variable-1**variable-2**…, where any number of these requests can be made in a single TABLES statement. For two-way crosstabulations, the first variable represents the rows and the second variable represents the columns.

✎ PROC FREQ can generate large volumes of output if you have many variables or variables with many distinct values.

Examining Categorical Distributions

m1demo06.sas

Example: Invoke PROC FREQ and create one-way frequency tables for the variables **gender**, **age**, **income**, and **purchase**. Create two-way frequency tables for the variables **purchase** and **gender**, and **purchase** and **income**. Also, use the FORMAT procedure to format the values of **purchase**.

```
proc format lib=sasuser;
   value purfmt 1 = "$100 +"
                0 = "< $100"
                  ;
run;

proc freq data=sasuser.b_sales;
   tables gender age income purchase
          gender*purchase income*purchase;
   format purchase purfmt.;
run;
```

PROC FREQ Output

```
                          The FREQ Procedure

                                                   Cumulative    Cumulative
       gender     Frequency      Percent      Frequency       Percent
       ────────────────────────────────────────────────────────────────────
       Female          240        55.68            240         55.68
       Male            191        44.32            431        100.00
```

PROC FREQ Output (continued)

```
                                       Cumulative    Cumulative
 age     Frequency       Percent        Frequency      Percent
 ________________________________________________________________

  23            1          0.23                1          0.23
  24            1          0.23                2          0.46
  25            2          0.46                4          0.93
  26            5          1.16                9          2.09
  28            3          0.70               12          2.78
  29            6          1.39               18          4.18
  30            6          1.39               24          5.57
  31           11          2.55               35          8.12
  32           11          2.55               46         10.67
  33           25          5.80               71         16.47
  34           23          5.34               94         21.81
  35           28          6.50              122         28.31
  36           19          4.41              141         32.71
  37           29          6.73              170         39.44
  38           37          8.58              207         48.03
  39           30          6.96              237         54.99
  40           31          7.19              268         62.18
  41           35          8.12              303         70.30
  42           19          4.41              322         74.71
  43           18          4.18              340         78.89
  44           19          4.41              359         83.29
  45           17          3.94              376         87.24
  46           12          2.78              388         90.02
  47           13          3.02              401         93.04
  48            8          1.86              409         94.90
  49            7          1.62              416         96.52
  50            5          1.16              421         97.68
  51            4          0.93              425         98.61
  52            2          0.46              427         99.07
  55            2          0.46              429         99.54
  56            1          0.23              430         99.77
  58            1          0.23              431        100.00

                                         Cumulative    Cumulative
income     Frequency       Percent        Frequency      Percent
__________________________________________________________________

High             155         35.96              155         35.96
Low              132         30.63              287         66.59
Medium           144         33.41              431        100.00

                                           Cumulative    Cumulative
purchase     Frequency       Percent        Frequency      Percent
____________________________________________________________________

  < $100           269         62.41              269         62.41
  $100 +           162         37.59              431        100.00
```

There does not appear to be any unusual data values, for any of the variables, that could have been caused by coding errors.

The requested two-way frequency tables are shown below. You can get a preliminary idea whether there are associations between the outcome variable, **`purchase`**, and the predictor variables, **`gender`** and **`income`**, by examining the distribution of **`purchase`** for each value of the predictors.

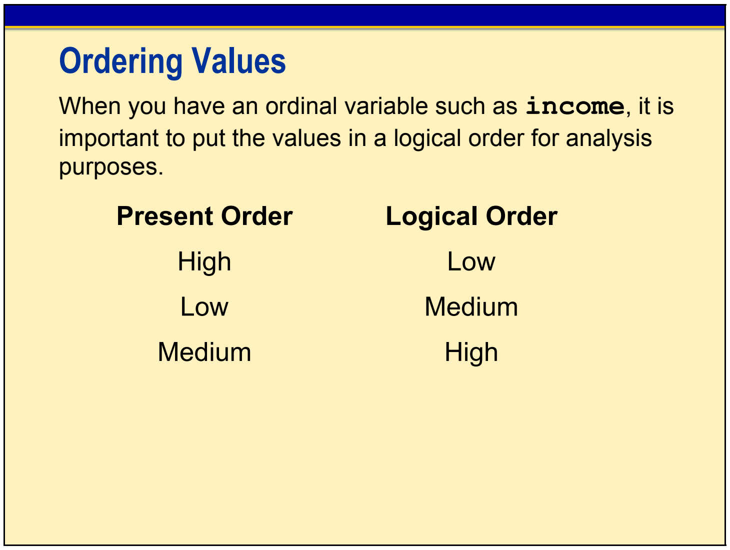
PROC FREQ Output (continued)

```
 Table of gender by purchase

gender     purchase

Frequency|
Percent  |
Row Pct  |
Col Pct  |< $100   |$100 +   |  Total
---------+---------+---------+
Female   |    139  |    101  |    240
         |  32.25  |  23.43  |  55.68
         |  57.92  |  42.08  |
         |  51.67  |  62.35  |
---------+---------+---------+
Male     |    130  |     61  |    191
         |  30.16  |  14.15  |  44.32
         |  68.06  |  31.94  |
         |  48.33  |  37.65  |
---------+---------+---------+
Total         269       162       431
            62.41     37.59    100.00
```

PROC FREQ Output (continued)

```
    Table of income by purchase

income     purchase

Frequency|
Percent  |
Row Pct  |
Col Pct  |< $100  |$100 +  |  Total
---------+--------+--------+
High     |     81 |     74 |    155
         |  18.79 |  17.17 |  35.96
         |  52.26 |  47.74 |
         |  30.11 |  45.68 |
---------+--------+--------+
Low      |     90 |     42 |    132
         |  20.88 |   9.74 |  30.63
         |  68.18 |  31.82 |
         |  33.46 |  25.93 |
---------+--------+--------+
Medium   |     98 |     46 |    144
         |  22.74 |  10.67 |  33.41
         |  68.06 |  31.94 |
         |  36.43 |  28.40 |
---------+--------+--------+
Total         269      162      431
            62.41    37.59   100.00
```

When you examine the row percentages, it appears that **`purchase`** is associated with **`gender`** and **`income`**. For example, 48% of the high-income customers made purchases of 100 dollars or more compared to 32% of the low-income customers and 32% of the medium-income customers.



Ordering Values

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

Present Order	Logical Order
High	Low
Low	Medium
Medium	High

Treating an ordinal variable as nominal can reduce the power of your statistical tests. In other words, statistical tests that detect linear associations have more power than statistical tests that detect general associations.

Reordering Values

m1demo07.sas

Example: Obtain a logical order in a frequency table for the values in the variable **`income`**.

1. Create a new variable called **`inclevel`** so that the sort order corresponds to its logical order.

```
data sasuser.b_sales_inc;
   set sasuser.b_sales;
   inclevel=1*(income='Low') + 2*(income='Medium')
            + 3*(income='High');
run;
```

✎ An expression enclosed in parentheses is a logical operator that returns the value 1 if the expression is true and 0 if the expression is false.

2. Use PROC FORMAT to create user-defined formats.

```
proc format lib=sasuser;
   value incfmt 1='Low Income'
                2='Medium Income'
                3='High Income';
run;
```

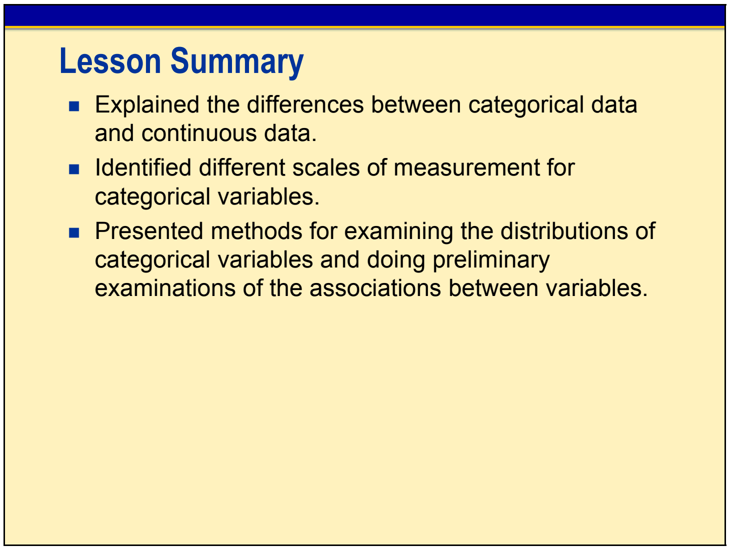
3. Use PROC FREQ with a FORMAT statement.

```
proc freq data=sasuser.b_sales_inc;
   tables inclevel*purchase;
   format inclevel incfmt. purchase purfmt.;
   title1 'Create variable INCLEVEL to correct INCOME';
run;
```

✎ If your data is in a logical order in a data set, you can use the ORDER=DATA option in PROC FREQ.

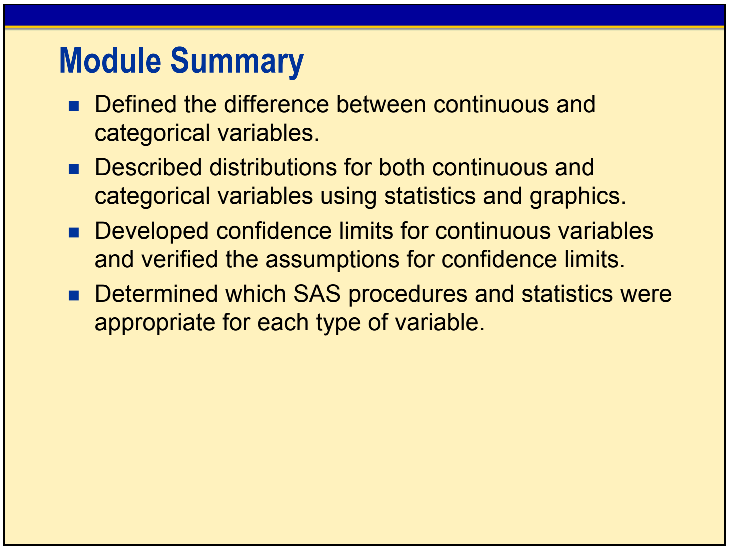
The crosstabulation of **`inclevel*purchase`** is shown below. The values of **`inclevel`** are now in a logical order.

```
Create variable INCLEVEL to correct INCOME

            The FREQ Procedure

       Table of inclevel by purchase

 inclevel       purchase

 Frequency     |
 Percent       |
 Row Pct       |
 Col Pct       |< $100   |$100 +   |  Total
 --------------+---------+---------+
 Low Income    |     90  |     42  |    132
               |  20.88  |   9.74  |  30.63
               |  68.18  |  31.82  |
               |  33.46  |  25.93  |
 --------------+---------+---------+
 Medium Income |     98  |     46  |    144
               |  22.74  |  10.67  |  33.41
               |  68.06  |  31.94  |
               |  36.43  |  28.40  |
 --------------+---------+---------+
 High Income   |     81  |     74  |    155
               |  18.79  |  17.17  |  35.96
               |  52.26  |  47.74  |
               |  30.11  |  45.68  |
 --------------+---------+---------+
 Total              269       162      431
                  62.41     37.59   100.00
```



Lesson 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 and doing preliminary examinations of the associations between variables.

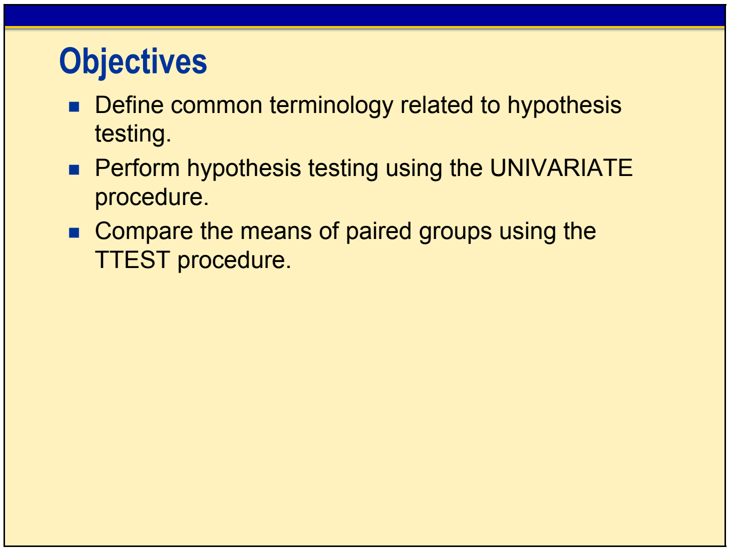


Module Summary

- Defined the difference between continuous and categorical variables.
- Described distributions for both continuous and categorical variables using statistics and graphics.
- Developed confidence limits for continuous variables and verified the assumptions for confidence limits.
- Determined which SAS procedures and statistics were appropriate for each type of variable.

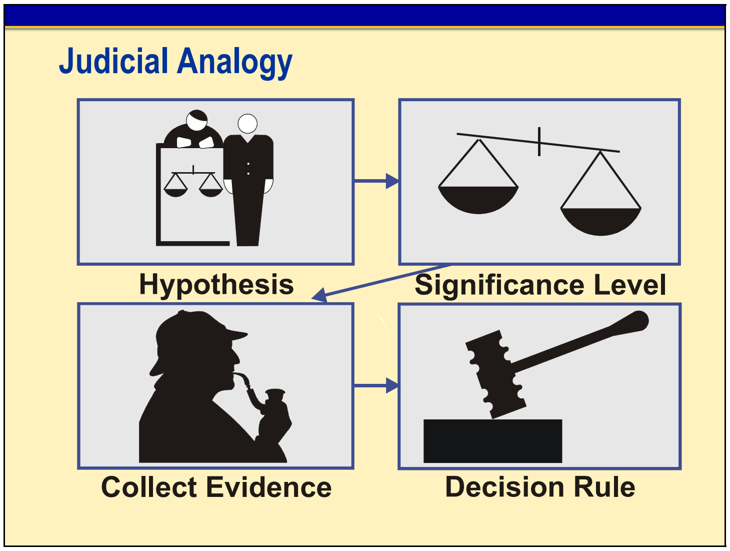
Module 2 Hypothesis Testing and Analysis of Variance

2.1 Hypothesis Testing



Objectives

- Define common terminology related to hypothesis testing.
- Perform hypothesis testing using the UNIVARIATE procedure.
- Compare the means of paired groups using the TTEST procedure.



Judicial Analogy

Hypothesis

Significance Level

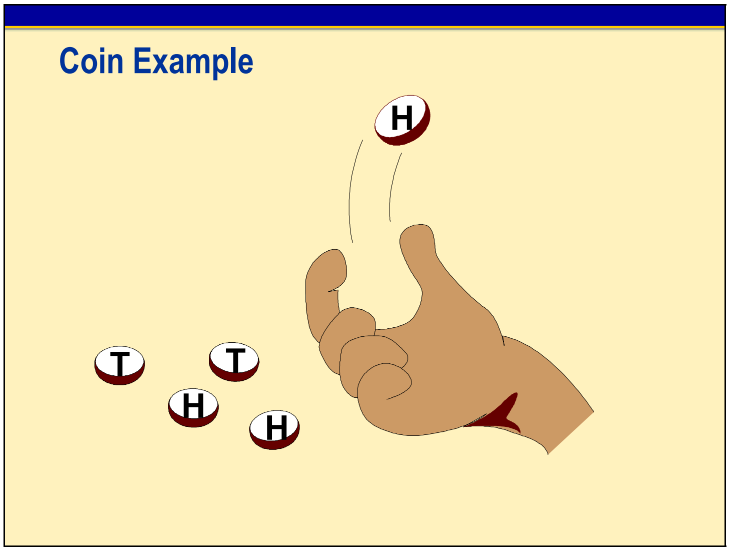
Collect Evidence

Decision Rule

In a criminal court, you put defendants on trial because you suspect they are guilty of a crime. But how does the trial proceed?

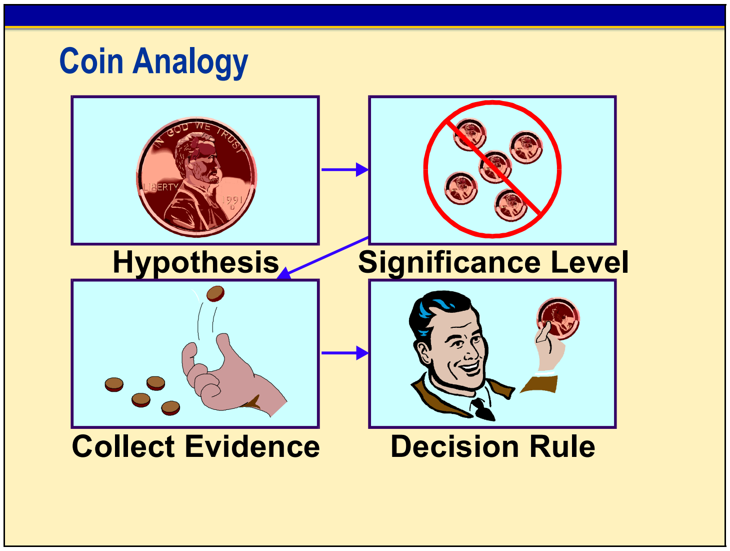
1. Determine the alternative and null hypotheses. The *alternative* hypothesis is your initial research hypothesis (the defendant is guilty). The *null* is the logical opposite of the alternative hypothesis (the defendant is not guilty).
2. Select a *significance level*, the amount of evidence needed to convict. In a court of law, the evidence must prove guilt "beyond a reasonable doubt."
3. Collect evidence.
4. Use a *decision rule* to make a judgment. If the evidence is
   - sufficiently strong, reject the null hypothesis.
   - not strong enough, fail to reject the null hypothesis. (Failing to prove guilt does not prove that the defendant is innocent.)

Statistical hypothesis testing follows this same basic path.



Coin Example

Suppose you want to know whether a coin is fair. You cannot flip it forever, so you decide to take a sample. Flip it five times and count the number of heads and tails.



Coin Analogy

Hypothesis

Significance Level

Collect Evidence

Decision Rule

Test whether a coin is fair.

1. You suspect the coin is **not** fair. However, recall the legal example and begin by assuming the coin is fair.

2. You select a significance level. If you observe five heads in a row or five tails in a row, you conclude the coin is not fair; otherwise, you decide there is **not** enough evidence to show that the coin is not fair.

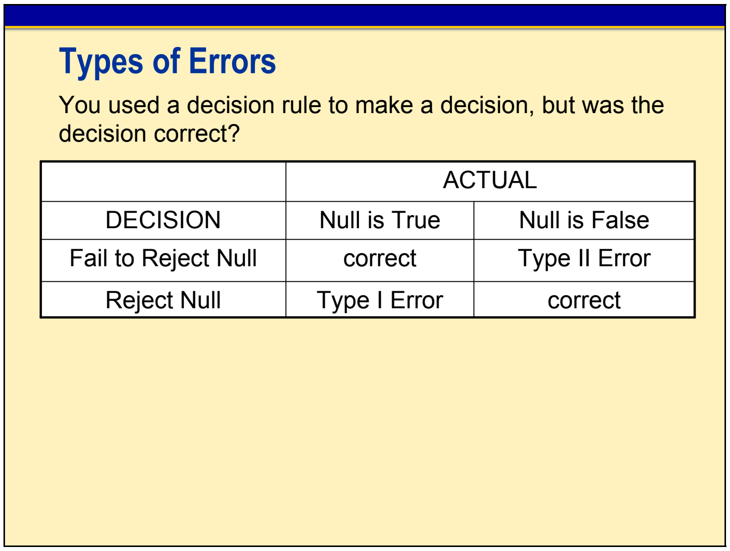
3. You flip the coin five times and count the number of heads and tails.

4. You evaluate the data using your decision rule and make a decision that there either is

   - enough evidence to reject the assumption that the coin is fair

   or

   - not enough evidence to reject the assumption that the coin is fair.



Types of Errors

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

	ACTUAL	
DECISION	Null is True	Null is False
Fail to Reject Null	correct	Type II Error
Reject Null	Type I Error	correct

Recall that you start by assuming the null is true.

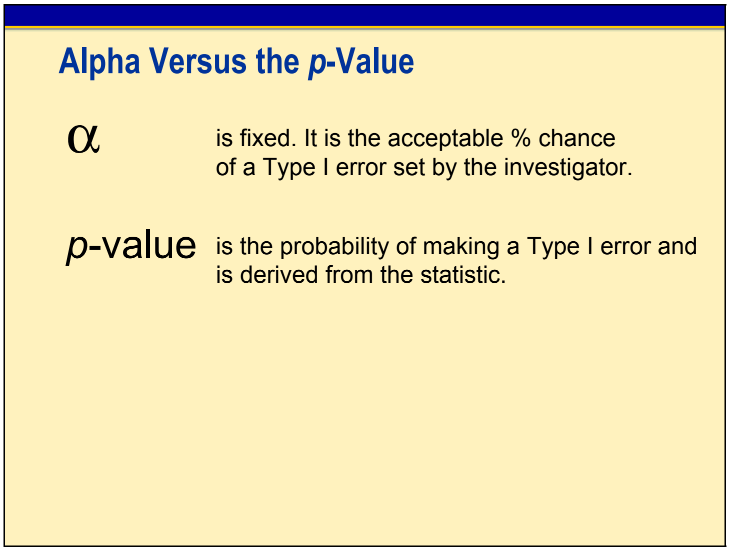
The probability of a Type I error, often denoted α, is the probability that you reject the null hypothesis when it is true. It is also called the significance level of a test.

- In the legal example, it is the probability that you conclude the person is guilty when he or she is innocent.
- In the coin example, it is the probability that you conclude the coin is not fair when it is fair.
- In the cereal example, it is the probability that you conclude that the mean ounces of cereal is not fifteen when it actually is.

The probability of a Type II error, often denoted β, is the probability you fail to reject the null hypothesis when it is false.

- In the legal example, it is the probability that you fail to find the person guilty when he or she is guilty
- In the coin example, it is the probability that you fail to find the coin is not fair when it is not fair.
- In the cereal example, it is the probability that you fail to say the mean is different from fifteen when it is different from fifteen.

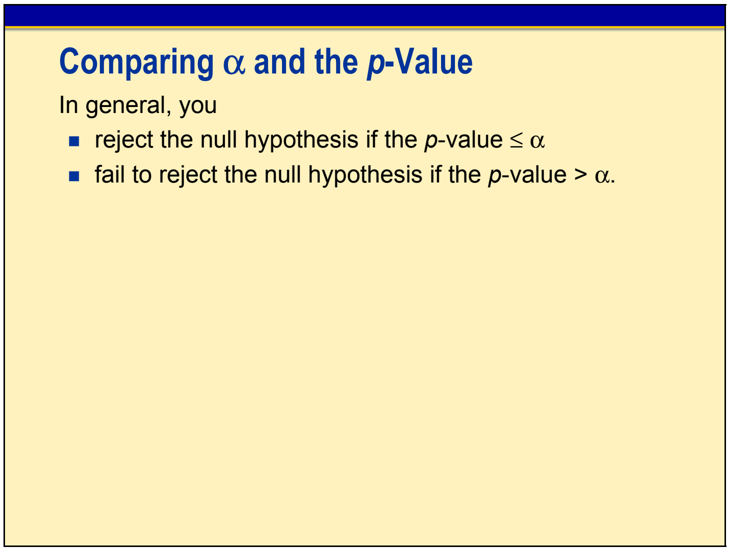
The power of a statistical test is equal to 1–β, where β is the Type II error rate. This is the probability that you correctly reject the null hypothesis.



Alpha Versus the *p*-Value

α is fixed. It is the acceptable % chance of a Type I error set by the investigator.

p-value is the probability of making a Type I error and is derived from the statistic.



Comparing α and the *p*-Value

In general, you

- reject the null hypothesis if the *p*-value $\leq \alpha$
- fail to reject the null hypothesis if the *p*-value $> \alpha$.

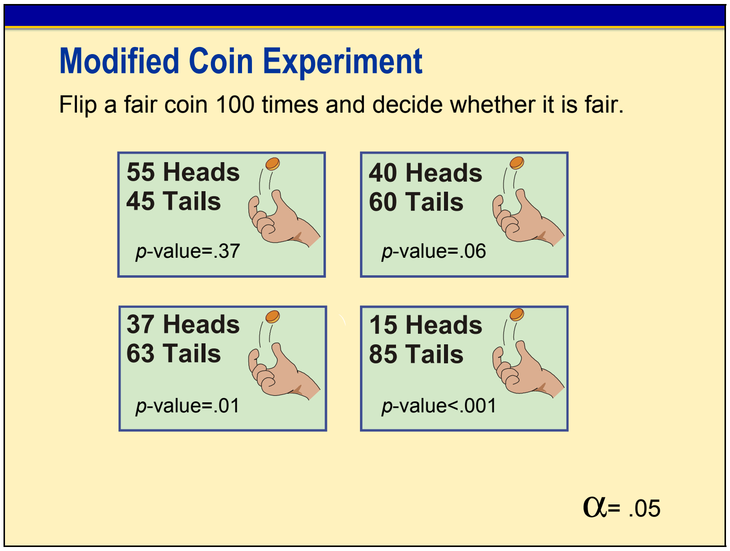
It is important to clarify that

- α, the probability of Type I error, is specified by the experimenter before collecting data
- the *p*-value is calculated from the collected data.

In most statistical hypothesis tests, you compare α and the associated *p*-value to make a decision.

Remember, α is set ahead of time based on the circumstances of the experiment. The level of α is chosen based on what it costs to make a mistake. It is also a function of your knowledge of the data and theoretical considerations.

For the cereal example, α was set to 0.05 based on the consequences of making a Type I error (if you conclude that the mean cereal weight is not 15 ounces when it really is 15 ounces). For example, if making a Type I error causes serious problems, you might want to lower your significance level.



Modified Coin Experiment

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

55 Heads
45 Tails
p-value=.37

40 Heads
60 Tails
p-value=.06

37 Heads
63 Tails
p-value=.01

15 Heads
85 Tails
p-value<.001

α= .05

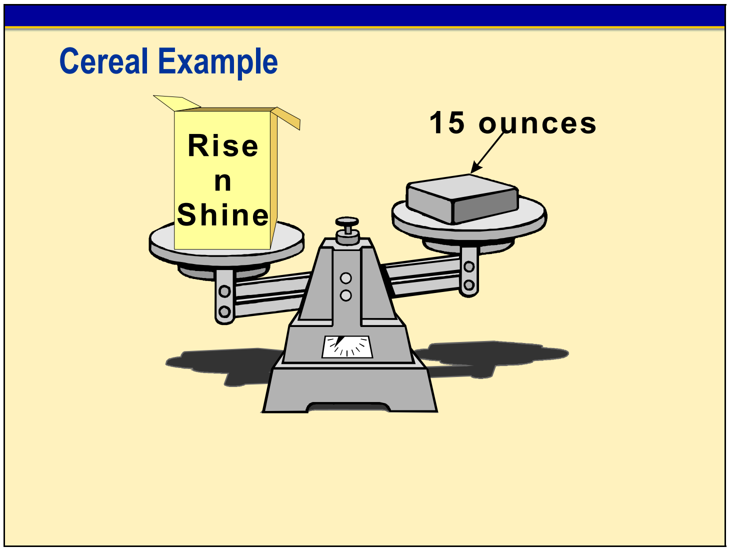
If you flip a coin 100 times and count the number of heads, you do not doubt the coin is fair if you observe exactly 50 heads. However, you might be

- somewhat skeptical that the coin is fair if you observe 40 or 60 heads
- even more skeptical that the coin is fair if you observe 37 or 63 heads
- highly skeptical that the coin is fair if you observe 15 or 85 heads.

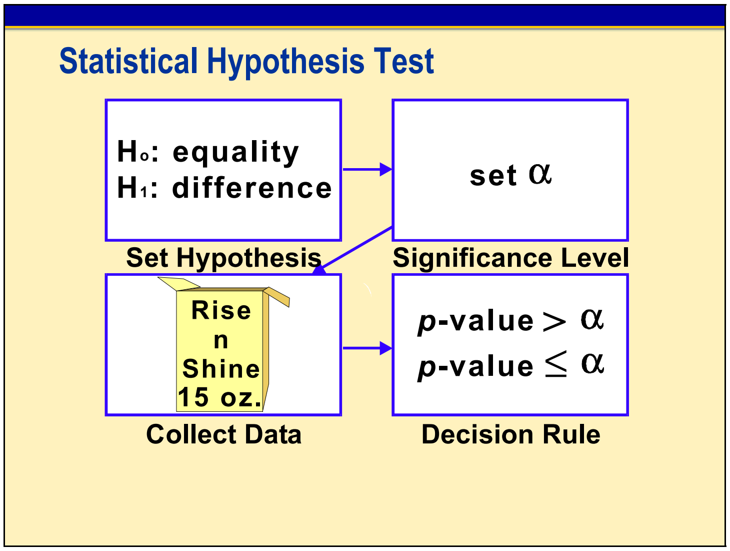
In this example, the greater the difference between the number of heads and tails, the more evidence you have that the coin is not fair.

A *p-value* measures the probability of observing a value as extreme or more extreme than the one observed if the null hypothesis is true. For example, if your null hypothesis is that the coin is fair and you observe 40 heads (60 tails), the *p*-value is the probability of observing a difference in the number of heads and tails of 20 or more from a fair coin tossed 100 times.

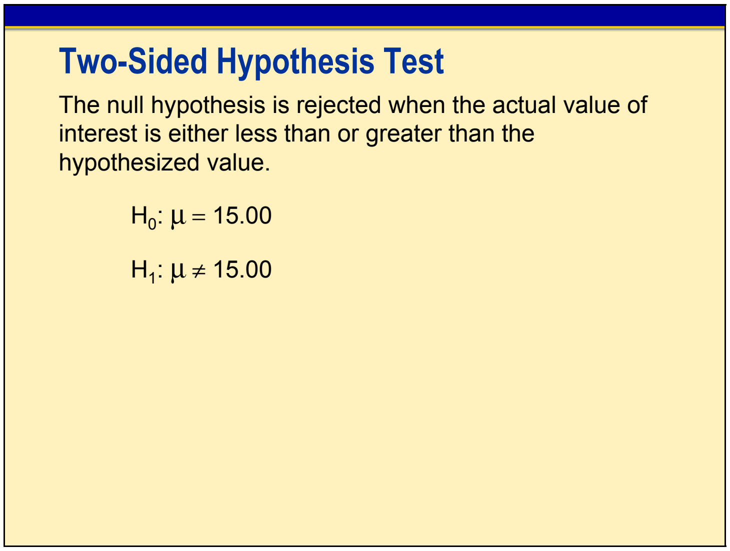
If the *p*-value is large, you often see a difference this large in experiments with a fair coin. If the *p*-value is small, however, you rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair.



Cereal Example



Statistical Hypothesis Test



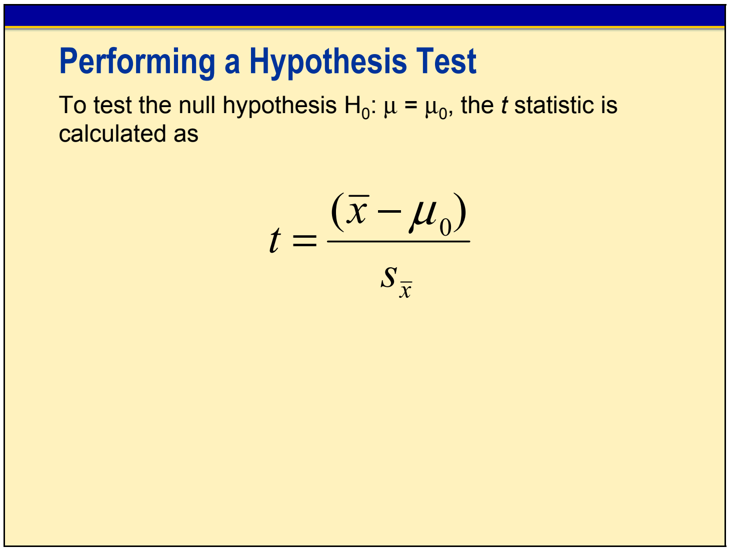
Two-Sided Hypothesis Test

The null hypothesis is rejected when the actual value of interest is either less than or greater than the hypothesized value.

H_0: $\mu = 15.00$

H_1: $\mu \neq 15.00$

In this hypothesis test, it is immaterial whether the mean is greater than or less than the hypothesized mean in rejecting the null hypothesis.



Performing a Hypothesis Test

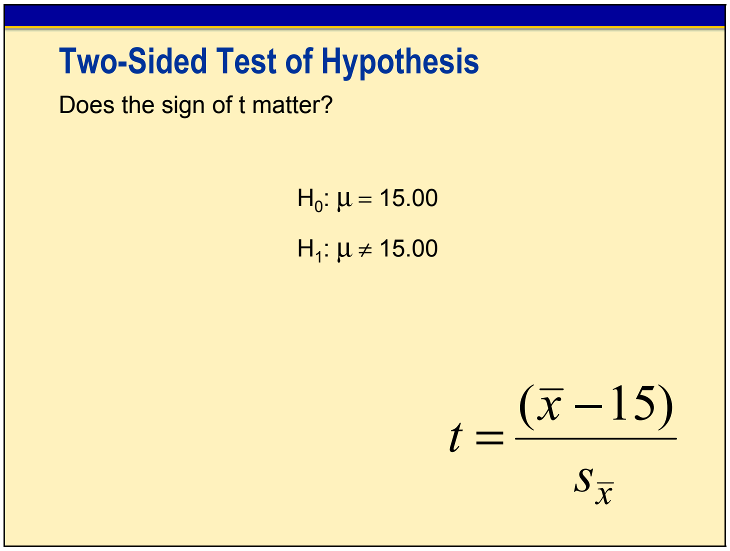
To test the null hypothesis H_0: $\mu = \mu_0$, the *t* statistic is calculated as

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

For the cereal example, μ_0 is the hypothesized value of 15 ounces, $\bar{x}$ is the sample mean weight of the cereal, and $s_{\bar{x}}$ is the standard error of the mean.

- This statistic measures how far $\bar{x}$ is from the hypothesized mean.
- To reject a two-sided test with this statistic, the *t* statistic should be much higher or lower than 0 and have a small corresponding *p*-value.
- The results of this test are valid if the distribution of sample means is normally distributed.

To reject a one-sided test with this statistic, the *t* statistic should have a small corresponding *p*-value and a sign (either positive or negative) that supports the alternative hypothesis. SAS generally reports *p*-values for two-sided tests. Therefore, in SAS, the *p*-value should be divided by 2 if the test statistic has the desired sign.



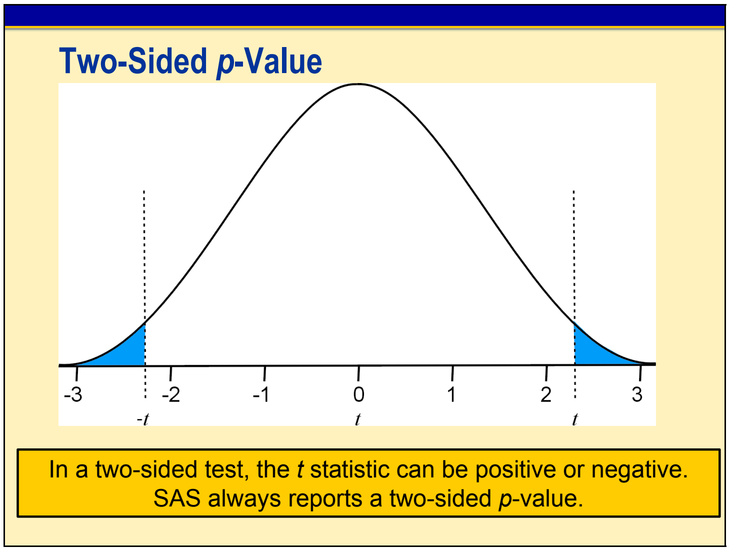
Two-Sided Test of Hypothesis

Does the sign of t matter?

H_0: $\mu = 15.00$

H_1: $\mu \neq 15.00$

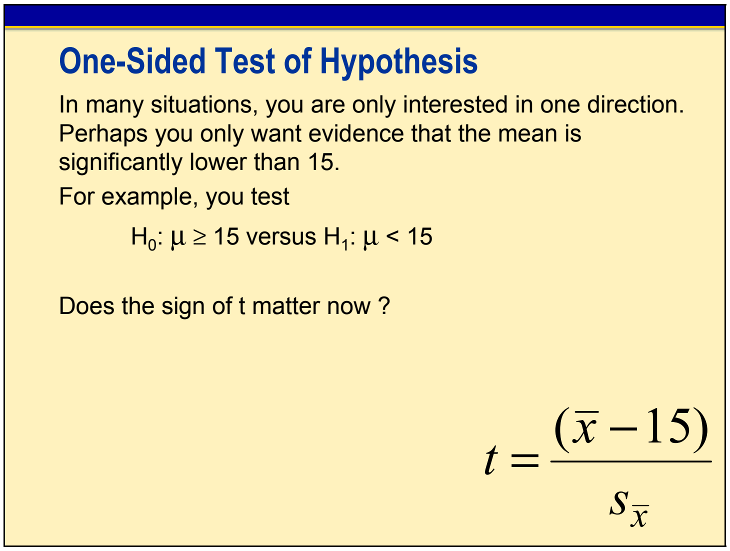
$$t = \frac{(\bar{x} - 15)}{s_{\bar{x}}}$$



Two-Sided *p*-Value

In a two-sided test, the *t* statistic can be positive or negative.
SAS always reports a two-sided *p*-value.

If the researcher is doing a two-sided test, the value of t could be negative or positive. Therefore, the researcher must measure the area under the curve for both a negative and a positive value of the calculated *t* statistic. Because the *t* distribution is symmetric, a two-sided *p*-value is always twice the size of a one-sided *p*-value.



One-Sided Test of Hypothesis

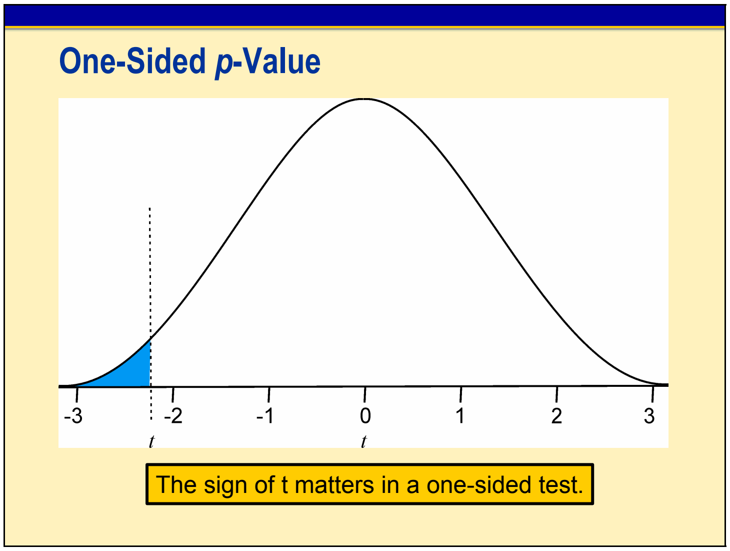
In many situations, you are only interested in one direction. Perhaps you only want evidence that the mean is significantly lower than 15.

For example, you test

H_0: $\mu \geq 15$ versus H_1: $\mu < 15$

Does the sign of t matter now ?

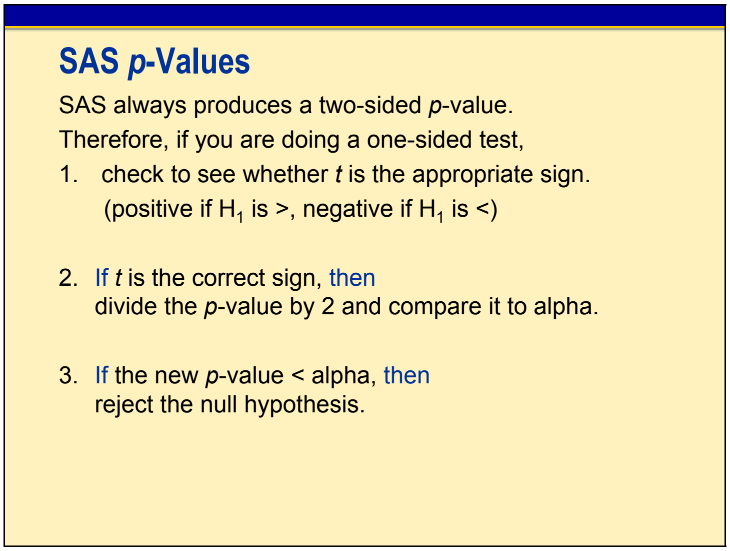
$$t = \frac{(\bar{x} - 15)}{s_{\bar{x}}}$$



One-Sided *p*-Value

The sign of t matters in a one-sided test.

The *t* statistic plotted above is the *t* statistic calculated from the data. The area under the curve between that value of t and the end of the curve represents the *p*-value.

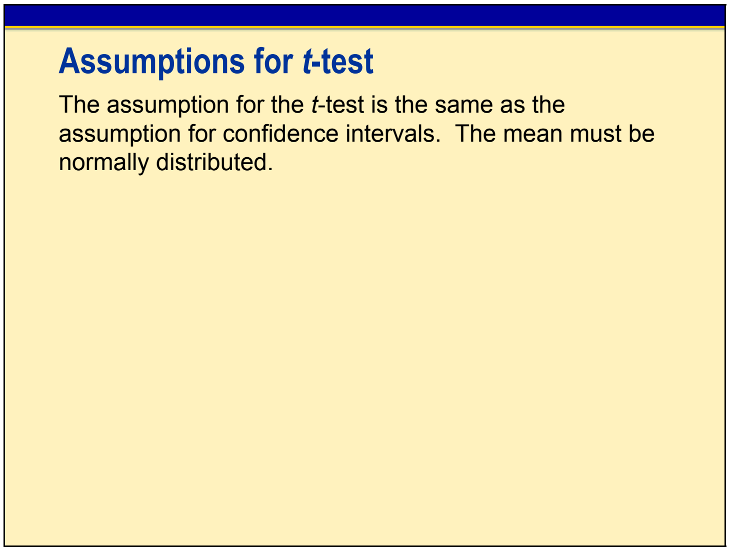


SAS *p*-Values

SAS always produces a two-sided *p*-value.

Therefore, if you are doing a one-sided test,

1. check to see whether *t* is the appropriate sign. (positive if H_1 is >, negative if H_1 is <)
2. If *t* is the correct sign, then divide the *p*-value by 2 and compare it to alpha.
3. If the new *p*-value < alpha, then reject the null hypothesis.



Assumptions for *t*-test

The assumption for the *t*-test is the same as the assumption for confidence intervals. The mean must be normally distributed.

Hypothesis Testing

m2demo01.sas

Example: Use the MU0= option in the UNIVARIATE procedure to test the hypothesis that the mean of the cereal example is equal to 15 ounces.

```
proc univariate data=sasuser.b_rise mu0=15;
   var weight;
   title 'Testing Whether the Mean of Cereal = 15 Ounces';
run;
```

Selected PROC UNIVARIATE statement option:

MU0 = specifies the value of the mean or location parameter in the null hypothesis for tests of location.

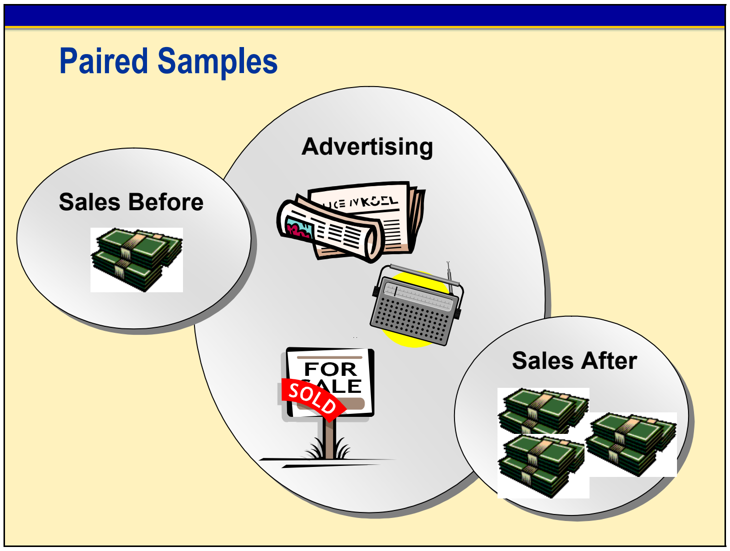
Partial PROC UNIVARIATE Output

```
                        Tests for Location: Mu0=15

            Test           -Statistic-    -----p Value------

            Student's t    t  8.566258    Pr > |t|    <.0001
            Sign           M        17    Pr >= |M|   <.0001
            Signed Rank    S       396    Pr >= |S|   <.0001
```

The *t* statistic and *p*-value are labeled Student’s t and Pr > |t|, respectively.

- The *t* statistic value is 8.566258 and the *p*-value is < .0001.

Therefore, you can reject the null hypothesis at the 0.05 level. Thus, there is enough evidence to conclude the mean is not equal to 15 ounces.



Paired Samples

Sales Before

Advertising

Sales After

For many types of data, repeated measurements are taken on the same subject throughout a study. The simplest form of this study is often referred to as the *paired t-test*.

In this study design,

- subjects are exposed to a treatment, for example, an advertising strategy
- a measurement is taken on the subjects before and after the treatment
- the subjects, on average, respond the same way to the treatment, although there can be differences between the subjects.

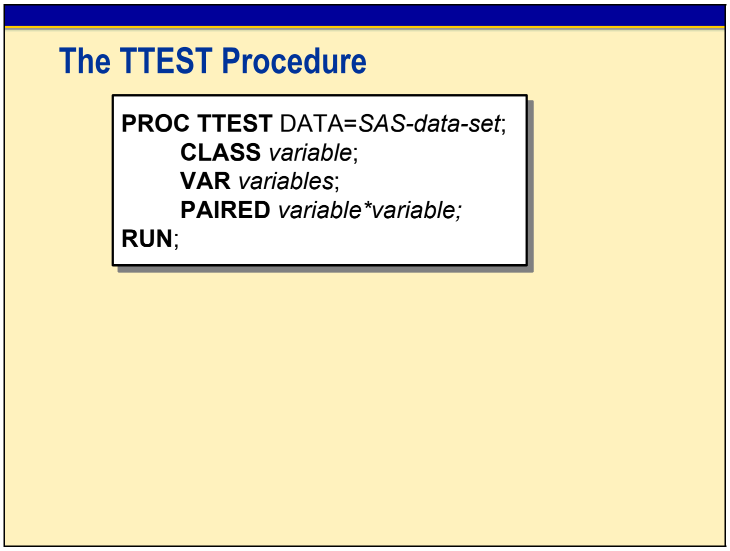
The assumptions of this test are

- the subjects are selected randomly
- the distribution of the sample mean differences is normal.

The hypotheses of this test are

$$H_0: \mu_{POST} \leq \mu_{PRE}$$

$$H_1: \mu_{POST} > \mu_{PRE}$$



The TTEST Procedure

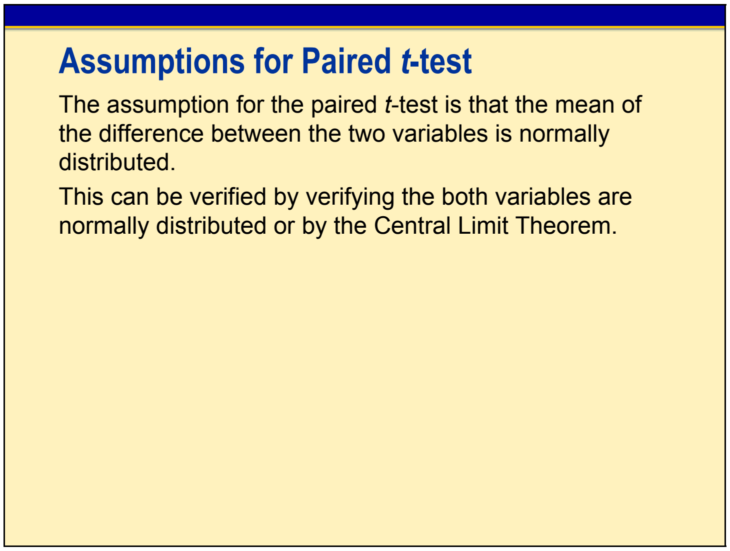
```
PROC TTEST DATA=SAS-data-set;
        CLASS variable;
        VAR variables;
        PAIRED variable*variable;
RUN;
```

Selected TTEST procedure statements:

CLASS specifies the two-level variable for the analysis. Only one variable is allowed in the CLASS statement.

VAR specifies numeric response variables for the analysis. If the VAR statement is not specified, PROC TTEST analyzes all numeric variables in the input data set that are not listed in a CLASS (or BY) statement.

PAIRED identifies the variables to be compared in paired comparisons. Variables are separated by an asterisk (*). The asterisk requests comparisons between each variable on the left with each variable on the right. The differences are calculated by taking the variable on the left minus the variable on the right of the asterisk.



Assumptions for Paired *t*-test

The assumption for the paired *t*-test is that the mean of the difference between the two variables is normally distributed.

This can be verified by verifying the both variables are normally distributed or by the Central Limit Theorem.

Paired *t*-Test

m2demo02.sas, m2demo03.sas

Example: Dollar values of sales have been collected both before and after a particular advertising campaign. You are interested in determining the effect of the campaign on sales. You have collected data from 30 different randomly selected regions. The level of sales both before (**pre**) and after (**post**) the campaign are recorded and are shown below.

```
proc print data=sasuser.b_market (obs=20);
   title;
run;
```

```
                       OBS      PRE      POST

                         1      9.52    10.28
                         2      9.63    10.45
                         3      7.71     8.51
                         4      7.83     8.62
                         5      8.97    10.03
                         6      8.62     9.45
                         7     10.11     9.68
                         8      9.96     9.62
                         9      8.50    11.84
                        10      9.62    11.95
                        11     10.29    10.52
                        12     10.13    10.67
                        13      9.11    11.03
                        14      8.95    10.53
                        15     10.86    10.70
                        16      9.31    10.24
                        17      9.59    10.82
                        18      9.27    10.16
                        19     11.86    12.12
                        20     10.15    11.28
```

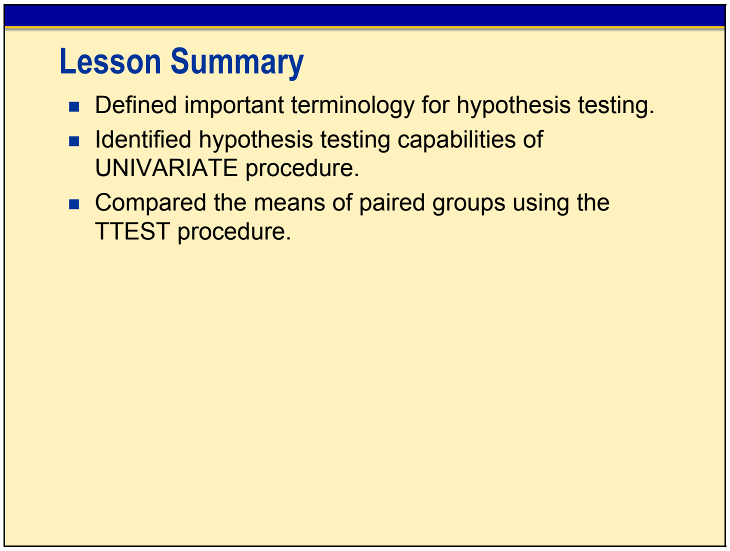
The PAIRED statement used below is testing whether the mean of the post sales is significantly different from the mean of the presales, because **`post`** is on the left of the asterisk and **`pre`** is on the right.

```
proc ttest data=sasuser.b_market;
   paired post*pre;
   title 'Testing the Difference Before and After a Sales '
         'Campaign';
run;
```

```
        Testing the Difference Before and After a Sales Campaign

                       The TTEST Procedure

                           Statistics

                 Lower CL         Upper CL  Lower CL          Upper CL
Difference       N      Mean    Mean      Mean   Std Dev  Std Dev   Std Dev

post - pre      30    0.6001  0.9463    1.2925    0.7384   0.9271    1.2464

                           Statistics

                Difference   Std Err    Minimum    Maximum

                post - pre    0.1693      -0.48       3.34


                            T-Tests

                Difference       DF    t Value    Pr > |t|

                post - pre       29       5.59      <.0001
```

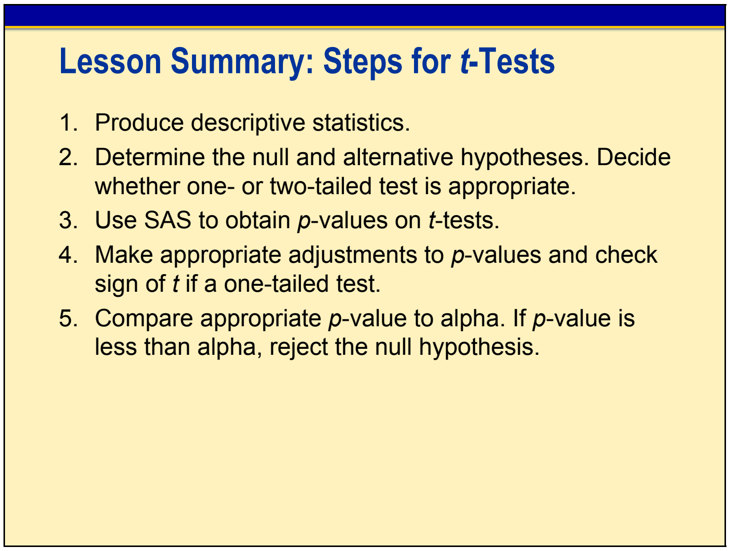
The T-Tests table provides the requested analysis. The *p*-value for the difference **`post`**–**`pre`** is less than 0.0001. Assuming that you want 0.01 level of significance, you reject the null hypothesis and conclude there is a change in the average sales after the advertising campaign. Also, based on the fact that the mean is positive 0.9463, there appears to be an increase in the average sales after the advertising campaign.

Exercise: Refer to your course workbook.



Lesson Summary

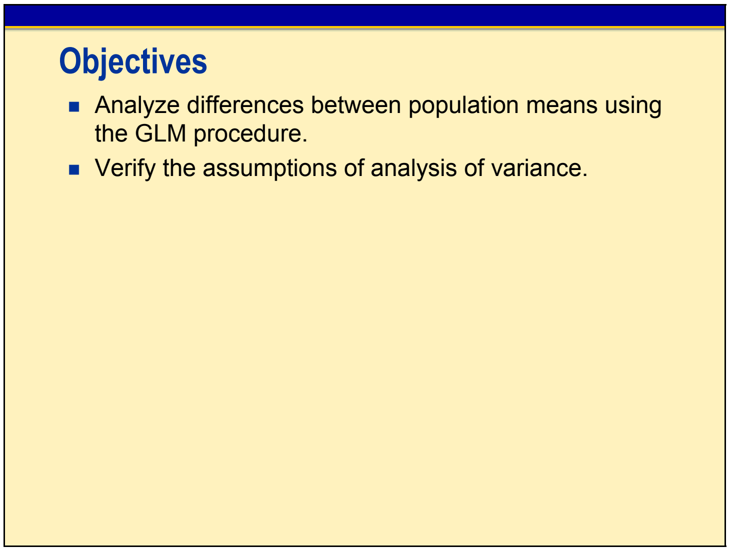
- Defined important terminology for hypothesis testing.
- Identified hypothesis testing capabilities of UNIVARIATE procedure.
- Compared the means of paired groups using the TTEST procedure.



Lesson Summary: Steps for *t*-Tests

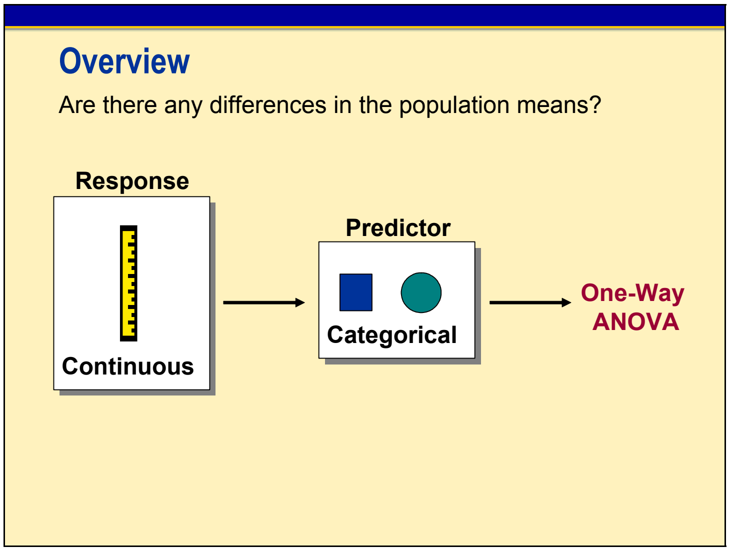
1. Produce descriptive statistics.
2. Determine the null and alternative hypotheses. Decide whether one- or two-tailed test is appropriate.
3. Use SAS to obtain *p*-values on *t*-tests.
4. Make appropriate adjustments to *p*-values and check sign of *t* if a one-tailed test.
5. Compare appropriate *p*-value to alpha. If *p*-value is less than alpha, reject the null hypothesis.

2.2 One-Way ANOVA: Two Populations



Objectives

- Analyze differences between population means using the GLM procedure.
- Verify the assumptions of analysis of variance.

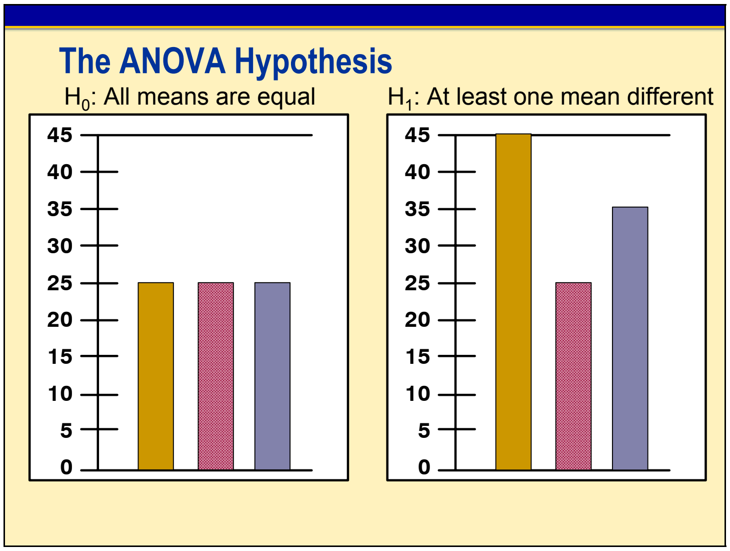


Overview

Are there any differences in the population means?

Analysis of variance (ANOVA) is a statistical technique used to compare the means of two or more groups of observations, or treatments. In this lesson, you apply analysis of variance to examine problems where there are two treatments. For this type of problem, you have a

- continuous dependent, or *response*, variable
- categorical independent variable, also called a *predictor* or *explanatory* variable.



The ANOVA Hypothesis

H_0: All means are equal

H_1: At least one mean different

Small differences between sample means are usually present. The objective is to determine whether these differences are significant. In other words, is the difference more than what might be expected to occur by chance?

The assumptions for ANOVA are

- independent observations
- normally distributed error terms for each treatment
- approximately equal error variances for each treatment.

Descriptive Statistics for Comparing Groups

m2demo04.sas, m2demo05.sas

Example: Print the data in the **`sasuser.b_cereal`** data set and do an initial check of the assumptions of the *t*-test and the *F* test using the UNIVARIATE procedure. Then invoke PROC GLM to test the hypothesis that the means are equal for the two groups.

```
proc print data=sasuser.b_cereal (obs=15);
   title 'Partial Listing of Cereal Data';
run;
```

Part of the data is shown below.

```
                           Partial Listing of Cereal Data

                    OBS    BRAND            WEIGHT       ID

                      1    Morning         14.9982    61469897
                      2    Rise n Shine    15.0136    33081197
                      3    Morning         15.0100    68137597
                      4    Rise n Shine    14.9982    37070397
                      5    Morning         15.0052    64608797
                      6    Rise n Shine    14.9930    60714297
                      7    Morning         14.9733    16907997
                      8    Rise n Shine    15.0812     9589297
                      9    Morning         15.0037    93891897
                     10    Rise n Shine    15.0418    85859397
                     11    Morning         14.9957    38152597
                     12    Rise n Shine    15.0639    99108497
                     13    Morning         15.0099    59666697
                     14    Rise n Shine    15.0613    70847197
                     15    Morning         14.9943    47613397
```

```
proc sort data=sasuser.b_cereal out=sorted_cereal;
   by brand;
run;

options ps=40;
proc univariate data=sorted_cereal plot;
   var weight;
   by brand;
   probplot weight / normal (mu=est sigma=est
                             color=blue w=1);
   title 'Univariate Analysis of the Cereal Data';
run;

options ps=50;
```

In order to generate the analysis for each brand, the data must be sorted by the variable **`brand`**. The SORT procedure step is needed before PROC UNIVARIATE, and the same BY variable used in PROC SORT is needed in PROC UNIVARIATE.

PLOT produces a stem-and-leaf plot, a box-and-whisker plot, and a normal probability plot. When a BY statement is used in combination with the PLOT option, side-by-side box-and-whisker plots are produced.

Partial PROC UNIVARIATE Output

```
                   Univariate Analysis of the Cereal Data

------------------------------ brand=Morning ------------------------------

                          The UNIVARIATE Procedure
                              Variable:  weight

                                   Moments

      N                         40    Sum Weights                 40
      Mean              14.9970125    Sum Observations      599.8805
      Std Deviation     0.02201048    Variance            0.00048446
      Skewness          0.87481049    Kurtosis            2.07993397
      Uncorrected SS    8996.43425    Corrected SS        0.01889398
      Coeff Variation   0.14676575    Std Error Mean      0.00348016


                         Basic Statistical Measures

              Location                    Variability

          Mean     14.99701     Std Deviation           0.02201
          Median   14.99490     Variance              0.0004845
          Mode     14.97790     Range                   0.12010
                                Interquartile Range     0.03095

  NOTE: The mode displayed is the smallest of 2 modes with a count of 2.


                         Tests for Location: Mu0=0

             Test           -Statistic-    -----p Value------

             Student's t    t  4309.286    Pr > |t|    <.0001
             Sign           M        20    Pr >= |M|   <.0001
             Signed Rank    S       410    Pr >= |S|   <.0001


                            Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W      0.95094    Pr < W      0.0817
        Kolmogorov-Smirnov    D     0.078487    Pr > D     >0.1500
        Cramer-von Mises      W-Sq  0.049936    Pr > W-Sq  >0.2500
        Anderson-Darling      A-Sq  0.414338    Pr > A-Sq  >0.2500
```

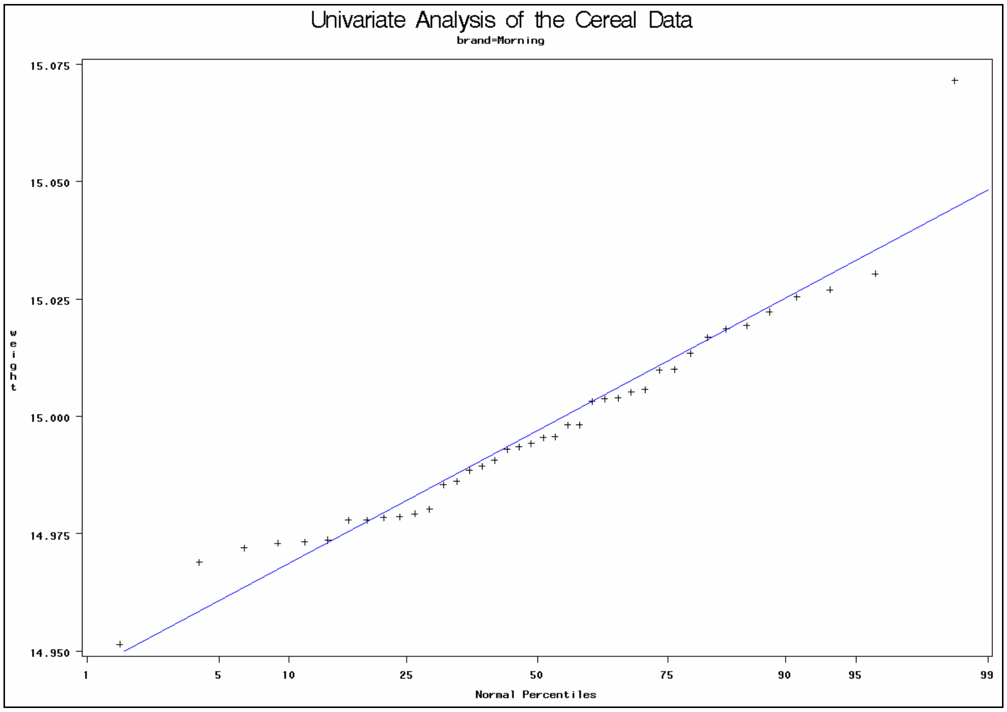
Examine the Tests for Normality table above. The null hypothesis is that the data is normally distributed. Because all the observed *p*-values are greater than 0.05, there is insufficient evidence to reject the null hypothesis.

PROC UNIVARIATE Output (continued)

```
Stem Leaf                      #            Boxplot
1507 2                         1               0
1506
1505
1504
1503 0                         1               |
1502 257                       3               |
1501 003799                    6            +-----+
1500 34456                     5            |     |
1499 13446688                  8            *--+--*
1498 06689                     5            |     |
1497 233488899                 9            +-----+
1496 9                         1               |
1495 2                         1               |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-2
```

The stem-and-leaf plot and the box-and-whisker plot show one extreme value. Otherwise, the data for Morning appears to be symmetric.

PROC UNIVARIATE Output (continued)



The normal probability plot shows no serious departures from normality, allowing for the one extreme point previously noted. There appears to be no pattern for the data that reflects skewness or kurtosis.

PROC UNIVARIATE Output (continued)

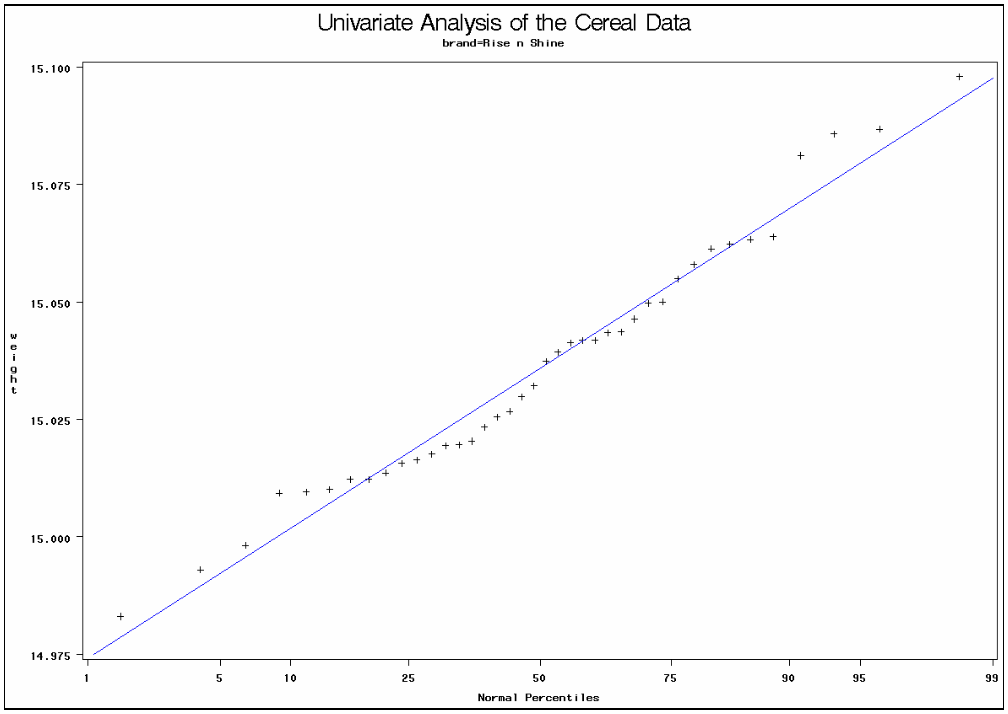
```
                 Univariate Analysis of the Cereal Data

--------------------------- brand=Rise n Shine ----------------------------

                          The UNIVARIATE Procedure
                              Variable:  weight

                                   Moments

      N                            40    Sum Weights                 40
      Mean                   15.03596    Sum Observations      601.4384
      Std Deviation        0.02654963    Variance            0.00070488
      Skewness             0.39889232    Kurtosis            -0.1975717
      Uncorrected SS       9043.23122    Corrected SS        0.02749044
      Coeff Variation      0.17657424    Std Error Mean      0.00419787


                         Basic Statistical Measures

              Location                    Variability

          Mean     15.03596     Std Deviation           0.02655
          Median   15.03480     Variance              0.0007049
          Mode     15.01220     Range                   0.11490
                                Interquartile Range     0.03650

  NOTE: The mode displayed is the smallest of 2 modes with a count of 2.


                            Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W     0.974477    Pr < W      0.4926
        Kolmogorov-Smirnov    D     0.096086    Pr > D     >0.1500
        Cramer-von Mises      W-Sq  0.059304    Pr > W-Sq  >0.2500
        Anderson-Darling      A-Sq  0.387763    Pr > A-Sq  >0.2500
```

The tests for normality for the brand Rise n Shine are not significant. Therefore, there is insufficient evidence to conclude the data is not normally distributed.

The stem-and-leaf plot and the box-and-whisker plot illustrate that the data is fairly symmetric. There are also no extreme values. The normal probability plot shows no serious departures from normality.

PROC UNIVARIATE Output (continued)

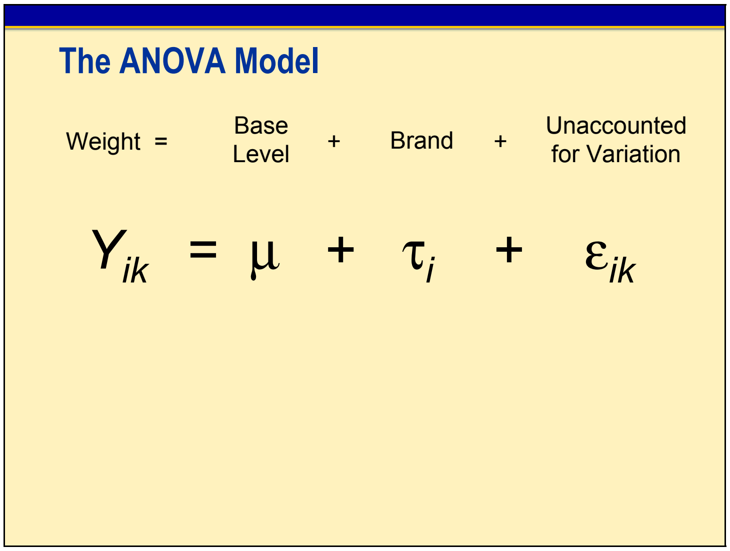
```
Stem Leaf                     #             Boxplot
1509 8                        1                |
1508 167                      3                |
1507                                           |
1506 1234                     4                |
1505 0058                     4             +-----+
1504 122446                   6             |     |
1503 0279                     4             *--+--*
1502 00367                    5             |     |
1501 002246689                9             +-----+
1500 9                        1                |
1499 38                       2                |
1498 3                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-2
```



PROC UNIVARIATE Output (continued)

```
 Univariate Analysis of the Cereal Data

        The UNIVARIATE Procedure
             Schematic Plots

      |
 15.1 +                        |
      |                        |
      |                        |
      |                        |
15.08 +                        |
      |                        |
      |            0           |
      |                        |
15.06 +                        |
      |                     +-----+
      |                     |     |
      |                     |     |
15.04 +                     |     |
      |                     *--+--*
      |            |        |     |
      |            |        |     |
15.02 +            |        |     |
      |            |        +-----+
      |         +-----+        |
      |         |     |        |
   15 +         |     |        |
      |         *--+--*        |
      |         |     |        |
      |         |     |        |
14.98 +         +-----+
      |            |
      |            |
      |            |
14.96 +            |
      |            |
      |            |
      |
14.94 +
       ------------+-----------+-----------
brand           Morning    Rise n S
```

The comparative box-and-whisker plots show that the weights of the brand Rise n Shine have a larger mean and more variability than Morning cereal weights.



The ANOVA Model

Weight = Base Level + Brand + Unaccounted for Variation

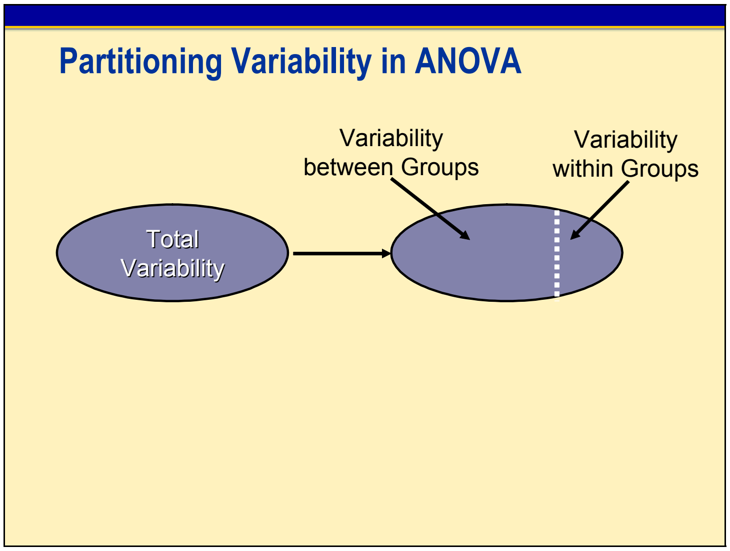
$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

Y_{ik} the k^{th} value of the response variable for the i^{th} treatment.

μ the overall population mean of the response—for instance, cereal weight.

τ_i the difference between the population mean of the i^{th} treatment and the overall mean, μ. This is referred to as the *effect* of treatment *i*.

ε_{ik} error term.



Partitioning Variability in ANOVA

Variability between Groups

Variability within Groups

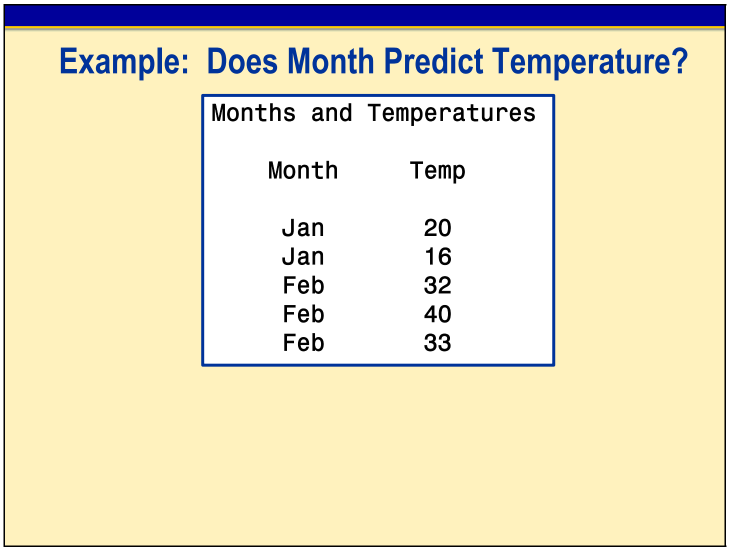
Total Variability

In ANOVA, the corrected total sum of squares is partitioned into two parts, the model sum of squares and the error sum of squares.

model sum of squares (SSM) the variability explained by the independent variable and therefore represented by the **between** treatment sums of squares.

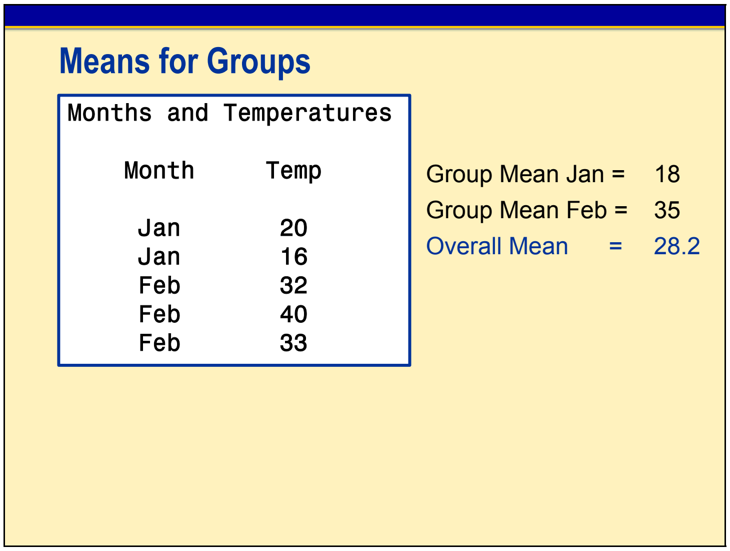
error sum of squares (SSE) the variability not explained by the independent variable. Also referred to as **within** treatment variability.

total sum of squares (SST) the **overall** variability in the response variable. SST=SSM + SSE.



Example: Does Month Predict Temperature?

```
Months and Temperatures

    Month      Temp

     Jan        20
     Jan        16
     Feb        32
     Feb        40
     Feb        33
```



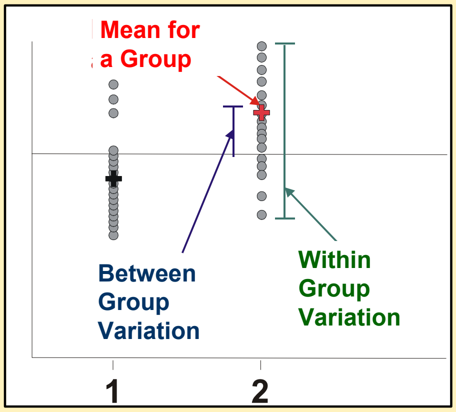
Means for Groups

```
Months and Temperatures

    Month      Temp

     Jan        20
     Jan        16
     Feb        32
     Feb        40
     Feb        33
```

Group Mean Jan = 18
Group Mean Feb = 35
Overall Mean = 28.2

ANOVA is described by some as analysis of variance *from the mean*. In order to find the model sum of squares and the error sum of squares, the mean for each group and the overall mean for the sample must first be calculated.

Sums of Squares

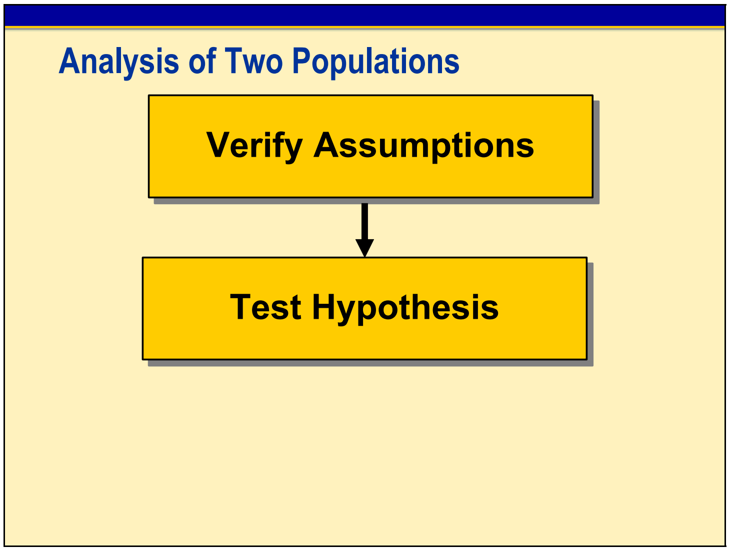


As its name implies, analysis of variance analyzes the variances of the data to determine whether there is a difference between the group means.

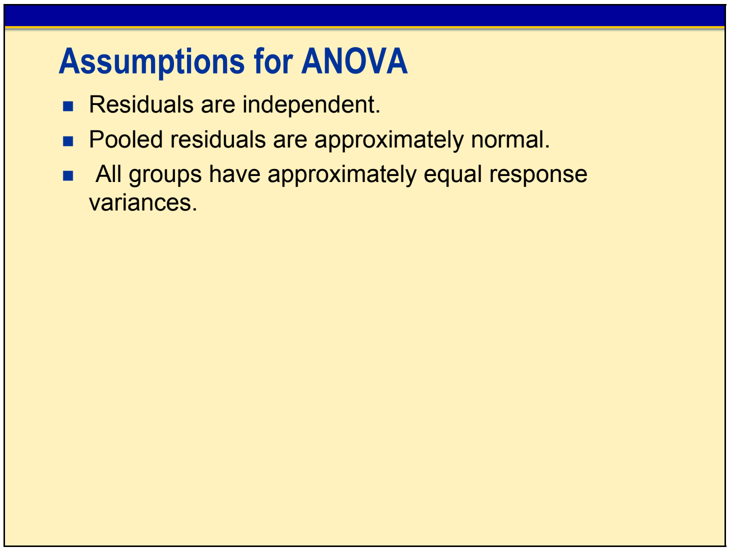
between group variation — the sum of the squared differences between the mean for each group and the overall mean, $\Sigma n_i(\tau_i)^2$.

within group variation — the sum of the squared differences between each observed value and the mean for its group, $\Sigma\Sigma(Y_{ij}-(\mu+\tau_i))^2$.

total variation — the sum of the squared differences between each observed value and the overall mean, $\Sigma\Sigma(Y_{ij}-\mu)^2$.



Analysis of Two Populations

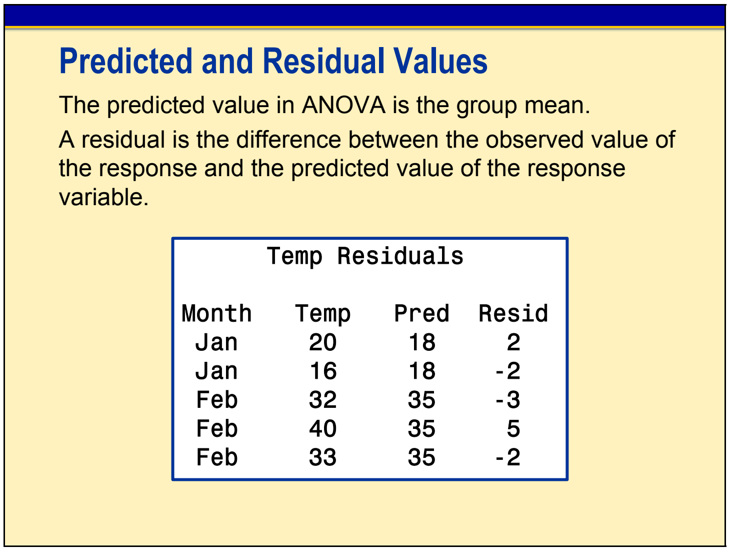


Assumptions for ANOVA

- Residuals are independent.
- Pooled residuals are approximately normal.
- All groups have approximately equal response variances.

One assumption of ANOVA is approximately equal error variances for each group. Although you can get an idea about the equality of variances by looking at the descriptive statistics and plots of the data, you should also consider a formal test for homogeneity of variances. The GLM procedure provides several homogeneity of variance tests with the HOVTEST option, most of which do not require additional assumptions to be valid.

The other assumptions of ANOVA, independent error terms and normally distributed error terms, can be verified by analyzing the residuals. Although the assumption is that the residuals are normally distributed for each group, if the variances are approximately equal, the residuals can be combined into a single group to check for normality.



Predicted and Residual Values

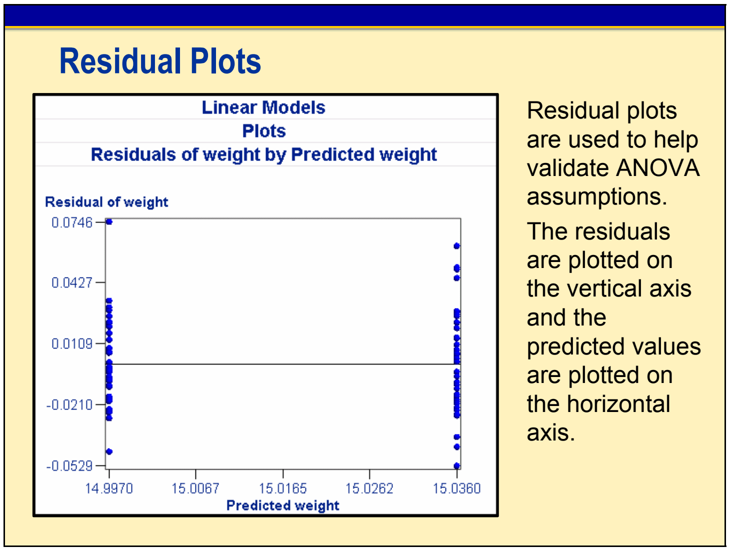
The predicted value in ANOVA is the group mean.

A residual is the difference between the observed value of the response and the predicted value of the response variable.

```
       Temp Residuals

Month   Temp    Pred  Resid
 Jan     20      18     2
 Jan     16      18    -2
 Feb     32      35    -3
 Feb     40      35     5
 Feb     33      35    -2
```

The group means are the predicted values in ANOVA.

Residuals are statistics, implying that they have variability. They can either be positive or negative. Residuals have a unique property in that their sum is zero. A residual is the difference between the observed value of the response and the predicted value of the response variable.



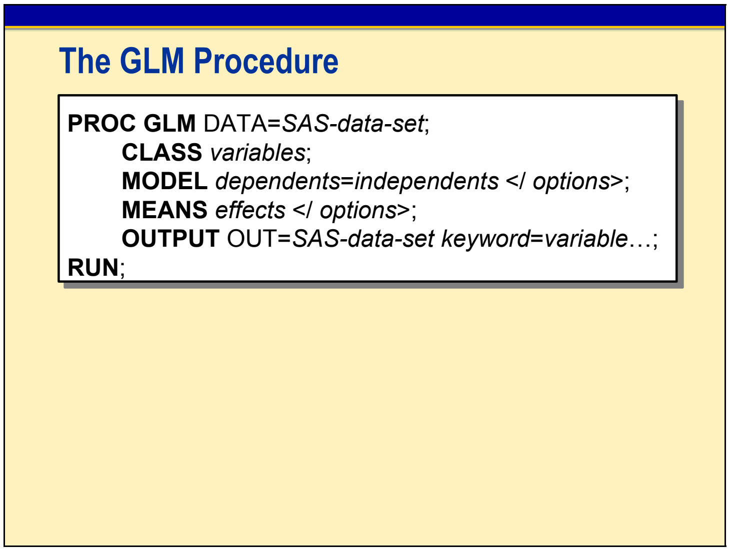
Residual Plots

Residual plots are used to help validate ANOVA assumptions.

The residuals are plotted on the vertical axis and the predicted values are plotted on the horizontal axis.

In a residual plot, the vertical axis represents the residuals, and the horizontal axis represents the predicted values of the dependent variable. The horizontal reference line at 0 represents the average of the residuals.

The residual plot is an important tool in verifying the assumptions of ANOVA. Examine the plot and determine whether the spread of the points for each group is about the same. You are hoping to validate the assumption of equal variances.



The GLM Procedure

```
PROC GLM DATA=SAS-data-set;
      CLASS variables;
      MODEL dependents=independents </ options>;
      MEANS effects </ options>;
      OUTPUT OUT=SAS-data-set keyword=variable…;
RUN;
```

Selected GLM procedure statements:

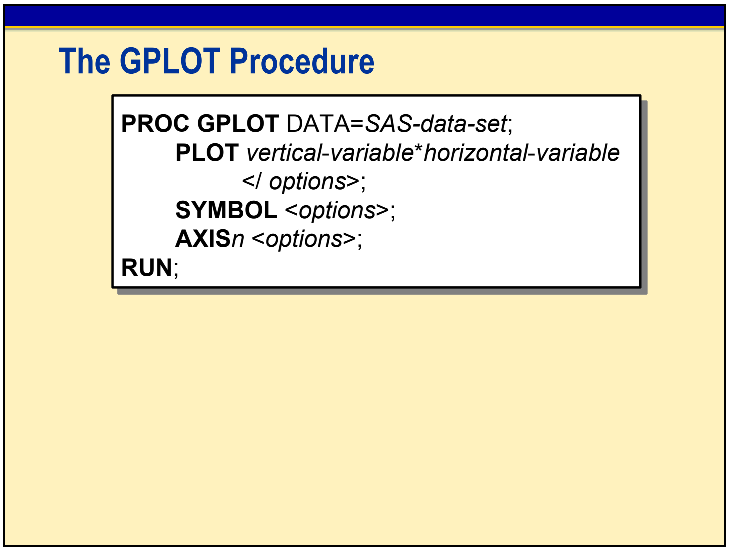
CLASS specifies classification variables for the analysis.

MODEL specifies dependent and independent variables for the analysis.

MEANS computes means of the dependent variable for each value of the specified effect.

OUTPUT specifies an output data set that contains all variables from the input data set and variables representing statistics from the analysis.

PROC GLM supports RUN-group processing, which means the procedure stays active until a PROC, DATA, or QUIT statement is encountered. This enables you to submit additional statements followed by another RUN statement without resubmitting the PROC statement.



The GPLOT Procedure

```
PROC GPLOT DATA=SAS-data-set;
     PLOT vertical-variable*horizontal-variable
          </ options>;
     SYMBOL <options>;
     AXISn <options>;
RUN;
```

The GPLOT procedure is a SAS/GRAPH procedure that produces scatter plots.

Selected GPLOT procedure statements:

PLOT specifies the vertical axis variable and the horizontal axis variable.

SYMBOL defines the appearance of the plotting symbol and plot lines, and optionally specifies the type and additional characteristics of the plot line.

AXIS*n* specifies detailed definitions of individual axis characteristics including the range of values and scaling for the axis, and the number of major and minor tick marks. The value of *n* can range from 1 to 99.

✎ PROC GPLOT supports RUN-group processing.

Verifying ANOVA Assumptions for Two Groups

m2demo06.sas

Example: Test the equality of means for the **sasuser.b_cereal** data set using PROC GLM. Also test for equality of variances and output the residuals for plotting.

```
proc glm data=sasuser.b_cereal;
   class brand;
   model weight=brand;
   means brand / hovtest;
   output out=check r=resid p=pred; /*Output statement*/
                                    /*Creates residuals and */
                                    /*Predicted values for*/
                                    /*assumption validation*/
   title 'Testing for Equality of Means with PROC GLM';
   title2 'HOVTEST Option Tests Equal Variances Using Levene Method';
   title3 'Null hypothesis for Levene is Variances Are Equal';
run;
quit;
```

Selected MEANS statement option:

HOVTEST performs Levene’s test for homogeneity (equality) of variances. The null hypothesis for this test is that the variances are equal. Levene’s test is the default test.

```
goptions reset=all;

proc gplot data= check;
   plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
   symbol v=star h=3pct;
   axis1 w=2 major=(w=2) minor=none offset=(10pct);
   axis2 w=2 major=(w=2) minor=none;
   title 'Plot of Residuals vs. Predicted Values for '
         'Cereal Data Set';
   title2 'Helps Verify Independence and Equal Variance Assumptions';
run;
quit;
```

Selected PLOT statement options:

HAXIS= associates an axis statement with the horizontal axis.

VAXIS= associates an axis statement with the vertical axis.

Selected SYMBOL statement options:

V= specifies the plotting symbol.

H= specifies the height of the plotting symbol in CELLS (the default), CM (centimeters), IN (inches), or PCT (percent)

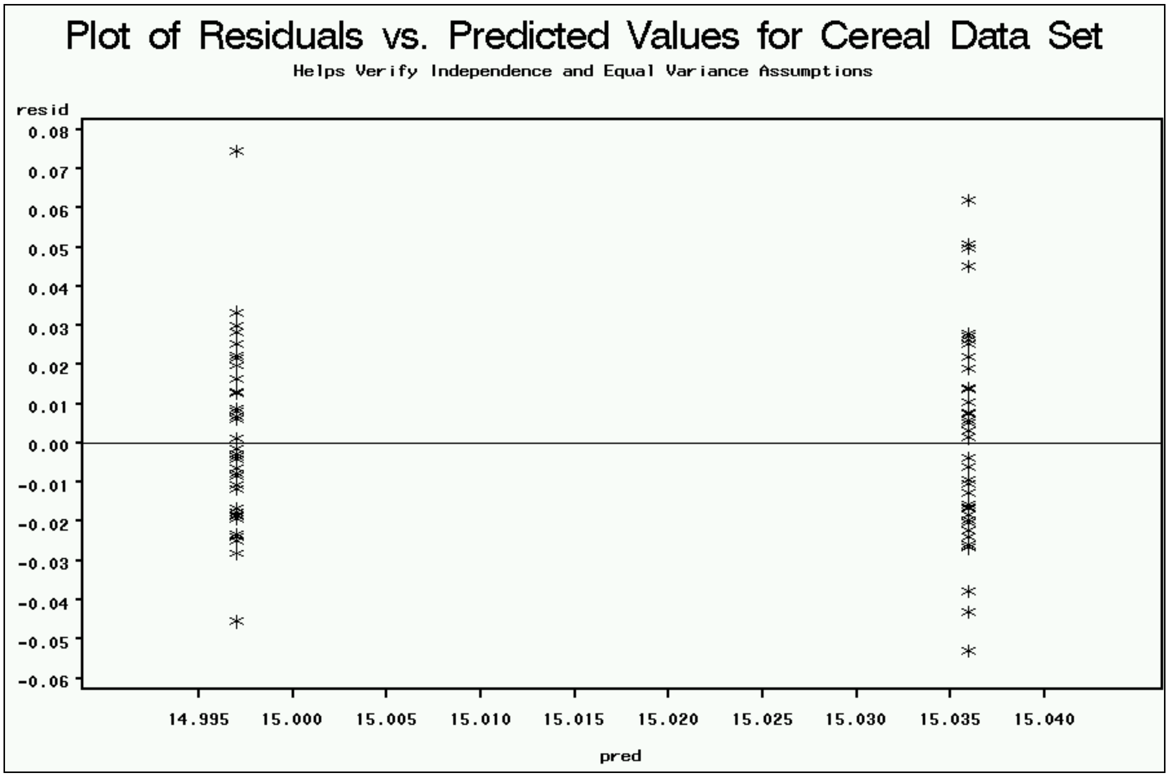
Selected AXIS statement options:

W= specifies the thickness of the axis line.

MAJOR= defines the appearance of major tick marks.

MINOR= defines the appearance of minor tick marks.

PROC GPLOT Output



The graph above is a plot of the residuals versus the fitted values from the ANOVA model. Essentially, you are looking for a random scatter about the zero reference line for each of the fitted values. Any patterns or trends in this plot can indicate model assumption violations.

A similar graph can be generated using the PLOT procedure to provide the same information as PROC GPLOT.

```
proc plot data= check;
   plot resid*pred / vref=0;
   title 'PROC PLOT is an Alternative if GRAPH is Unavailable';
run;
quit;
```

✎ The letters in the plot symbolize the number of points at that specific value in the plot.

PROC PLOT Output

```
   PROC PLOT is an Alternative if GRAPH is Unavailable
 Plot of resid*pred.  Legend: A = 1 obs, B = 2 obs, etc.

resid |
 0.08 +
      |  A
      |
      |
      |                                               A
 0.06 +
      |
      |                                               A
      |                                               A
      |                                               A
 0.04 +
      |
      |  A
      |  B                                            C
      |  B                                            B
 0.02 +  B                                            A
      |  A                                            A
      |  B                                            B
      |  E                                            B
      |                                               D
 0.00 +--D--------------------------------------------A--
      |  C                                            A
      |  C                                            B
      |  B                                            B
      |  B                                            C
-0.02 +  D                                            C
      |  D                                            D
      |  A                                            B
      |
      |                                               A
-0.04 +
      |  A                                            A
      |
      |                                               A
      |
-0.06 +
      ---+--------------------------------------------+--
      14.997                                       15.036

                              pred
```

```
proc univariate data=check normal plot;
   var resid;
   histogram / normal;
   title 'Verify Normality of Errors Assumption';
run;
```

Selected PROC UNIVARIATE statement options:

NORMAL produces four test statistics and their corresponding *p*-values for testing

H_0: Normal Distribution

versus

H_1: Nonnormal Distribution

Output from the UNIVARIATE procedure helps to verify the assumption of normality of the residuals. The box plot, stem-and-leaf plot, normal probability plot, the histogram with the normal curve superimposed on it, and the statistics found in the Goodness-of-Fit Tests for Normal Distribution table do not indicate any major departures from normality.

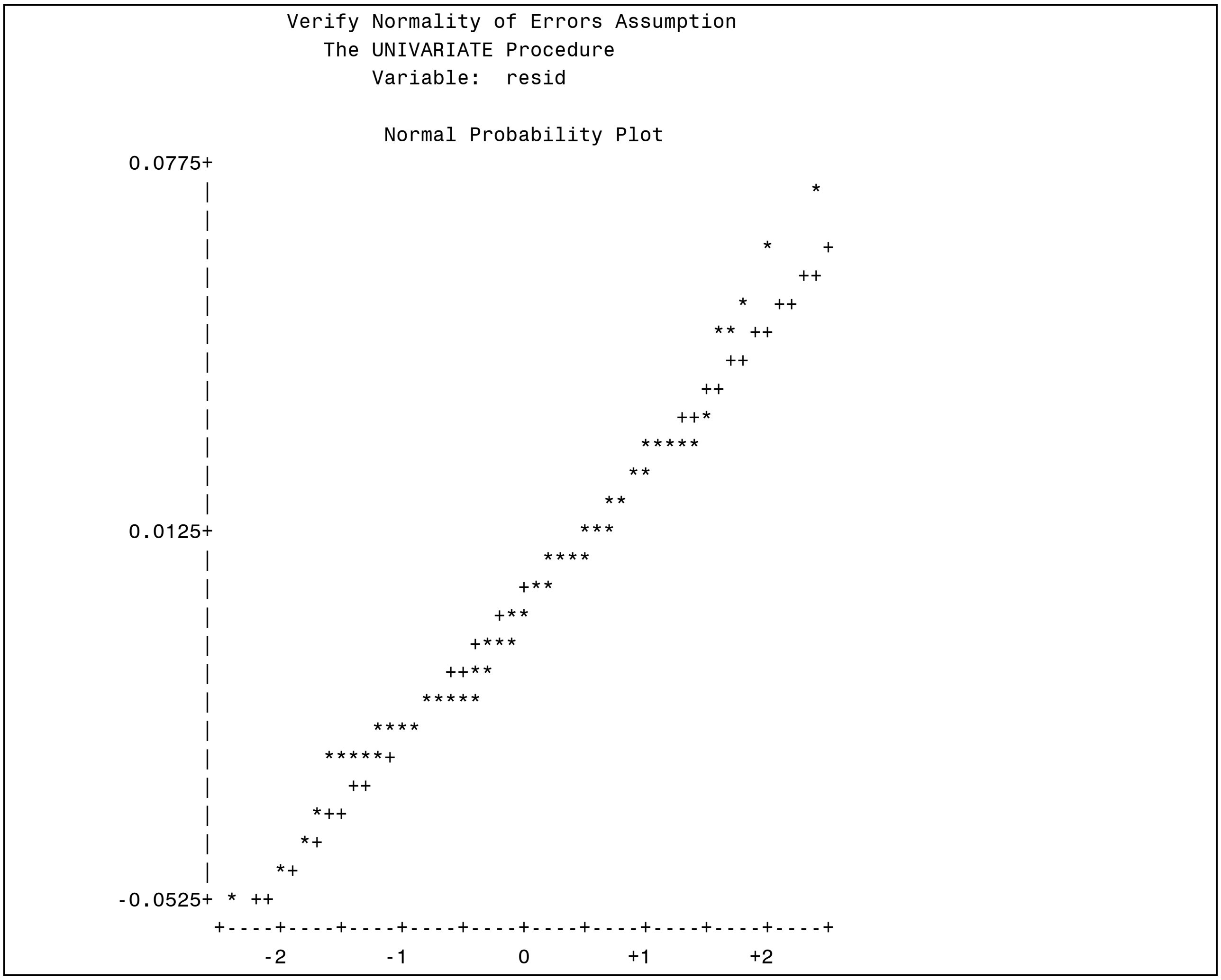
Partial PROC UNIVARIATE Output

```
                    Verify Normality of Errors Assumption
                         The UNIVARIATE Procedure
                             Variable:  resid

                                  Moments

      N                          80    Sum Weights                 80
      Mean                        0    Sum Observations             0
      Std Deviation      0.02423107    Variance            0.00058714
      Skewness           0.56777406    Kurtosis            0.54402045
      Uncorrected SS     0.04638442    Corrected SS        0.04638442
      Coeff Variation             .    Std Error Mean      0.00270912


                        Basic Statistical Measures

              Location                    Variability

          Mean      0.00000     Std Deviation           0.02423
          Median   -0.00211     Variance              0.0005871
          Mode     -0.02376     Range                   0.12745
                                Interquartile Range     0.03233

  NOTE: The mode displayed is the smallest of 4 modes with a count of 2.
```

PROC UNIVARIATE Output (continued)

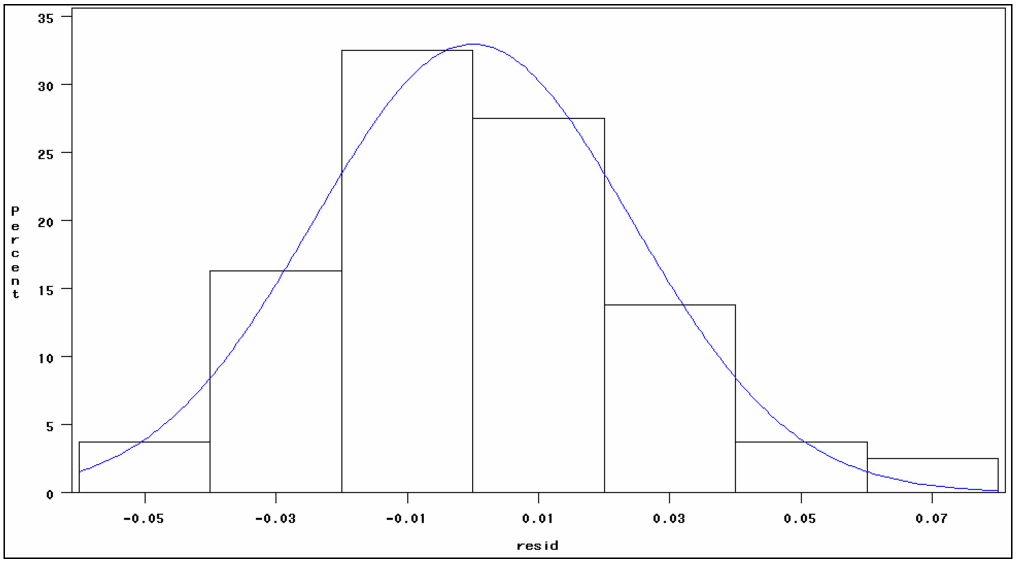
```
                    Verify Normality of Errors Assumption
                       The UNIVARIATE Procedure
                             Variable:  resid

             Stem Leaf                     #             Boxplot
                7 5                        1                0
                7
                6
                6 2                        1                |
                5                                           |
                5 01                       2                |
                4 5                        1                |
                4                                           |
                3                                           |
                3 03                       2                |
                2 556788                   6                |
                2 0222                     4                |
                1 69                       2                |
                1 03344                    5             +-----+
                0 5666778889              10             |     |
                0 1113                     4             |  +  |
               -0 444321                   6             *-----*
               -0 99866                    5             |     |
               -1 3210                     4             |     |
               -1 9998887766              10             +-----+
               -2 44443200                 8                |
               -2 87665                    5                |
               -3                                           |
               -3 8                        1                |
               -4 3                        1                |
               -4 6                        1                |
               -5 3                        1                |
                  ----+----+----+----+
              Multiply Stem.Leaf by 10**-2
```

PROC UNIVARIATE Output (continued)



```
                Verify Normality of Errors Assumption
                   The UNIVARIATE Procedure
                       Variable:  resid

                        Normal Probability Plot
    0.0775+
          |                                                *
          |
          |                                            *    +
          |                                               ++
          |                                          *  ++
          |                                        ** ++
          |                                         ++
          |                                       ++
          |                                     ++*
          |                                  *****
          |                                 **
          |                               **
    0.0125+                             ***
          |                          ****
          |                        +**
          |                      +**
          |                    +***
          |                  ++**
          |                *****
          |            ****
          |        *****+
          |          ++
          |       *++
          |      *+
          |    *+
   -0.0525+ * ++
          +----+----+----+----+----+----+----+----+----+----+
              -2        -1         0        +1        +2
```

PROC UNIVARIATE Output (continued)



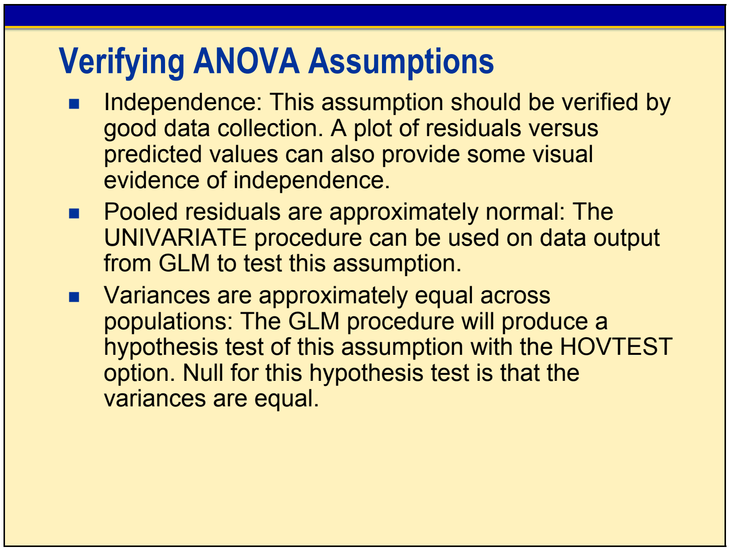
```
                     Verify Normality of Errors Assumption
                           The UNIVARIATE Procedure
                        Fitted Distribution for resid

                      Parameters for Normal Distribution

                        Parameter   Symbol   Estimate

                        Mean        Mu              0
                        Std Dev     Sigma    0.024231


               Goodness-of-Fit Tests for Normal Distribution

         Test                  ---Statistic----   -----p Value-----

         Kolmogorov-Smirnov    D     0.07711238   Pr > D     >0.150
         Cramer-von Mises      W-Sq  0.08755262   Pr > W-Sq   0.167
         Anderson-Darling      A-Sq  0.61754096   Pr > A-Sq   0.105
```

The tests for normality show no serious departure from the assumption that the residuals are normal.

The output below is the result of the HOVTEST option in the MEANS statement. Levene's test for homogeneity of variances is the default test. The null hypothesis is that the variances for the treatments are equal. The *p*-value indicates that you do not reject the null hypothesis. Therefore, the assumption of equal variances appears to be satisfied.

```
             Testing for Equality of Means with PROC GLM
        HOVTEST Option Tests Equal Variances Using Levene Method
            Null hypothesis for Levene is Variances Are Equal

                            The GLM Procedure

            Levene's Test for Homogeneity of weight Variance
              ANOVA of Squared Deviations from Group Means

                            Sum of        Mean
      Source        DF     Squares      Square    F Value    Pr > F

      brand          1    9.237E-7    9.237E-7       1.12    0.2942
      Error         78    0.000065    8.283E-7
```

If at this point you determine that the variances are not equal, you would add the WELCH option to the MEANS statement. This requests Welch's (1951) variance-weighted one-way ANOVA. This alternative to the usual ANOVA is robust to the assumption of equal variances. This is similar to the unequal variance *t*-test for two populations.



Verifying ANOVA Assumptions

- Independence: This assumption should be verified by good data collection. A plot of residuals versus predicted values can also provide some visual evidence of independence.
- Pooled residuals are approximately normal: The UNIVARIATE procedure can be used on data output from GLM to test this assumption.
- Variances are approximately equal across populations: The GLM procedure will produce a hypothesis test of this assumption with the HOVTEST option. Null for this hypothesis test is that the variances are equal.

ANOVA for Two Groups

m2demo06.sas

After you are satisfied that the assumptions are met, turn your attention to the first page of the PROC GLM output, which specifies the number of levels, the values of the class variable, and the number of observations.

```
                  Testing for Equality of Means with PROC GLM
         HOVTEST Option Tests Equal Variances Using Levene Method
             Null hypothesis for Levene is Variances Are Equal

                             The GLM Procedure

                          Class Level Information

               Class         Levels    Values

               brand              2    Morning Rise n Shine


                       Number of observations    80
```

The second page of the output contains all of the information needed to test the equality of the treatment means.

```
              Testing for Equality of Means with PROC GLM
        HOVTEST Option Tests Equal Variances Using Levene Method
            Null hypothesis for Levene is Variances Are Equal

                            The GLM Procedure

Dependent Variable: weight

                                    Sum of
 Source                      DF       Squares   Mean Square  F Value  Pr > F

 Model                        1    0.03033816    0.03033816    51.02  <.0001

 Error                       78    0.04638442    0.00059467

 Corrected Total             79    0.07672257


           R-Square     Coeff Var      Root MSE    weight Mean

           0.395427      0.162394      0.024386       15.01649


 Source                      DF     Type I SS   Mean Square  F Value  Pr > F

 brand                        1    0.03033816    0.03033816    51.02  <.0001


 Source                      DF   Type III SS   Mean Square  F Value  Pr > F

 brand                        1    0.03033816    0.03033816    51.02  <.0001
```

This output is divided into three parts:

- the analysis of variance table
- descriptive information
- information about the class variable in the model.

Look at each of these parts separately.

```
                                     Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F

Model                       1    0.03033816    0.03033816    51.02  <.0001

Error                      78    0.04638442    0.00059467

Corrected Total            79    0.07672257
```

In general, *degrees of freedom* (DF) can be thought of as the number of independent pieces of information.

- Model DF is the number of treatments (or groups) minus 1.
- Corrected total DF is the sample size minus 1.

Mean squares are calculated by taking sums of squares and dividing by the corresponding degrees of freedom.

- Mean square for error (MSE) is an estimate of σ^2, the constant variance assumed for all treatments.
- If $\mu_1 = \mu_2$, the mean square for the model (MSM) is also an estimate of σ^2.
- If $\mu_1 \neq \mu_2$, MSM estimates σ^2 plus a positive constant.
- $F = \dfrac{MSM}{MSE}$.

Based on the above, if the *F* statistic is significantly larger than 1, it supports rejecting the null hypothesis, concluding that the treatment means are not equal.

The *F* statistic and corresponding *p*-value are reported in the analysis of variance table. Because the reported *p*-value is less than 0.0001, you conclude that there is a statistical difference between the means.

```
                          R-Square     Coeff Var      Root MSE    weight Mean

                          0.395427      0.162394      0.024386       15.01649
```

The *coefficient of determination*, R^2 (denoted in this table as R-Square), is a measure of the proportion of variability explained by the independent variables in the analysis. This statistic is calculated as

$$R^2 = \frac{SSM}{SST}$$

The value of R^2 is between 0 and 1. The value is

- close to 0, if the independent variables do not explain much variability in the data
- close to 1, if the independent variables explain a relatively large proportion of variability in the data.

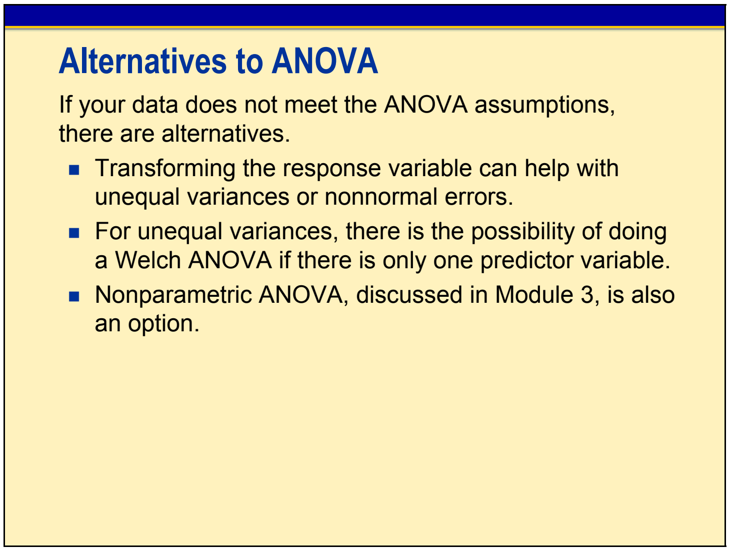
Although values of R^2 closer to 1 are preferred, judging the magnitude of R^2 depends on the context of the problem.

The coefficient of variation (denoted Coeff Var) expresses the root MSE (the estimate of the standard deviation for all treatments) as a percent of the mean. It is a unitless measure that is useful in comparing the variability of two sets of data with different units of measure.

The weight Mean is the mean of all of the data values in the variable **`weight`** without regard to **`brand`**.

```
Source                     DF     Type I SS   Mean Square  F Value  Pr > F

 brand                       1    0.03033816    0.03033816    51.02  <.0001


 Source                     DF   Type III SS   Mean Square  F Value  Pr > F

 brand                       1    0.03033816    0.03033816    51.02  <.0001
```

For a one-way analysis of variance (only one classification variable), the information about the class variable in the model is an exact duplicate of the model line of the analysis of variance table.

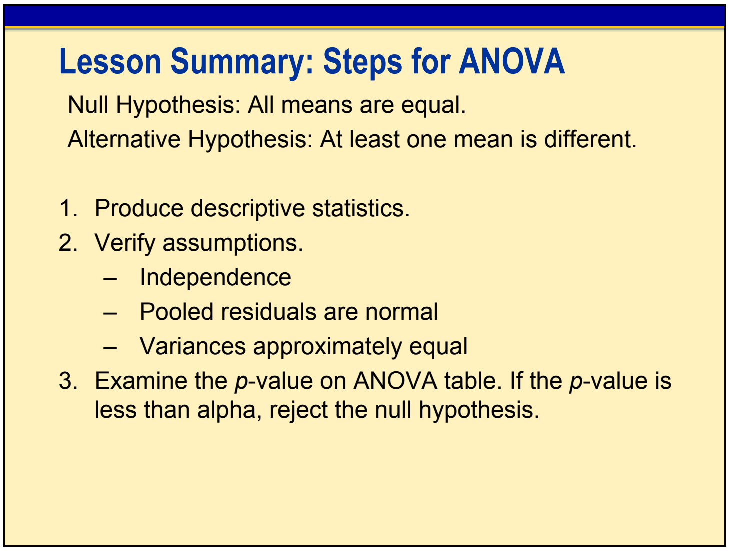


Alternatives to ANOVA

If your data does not meet the ANOVA assumptions, there are alternatives.

- Transforming the response variable can help with unequal variances or nonnormal errors.
- For unequal variances, there is the possibility of doing a Welch ANOVA if there is only one predictor variable.
- Nonparametric ANOVA, discussed in Module 3, is also an option.

Exercise: Refer to your course workbook.



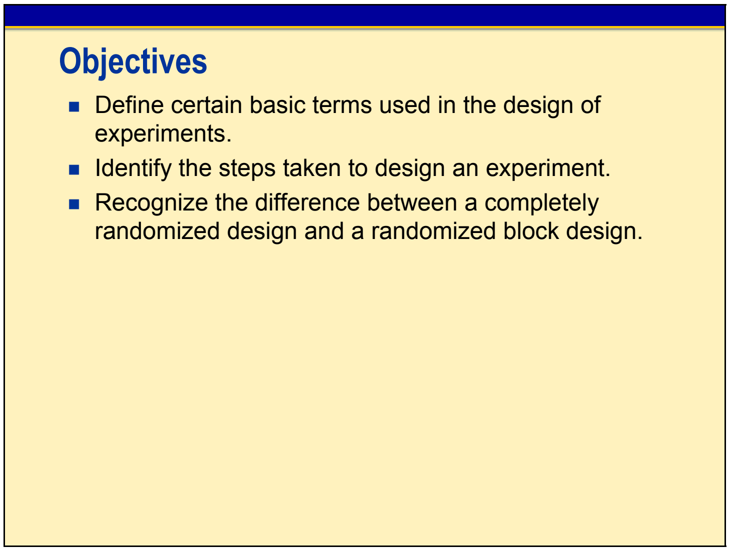
Lesson Summary: Steps for ANOVA

Null Hypothesis: All means are equal.

Alternative Hypothesis: At least one mean is different.

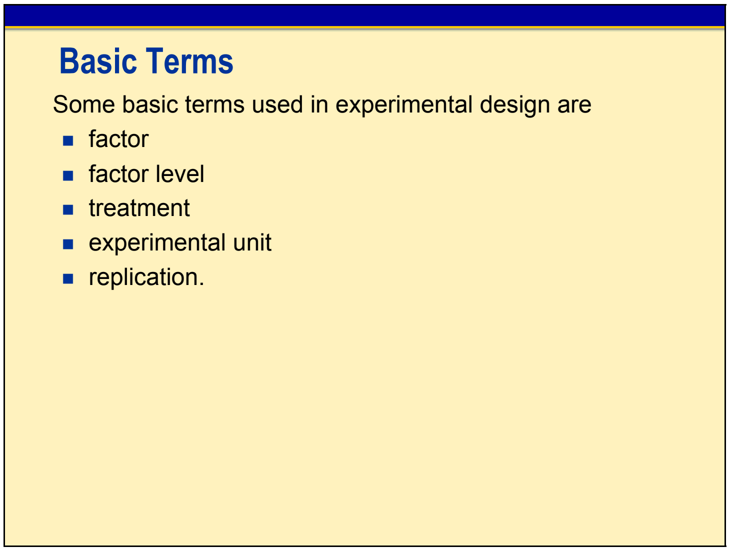
1. Produce descriptive statistics.
2. Verify assumptions.
   - Independence
   - Pooled residuals are normal
   - Variances approximately equal
3. Examine the *p*-value on ANOVA table. If the *p*-value is less than alpha, reject the null hypothesis.

2.3 Design of Experiments



Objectives

- Define certain basic terms used in the design of experiments.
- Identify the steps taken to design an experiment.
- Recognize the difference between a completely randomized design and a randomized block design.

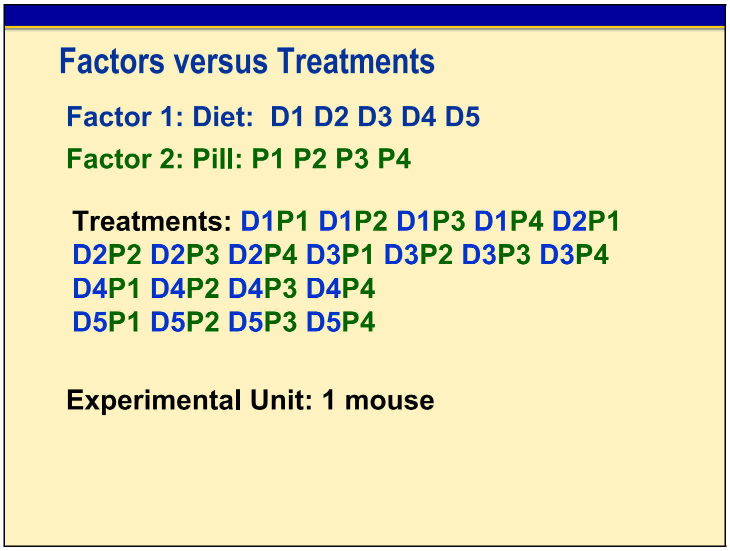


Basic Terms

Some basic terms used in experimental design are

- factor
- factor level
- treatment
- experimental unit
- replication.

These terms are defined throughout this lesson.



Factors versus Treatments

Factor 1: Diet: D1 D2 D3 D4 D5

Factor 2: Pill: P1 P2 P3 P4

Treatments: D1P1 D1P2 D1P3 D1P4 D2P1 D2P2 D2P3 D2P4 D3P1 D3P2 D3P3 D3P4 D4P1 D4P2 D4P3 D4P4 D5P1 D5P2 D5P3 D5P4

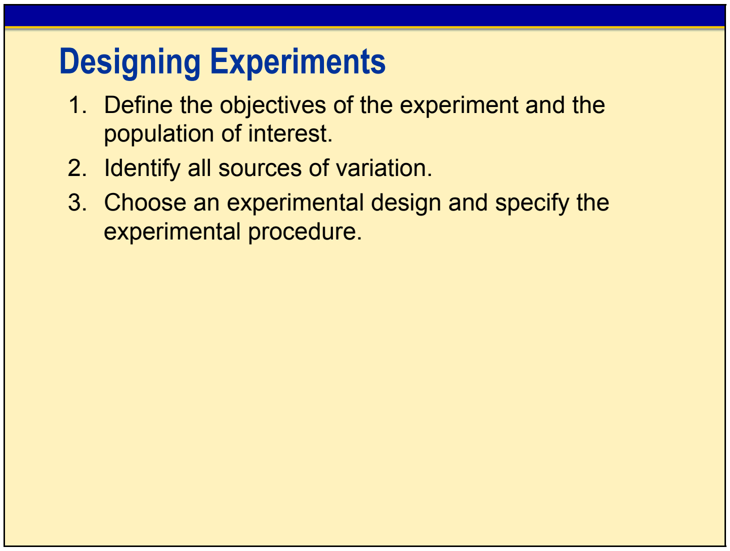
Experimental Unit: 1 mouse

A *factor* is an independent or predictor variable that is a possible source of variation in the response variable. Many factors cause variability in the response variable. Some of these causes are of interest to the investigator and others are not.

A *factor level* is a particular value of a factor, or the specific types or amounts of the factor used in the experiment. In this experiment there are five different diets, so there are five different factor levels for diet. There are four different drugs, so there are four factor levels of drug.

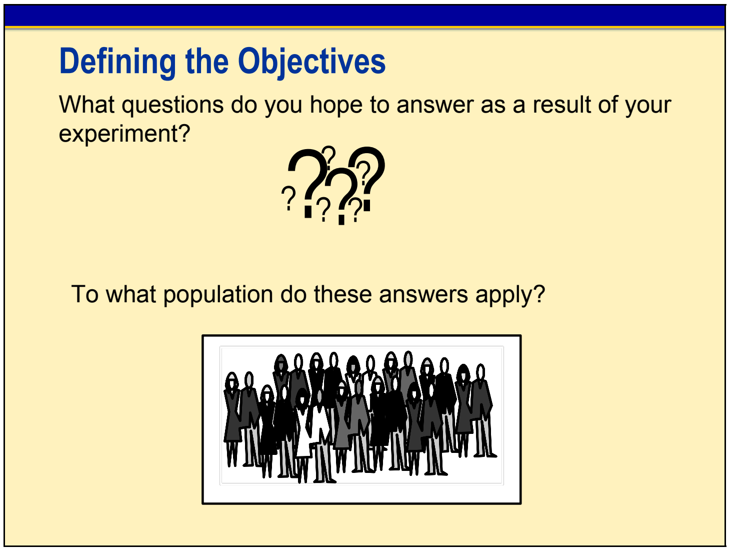
A *treatment* is a combination of factor levels used in the experiment. In single factor studies, a treatment is the same as a factor level.

An *experimental unit* is the smallest object to which a treatment is applied. It is the smallest part of experimental material where any two experimental units can receive different treatments. In many experiments involving people, the experimental unit is called a *subject*.



Designing Experiments

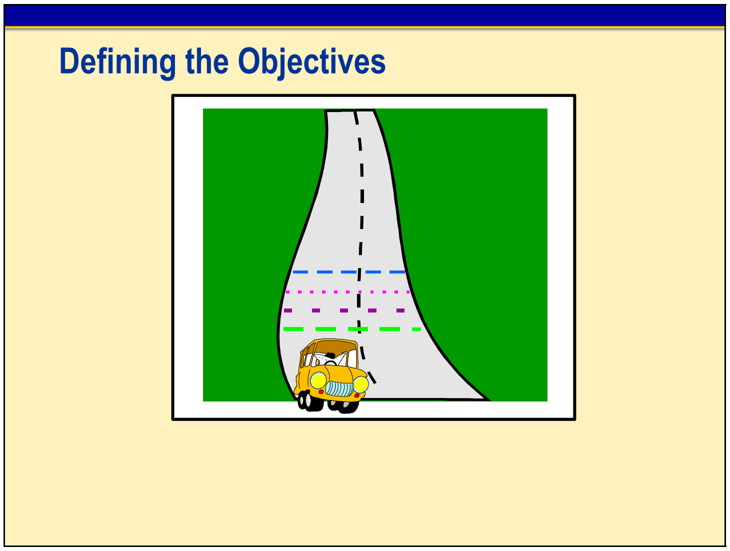
1. Define the objectives of the experiment and the population of interest.
2. Identify all sources of variation.
3. Choose an experimental design and specify the experimental procedure.



Defining the Objectives

What questions do you hope to answer as a result of your experiment?

To what population do these answers apply?



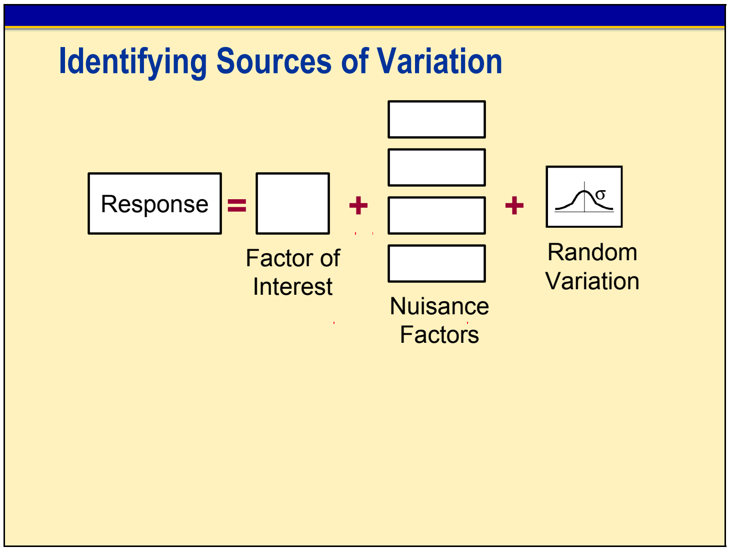
Defining the Objectives

Question: Which paint formula is the brightest on the town roads?

Population: The seven busiest roads in town.

The question is specific. It indicates that you are interested only in the effect paint formula has on brightness.

The target population is also specific. It indicates that inferences are only to be drawn on the seven busiest roads in the town.



Identifying Sources of Variation

Response = + +

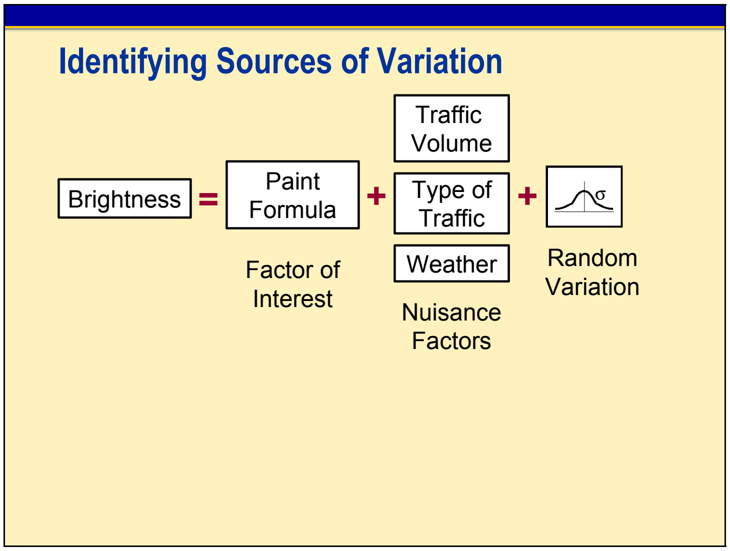
Factor of Interest

Nuisance Factors

Random Variation

Variability is inherent to any process. In an experiment, the variable that measures the outcome of interest is called the *response variable.*

A properly designed experiment enables you to identify the variability explained by the variables of interest, called factors or *effects*, and the variables not of interest, called *nuisance factors*.



Identifying Sources of Variation

Brightness = Paint Formula + Traffic Volume / Type of Traffic / Weather +

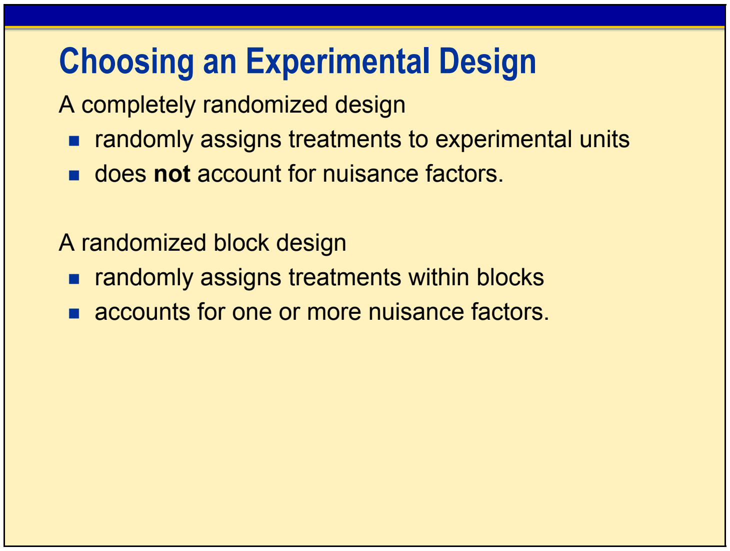
Factor of Interest

Nuisance Factors

Random Variation

You want to measure the brightness of the paint after one month of wear to determine the best paint formula. To begin, list the sources of variability.

It is good practice to list all potential sources of variability, even those you cannot control. In the design of the experiment, you control the sources of variability that you are able to control and randomize over those you cannot control.



Choosing an Experimental Design

A completely randomized design

- randomly assigns treatments to experimental units
- does **not** account for nuisance factors.

A randomized block design

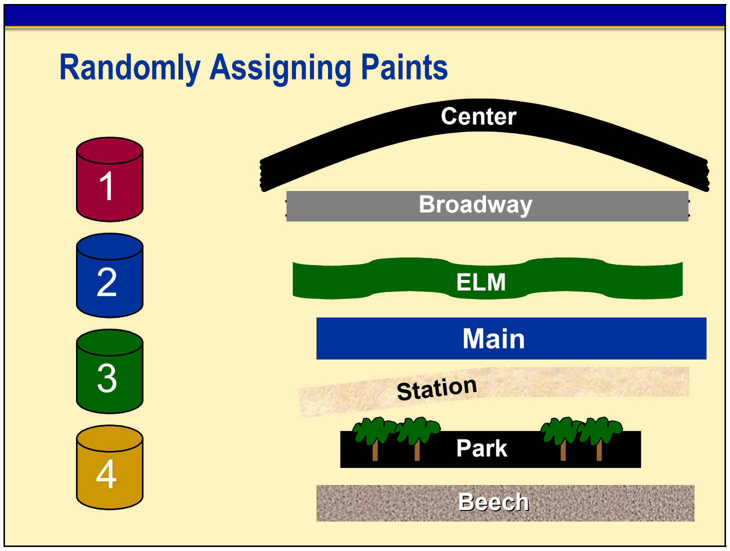
- randomly assigns treatments within blocks
- accounts for one or more nuisance factors.

Both of the designs above mention randomly assigning the treatments. Randomization is necessary to remove systematic and personal biases that might otherwise be introduced into the experiment.

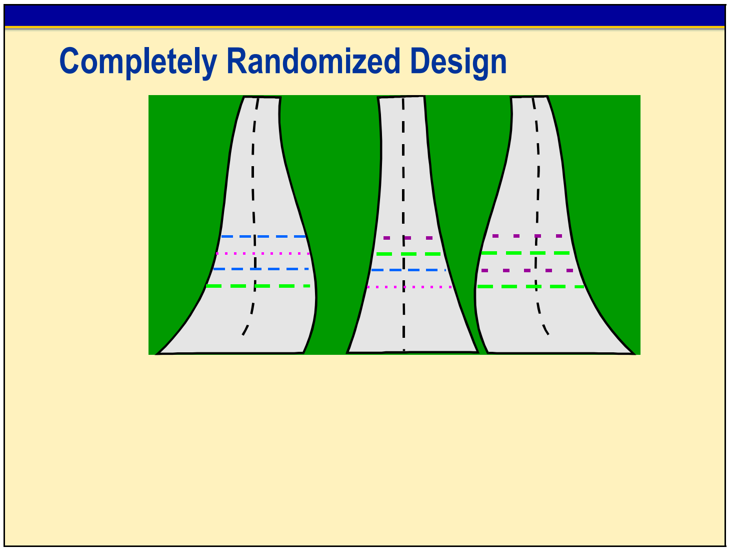
In a *completely randomized design*, presume you can control nuisance factors or randomize over them.

In a *randomized block design*, presume a known nuisance factor is a source of variability that you cannot control. You include the nuisance factor in the model even though it is not necessarily of interest. The nuisance factor included in the model is the *blocking factor*. A *block* is a group of similar experimental units. Blocks are designed so that an experimenter can isolate variability due to extraneous causes. These extraneous causes, or nuisance factors, can be characteristics associated with the experimental units or with the experimental setting.

There are many other experimental designs. This course addresses only a few of the possible designs you can use.



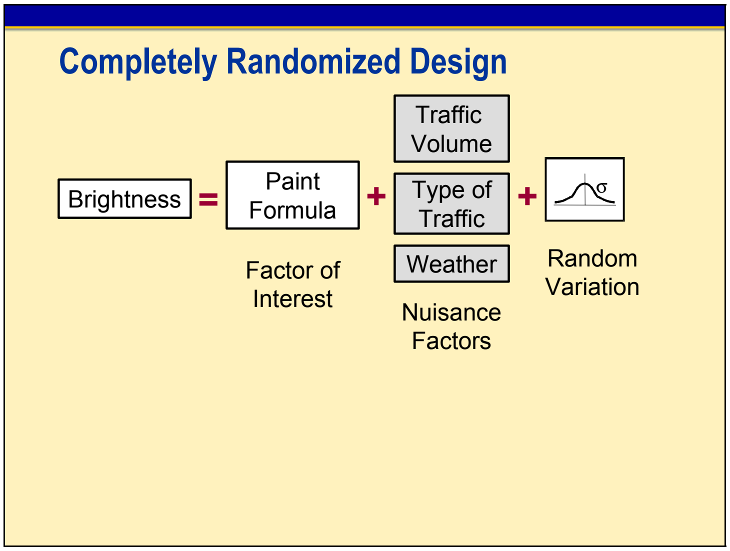
Randomly Assigning Paints



Completely Randomized Design

Example: You have identified the 7 roads to paint and the 4 paint formulas to test. You plan to paint 4 stripes of paint on each road, for a total of 28 stripes. One paint formula is randomly assigned to each of the 28 stripes.

Careful planning is required to ensure that the paints are randomly assigned to each of the 28 stripes. Appendix E, “Randomization Techniques,” contains a possible program to accomplish this task.



Completely Randomized Design

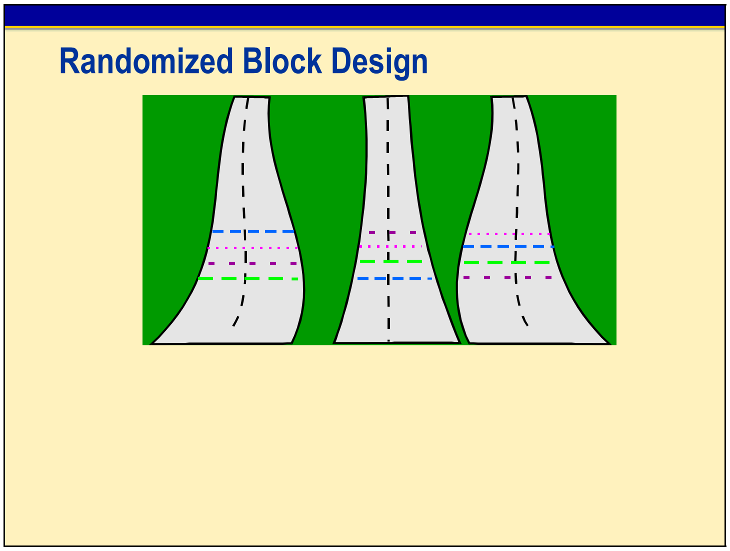
Brightness = Paint Formula + Traffic Volume / Type of Traffic / Weather + Random Variation

Factor of Interest

Nuisance Factors

After you have identified your factor of interest—in this example, paint formulas—you need to control all of the nuisance factors that you can.

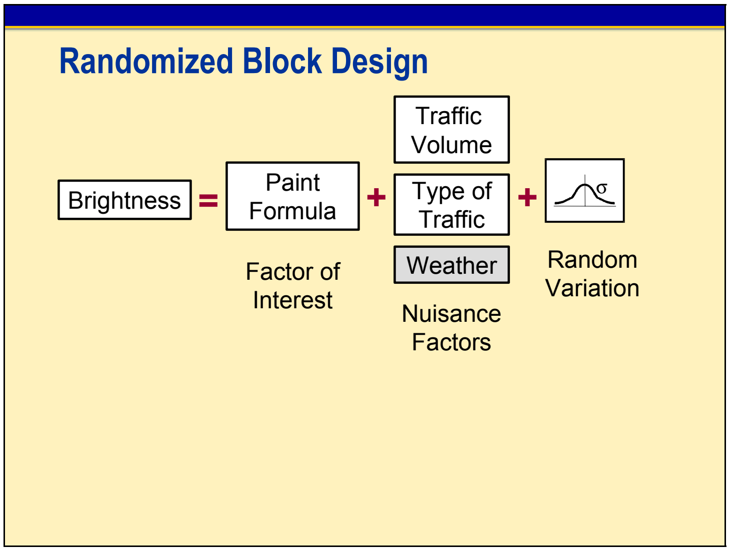
In this case, the nuisance factors identified cannot be controlled. Variability due to these factors is part of the random variation.



Randomized Block Design

In this experiment, many of the nuisance factors are a function of the different roads. One way to control for these nuisance factors is to use **`road`** as a blocking factor.

Rather than randomly assigning paint formulas to the 28 stripes, you use each paint formula once on every road. You randomly assign the paint formulas to one position on each road.



Randomized Block Design

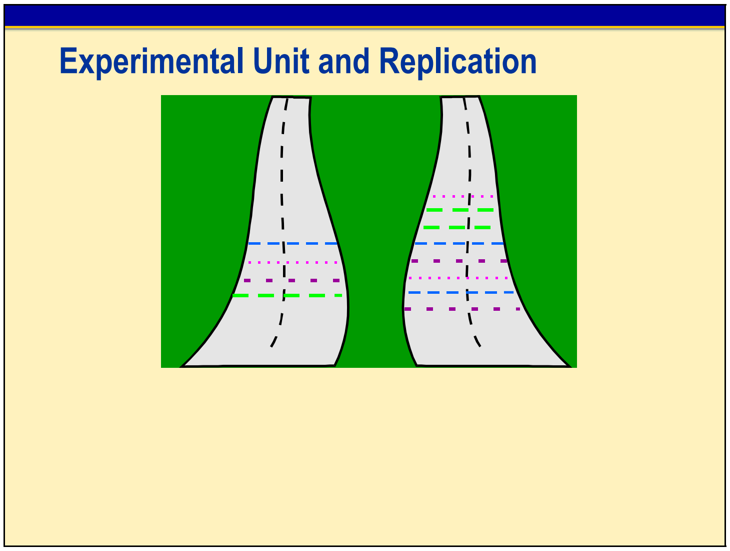
Brightness = Paint Formula + Traffic Volume / Type of Traffic / Weather + σ

Paint Formula: Factor of Interest

Traffic Volume, Type of Traffic, Weather: Nuisance Factors

σ: Random Variation

By blocking on **`road`**, you are accounting for both traffic volume and type of traffic variations. Variability due to these factors is no longer part of the random variation.

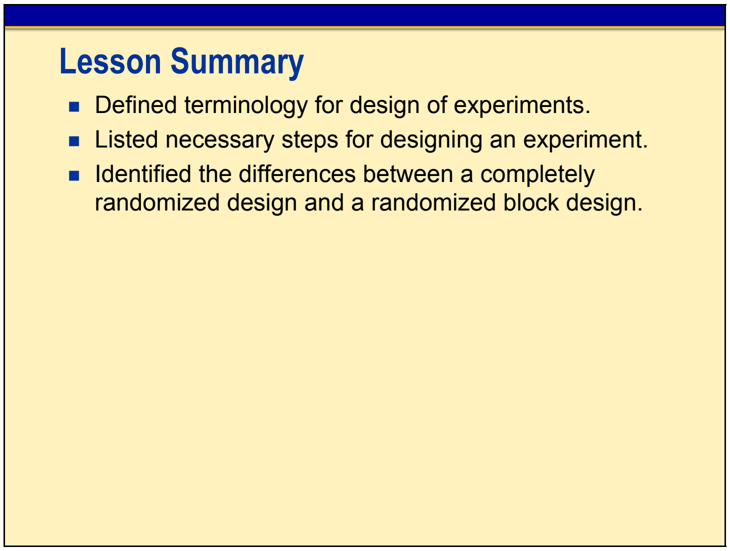


Experimental Unit and Replication

A *replication* occurs when you assign each treatment to more than one experimental unit.

In the picture on the left, there is one stripe of each paint formula applied to the road. If you are concerned that a sample size of one for each treatment is insufficient, you might consider dividing each stripe into two pieces and measuring the brightness of each piece. You reason that this gives you two observations, or replicates, for each treatment. What is wrong with this approach?

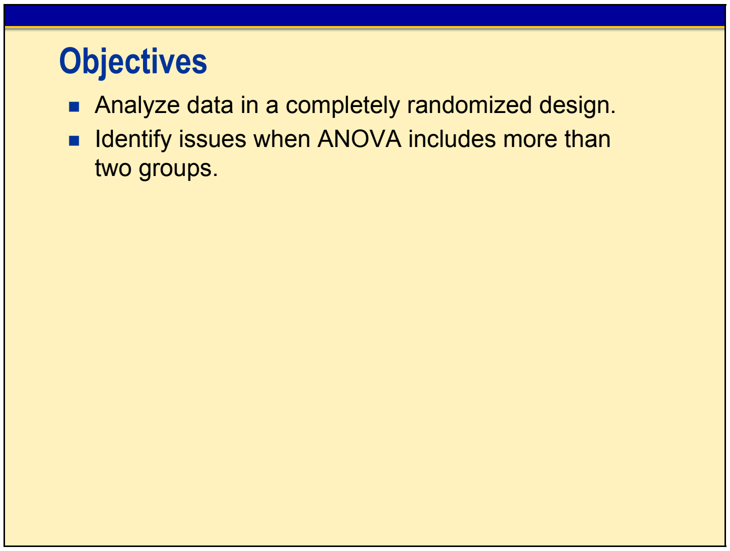
You cannot apply different treatments (paint formulas) to part of a stripe of paint. You can only do this to each stripe. By dividing the stripes and using each piece as an experimental unit, you have done pseudo-replication, not true replication. To have true replication, you would have to paint more stripes as shown in the picture on the right.



Lesson Summary

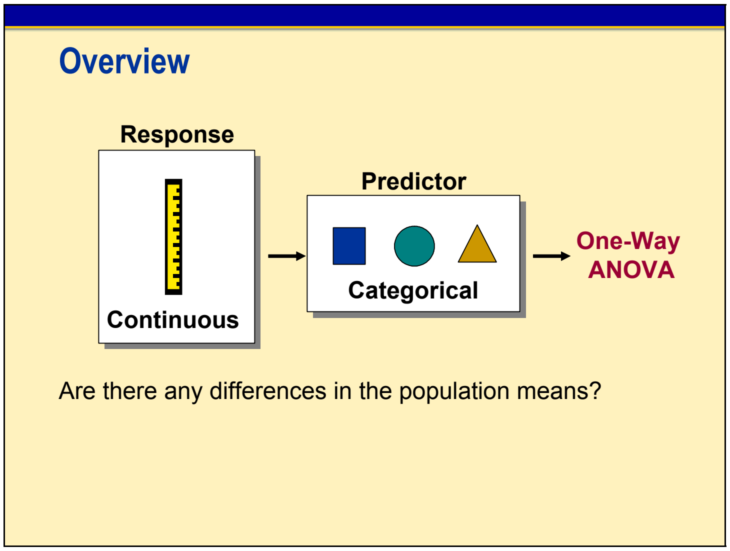
- Defined terminology for design of experiments.
- Listed necessary steps for designing an experiment.
- Identified the differences between a completely randomized design and a randomized block design.

2.4 One-Way ANOVA: More than Two Populations



Objectives

- Analyze data in a completely randomized design.
- Identify issues when ANOVA includes more than two groups.



Overview

Response

Continuous

Predictor

Categorical

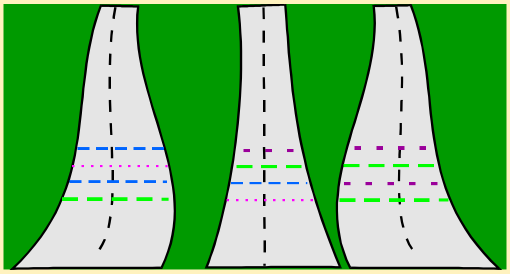
One-Way ANOVA

Are there any differences in the population means?

In this lesson, you apply analysis of variance to examine problems where there are more than two treatments. For this type of problem, you have a

- continuous dependent, or response, variable
- categorical independent, or classification, variable.

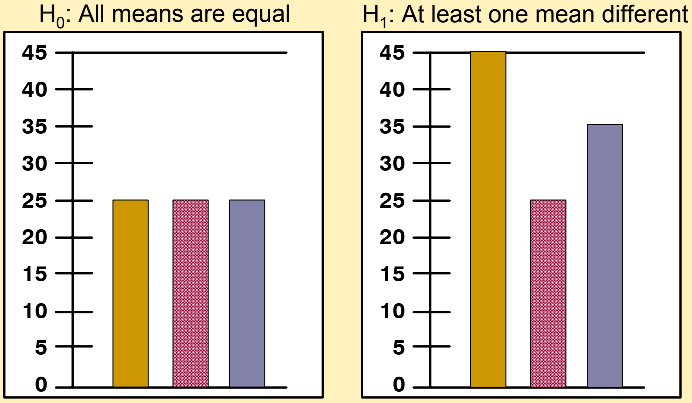
Analyzing More than Two Populations



The same basic concepts that apply when you analyze two populations are also true when you analyze more than two populations. The model and its assumptions are identical.

Consider the experiment to determine the best paint formula for roads with a completely randomized design. You want to determine whether the brightness of the paint is significantly different for the various paint formulas. There are seven roads, and four paint formulas are randomly assigned to each road.

The ANOVA Hypothesis



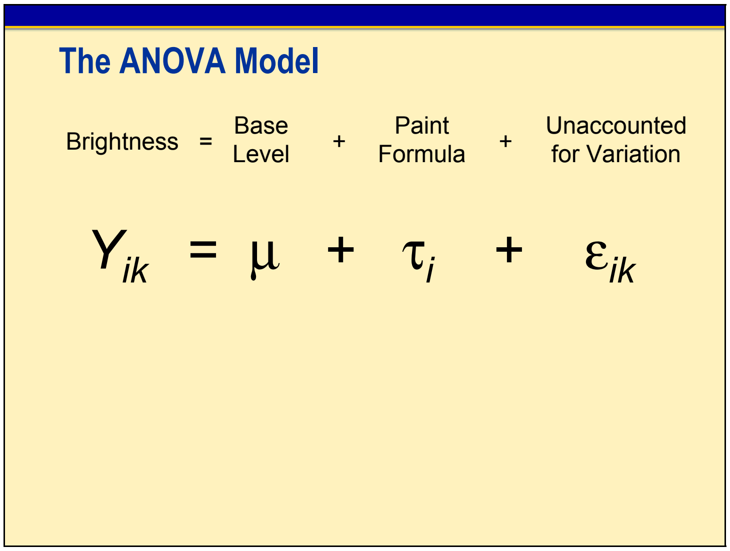
H_0: All means are equal

H_1: At least one mean different

Recall the objective is to determine whether there are differences between population means. Now, with more than two populations, you are testing the hypothesis

H_0: all means are equal

H_1: at least one mean is different from one of the other means.



The ANOVA Model

Brightness = Base Level + Paint Formula + Unaccounted for Variation

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

The model is the same as ANOVA for two groups.

Analysis of Variance: More Than Two Populations

m2demo07.sas, m2demo08.sas, m2demo09.sas

Example: Analyze the road paint data stored in the **`sasuser.b_roads`** data set.

The variables in the data set are

`road` the name of the road

`paint` the paint formula used

`bright` the brightness of the paint after one month on the road (candellas/m^2).

Print the data set.

```
proc print data=sasuser.b_roads;
   title 'Paint Experiment Data - Completely Randomized Design';
run;
```

```
              Paint Experiment Data - Completely Randomized Design

                    Obs    road           paint    bright

                      1    Center St.       1        43
                      2    Broadway         1        46
                      3    Main St.         1        47
                      4    Main St.         3        54
                      5    Elm St.          1        55
                      6    Station Rd.      1        56
                      7    Center St.       1        59
                      8    Center St.       4        61
                      9    Main St.         3        62
                     10    Center St.       4        62
                     11    Park Dr.         3        63
                     12    Main St.         2        64
                     13    Park Dr.         1        64
                     14    Broadway         4        64
                     15    Broadway         2        64
                     16    Broadway         3        65
                     17    Station Rd.      3        67
                     18    Station Rd.      3        67
                     19    Elm St.          3        68
                     20    Beech St.        4        71
                     21    Elm St.          4        72
                     22    Beech St.        2        75
                     23    Beech St.        4        75
                     24    Beech St.        2        76
                     25    Park Dr.         4        77
                     26    Elm St.          2        79
                     27    Station Rd.      2        79
                     28    Park Dr.         2        84
```

Initially, you want to examine the data to identify any unusual values and get a general idea about the distribution of the data. The UNIVARIATE procedure provides much of the information needed, including side-by-side box plots.

```
proc sort data=sasuser.b_roads out=sorted_roads;
   by paint;
run;

proc univariate data=sorted_roads plot;
   by paint;
   var bright;
   title 'Paint Experiment - Investigate Data';
run;
```

Selected PROC SORT statement option:

OUT= specifies a name for the output data set. If the OUT= option is omitted, the DATA= data set is sorted and the sorted version replaces the original data set.

Partial PROC UNIVARIATE Output

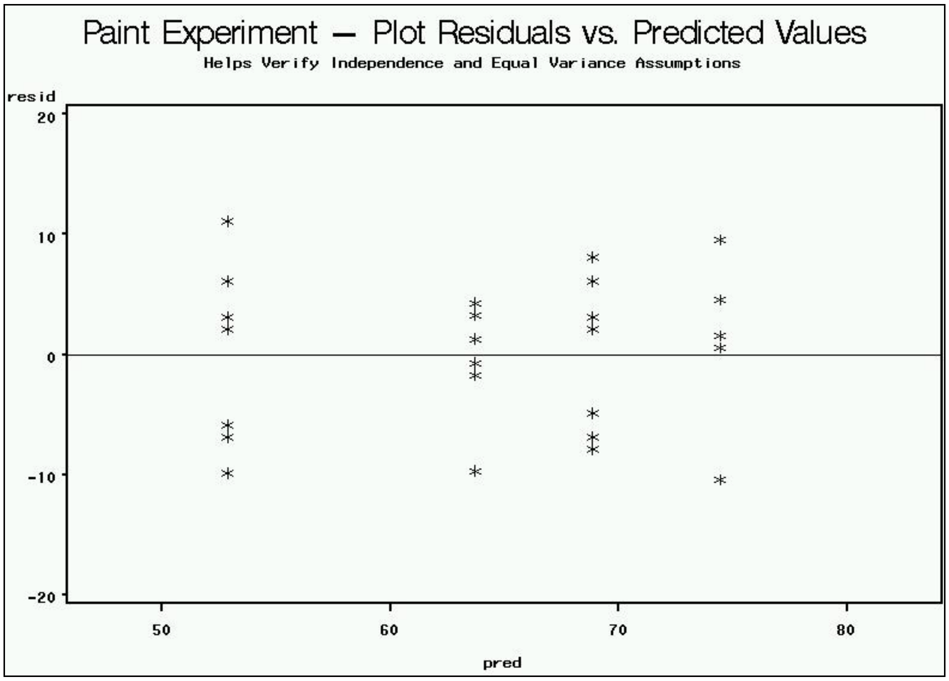
```
                    Paint Experiment - Investigate Data

                         The UNIVARIATE Procedure
                              Schematic Plots

           |
        90 +
           |
           |                        |
           |                        |
        80 +                     +-----+
           |                     |     |                    |
           |                     *--+--*                 +-----+
           |                     |     |                 |     |
        70 +                     |     |                 *--+--*
           |                     |     |     +-----+     |     |
           |            |        +-----+     *-----*     |     |
           |            |                    +--+--+     +-----+
        60 +         +-----+                                |
           |         |     |
           |         *-----*                    0
           |         |  +  |
        50 +         |     |
           |         |     |
           |         +-----+
           |            |
        40 +
            ------------+-----------+-----------+-----------+-----------
     paint                 1           2           3           4
```

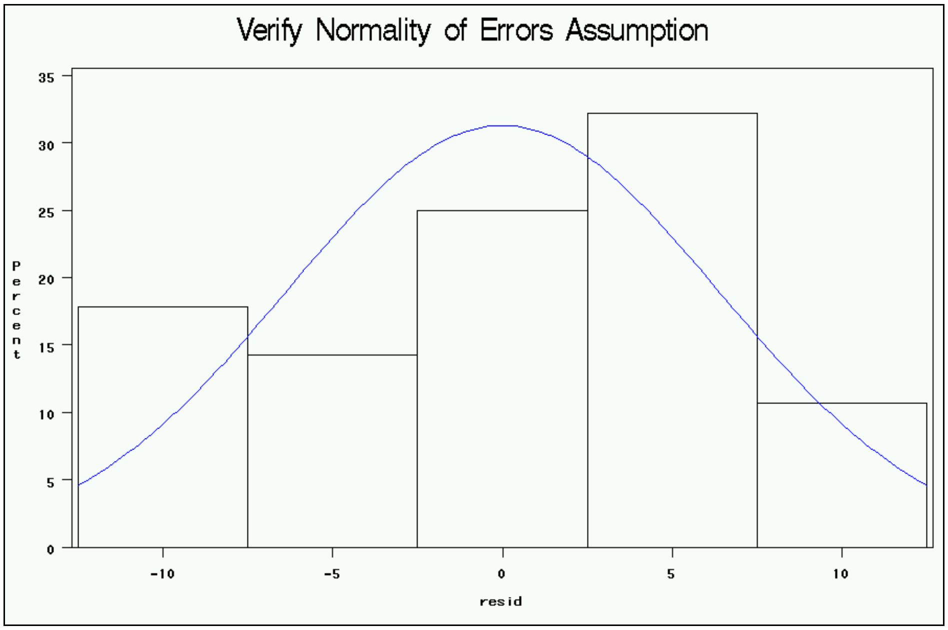
There do not appear to be any unusual data values, although paint formula 3 does have one outlier. There do appear to be differences between the mean brightness for the different types of paint. Specifically, paint formula 1 seems to have lower brightness than the other paint formulas. However, are the differences more than could reasonably occur by chance alone? In other words, are the differences statistically significant?

You can use the GLM procedure to test the null hypothesis that the means are equal. This program runs PROC GLM and also uses the UNIVARIATE and GPLOT procedures to check the assumptions of the ANOVA model.

```
proc glm data=sasuser.b_roads;
   class paint;
   model bright=paint;
   means paint / hovtest;
   output out=check r=resid p=pred;
   title 'Paint Experiment - Completely Randomized Design';
   title2 'HOVTEST Option Tests Equal Variances Using Levene Method';
   title3 'Null hypothesis for Levene is Variances Are Equal';
run;

proc univariate data=check plot;
   var resid;
   histogram / normal;
   title 'Verify Normality of Errors Assumption';
run;

proc gplot data=check;
   plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
   symbol v=star h=3pct;
   axis1 w=2 major=(w=2) minor=none offset=(10pct);
   axis2 w=2 major=(w=2) minor=none;
   title 'Paint Experiment - Plot Residuals vs. Predicted '
         'Values';
   title2 'Helps Verify Independence and Equal Variance Assumptions';
run;
quit;
```

Based on the plot of the residuals, there do not appear to be any extreme violations of the assumptions.





The normal probability plot and the stem-and-leaf and box plots shown below do not indicate any severe departure from the normality assumption. When you examine the Tests for Normality table, you are given what appears to be mixed signals. Two of the three normality tests are significant (0.041 and 0.049) at the 5 percent level of significance. Remember that these tests for normality should not be used exclusively to validate the normality assumption.

Partial PROC UNIVARIATE Output

```
                 Verify Normality of Errors Assumption

                          The UNIVARIATE Procedure
                              Variable:  resid

                                   Moments

       N                          28    Sum Weights                 28
       Mean                        0    Sum Observations             0
       Std Deviation      6.37953076    Variance            40.6984127
       Skewness           -0.2920394    Kurtosis            -0.9980851
       Uncorrected SS     1098.85714    Corrected SS        1098.85714
       Coeff Variation             .    Std Error Mean      1.20561799

                         Basic Statistical Measures

               Location                    Variability
           Mean       0.0000     Std Deviation            6.37953
           Median     1.8571     Variance                40.69841
           Mode     -10.4286     Range                   21.57143
                                 Interquartile Range     10.78571

   NOTE: The mode displayed is the smallest of 3 modes with a count of 2.


              Stem Leaf                    #             Boxplot
                10 1                       1                |
                 8 16                      2                |
                 6 11                      2                |
                 4 366                     3             +-----+
                 2 111133                  6             |     |
                 0 636                     3             *--+--*
                -0 77                      2             |     |
                -2                                       |     |
                -4 99                      2             |     |
                -6 999                     3             +-----+
                -8 97                      2                |
               -10 44                      2                |
                   ----+----+----+----+
                       Normal Probability Plot
      11+                                         +++ *
        |                                      *+++*
        |                                   **+
        |                               **++
        |                         *****++
        |                       ***+++
        |                     **++
        |                   +++
        |                +++**
        |             ++* **
        |          ++* *
     -11+     * +++*
         +----+----+----+----+----+----+----+----+----+----+
             -2        -1         0        +1        +2


     Goodness-of-Fit Tests for Normal Distribution

 Test                  ---Statistic----   -----p Value-----

 Kolmogorov-Smirnov    D     0.15128939   Pr > D      0.097
 Cramer-von Mises      W-Sq  0.13197489   Pr > W-Sq   0.041
 Anderson-Darling      A-Sq  0.73596819   Pr > A-Sq   0.049
```

After reviewing this information regarding the residuals, look at the part of the PROC GLM output that shows Levene's test for equality of variances.

```
      Paint Experiment - Completely Randomized Design
    HOVTEST Option Tests Equal Variances Using Levene Method
      Null hypothesis for Levene is Variances Are Equal
                      The GLM Procedure

      Levene's Test for Homogeneity of bright Variance
        ANOVA of Squared Deviations from Group Means

                      Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F

paint          3      4505.0      1501.7       0.97    0.4224
Error         24     37090.8      1545.5
```

The *p*-value of 0.4224 indicates that you do not reject the null hypothesis that the variances for the treatments are equal. Therefore, the analysis of variance procedure appears to be appropriate.

Now that you are reasonably sure the assumptions of the ANOVA model have been met, turn your attention to the class level information and the ANOVA table.

The first page of PROC GLM output, shown below, specifies the number of levels and the values of the class variable, as well as the number of observations.

```
              Paint Experiment - Completely Randomized Design
           HOVTEST Option Tests Equal Variances Using Levene Method
             Null hypothesis for Levene is Variances Are Equal

                             The GLM Procedure

                         Class Level Information

                     Class         Levels    Values

                     paint              4    1 2 3 4


                       Number of observations    28
```

Part of the second page of the PROC GLM output is shown below.

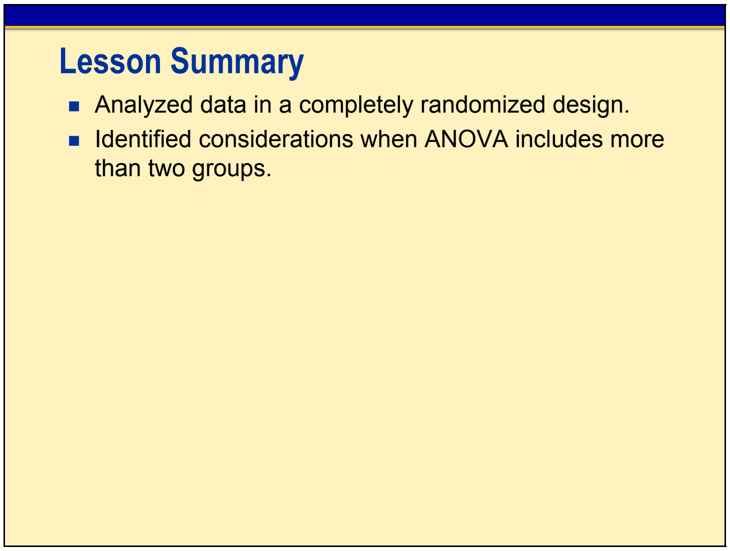
```
           Paint Experiment - Completely Randomized Design
           HOVTEST Option Tests Equal Variances Using Levene Method
             Null hypothesis for Levene is Variances Are Equal

                             The GLM Procedure

Dependent Variable: bright

                                     Sum of
 Source                     DF      Squares   Mean Square  F Value  Pr > F

 Model                       3   1770.107143    590.035714    12.89  <.0001

 Error                      24   1098.857143     45.785714

 Corrected Total            27   2868.964286
```

With a *p*-value less than or equal to 0.0001, you reject the null hypothesis that all treatment means are equal.

At this point, you know there is **at least** one treatment mean that is different from one other treatment mean, but you cannot be sure which one(s) are different. Some insight can be gained by looking at the side-by-side box plots from PROC UNIVARIATE and the page of the PROC GLM output produced by the MEANS statement.

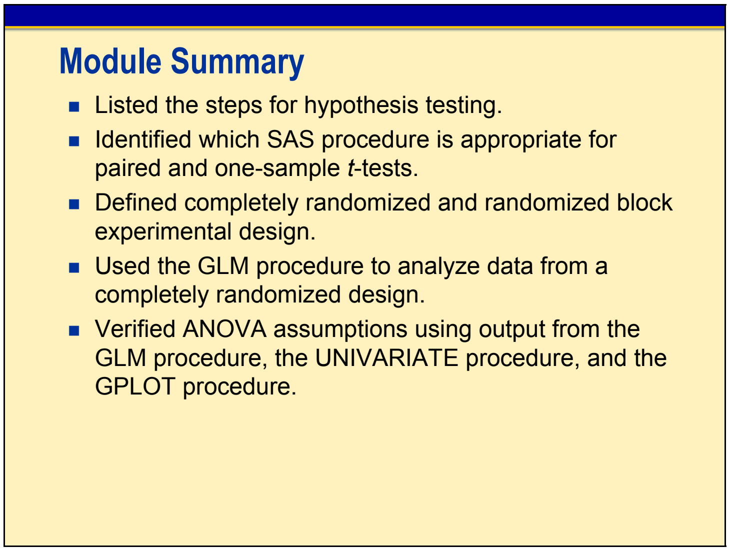
```
                 Paint Experiment - Completely Randomized Design
            HOVTEST Option Tests Equal Variances Using Levene Method
              Null hypothesis for Levene is Variances Are Equal

                              The GLM Procedure

              Level of           ------------bright-----------
              paint        N             Mean          Std Dev

              1            7       52.8571429       7.69043933
              2            7       74.4285714       7.67804539
              3            7       63.7142857       4.82059076
              4            7       68.8571429       6.46602844
```

It appears paint formula 1 has lower brightness than the other formulas and paint formula 2 results in the highest brightness. Multiple comparison techniques can be used to determine whether these are statistically significant differences.



Lesson Summary

- Analyzed data in a completely randomized design.
- Identified considerations when ANOVA includes more than two groups.



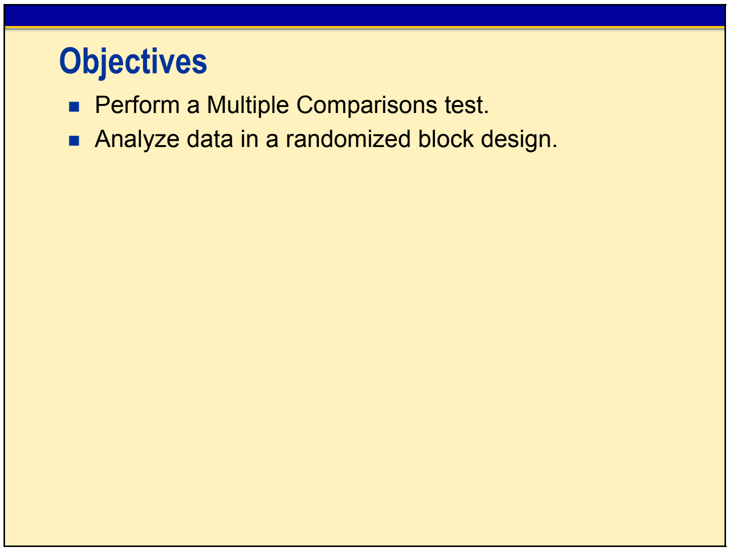
Module Summary

- Listed the steps for hypothesis testing.
- Identified which SAS procedure is appropriate for paired and one-sample *t*-tests.
- Defined completely randomized and randomized block experimental design.
- Used the GLM procedure to analyze data from a completely randomized design.
- Verified ANOVA assumptions using output from the GLM procedure, the UNIVARIATE procedure, and the GPLOT procedure.

Exercise: Refer to your course workbook.

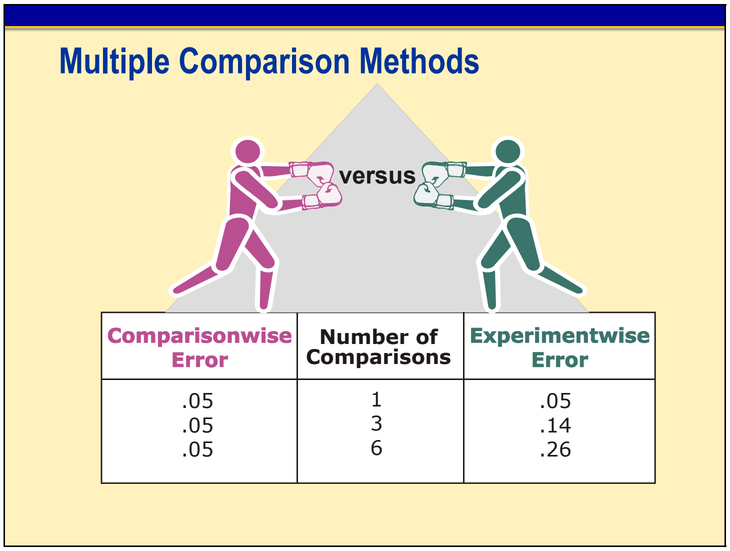
Module 3 Multiple Comparisons, Nonparametric ANOVA, and Regression

3.1 Multiple Comparisons



Objectives

- Perform a Multiple Comparisons test.
- Analyze data in a randomized block design.



Multiple Comparison Methods

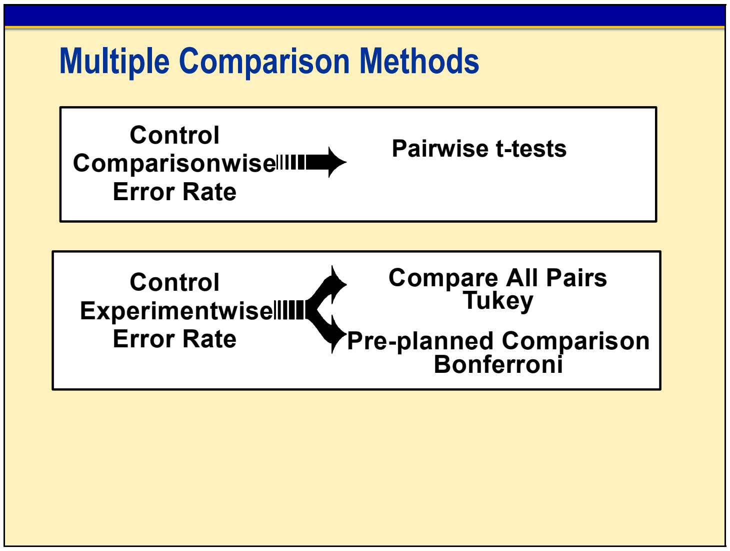
Comparisonwise Error	Number of Comparisons	Experimentwise Error
.05	1	.05
.05	3	.14
.05	6	.26

When you control the comparisonwise error rate, you fix the level of alpha for a single comparison, without taking into consideration all the pairwise comparisons you are making.

The experimentwise error rate uses an alpha that takes into consideration all the pairwise comparisons you are making. If you make 10 comparisons, each with a comparisonwise error rate of alpha=0.05, the experimentwise error rate is less than or equal to 0.401 ($EER \le 1-(1-\alpha)^{10}$). Presuming no differences exist, the chance you falsely conclude that at least one difference exists is much higher when you consider all 10 comparisons. In our example, the experimentwise error rate would be calculated as $EER \le 1-(0.95)^{6}$, or approximately 0.26491.

If you want to make sure the error rate is 0.05 for the entire set of comparisons, use a method that controls the experimentwise error rate at 0.05.

✐ There is some disagreement among statisticians about the need to control the experimentwise error rate.



Multiple Comparison Methods

Control Comparisonwise Error Rate → Pairwise t-tests

Control Experimentwise Error Rate → Compare All Pairs Tukey; Pre-planned Comparison Bonferroni

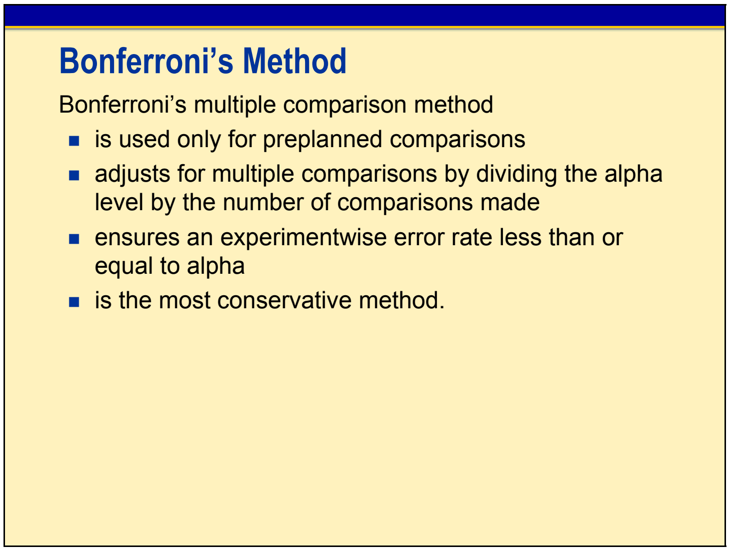
All of these multiple comparison methods are requested with options in the MEANS statement of PROC GLM.

This course addresses these options:

Comparisonwise Control	LSD
Experimentwise Control	TUKEY and BONFERRONI.

There are many other options available that control the experimentwise error rate. These include REGWQ, REGWF, WALLER, DUNCAN, DUNNETT, GABRIEL, and SCHEFFE.

For information about these options, see the *SAS/STAT® User's Guide*, *Version 8, Volume 2*.



Bonferroni’s Method

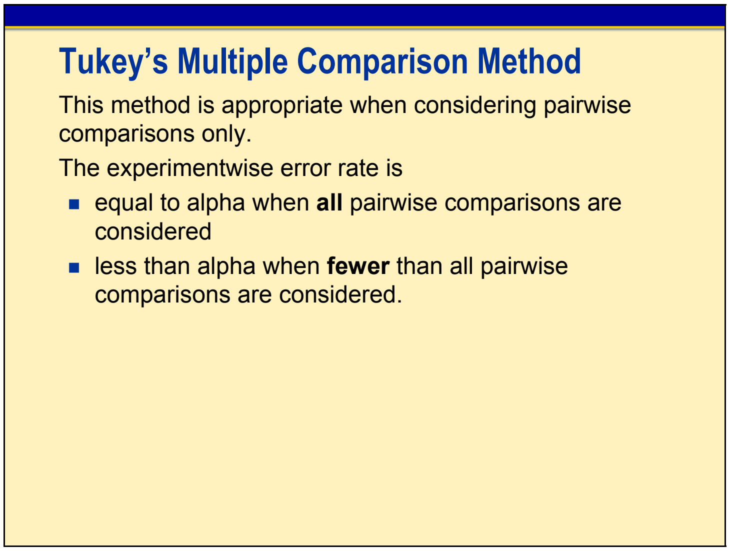
Bonferroni’s multiple comparison method

- is used only for preplanned comparisons
- adjusts for multiple comparisons by dividing the alpha level by the number of comparisons made
- ensures an experimentwise error rate less than or equal to alpha
- is the most conservative method.

Bonferroni’s method is not generally considered appropriate for comparisons made after looking at the data because the adjustment is made based on the number of comparisons you intend to do. If you look at the data to determine what comparisons to make and how many, you are using the data to determine the adjustment.

A conservative method tends to find fewer significant differences than might otherwise be found.

While Bonferroni’s method can be used for all pairwise comparisons, Tukey’s method is generally less conservative and more appropriate.



Tukey’s Multiple Comparison Method

This method is appropriate when considering pairwise comparisons only.

The experimentwise error rate is

- equal to alpha when **all** pairwise comparisons are considered
- less than alpha when **fewer** than all pairwise comparisons are considered.

A pairwise comparison examines the difference between two treatment means. All pairwise comparisons are all possible combinations of two treatment means.

Tukey’s multiple comparison adjustment is based on conducting all pairwise comparisons, and it guarantees the Type I experimentwise error rate is equal to alpha for this situation. If you choose to do fewer than all pairwise comparisons, this method is more conservative.

Multiple Comparison Methods

m3demo01.sas, m3demo02.sas

Example: Use the LSD option in the MEANS statement of PROC GLM to produce comparison information on the means of the treatments. Examine the output produced by the CLDIFF and LINES options.

```
proc glm data=sasuser.b_roads;
   class paint;
   model bright=paint;
   means paint / lsd cldiff lines;
   title 'Paint Experiment - Completely Randomized - '
         'Control CER';
run;
```

Selected MEANS statement options:

LSD performs pairwise *t*-tests for all means in the MEANS statement.

CLDIFF produces confidence limits for the difference between pairs of means. This option marks those differences that are found to be significantly different from zero.

LINES presents results in the form of a listing that provides the means in descending order and indicates nonsignificant subsets by letters beside the corresponding means. This option should be used only if the sample sizes for the treatments are equal.

ALPHA= gives the significance for comparisons among the means. By default, ALPHA=0.05.

Partial PROC GLM Output

```
                   Paint Experiment - Completely Randomized - Control CER

                                   The GLM Procedure

                                 t Tests (LSD) for bright

             NOTE: This test controls the Type I comparisonwise error rate, not the
                                 experimentwise error rate.


                      Alpha                            0.05
                      Error Degrees of Freedom           24
                      Error Mean Square            45.78571
                      Critical Value of t           2.06390
                      Least Significant Difference   7.4648

                Comparisons significant at the 0.05 level are indicated by ***.


                                    Difference
                         paint        Between      95% Confidence
                      Comparison       Means           Limits

                        2    - 4        5.571      -1.893   13.036
                        2    - 3       10.714       3.249   18.179  ***
                        2    - 1       21.571      14.107   29.036  ***
                        4    - 2       -5.571     -13.036    1.893
                        4    - 3        5.143      -2.322   12.608
                        4    - 1       16.000       8.535   23.465  ***
                        3    - 2      -10.714     -18.179   -3.249  ***
                        3    - 4       -5.143     -12.608    2.322
                        3    - 1       10.857       3.392   18.322  ***
                        1    - 2      -21.571     -29.036  -14.107  ***
                        1    - 4      -16.000     -23.465   -8.535  ***
                        1    - 3      -10.857     -18.322   -3.392  ***
```

The output above is from the CLDIFF option. Note that the message stating this test does not control the experimentwise error rate.

The pairs of treatments that are significantly different from one another at the 0.05 level of significance are marked with asterisks (***). For the paint formula data, it appears paint formula 1 is different from all other formulas. Paint formula 2 is different from formula 3.

Partial PROC GLM Output (continued)

```
                      Paint Experiment - Completely Randomized - Control CER

                                 The GLM Procedure

                               t Tests (LSD) for bright

                NOTE: This test controls the Type I comparisonwise error rate, not the
                              experimentwise error rate.


                          Alpha                              0.05
                          Error Degrees of Freedom             24
                          Error Mean Square              45.78571
                          Critical Value of t             2.06390
                          Least Significant Difference   7.4648


                    Means with the same letter are not significantly different.


                         t Grouping          Mean      N    paint

                                   A       74.429      7    2
                                   A
                              B    A       68.857      7    4
                              B
                              B            63.714      7    3

                                   C       52.857      7    1
```

With the LINES option, you draw the same conclusions as with the CLDIFF option.

Your choice of which output option you prefer depends on what information you want to obtain. In summary,

- the CLDIFF option provides confidence limits and indicates significant differences
- the LINES option, which should be used only when treatments have equal sample sizes, indicates significant differences but does not give confidence limits.

The LSD option controls only the comparisonwise error rate. In order to control the experimentwise error rate, you can use Bonferroni's or Tukey's method.

Example: Use the Bonferroni and Tukey methods for multiple comparisons to test differences between the treatment means for the variable **`bright`** in the **`sasuser.b_roads`** data set.

```
proc glm data=sasuser.b_roads;
   class paint;
   model bright=paint;
   means paint / bon;
   title 'Paint Experiment - Completely Randomized'
         ' - Control EER (bon)';
run;
quit;

proc glm data=sasuser.b_roads;
   class paint;
   model bright=paint;
   means paint / tukey;
   title 'Paint Experiment - Completely Randomized'
         ' - Control EER (tukey)';
run;
quit;
```

Selected MEANS statement options:

BON performs Bonferroni *t*-tests of differences between means for all main-effect means.

TUKEY performs Tukey's studentized range test on all main-effect means.

Like the LSD option, the BON and TUKEY options can be used with the CLDIFF or LINES options. When the sample sizes are equal for each treatment, the LINES option is the default.

Partial PROC GLM Output

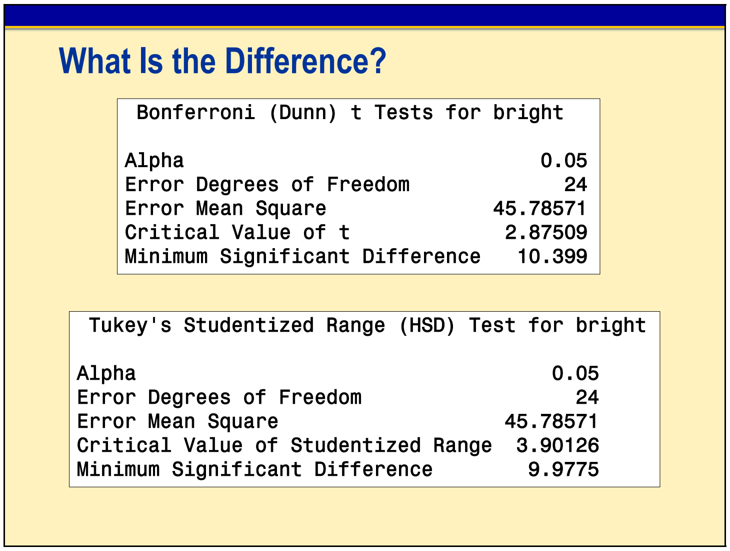
```
                    Paint Experiment - Completely Randomized - Control EER (bon)

                                   The GLM Procedure

                            Bonferroni (Dunn) t Tests for bright

  NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                             Type II error rate than REGWQ.

                         Alpha                                0.05
                         Error Degrees of Freedom               24
                         Error Mean Square              45.78571
                         Critical Value of t             2.87509
                         Minimum Significant Difference   10.399


                 Means with the same letter are not significantly different.


                          Bon Grouping          Mean      N    paint

                                    A        74.429      7    2
                                    A
                               B    A        68.857      7    4
                               B
                               B             63.714      7    3

                                    C        52.857      7    1
```

The output shows the same findings as with the LSD option. However, note that the least significant difference with the LSD option was 7.46487 and the minimum significant difference here is 10.399. Also, if you choose to use the CLDIFF option here, you see the confidence intervals are wider when the experimentwise error rate is controlled.

Partial PROC GLM Output (continued)

```
                  Paint Experiment - Completely Randomized - Control EER (tukey)

                                    The GLM Procedure

                      Tukey's Studentized Range (HSD) Test for bright

   NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                               Type II error rate than REGWQ.


                               Alpha                                   0.05
                               Error Degrees of Freedom                  24
                               Error Mean Square                   45.78571
                               Critical Value of Studentized Range  3.90126
                               Minimum Significant Difference        9.9775


                      Means with the same letter are not significantly different.


                             Tukey Grouping          Mean      N    paint

                                          A        74.429      7    2
                                          A
                                     B    A        68.857      7    4
                                     B
                                     B             63.714      7    3

                                          C        52.857      7    1
```

The significant differences using Tukey's method are the same as those with Bonferroni's method in this case. This might not always be true. Note the minimum significant difference of 9.9775, which is larger than the least significant difference but smaller than the minimum significant difference with Bonferroni's method.

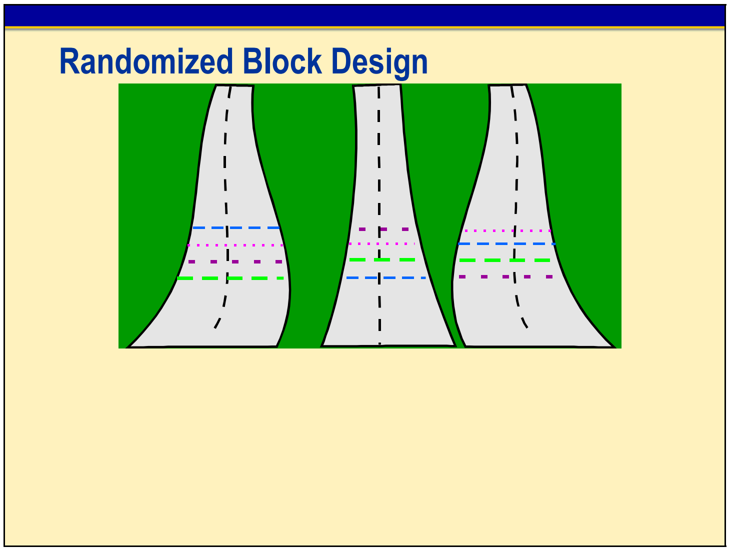


What Is the Difference?

```
 Bonferroni (Dunn) t Tests for bright

Alpha                               0.05
Error Degrees of Freedom              24
Error Mean Square               45.78571
Critical Value of t              2.87509
Minimum Significant Difference    10.399
```

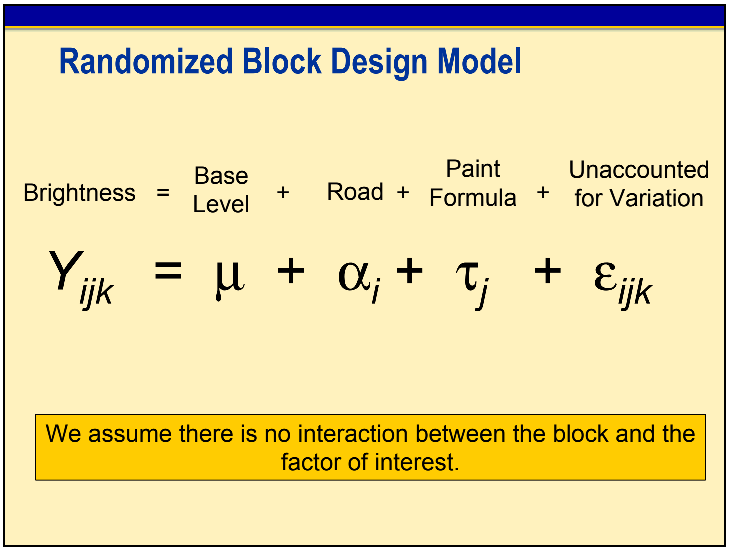
```
 Tukey's Studentized Range (HSD) Test for bright

Alpha                                      0.05
Error Degrees of Freedom                     24
Error Mean Square                      45.78571
Critical Value of Studentized Range     3.90126
Minimum Significant Difference           9.9775
```

Though the results of the statistical tests are the same in the example, the Bonferroni and Tukey tests are different. Under Bonferroni, two means must be farther apart to be considered statistically significant. This is why Bonferroni is considered the more conservative test.



Randomized Block Design

An experienced road paint expert might anticipate that there would be so much variability in brightness caused by the nuisance factors that the statistical test would not detect differences caused by paint formulas alone. By using a block design, variability due to one or more nuisance factors can be isolated, which enables the test to be more sensitive to differences caused by paint formulas.



Randomized Block Design Model

Brightness = Base Level + Road + Paint Formula + Unaccounted for Variation

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$$

We assume there is no interaction between the block and the factor of interest.

In addition to the normal ANOVA assumptions, the assumptions of this model are

- paint formulas are randomly assigned within each road.
- the effects of the roads are additive (there is no interaction between road and paint formula). In other words, different formulas are not better for different roads.

Randomized Block Design

m3demo03.sas, m3demo04.sas

Example: The data set **sasuser.b_roads1** is a fabricated example of data collected as a randomized block design. Note that each paint formula appears exactly once on each road.

```
proc print data=sasuser.b_roads1;
   title 'Paint Experiment - Randomized Block Design';
run;
```

```
                Paint Experiment - Randomized Block Design

                   Obs    road           paint    bright

                     1    Broadway         1        48
                     2    Main St.         1        49
                     3    Center St.       1        49
                     4    Center St.       3        56
                     5    Elm St.          1        57
                     6    Main St.         3        57
                     7    Station Rd.      1        58
                     8    Broadway         3        59
                     9    Beech St.        1        60
                    10    Park Dr.         1        61
                    11    Broadway         2        62
                    12    Center St.       4        62
                    13    Main St.         4        63
                    14    Station Rd.      3        65
                    15    Main St.         2        66
                    16    Broadway         4        66
                    17    Center St.       2        68
                    18    Elm St.          3        68
                    19    Beech St.        3        69
                    20    Park Dr.         3        70
                    21    Station Rd.      2        72
                    22    Beech St.        2        73
                    23    Park Dr.         2        73
                    24    Elm St.          4        73
                    25    Station Rd.      4        73
                    26    Elm St.          2        74
                    27    Beech St.        4        75
                    28    Park Dr.         4        78
```

Example: To include the blocking factor in the model, add the variable name to the CLASS and MODEL statements.

```
proc glm data=sasuser.b_roads1;
   class paint road;
   model bright=paint road;
   means paint / tukey;
   title 'Paint Experiment - Randomized Block - '
         'Control EER (tukey)';
run;
```

```
         Paint Experiment - Randomized Block - Control EER (tukey)

                             The GLM Procedure

                          Class Level Information

Class       Levels  Values

paint            4  1 2 3 4

road             7  Beech St. Broadway Center St. Elm St. Main St. Park Dr.
                    Station Rd.


                       Number of observations    28
```

PROG GLM Output (continued)

```
         Paint Experiment - Randomized Block - Control EER (tukey)

                             The GLM Procedure

Dependent Variable: bright

                                     Sum of
 Source                     DF      Squares   Mean Square  F Value  Pr > F

 Model                       9   1804.857143   200.539683    60.16  <.0001

 Error                      18     60.000000     3.333333

 Corrected Total            27   1864.857143


            R-Square     Coeff Var      Root MSE    bright Mean

            0.967826      2.833746      1.825742       64.42857


 Source                     DF     Type I SS   Mean Square  F Value  Pr > F

 paint                       3   1100.000000   366.666667   110.00  <.0001
 road                        6    704.857143   117.476190    35.24  <.0001


 Source                     DF   Type III SS   Mean Square  F Value  Pr > F

 paint                       3   1100.000000   366.666667   110.00  <.0001
 road                        6    704.857143   117.476190    35.24  <.0001
```

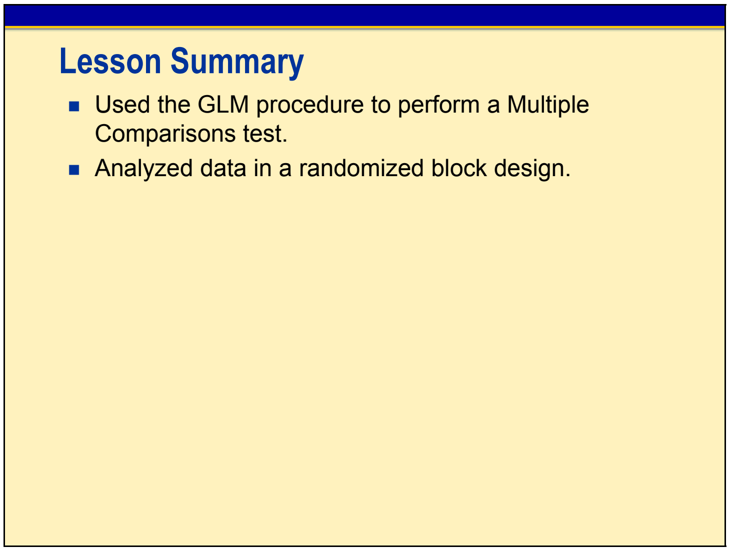
As expected, the overall *F*-test indicates that there are significant differences between the means of the different types of paint formula.

What have you gained by using the block design over the completely randomized design? If you compare the estimate of the experimental error variance (MSE), you note this has decreased in the block design (3.33333 versus 45.78571). Depending on the magnitude of the decrease, this could affect the comparisons between the treatment means by finding more significant differences than without the blocking factor.

```
                  Paint Experiment - Randomized Block - Control EER (tukey)

                                 The GLM Procedure

                  Tukey's Studentized Range (HSD) Test for bright

 NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher
                          Type II error rate than REGWQ.


                        Alpha                                  0.05
                        Error Degrees of Freedom                 18
                        Error Mean Square                  3.333333
                        Critical Value of Studentized Range  3.99698
                        Minimum Significant Difference       2.7582


               Means with the same letter are not significantly different.

                  Tukey Grouping          Mean       N    paint

                               A       70.0000      7     4
                               A
                               A       69.7143      7     2

                               B       63.4286      7     3

                               C       54.5714      7     1
```

In this case, with the blocking factor in the model, paint formulas 2 and 4 are the only ones found not to be significantly different. Also note that the minimum significant difference has been reduced to 2.7582.

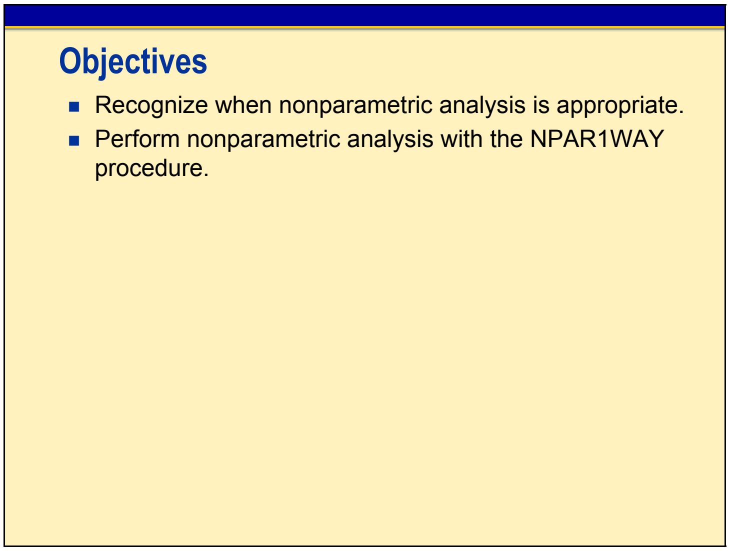
In determining the usefulness of having a block effect included in the model, you can consider the F Value for the block. Some statisticians suggest that if this ratio is greater than 1, then creation of the blocks is useful; however, if the ratio is less than 1, then creation of the blocks is detrimental to the analysis. If the creation of the blocks is found to be detrimental to the analysis, the block would not be included in future studies.



Lesson Summary

- Used the GLM procedure to perform a Multiple Comparisons test.
- Analyzed data in a randomized block design.

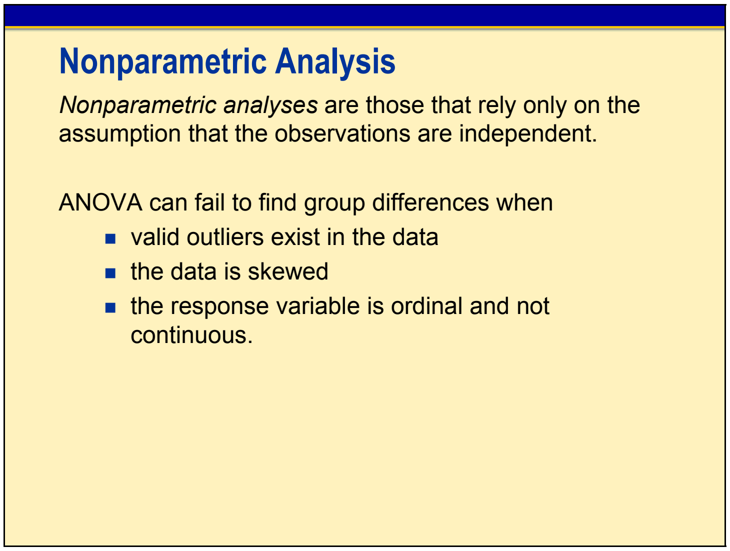
3.2 Nonparametric ANOVA



Objectives

- Recognize when nonparametric analysis is appropriate.
- Perform nonparametric analysis with the NPAR1WAY procedure.

This lesson addresses nonparametric options within the NPAR1WAY procedure. Nonparametric one-sample tests are also available in the UNIVARIATE procedure.



Nonparametric Analysis

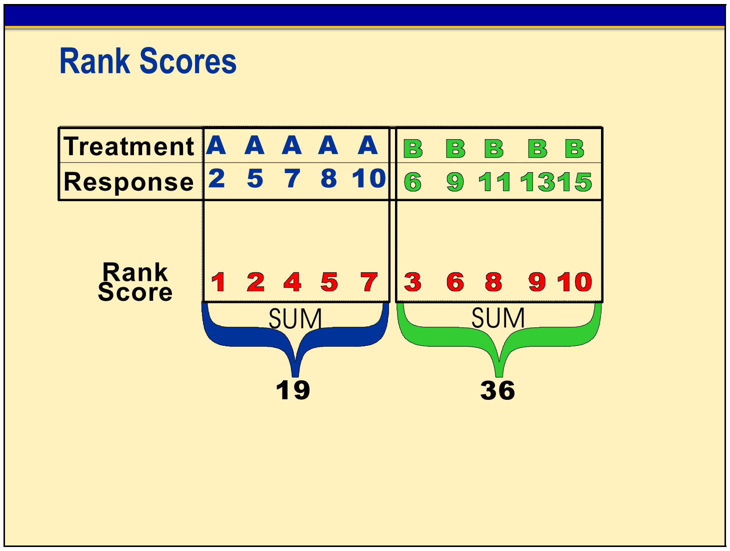
Nonparametric analyses are those that rely only on the assumption that the observations are independent.

ANOVA can fail to find group differences when

- valid outliers exist in the data
- the data is skewed
- the response variable is ordinal and not continuous.

Nonparametric tests are most often used when the normality assumption required for analysis of variance is in question. Although ANOVA is robust against minor departures from its normality assumption, extreme departures from normality can make the test less sensitive to differences between means. Therefore, when the data is very skewed or there are extreme outliers, nonparametric methods might be more appropriate. In addition, when the data follows a count measurement scale instead of an interval scale, nonparametric methods should be used.

When the normality assumption is met, nonparametric tests are almost as good as parametric tests.



Rank Scores

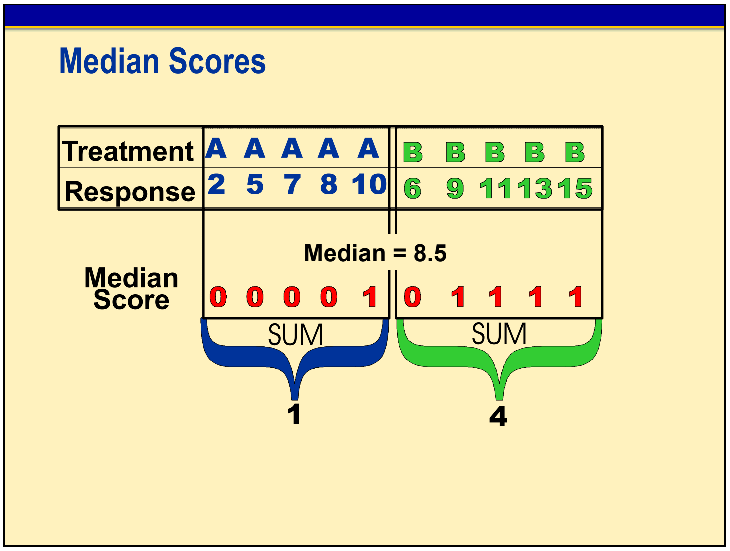
Treatment	A	A	A	A	A	B	B	B	B	B
Response	2	5	7	8	10	6	9	11	13	15
Rank Score	1	2	4	5	7	3	6	8	9	10

SUM 19

SUM 36

In nonparametric analysis, the rank of each data point is used instead of the raw data.

The illustrated ranking system ranks the data from smallest to largest. In the case of ties, the ranks are averaged. The sums of the ranks for each of the treatments are used to test the hypothesis that the populations are identical. For two populations, the Wilcoxon rank-sum test is performed. For any number of populations, the Kruskal-Wallis test is used.



Median Scores

Treatment	A	A	A	A	A	B	B	B	B	B
Response	2	5	7	8	10	6	9	11	13	15
Median Score	0	0	0	0	1	0	1	1	1	1

Median = 8.5

SUM 1

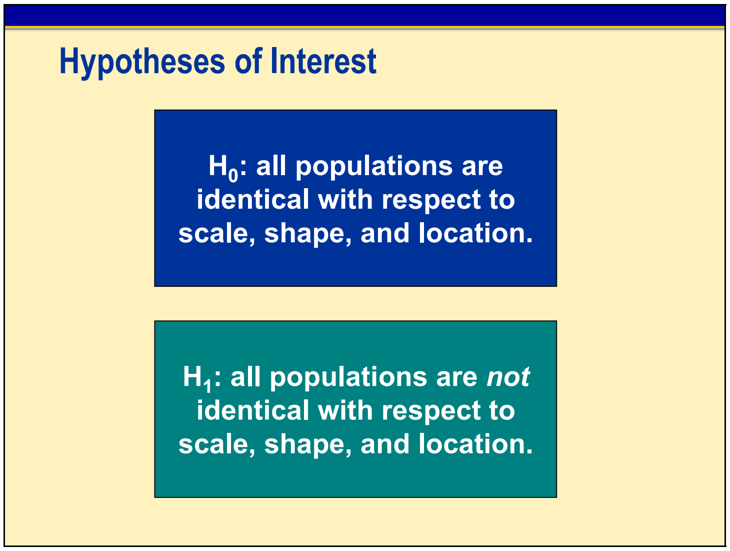
SUM 4

Recall that the median is the 50th percentile, which is the middle of your data values.

When calculating median scores, a score of

- 0 is assigned, if the data value is less than or equal to the median
- 1 is assigned, if the data value is above the median.

The sums of the median scores are used to conduct the Median test for two populations or the Brown-Mood test for any number of populations.



Hypotheses of Interest

H_0: all populations are identical with respect to scale, shape, and location.

H_1: all populations are *not* identical with respect to scale, shape, and location.

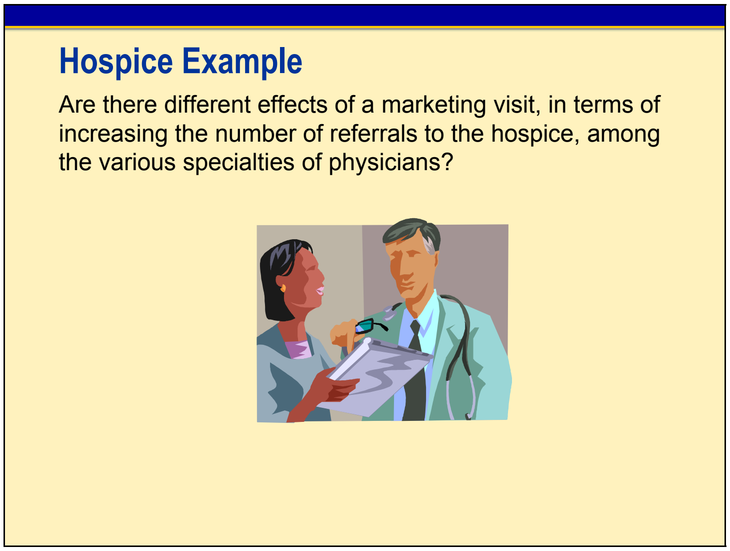
Nonparametric tests compare the probability distributions of sampled populations rather than specific parameters of these populations.

In general, with no assumptions about the distributions of the data, you are testing these hypotheses:

H_0: all populations are identical with respect to shape and location

H_1: all populations are **not** identical with respect to shape and location.

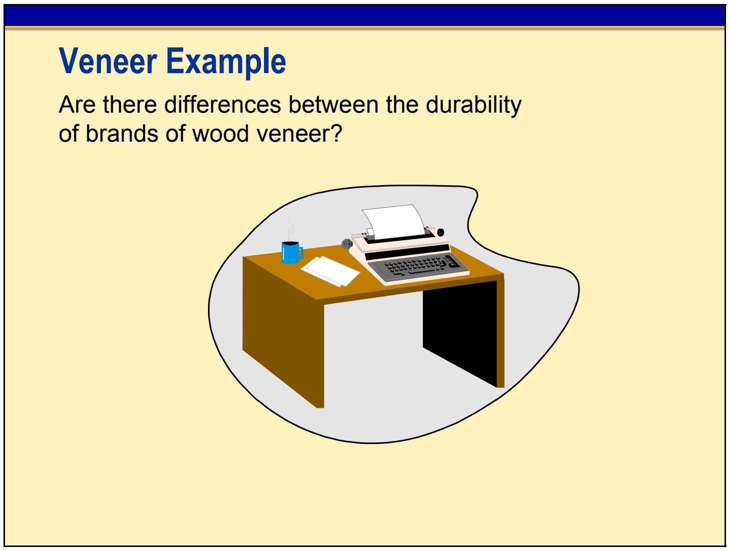
Thus, if you reject the null hypothesis, you conclude that the population distributions are different, but you have not identified the reason for the difference. The difference could be because of different variances, skewness, kurtosis, or means.



Hospice Example

Are there different effects of a marketing visit, in terms of increasing the number of referrals to the hospice, among the various specialties of physicians?

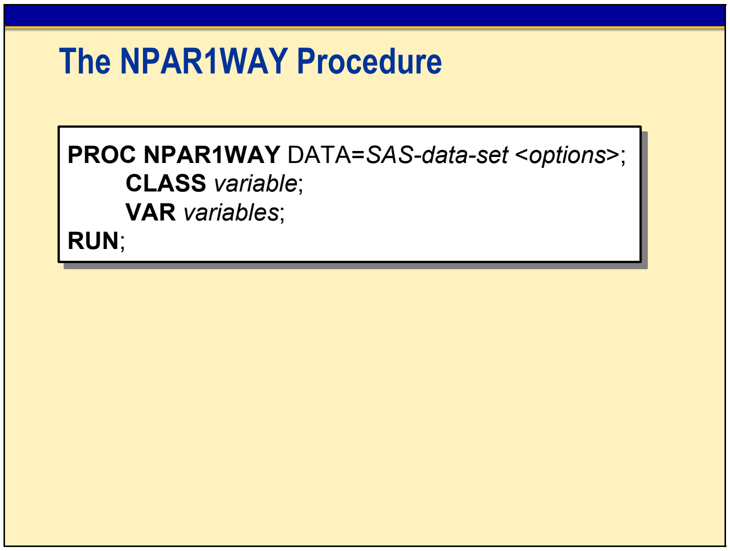
Consider a study done by Kathryn Skarzynski to determine whether there was a change in the number of referrals received from physicians after a visit by a hospice marketing nurse. One of her study questions was "Are there different effects of the marketing visits, in terms of increasing the number of referrals, among the various specialties of physicians?"



Veneer Example

Are there differences between the durability of brands of wood veneer?

Consider another experiment where the goal is to compare the durability of three brands of synthetic wood veneer. This type of veneer is often used in office furniture and on kitchen countertops. To determine durability, four samples of each of three brands are subjected to a friction test. The amount of veneer material that is worn away due to the friction is measured. The resulting wear measurement is recorded for each sample. Brands that have a small wear measurement are desirable.



The NPAR1WAY Procedure

```
PROC NPAR1WAY DATA=SAS-data-set <options>;
      CLASS variable;
      VAR variables;
RUN;
```

Selected NPAR1WAY procedure statements:

CLASS specifies a classification variable for the analysis. You must specify exactly one variable, although this variable can have any number of values.

VAR specifies the numeric analysis variables.

Nonparametric One-Way ANOVA

m3demo05.sas, m3demo06.sas, m3demo7.sas , m3demo8.sas, m3demo09.sas

Example: A portion of the data about the hospice marketing visits is in the data set **sasuser.b_hosp**. The variables in the data set are

ID the ID number of the physician's office visited

visit the type of visit, to the physician or to the physician's staff

code the medical specialty of the physician

ref3p the number of referrals three months prior to the visit

ref2p the number of referrals two months prior to the visit

ref1p the number of referrals one month prior to the visit

ref3a the number of referrals three months after the visit

ref2a the number of referrals two months after the visit

ref1a the number of referrals one month after the visit.

In addition, these variables have been calculated:

avgprior the average number of referrals per month for the three months prior to the visit

diff1 the difference between the number of referrals one month after the visit and the average number of referrals prior to the visit

diff2 the difference between the number of referrals two months after the visit and the average number of referrals prior to the visit

diff3 the difference between the number of referrals three months after the visit and the average number of referrals prior to the visit

diffbys1 the difference between the number of referrals one month after the visit and the number of referrals three months prior to the visit

diffbys2 the difference between the number of referrals two months after the visit and the number of referrals three months prior to the visit

diffbys3 the difference between the number of referrals three months after the visit and the number of referrals three months prior to the visit.

Print a subset of the variables for the first 15 observations in the data set.

```
proc print data=sasuser.b_hosp (obs=15);
   var visit code diffbys3;
run;
```

```
               Obs      visit          code        diffbys3

                 1    physician    family prac        0
                 2    physician    family prac        1
                 3    physician    oncologist        -1
                 4    physician    family prac       -3
                 5    physician    oncologist         1
                 6    physician    family prac        0
                 7    physician    oncologist        -1
                 8    physician    oncologist        -1
                 9    physician    internal med       1
                10    physician    oncologist         1
                11    physician    internal med       0
                12    physician    oncologist         0
                13    physician    oncologist         0
                14    physician    internal med       1
                15    physician    oncologist        -7
```

One of the analyses to answer the research question is to compare **diffbys3** (the number of referrals three months after the visit minus the number of referrals three months before the visit) for the different specialties.

Initially, you want to examine the distribution of the data. PROC UNIVARIATE provides much of the information needed, including stem-and-leaf and box plots as well as side-by-side box plots. Be sure to sort the data first.

```
proc sort data=sasuser.b_hosp out=hosp;
   by code;
run;

proc univariate data=hosp normal plot;
   by code;
   var diffbys3;
run;
```

Partial PROC UNIVARIATE Output

```
                         The UNIVARIATE Procedure
                              Schematic Plots

                 |
               6 +
                 |
                 |            0
                 |
               4 +
                 |
                 |
                 |
               2 +                        |           *
                 |                        |
                 |         +-----+     +-----+        *
                 |         |     |     |  +  |
               0 +         *--+--*     *-----*     *--+--*
                 |         |     |
                 |         +-----+                    *
                 |            |
              -2 +            |
                 |
                 |                                    *
                 |
              -4 +
                 |
                 |
                 |
              -6 +
                 |
                 |            0
                 |
              -8 +
                  ------------+-----------+-----------+-----------
            code          oncologi    internal    family p
```

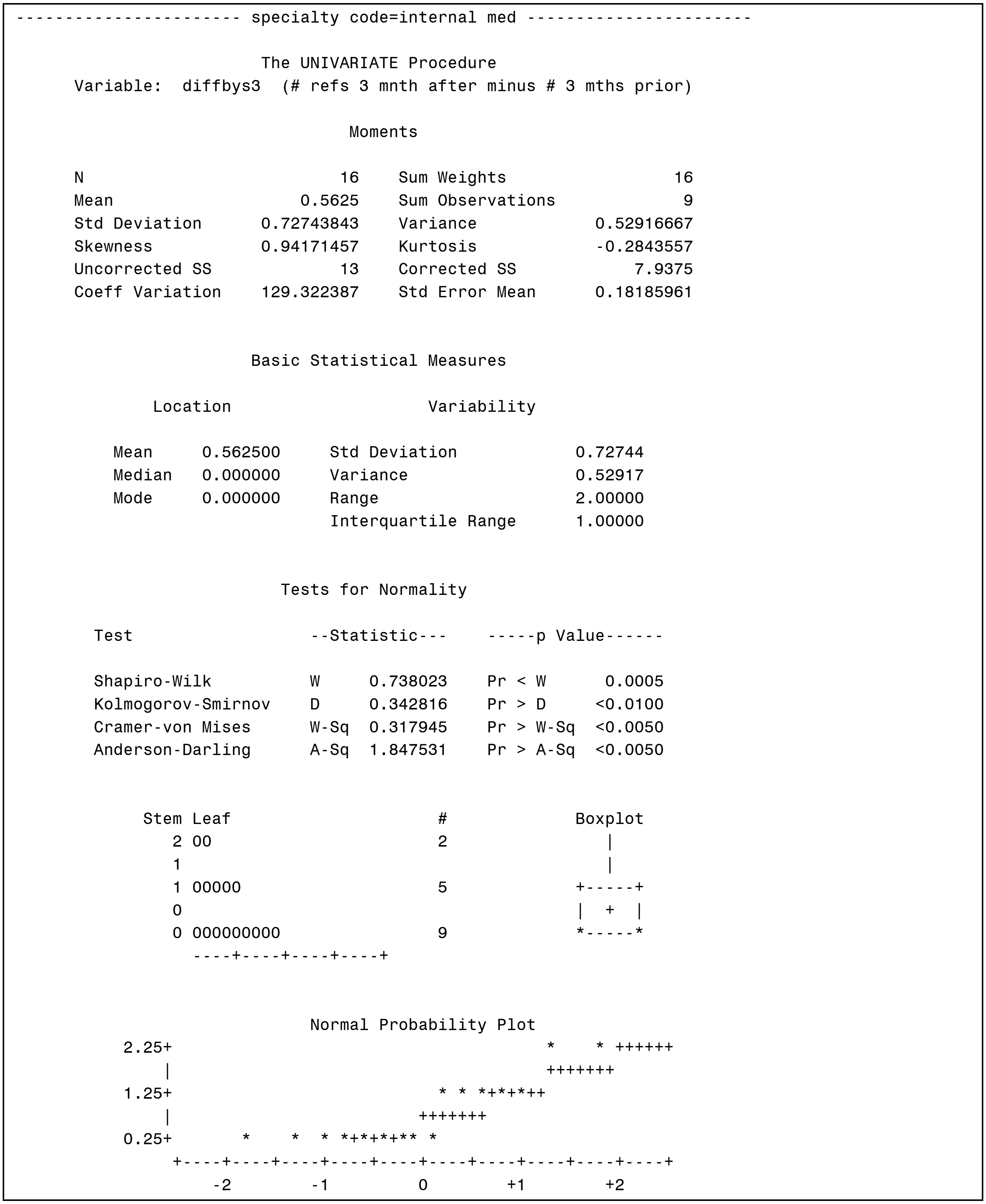
The stem-and-leaf plots show a large concentration of data at the single value zero. The data for internal medicine doctors is skewed and there are a few outliers. These characteristics, combined with the fact that the data values are actually counts and therefore ordinal, suggest that a nonparametric analysis would be more appropriate.

Examining the following PROC UNIVARIATE output for each type of specialist indicates that the data for each group is not normal. The tests for normality, the stem-and-leaf plots, and the normal probability plots all provide strong support that the data is not normal.

Partial PROC UNIVARIATE Output

```
------------------------ specialty code=oncologist ------------------------

                         The UNIVARIATE Procedure
      Variable:  diffbys3  (# refs 3 mnth after minus # 3 mths prior)

                                  Moments

      N                          19    Sum Weights                 19
      Mean               -0.2105263    Sum Observations            -4
      Std Deviation      2.22558226    Variance            4.95321637
      Skewness            -0.988574    Kurtosis            5.58306776
      Uncorrected SS             90    Corrected SS        89.1578947
      Coeff Variation    -1057.1516    Std Error Mean      0.51058359


                        Basic Statistical Measures

              Location                    Variability

          Mean     -0.21053     Std Deviation            2.22558
          Median    0.00000     Variance                 4.95322
          Mode      0.00000     Range                   12.00000
                                Interquartile Range      2.00000


                           Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W     0.810435    Pr < W      0.0016
        Kolmogorov-Smirnov    D     0.240619    Pr > D     <0.0100
        Cramer-von Mises      W-Sq  0.255644    Pr > W-Sq  <0.0050
        Anderson-Darling      A-Sq  1.459545    Pr > A-Sq  <0.0050


             Stem Leaf                     #             Boxplot
                4 0                        1                0
                2
                0 000000000000            12             +-----+
               -0 000                      3             +--+--+
               -2 00                       2                |
               -4
               -6 0                        1                0
                  ----+----+----+----+


                              Normal Probability Plot
              5+                                           *+++++++
               |                                   +++++++++
               |                     ** ***+***+++*    *
             -1+           *  * *+++*+++++
               |        +++++++++
               |++++++++
             -7+       *
                +----+----+----+----+----+----+----+----+----+----+
```

Partial PROC UNIVARIATE Output (continued)



```
----------------------- specialty code=internal med -----------------------

                         The UNIVARIATE Procedure
      Variable:  diffbys3  (# refs 3 mnth after minus # 3 mths prior)

                                  Moments

      N                          16    Sum Weights             16
      Mean                   0.5625    Sum Observations         9
      Std Deviation      0.72743843    Variance        0.52916667
      Skewness           0.94171457    Kurtosis        -0.2843557
      Uncorrected SS             13    Corrected SS        7.9375
      Coeff Variation    129.322387    Std Error Mean  0.18185961


                        Basic Statistical Measures

              Location                    Variability

          Mean     0.562500     Std Deviation            0.72744
          Median   0.000000     Variance                 0.52917
          Mode     0.000000     Range                    2.00000
                                Interquartile Range      1.00000


                           Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W     0.738023    Pr < W      0.0005
        Kolmogorov-Smirnov    D     0.342816    Pr > D     <0.0100
        Cramer-von Mises      W-Sq  0.317945    Pr > W-Sq  <0.0050
        Anderson-Darling      A-Sq  1.847531    Pr > A-Sq  <0.0050


             Stem Leaf                    #             Boxplot
                2 00                      2                |
                1                                          |
                1 00000                   5             +-----+
                0                                       |  +  |
                0 000000000               9             *-----*
                  ----+----+----+----+


                              Normal Probability Plot
           2.25+                                     *    * ++++++
               |                                     +++++++
           1.25+                           * * *+*+*++
               |                         +++++++
           0.25+       *    *  * *+*+*+** *
                +----+----+----+----+----+----+----+----+----+----+
                    -2        -1         0        +1        +2
```

Partial PROC UNIVARIATE Output (continued)

```
----------------------- specialty code=family prac -----------------------

                         The UNIVARIATE Procedure
      Variable:  diffbys3  (# refs 3 mnth after minus # 3 mths prior)

                                  Moments

      N                          19    Sum Weights                 19
      Mean                        0    Sum Observations             0
      Std Deviation      0.94280904    Variance            0.88888889
      Skewness           -1.3336242    Kurtosis            6.24954044
      Uncorrected SS             16    Corrected SS                16
      Coeff Variation             .    Std Error Mean      0.21629523


                        Basic Statistical Measures

              Location                    Variability

          Mean            0     Std Deviation            0.94281
          Median          0     Variance                 0.88889
          Mode            0     Range                    5.00000
                                Interquartile Range            0


                           Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W     0.683337    Pr < W     <0.0001
        Kolmogorov-Smirnov    D     0.394737    Pr > D     <0.0100
        Cramer-von Mises      W-Sq  0.653756    Pr > W-Sq  <0.0050
        Anderson-Darling      A-Sq  3.003314    Pr > A-Sq  <0.0050


             Stem Leaf                     #            Boxplot
                2 0                        1               *
                1 00                       2               *
                0 00000000000000          14            +--+--+
               -0
               -1 0                        1               *
               -2
               -3 0                        1               *
                  ----+----+----+----+

                              Normal Probability Plot
            2.5+                                           *   ++++
               |                                    *++*+++++++
               |              * * * *** ***+***+*+*+
           -0.5+           *   ++++++++++
               |    +++++++++++
               |++++   *
           -3.5+
                +----+----+----+----+----+----+----+----+----+----+
                    -2        -1         0       +1        +2
```

For illustrative purposes, use the WILCOXON option to perform a rank sum test and the MEDIAN option to perform the median test. This data was actually analyzed using the rank sum test.

```
proc npar1way data=hosp wilcoxon median;
   class code;
   var diffbys3;
run;
```

Selected PROC NPAR1WAY statement options:

WILCOXON requests an analysis of the rank scores. The output includes the Wilcoxon 2-sample test and the Kruskal-Wallis test for two or more populations.

MEDIAN requests an analysis of the median scores. The output includes the median 2-sample test and the median 1-way analysis test for two or more populations.

```
                       The NPAR1WAY Procedure

         Wilcoxon Scores (Rank Sums) for Variable diffbys3
                      Classified by Variable code

                          Sum of     Expected      Std Dev         Mean
code                N     Scores     Under H0     Under H0        Score
_______________________________________________________________________

oncologist         19     468.50       522.50    49.907208    24.657895
internal med       16     538.00       440.00    47.720418    33.625000
family prac        19     478.50       522.50    49.907208    25.184211

                  Average scores were used for ties.


                          Kruskal-Wallis Test

                       Chi-Square         4.2304
                       DF                      2
                       Pr > Chi-Square    0.1206
```

The PROC NPAR1WAY output from the WILCOXON option shows the actual sums of the rank scores and the expected sums of the rank scores if the null hypothesis is true. From the Kruskal-Wallis test (chi-square approximation), the *p*-value is .1206. Therefore, at the 5% level of significance, you do not reject the null hypothesis. There is not enough evidence to conclude that the distributions of change in hospice referrals for the different groups of physicians are significantly different.

PROC NPAR1WAY Output (continued)

```
                          The NPAR1WAY Procedure

     Median Scores (Number of Points Above Median) for Variable diffbys3
                        Classified by Variable code

                              Sum of      Expected       Std Dev          Mean
  code                N       Scores      Under H0      Under H0         Score
  ―――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――――
  oncologist         19     8.566667          9.50      1.232093      0.450877
  internal med       16    10.300000          8.00      1.178106      0.643750
  family prac        19     8.133333          9.50      1.232093      0.428070

                      Average scores were used for ties.


                          Median One-Way Analysis

                         Chi-Square         3.8515
                         DF                      2
                         Pr > Chi-Square    0.1458
```

Again, based on the *p*-value of .1458, at the 5% level of significance, you do not reject the null hypothesis. There is not enough evidence to conclude that there are differences between specialists.

Example: Recall the experiment to compare the durability of three brands of synthetic wood veneer. The data is stored in the **`sasuser.b_ven`** data set.

```
proc print data=sasuser.b_ven;
   title 'Wood Veneer Wear Data';
run;
```

```
                         Wood Veneer Wear Data

                         Obs    brand    wear

                           1    Acme      2.3
                           2    Acme      2.1
                           3    Acme      2.4
                           4    Acme      2.5
                           5    Champ     2.2
                           6    Champ     2.3
                           7    Champ     2.4
                           8    Champ     2.6
                           9    Ajax      2.2
                          10    Ajax      2.0
                          11    Ajax      1.9
                          12    Ajax      2.1
```

Because there is only a sample size of 4 for each brand of veneer, the usual PROC NPAR1WAY Wilcoxon test *p*-values are inaccurate. Instead, the EXACT statement should be added to the PROC NPAR1WAY code. This provides exact *p*-values for the simple linear rank statistics based on the Wilcoxon scores rather than estimated *p*-values based on continuous approximations.

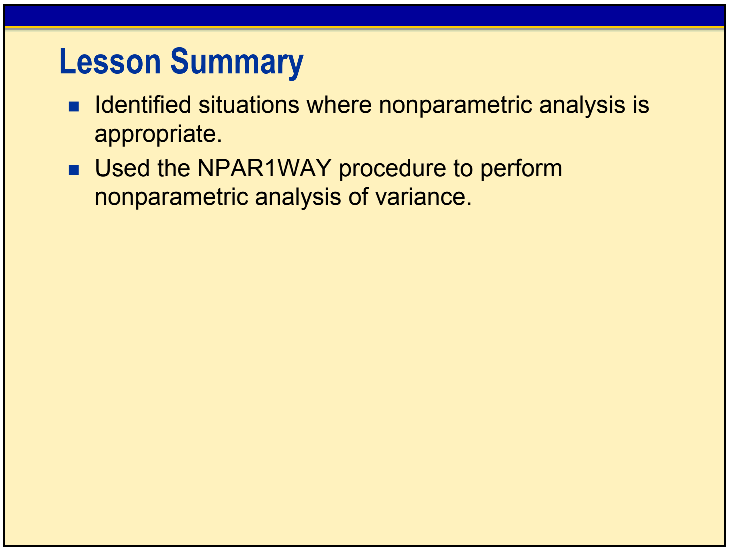
Exact analysis is available for both the WILCOXON and MEDIAN options in PROC NPAR1WAY. You can specify which of these scores you want to use to compute the exact *p*-values by adding either one or both of these options to the EXACT statement. If no options are listed in the EXACT statement, exact *p*-values are computed for all the linear rank statistics requested in the PROC NPAR1WAY statement.

You should exercise care when choosing to use the EXACT statement with PROC NPAR1WAY. Computational time can be prohibitive depending on the number of groups, the number of distinct response variables, the total sample size, and the speed and memory available on your computer. You can terminate exact computations and exit PROC NPAR1WAY at any time by pressing the system interrupt key and choosing to stop computations.

```
proc npar1way data=sasuser.b_ven wilcoxon;
   class brand;
   var wear;
   exact;
run;
```

```
                          Wood Veneer Wear Data

                          The NPAR1WAY Procedure

               Wilcoxon Scores (Rank Sums) for Variable wear
                       Classified by Variable brand

                        Sum of      Expected       Std Dev           Mean
   brand       N        Scores      Under H0      Under H0          Score
   ─────────────────────────────────────────────────────────────────────
   Acme        4         31.50          26.0      5.846522         7.8750
   Champ       4         34.50          26.0      5.846522         8.6250
   Ajax        4         12.00          26.0      5.846522         3.0000

                    Average scores were used for ties.


                            Kruskal-Wallis Test

                   Chi-Square                      5.8218
                   DF                                   2
                   Asymptotic Pr >  Chi-Square     0.0544
                   Exact      Pr >= Chi-Square     0.0480
```

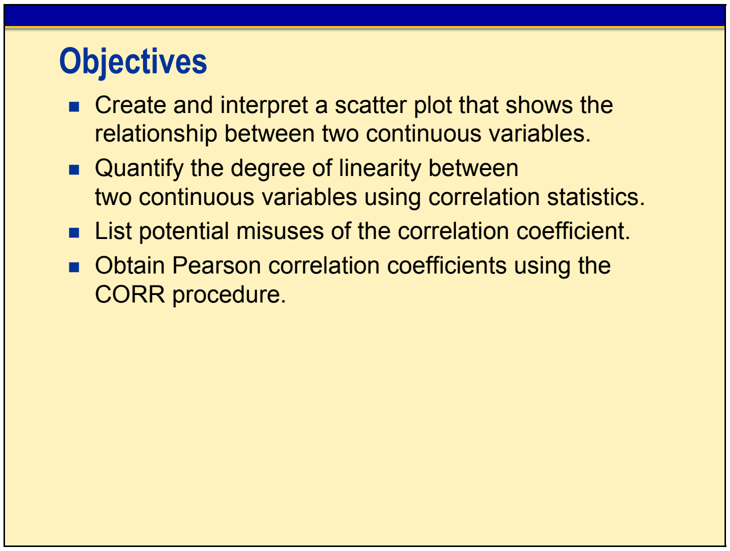
In the PROC NPAR1WAY output shown above, the exact *p*-value is .0480, which is significant at α=.05. Note the difference between the exact *p*-value and the *p*-value based on the chi-square approximation.



Lesson Summary

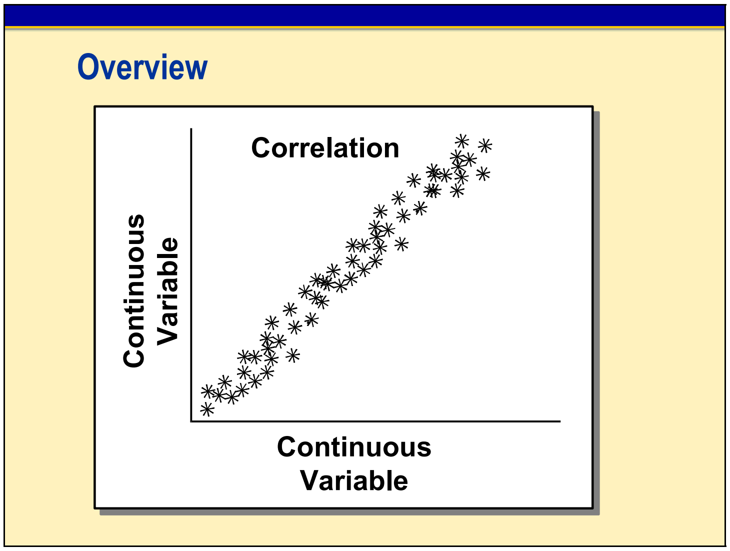
- Identified situations where nonparametric analysis is appropriate.
- Used the NPAR1WAY procedure to perform nonparametric analysis of variance.

3.3 Exploratory Data Analysis for Linear Regression



Objectives

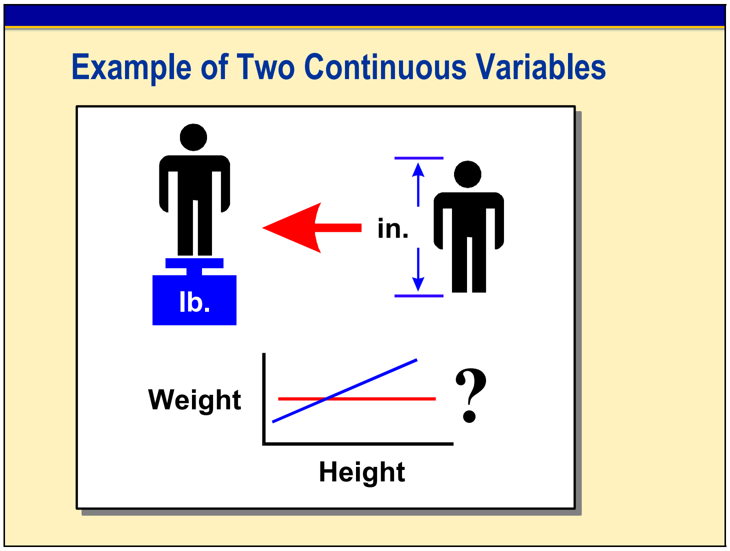
- Create and interpret a scatter plot that shows the relationship between two continuous variables.
- Quantify the degree of linearity between two continuous variables using correlation statistics.
- List potential misuses of the correlation coefficient.
- Obtain Pearson correlation coefficients using the CORR procedure.



Overview

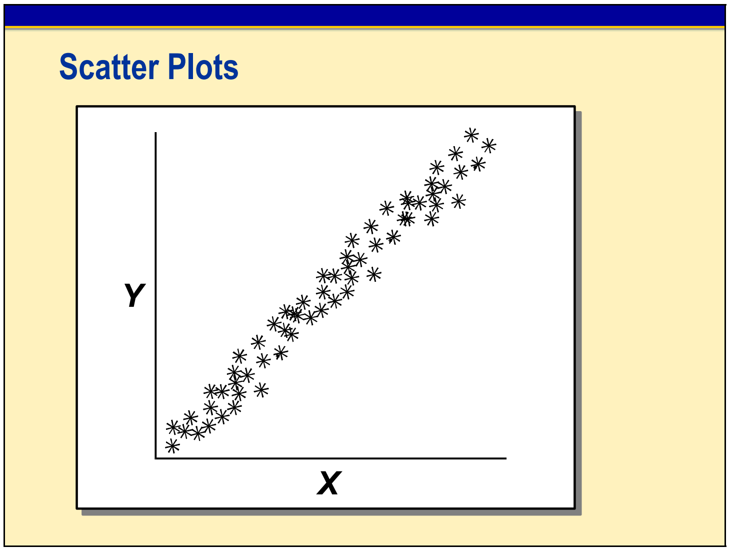
In Module 2, you learned that when you have a categorical predictor variable and a continuous outcome variable you use ANOVA to analyze your data. In this lesson, you have two continuous variables.

You use correlation analysis to examine and describe the relationship between two continuous variables. However, before you use correlation analysis, it is important to view the relationship between the two continuous variables using a scatter plot.



Example of Two Continuous Variables

Example: A random sample of high school students is selected to determine the relationship between a person’s height and weight. Height and weight are measured on a numeric scale. They have a large, potentially infinite number of possible values instead of only a few categories such as short, medium, and tall. Therefore, these variables are considered to be continuous.

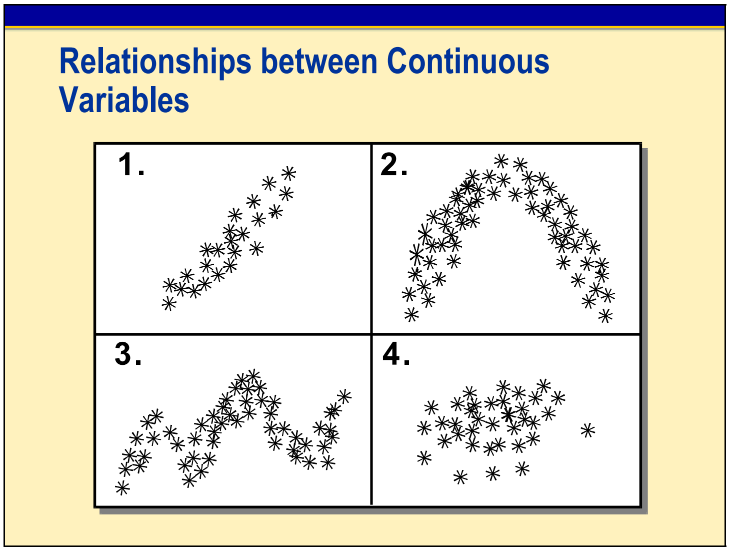


Scatter Plots

Scatter plots are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

Scatter plots are useful to

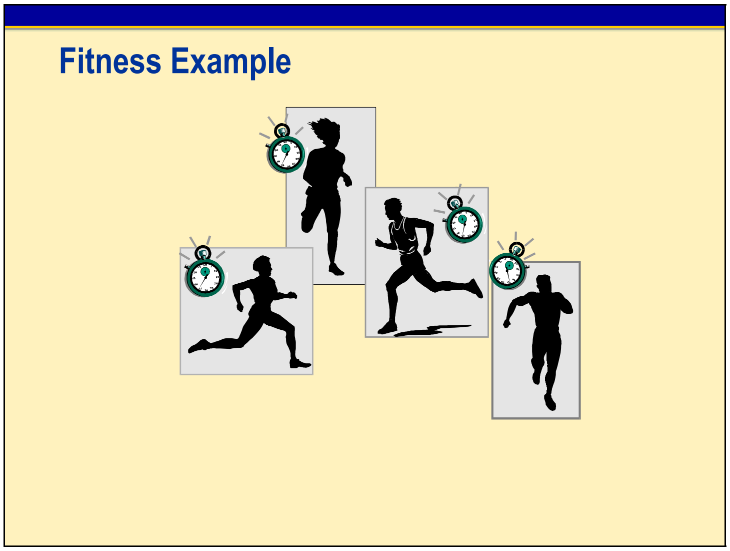
- explore the relationships between two variables
- locate outlying or unusual values
- identify possible trends
- communicate data analysis results.



Relationships between Continuous Variables

Describing the relationship between two continuous variables is an important first step in any statistical analysis. The scatter plot is the most important tool you have in describing these relationships. The diagrams above illustrate some possible relationships.

1. A straight line describes the relationship.
2. Curvature is present in the relationship.
3. There may be a cyclical pattern in the relationship. You might see this when the predictor is time.
4. There is no clear relationship between the variables.



Fitness Example

A club wants to evaluate the fitness of its members. One measure of fitness is oxygen consumption. The club measured the oxygen consumption as well as several other continuous measurements, such as age, pulse, and weight. They are interested in determining whether any of these other variables can help predict oxygen consumption.

The data set **`sasuser.b_fitness`** contains the following variables:

`name`	name of the member
`gender`	gender of the member
`runtime`	time to run 1.5 miles (in minutes)
`age`	age of the member (in years)
`weight`	weight of the member (in kilograms)
`oxygen_consumption`	a measure of the ability to use oxygen in the blood stream
`run_pulse`	pulse rate at the end of the run
`rest_pulse`	resting pulse rate
`maximum_pulse`	maximum pulse rate during the run
`performance`	a measure of overall fitness.

Generating Scatter Plots

m3demo10.sas, m3demo11.sas

You can view the data using the PRINT procedure.

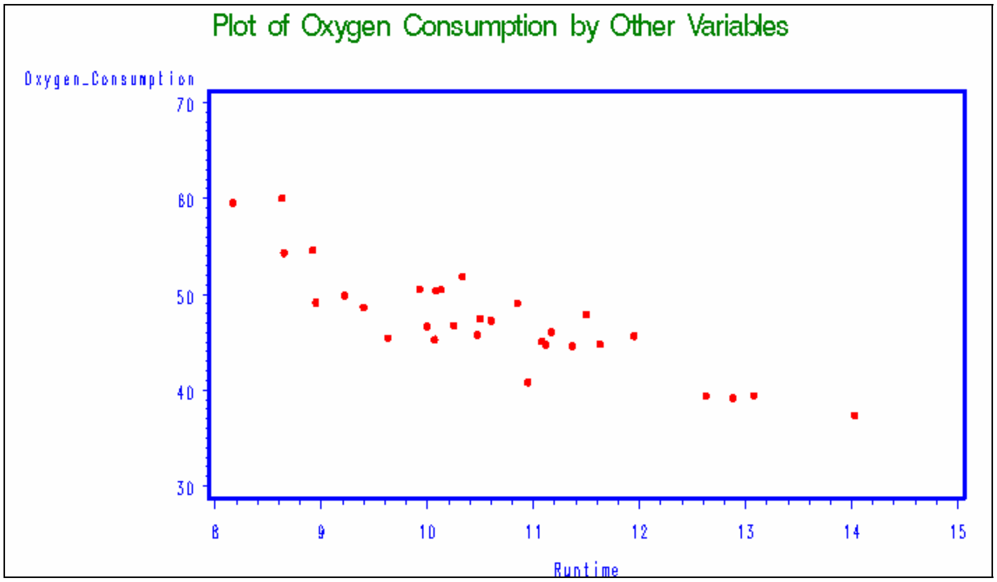
```
proc print data=sasuser.b_fitness;
   title 'Printout of the b_fitness data set';
run;
```

```
                    Printout of the b_fitness data set

                                              O
                                              x
                                              y
                                              g
                                              e
                                              n                 M
                                              _                 a
                                              C                 x     P
                                              o            R    i     e
                                              n      R     e    m     r
                                              s      u     s    u     f
                         R                    u      n     t    m     o
                   G     u            W       m      _     _    _     r
                   e     n            e       p      P     P    P     m
        N          n     t            i       t      u     u    u     a
    O   a          d     i      A     g       i      l     l    l     n
    b   m          e     m      g     h       o      s     s    s     c
    s   e          r     e      e     t       n      e     e    e     e

    1   Donna      F    8.17   42   68.15   59.57   166   40   172   14
    2   Gracie     F    8.63   38   81.87   60.06   170   48   186   13
    3   Luanne     F    8.65   43   85.84   54.30   156   45   168   13
    4   Mimi       F    8.92   50   70.87   54.63   146   48   155   11
    5   Chris      M    8.95   49   81.42   49.16   180   44   185   11
    6   Allen      M    9.22   38   89.02   49.87   178   55   180   12
    7   Nancy      F    9.40   49   76.32   48.67   186   56   188   10
    8   Patty      F    9.63   52   76.32   45.44   164   48   166   10
    9   Suzanne    F    9.93   57   59.08   50.55   148   49   155    9
   10   Teresa     F   10.00   51   77.91   46.67   162   48   168    9
   11   Bob        M   10.07   40   75.07   45.31   185   62   185    9
   12   Harriett   F   10.08   49   73.37   50.39   168   67   168    9
   13   Jane       F   10.13   44   73.03   50.54   168   45   168    9
   14   Harold     M   10.25   48   91.63   46.77   162   48   164    9
   15   Sammy      M   10.33   54   83.12   51.85   166   50   170    8
   16   Buffy      F   10.47   52   73.71   45.79   186   59   188    8
   17   Trent      M   10.50   52   82.78   47.47   170   53   172    8
   18   Jackie     F   10.60   47   79.15   47.27   162   47   164    8
   19   Ralph      M   10.85   43   81.19   49.09   162   64   170    7
   20   Jack       M   10.95   51   69.63   40.84   168   57   172    7
   21   Annie      F   11.08   51   67.25   45.12   172   48   172    7
   22   Kate       F   11.12   45   66.45   44.75   176   51   176    7
   23    Carl       M   11.17   54   79.38   46.08   156   62   165    7
   24    Don        M   11.37   44   89.47   44.61   178   62   182    6
   25    Effie      F   11.50   48   61.24   47.92   170   52   176    6
   26    George     M   11.63   47   77.45   44.81   176   58   176    6
   27    Iris       F   11.95   40   75.98   45.68   176   70   180    5
   28    Mark       M   12.63   57   73.37   39.41   174   58   176    4
   29    Steve      M   12.88   54   91.63   39.20   168   44   172    4
   30    Vaughn     M   13.08   44   81.42   39.44   174   63   176    2
   31    William    M   14.03   45   87.66   37.39   186   56   192    0
```

Examine the relationships between **oxygen_consumption** and the other continuous variables in the data set using the GPLOT procedure.

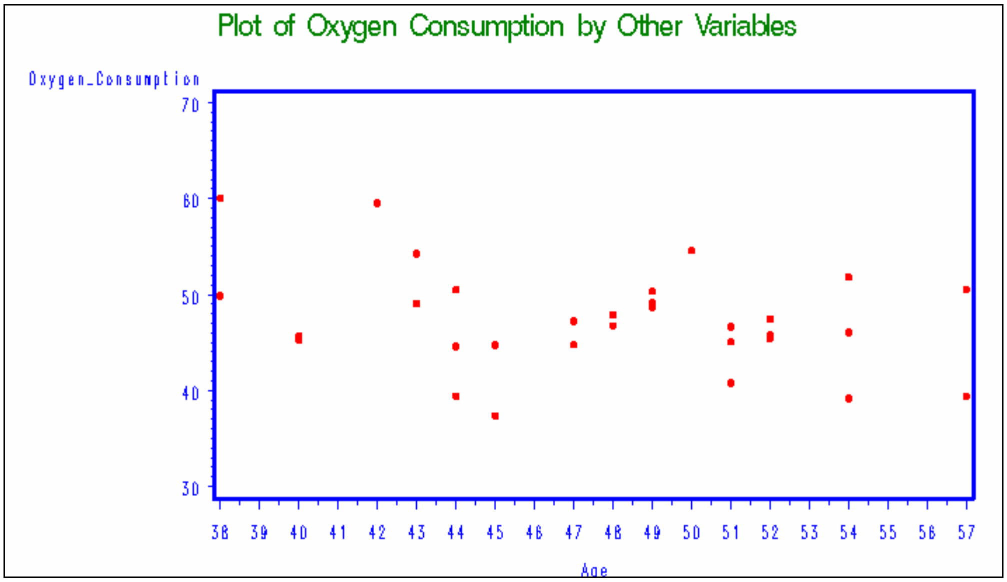
```
goptions reset=all gunit=pct border;

axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
axis2 length=70 w=3 color=blue label=(h=3) value=(h=3);

proc gplot data=sasuser.b_fitness;
   plot oxygen_consumption * (runtime age weight run_pulse
                              rest_pulse maximum_pulse performance)
        / vaxis=axis1 haxis=axis2;
   symbol1 v=dot h=2 w=4 color=red;
   title h=5 color=green
         'Plot of Oxygen Consumption by Other Variables';
run;
quit;
```

Partial PROC GPLOT Output



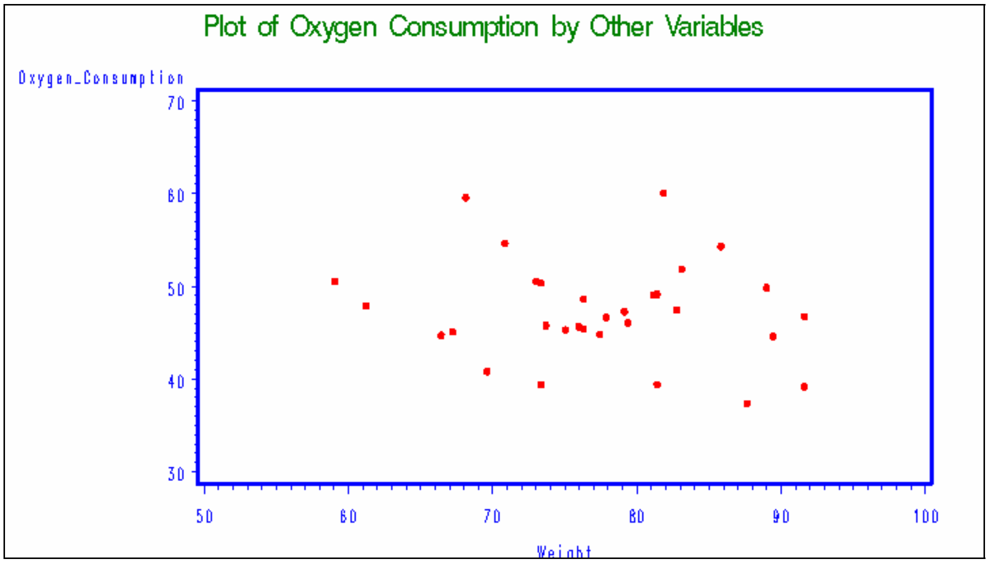
This plot suggests that the longer an individual takes to run 1.5 miles, the lower the oxygen consumption measurement.

PROG GPLOT Output (continued)



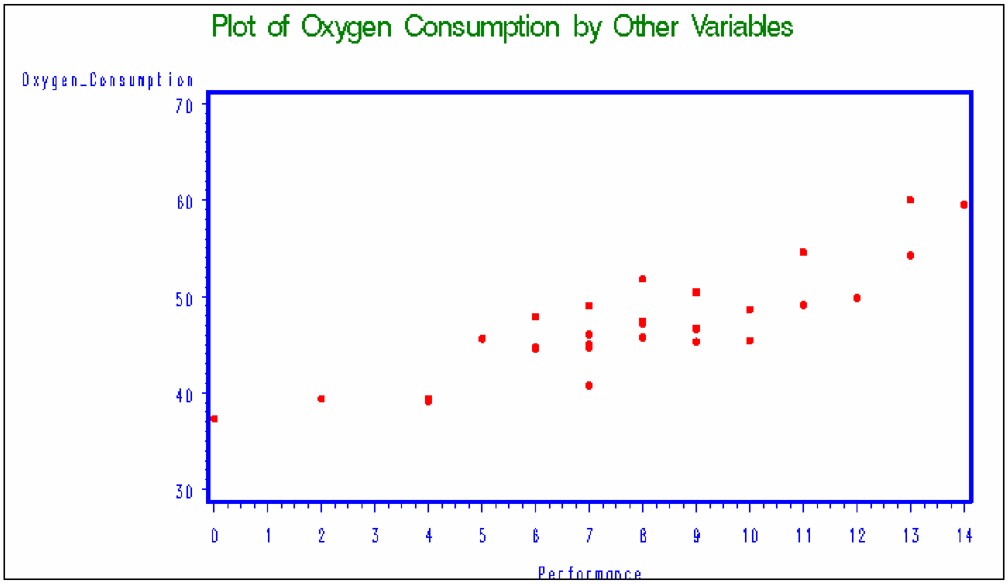
There appears to be a weak linear relationship between **`oxygen_consumption`** and **`age`**.

PROG GPLOT Output (continued)

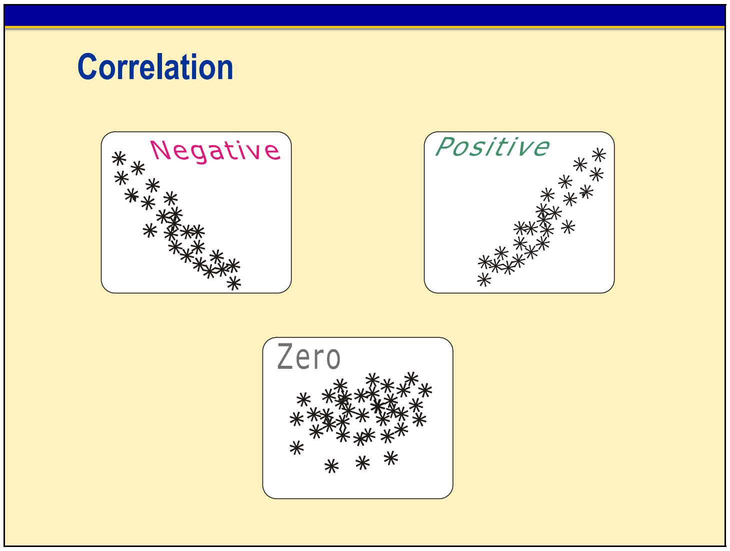


There does **not** appear to be a relationship between **`oxygen_consumption`** and **`weight`**.

PROG GPLOT Output (continued)



As **performance** increases, **oxygen_consumption** appears to increase slightly.



Correlation

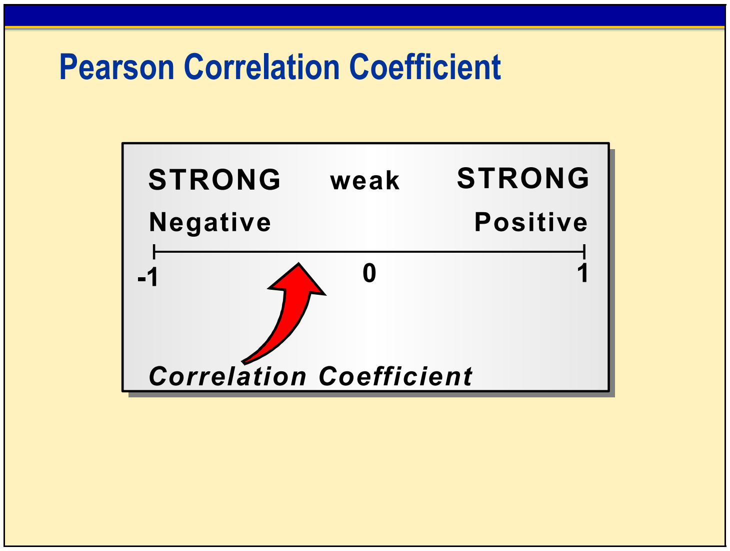
After you examine the scatter plot, you can quantify the relationship between two variables with correlation statistics. Two variables are correlated if there is a **linear** relationship between them. If not, the variables are uncorrelated.

You can classify correlated variables according to the type of correlation:

positive one variable tends to increase in value as the other variable increases in value

negative one variable tends to decrease in value as the other variable increases in value

zero no linear relationship between the two variables (uncorrelated).



Pearson Correlation Coefficient

STRONG Negative

weak

STRONG Positive

-1

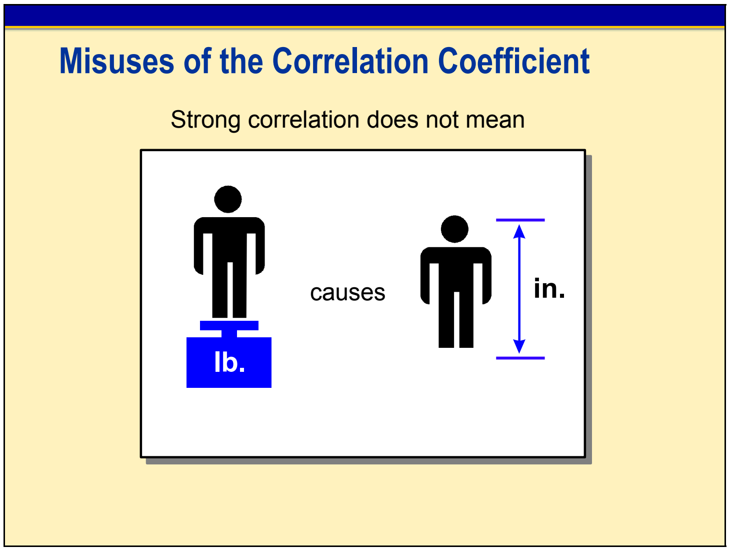
0

1

Correlation Coefficient

Correlation statistics measure the degree of linear relationship between two variables. A common correlation statistic used for continuous variables is the Pearson correlation coefficient. Values of correlation statistics are

- between −1 and 1
- closer to either extreme if there is a high degree of linear relationship between the two variables
- close to 0 if there is no linear relationship between the two variables
- close to 1 if there is a positive linear relationship
- close to −1 if there is a negative linear relationship.



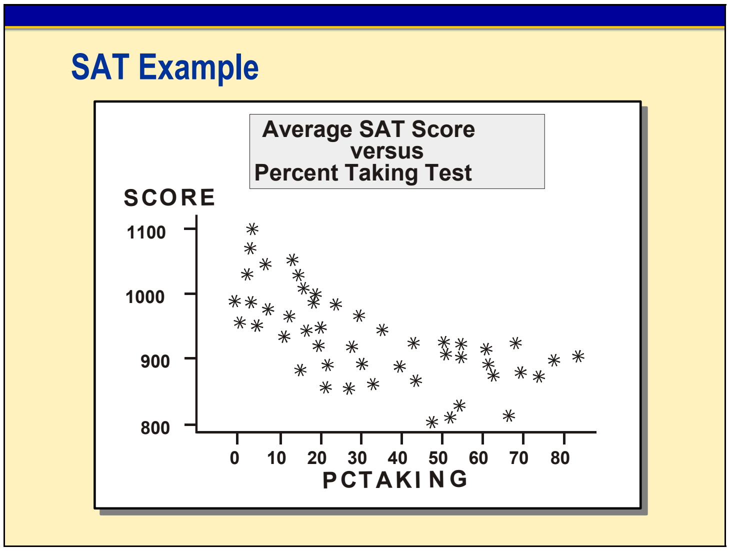
Misuses of the Correlation Coefficient

Strong correlation does not mean

Common errors can be made when interpreting the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

- A strong correlation between two variables does **not** mean that change in one variable causes the other variable to change, or vice versa.

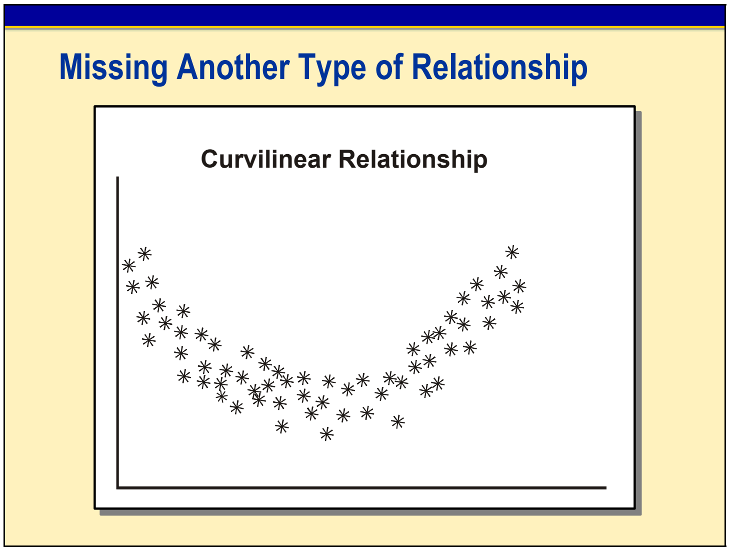
Sample correlation coefficients can be large because of chance or because both variables are affected by other variables.



SAT Example

An example of improperly concluding a cause-and-effect relationship is illustrated using data from the Scholastic Aptitude Test (SAT) from 1989. The scatter plot shown above plots each state's average total SAT score (`score`) versus the percent of eligible students in the state who took the SAT (`pctaking`). The correlation between `score` and `pctaking` is –0.86867. Looking at the plot and at this statistic, an eligible student for the next year can conclude, "If I am the only student in my state to take the SAT, I am guaranteed a good score."

Clearly this type of thinking is faulty. Can you think of possible explanations for this relationship?



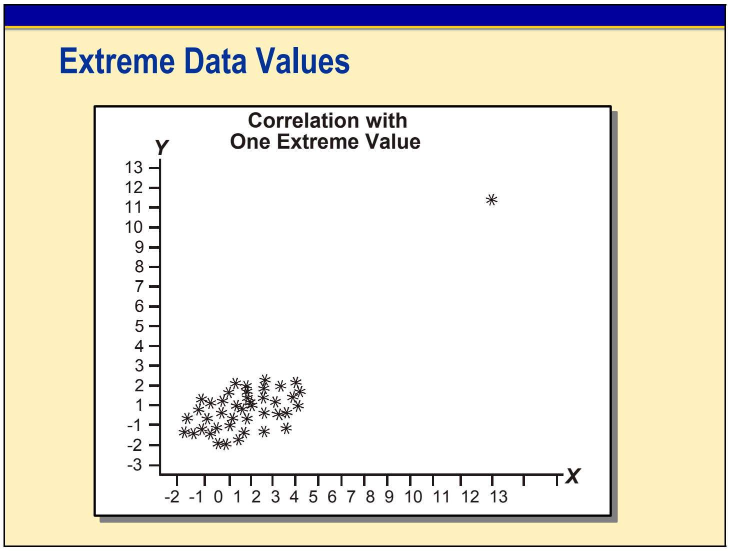
Missing Another Type of Relationship

Curvilinear Relationship

In the scatter plot above, the variables have a fairly low Pearson correlation coefficient. Why?

- Correlation coefficients measure linear relationships.
- A correlation coefficient close to 0 indicates that there is not a strong linear relationship between two variables.
- A correlation coefficient close to 0 does not mean that there is no relationship of any kind between the two variables.

In this example, there is a curvilinear relationship between the two variables.

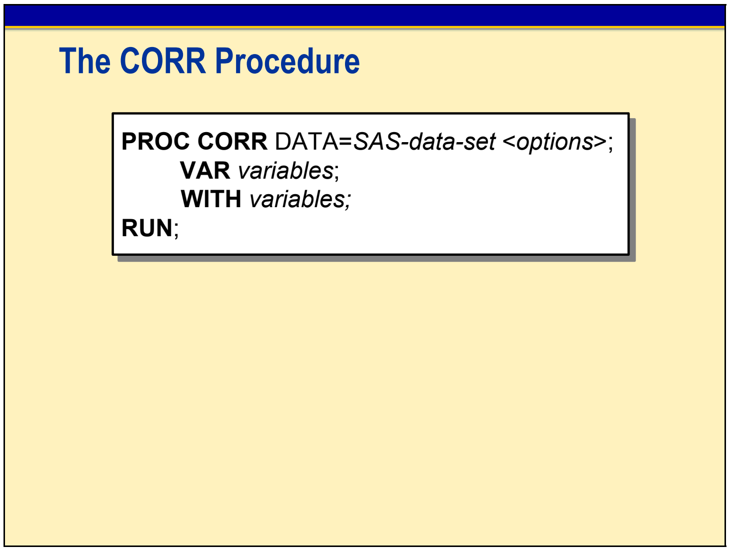


Extreme Data Values

Correlation coefficients are highly affected by a few extreme values of either variable. The scatter plot above shows the degree of linear relationship is mainly determined by one point. If you delete the unusual point from the data, the correlation is close to 0.

In this situation, follow these steps:

1. Investigate the unusual data point to make sure it is valid.
2. If the data point is valid, collect more data between the unusual data point and the group of data points to see whether a linear relationship unfolds.
3. Try to replicate the unusual data point by collecting data at a fixed value of *x* (in this case, *x*=11). This determines whether the data point is unusual.
4. Compute two correlation coefficients, one with the unusual data point and one without it. This shows how influential the unusual data point is in the analysis.



The CORR Procedure

```
PROC CORR DATA=SAS-data-set <options>;
        VAR variables;
        WITH variables;
RUN;
```

You can use the CORR procedure to produce correlation statistics for your data. By default, PROC CORR produces Pearson correlation statistics and corresponding *p*-values.

Selected CORR procedure statements:

VAR specifies variables for which to produce correlations. If a WITH statement is not specified, correlations are produced for each pair of variables in the VAR statement. If the WITH statement is specified, the VAR statement specifies the column variables in the correlation matrix.

WITH produces correlations for each variable in the VAR statement with all variables in the WITH statement. The WITH statement specifies the row variables in the correlation matrix.

Generating Correlation Coefficients

m3demo12.sas, m3demo13.sas

Use PROC CORR to produce a Pearson correlation coefficient for **oxygen_consumption** with the other continuous variables.

```
proc corr data=sasuser.b_fitness rank;
   var runtime age weight run_pulse rest_pulse
       maximum_pulse performance;
   with oxygen_consumption;
   title 'Example of CORR Procedure';
run;
```

Selected PROC CORR statement option:

RANK orders the correlations from highest to lowest in absolute value.

The output from PROC CORR is shown below. By default, the analysis generates univariate statistics for the analysis variables and a correlation statistic.

```
                           Example of CORR Procedure
                              The CORR Procedure

   1 With Variables:    Oxygen_Consumption
   7      Variables:    Runtime            Age                Weight
                        Run_Pulse          Rest_Pulse         Maximum_Pulse
                        Performance

                              Simple Statistics

 Variable                    N          Mean       Std Dev           Sum

 Oxygen_Consumption         31      47.37581       5.32777          1469
 Runtime                    31      10.58613       1.38741     328.17000
 Age                        31      47.67742       5.26236          1478
 Weight                     31      77.44452       8.32857          2401
 Run_Pulse                  31     169.64516      10.25199          5259
 Rest_Pulse                 31      53.45161       7.61944          1657
 Maximum_Pulse              31     173.77419       9.16410          5387
 Performance                31       8.00000       3.11983     248.00000

                              Simple Statistics

              Variable                 Minimum       Maximum

              Oxygen_Consumption      37.39000      60.06000
              Runtime                  8.17000      14.03000
              Age                     38.00000      57.00000
              Weight                  59.08000      91.63000
              Run_Pulse              146.00000     186.00000
              Rest_Pulse              40.00000      70.00000
              Maximum_Pulse          155.00000     192.00000
              Performance                    0      14.00000

                 Pearson Correlation Coefficients, N = 31
                        Prob > |r| under H0: Rho=0

Oxygen_Consumption      Performance        Runtime            Rest_Pulse
                              0.86377           -0.86219           -0.39935
                               <.0001             <.0001             0.0260

                 Pearson Correlation Coefficients, N = 31
                        Prob > |r| under H0: Rho=0

Oxygen_Consumption      Run_Pulse          Age                Maximum_Pulse
                             -0.39808           -0.31162           -0.23677
                               0.0266             0.0879             0.1997

                 Pearson Correlation Coefficients, N = 31
                        Prob > |r| under H0: Rho=0

                   Oxygen_Consumption      Weight
                                                -0.16289
                                                  0.3813
```

The correlation coefficient between **oxygen_consumption** and **performance** is 0.86377. The *p*-value is small, indicating that the population correlation coefficient (Rho) is significantly different from 0. The second largest correlation coefficient, in absolute value, is **runtime**, -0.86219.

The correlation analysis indicates that several variables could be good predictors for **oxygen_consumption**.

When you prepare to conduct a regression analysis, it is always good practice to examine the correlations between the potential predictor variables. PROC CORR can be used to generate a matrix of correlation coefficients.

```
proc corr data=sasuser.b_fitness nosimple;
   var runtime age weight run_pulse rest_pulse
       maximum_pulse performance;
   title;
run;
```

Selected PROC CORR statement option:

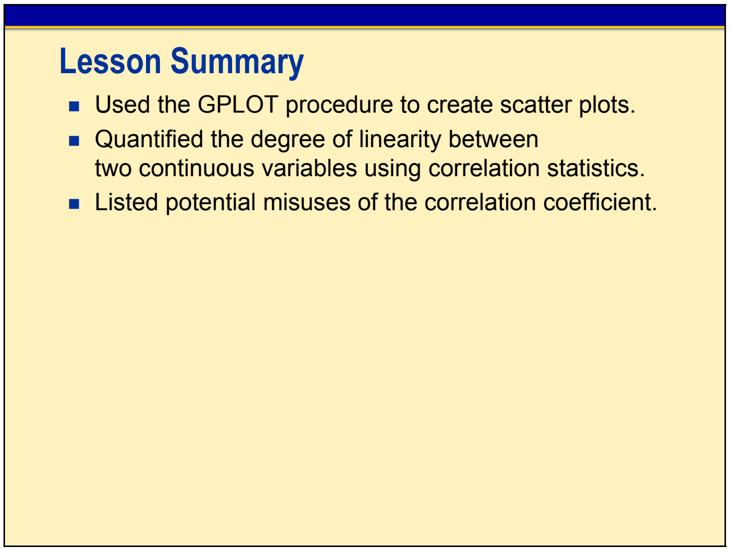
NOSIMPLE suppresses printing simple descriptive statistics for each variable.

```
                            The CORR Procedure

   7  Variables:    Runtime       Age           Weight        Run_Pulse
                    Rest_Pulse    Maximum_Pulse Performance



                Pearson Correlation Coefficients, N = 31
                        Prob > |r| under H0: Rho=0

                      Runtime           Age        Weight      Run_Pulse

  Runtime             1.00000       0.19523       0.14351        0.31365
                                     0.2926        0.4412         0.0858

  Age                 0.19523       1.00000      -0.24050       -0.31607
                       0.2926                      0.1925         0.0832

  Weight              0.14351      -0.24050       1.00000        0.18152
                       0.4412        0.1925                       0.3284

  Run_Pulse           0.31365      -0.31607       0.18152        1.00000
                       0.0858        0.0832        0.3284

  Rest_Pulse          0.45038      -0.15087       0.04397        0.35246
                       0.0110        0.4178        0.8143         0.0518
```

PROG CORR Output (continued)

```
               Pearson Correlation Coefficients, N = 31
                       Prob > |r| under H0: Rho=0

                             Rest_      Maximum_
                             Pulse         Pulse      Performance

       Runtime             0.45038       0.22610         -0.98841
                            0.0110        0.2213           <.0001

       Age                -0.15087      -0.41490         -0.22943
                            0.4178        0.0203           0.2144

       Weight              0.04397       0.24938         -0.10544
                            0.8143        0.1761           0.5724

       Run_Pulse           0.35246       0.92975         -0.31369
                            0.0518        <.0001           0.0857

       Rest_Pulse          1.00000       0.30512         -0.47957
                                          0.0951           0.0063

               Pearson Correlation Coefficients, N = 31
                       Prob > |r| under H0: Rho=0

                     Runtime           Age        Weight      Run_Pulse

 Maximum_Pulse       0.22610      -0.41490       0.24938        0.92975
                      0.2213        0.0203        0.1761         <.0001

 Performance        -0.98841      -0.22943      -0.10544       -0.31369
                      <.0001        0.2144        0.5724         0.0857

               Pearson Correlation Coefficients, N = 31
                       Prob > |r| under H0: Rho=0

                             Rest_      Maximum_
                             Pulse         Pulse      Performance

       Maximum_Pulse       0.30512       1.00000         -0.22035
                            0.0951                         0.2336

       Performance        -0.47957      -0.22035          1.00000
                            0.0063        0.2336
```

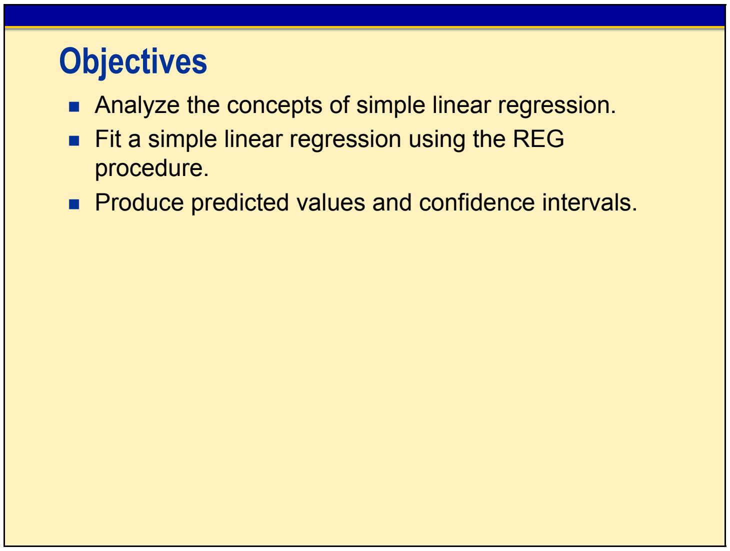
There are strong correlations between **runtime** and **performance** (-0.98841) and between **run_pulse** and **maximum_pulse** (0.92975).



Lesson Summary

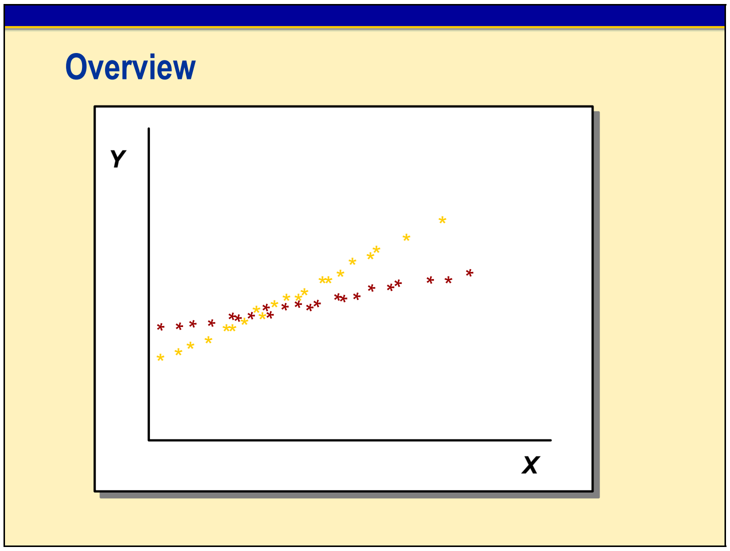
- Used the GPLOT procedure to create scatter plots.
- Quantified the degree of linearity between two continuous variables using correlation statistics.
- Listed potential misuses of the correlation coefficient.

3.4 Simple Linear Regression



Objectives

- Analyze the concepts of simple linear regression.
- Fit a simple linear regression using the REG procedure.
- Produce predicted values and confidence intervals.

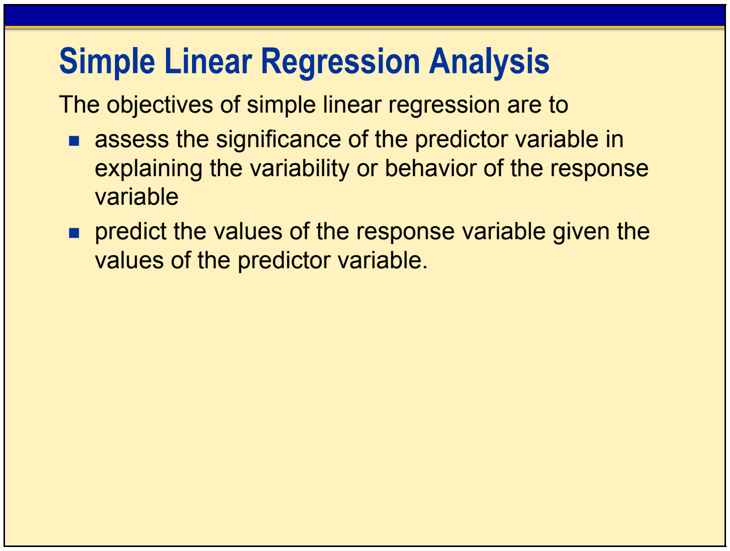


Overview

In the last lesson, you used correlation analysis to quantify the linear relationships between continuous response variables. Two pairs of variables can have the same correlation statistic, but the linear relationship can be different. In this lesson, you use simple linear regression to define the linear relationship between a response variable and a predictor variable.

The *response variable* is the variable of primary interest.

The *predictor variable* is used to explain the variability in the response variable.

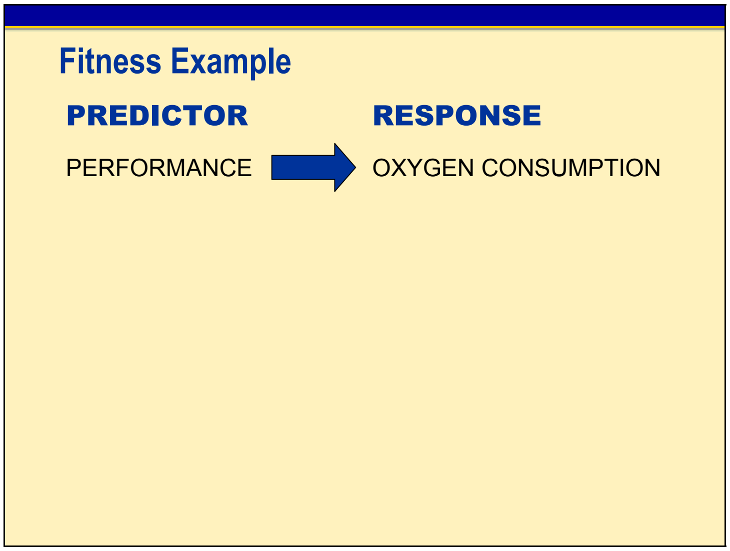


Simple Linear Regression Analysis

The objectives of simple linear regression are to

- assess the significance of the predictor variable in explaining the variability or behavior of the response variable
- predict the values of the response variable given the values of the predictor variable.

In simple linear regression, the values of the predictor variable are assumed fixed. Thus, you try to explain the variability of the response variable given the values of the predictor variable.

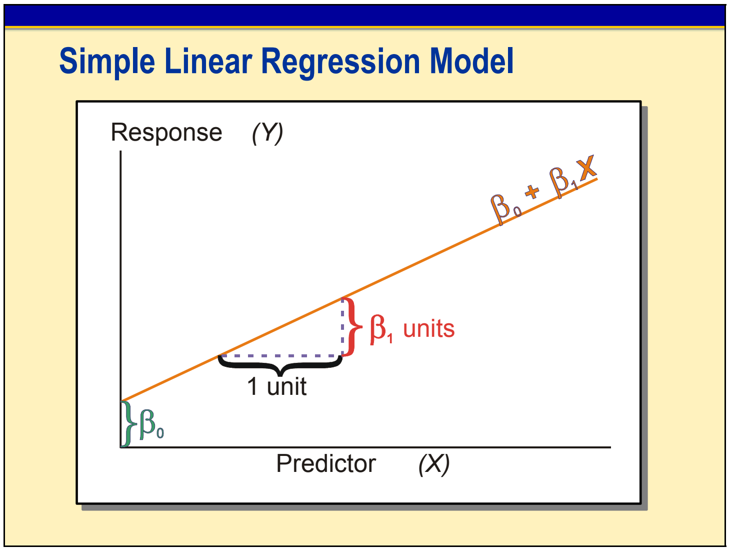


Fitness Example

PREDICTOR **RESPONSE**

PERFORMANCE → OXYGEN CONSUMPTION

You have noted that the performance measure has the highest correlation (-0.98841) with the oxygen consumption capacity of the club members. Consequently, you want to explore the relationship between **`oxygen_consumption`** and **`performance`** using simple linear regression.



Simple Linear Regression Model

The relationship between the response variable and the predictor variable can be characterized by the equation $Y = \beta_0 + \beta_1X + \varepsilon$

where

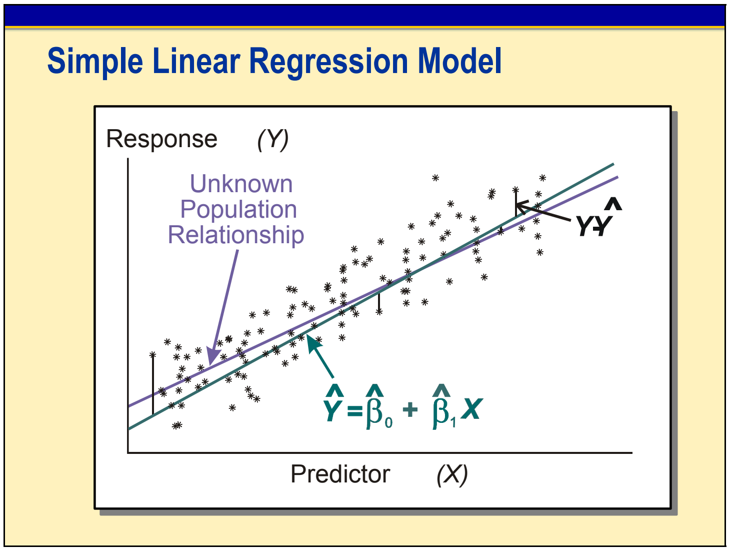
Y response variable

X predictor variable

β_0 intercept parameter, which corresponds to the value of the response variable when the predictor is 0

β_1 slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable

ε error term representing deviations of Y about $\beta_0 + \beta_1X$.



Simple Linear Regression Model

Because your goal in simple linear regression is usually to characterize the relationship between the response and predictor variables in your population, you begin with a sample of data. From this sample, you estimate the unknown population parameters (β_0, β_1) that define the assumed relationship between your response and predictor variables.

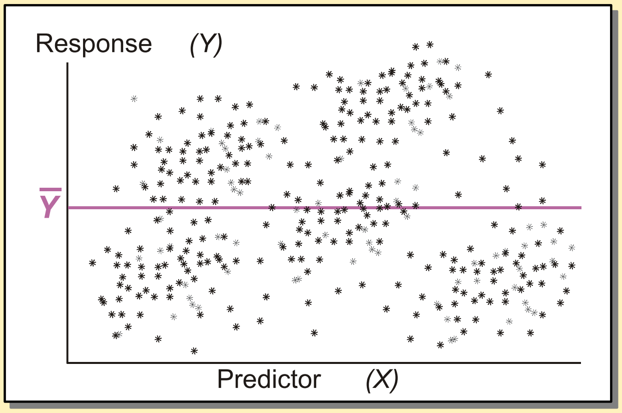
Estimates of the unknown population parameters β_0 and β_1 are obtained by the *method of least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

The method of least squares produces parameter estimates with certain optimum properties. If the assumptions of simple linear regression are valid, the least squares estimates are unbiased estimates of the population parameters and have minimum variance. The least squares estimators are often called BLUE (Best Linear Unbiased Estimators). The term *best* is used because of the minimum variance property.

Because of these optimum properties, the method of least squares is used by many data analysts to investigate the relationship between continuous predictor and response variables.

With a large and representative sample, the fitted regression line should be a good approximation of the relationship between the response and predictor variables in the population. The estimated parameters obtained using the method of least squares should be good approximations of the true population parameters.

The Baseline Model



To determine whether the predictor variable explains a significant amount of variability in the response variable, the simple linear regression model is compared to the baseline model. The fitted regression line in a baseline model is a horizontal line across all values of the predictor variable. The slope of the regression line is 0, and the intercept is the sample mean of the response variable, ($\bar{Y}$).

In a baseline model, there is no association between the response variable and the predictor variable. Knowing the mean of the response variable is as good in predicting values in the response variable as knowing the values of the predictor variable.

Model Hypothesis Test

Null Hypothesis:

The simple linear regression model does not fit the data better than the baseline model.

$$\beta_1 = 0$$

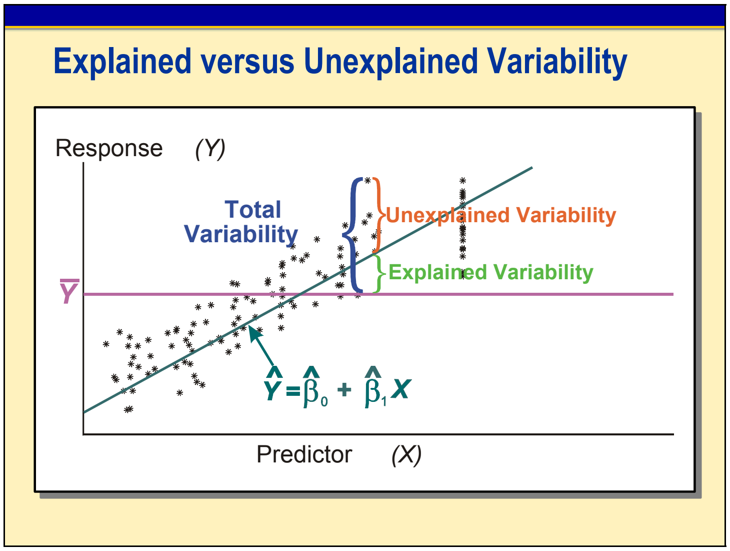
Alternative Hypothesis:

The simple linear regression model does fit the data better than the baseline model.

$$\beta_1 \neq 0$$

If the estimated simple linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you do **not** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.

If the estimated simple linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that the slope of the regression line in the population is **not** 0 and that the predictor variable explains a significant amount of variability in the response variable.



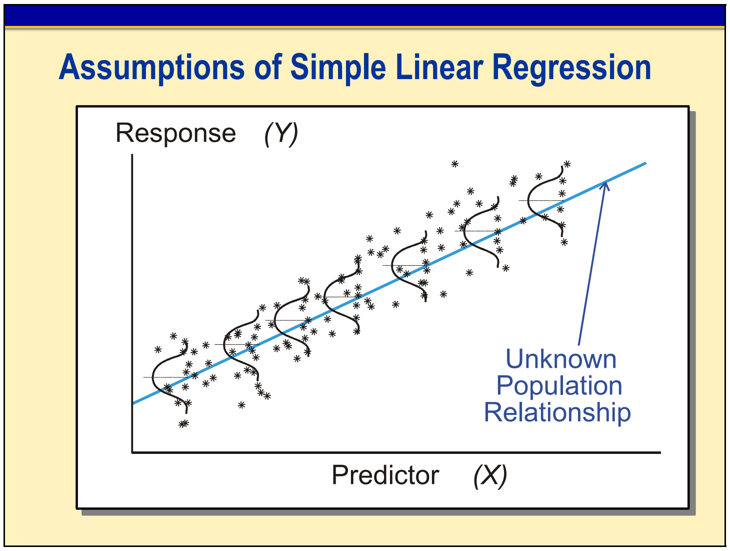
Explained versus Unexplained Variability

To determine whether a simple linear regression model is better than the baseline model, compare the explained variability to the unexplained variability.

Explained variability — is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSM) is the amount of variability explained by your model. The model sum of squares is equal to $\sum(\hat{Y}_i - \bar{Y})^2$.

Unexplained variability — is related to the difference between the observed values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by your model. The error sum of squares is equal to $\sum(Y_i - \hat{Y}_i)^2$.

Total variability — is related to the difference between the observed values and the mean of the response variable. The corrected total sum of squares is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to $\sum(Y_i - \bar{Y})^2$.

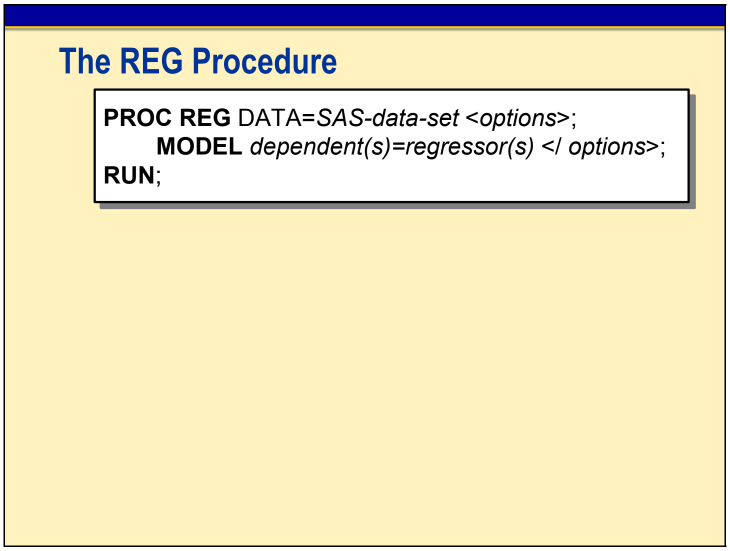


Assumptions of Simple Linear Regression

One of the assumptions of simple linear regression is that the mean of the response variable is linearly related to the value of the predictor variable. In other words, a straight line connects the means of the response variable at each value of the predictor variable.

The other assumptions are the same as the assumptions for ANOVA: the responses are normally distributed, have equal variances, and are independent at each value of the predictor variable.

The verification of these assumptions is discussed in a later module.



The REG Procedure

```
PROC REG DATA=SAS-data-set <options>;
      MODEL dependent(s)=regressor(s) </ options>;
RUN;
```

The REG procedure enables you to fit regression models to your data.

Selected REG procedure statements:

MODEL specifies the response and predictor variables. The variables must be numeric.

PROC REG supports RUN-group processing, which means that the procedure stays active until a PROC, DATA, or QUIT statement is encountered. This enables you to submit additional statements followed by another RUN statement without resubmitting the PROC statement.

Performing Simple Linear Regression

m3demo14.sas

Example: Because there is an apparent linear relationship between **oxygen_consumption** and **performance**, perform a simple linear regression analysis with **oxygen_consumption** as the response variable.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=performance;
   title 'Simple Linear Regression of Oxygen Consumption '
         'and Performance';
run;
quit;
```

```
      Simple Linear Regression of Oxygen Consumption and Performance

                           The REG Procedure
                             Model: MODEL1
                 Dependent Variable: Oxygen_Consumption

                         Analysis of Variance

                                 Sum of          Mean
 Source                  DF       Squares        Square   F Value   Pr > F

 Model                    1     635.34150     635.34150     85.22   <.0001
 Error                   29     216.21305       7.45562
 Corrected Total         30     851.55455


           Root MSE             2.73050    R-Square     0.7461
           Dependent Mean      47.37581    Adj R-Sq     0.7373
           Coeff Var            5.76349


                         Parameter Estimates

                           Parameter      Standard
 Variable             DF    Estimate         Error   t Value   Pr > |t|

 Intercept             1    35.57526       1.36917     25.98     <.0001
 Performance           1     1.47507       0.15979      9.23     <.0001
```

The Analysis of Variance (ANOVA) table provides an analysis of the variability observed in the data and the variability explained by the regression line.

```
                           Analysis of Variance

                                   Sum of           Mean
Source                  DF        Squares         Square    F Value    Pr > F

Model                    1      635.34150      635.34150      85.22    <.0001
Error                   29      216.21305        7.45562
Corrected Total         30      851.55455
```

The ANOVA table for simple linear regression is divided into six columns.

Source — labels the source of variability.

- Model — is the variability explained by your model.
- Error — is the variability unexplained by your model.
- Corrected Total — is the total variability in the data.

DF — is the degrees of freedom associated with each source of variability.

Sum of Squares — is the amount of variability associated with each source of variability.

Mean Square — is the ratio of the sum of squares and the degrees of freedom. This value corresponds to the amount of variability associated with each degree of freedom for each source of variation.

F Value — is the ratio of the mean square for the model and the mean square for the error. This ratio compares the variability explained by the regression line to the variability unexplained by the regression line.

Pr > F — is the *p*-value associated with the *F* value.

The *F* value is testing whether the slope of the predictor variable is equal to 0. The *p*-value is small (less than .05), so you have enough evidence at the .05 significance level to reject the null hypothesis. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, **`performance`** explains a significant amount of variability of **`oxygen_consumption`**.

The second part of the output provides summary measures of fit for the model.

```
            Root MSE             2.73050    R-Square     0.7461
            Dependent Mean      47.37581    Adj R-Sq     0.7373
            Coeff Var            5.76349
```

R-Square — the coefficient of determination, usually referred to as the R-square value. This value is

- between 0 and 1.
- the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.7461, which means that the regression line explains 75% of the total variation in the response values.
- the square of the Pearson correlation coefficient.

Root MSE — an estimate of the standard deviation of the response variable at each value of the predictor variable. It is the square root of the mean square error.

Dependent Mean — the overall mean of the response variable, $\overline{Y}$.

Coeff Var — the coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is

- calculated as $\left(\frac{RootMSE}{\overline{Y}}\right) * 100$
- a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.

Adj R-Sq — the adjusted R square is the R square that is adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in a later lesson.

The Parameter Estimates table defines the model for your data.

```
                            Parameter Estimates

                              Parameter      Standard
 Variable             DF       Estimate         Error   t Value   Pr > |t|

 Intercept             1       35.57526       1.36917     25.98     <.0001
 Performance           1        1.47507       0.15979      9.23     <.0001
```

DF — represents the degrees of freedom associated with each term in the model.

Parameter Estimate — is the estimated value of the parameters associated with each term in the model.

Standard Error — is the standard error of each parameter estimate.

t Value — is the *t* statistic, which is calculated by dividing the parameters by their corresponding standard errors.

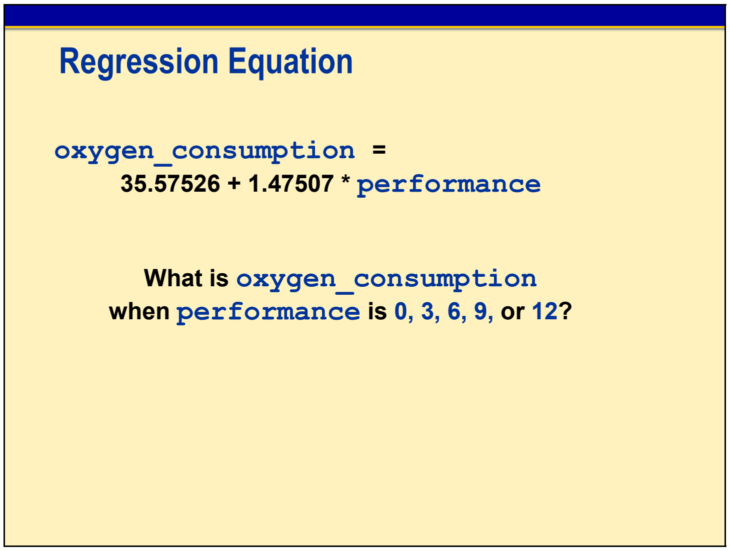
Pr > |t| — is the *p*-value associated with the *t* statistic. It tests whether the parameter associated with each term in the model is different from 0. For this example, the slope for the predictor variable is statistically different from 0. Thus, you can conclude that the predictor variable explains a significant portion of variability in the response variable.

Because the estimate of β_o=35.58 and β_1=1.48, the estimated regression equation is given by

Predicted **`oxygen_consumption`** = 35.58 + 1.48(**`performance`**)

The model indicates that an increase of one unit for **`performance`** amounts to a 1.48 increase in **`oxygen_consumption`**. However, this equation is appropriate only in the range of values you observed for the variable **`performance`**.

The Parameter Estimates table also shows that the intercept parameter is not equal to 0. However, the test for the intercept parameter only has practical significance when the range of values for the predictor variable includes 0. In this example, the test could have practical significance because **`performance`**=0 is inside the range of values you are considering. (**`Performance`** ranges from 0 to 14.)



Regression Equation

oxygen_consumption =
35.57526 + 1.47507 * performance

What is oxygen_consumption
when performance is 0, 3, 6, 9, or 12?

One objective in regression analysis is to predict values of the response variable given values of the predictor variables. You can obviously use the estimated regression equation to produce predicted values, but if you want a large number of predictions, this can be cumbersome.

To produce predicted values in PROC REG, follow these steps:

1. Create a data set with the values of the independent variable for which you want to make predictions.
2. Concatenate the data in the step above with the original data set.
3. Fit a simple linear regression model to the new data set and specify the P option in the MODEL statement. Because the observations added in the previous step contain missing values for the response variable, PROC REG does not include these observations when fitting the regression model. However, PROC REG does produce predicted values for these observations.

Producing Predicted Values

m3demo15.sas

Example: Produce predicted values of **oxygen_consumption** when **performance** is 0, 3, 6, 9, and 12.

```
data need_predictions;
   input performance @@;
   datalines;
0 3 6 9 12
;

data predoxy;
   set sasuser.b_fitness
       need_predictions;
run;

proc reg data=predoxy;
   model oxygen_consumption=performance / p;
   id performance;
   title 'Simple Linear Regression of Oxygen Consumption '
         'and Performance';
run;
quit;
```

Selected REG procedure statement:

ID specifies a variable to label observations in the output produced by certain MODEL statement options.

Selected MODEL statement option:

P prints the values of the response variable, the predicted values, and the residual values.

If you have a large data set and have already fitted the regression model, a more efficient way to produce predicted values is in a DATA step. You can either write the parameter estimates in the DATA step or use the OUTEST= option in PROC REG. Here is an example program:

```
data _null_;
   input performance @@;
   oxygen_consumption=35.57526+1.47507*performance;
   put performance= oxygen_consumption=;
   datalines;
0 3 6 9 12
;
run;
```

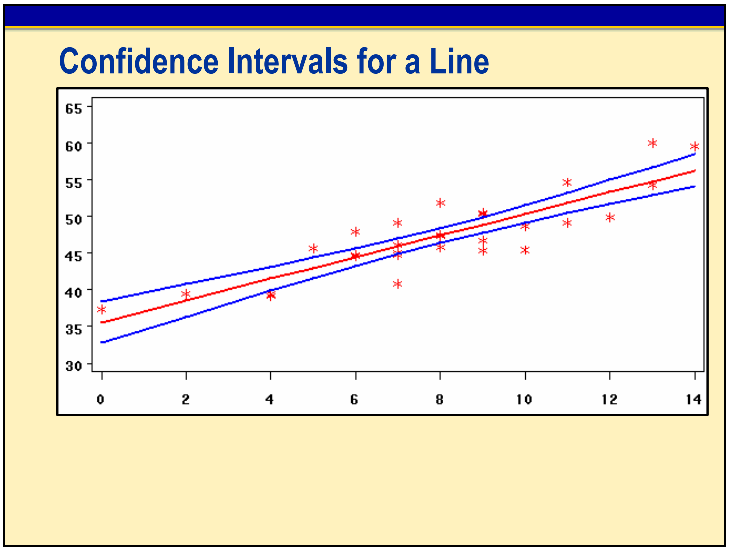
Partial PROC REG Output

```
                                   Dep Var    Predicted
   Obs    Performance    Oxygen_Consumption       Value     Residual
    32              0                     .     35.5753            .
    33              3                     .     40.0005            .
    34              6                     .     44.4257            .
    35              9                     .     48.8509            .
    36             12                     .     53.2761            .
```

Because you specified **performance** in the ID statement, the values of this variable appear in the first column.

The output shows that the estimated value of **oxygen_consumption** is 35.58 when **performance** equals 0. However, when the **performance** is 12, the predicted **oxygen_consumption** is 53.28.

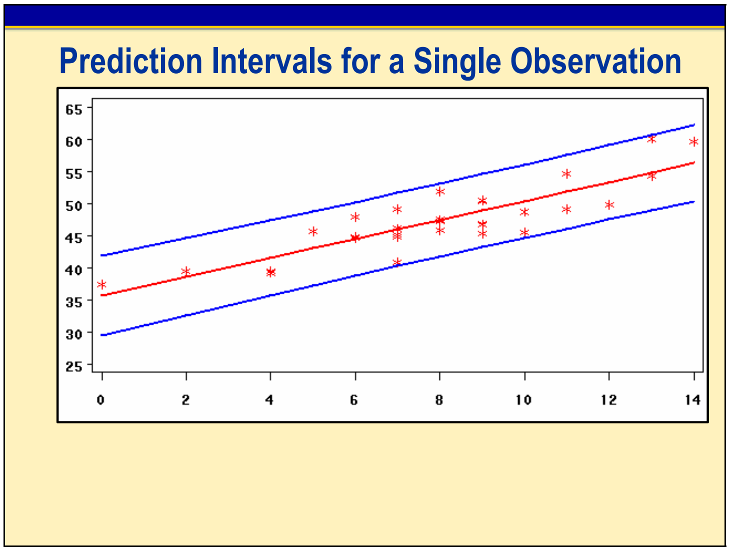
Choose only values within or near the range of the predictor variable when you are predicting new values for the response variable. For this example, the values of the variable **performance** range from 0 to 14. Therefore, it is unwise to predict the value of **oxygen_consumption** for a performance rating of 100. The reason is that the relationship between the predictor variable and the response variable can be different beyond the range of your data.



Confidence Intervals for a Line

To assess the level of precision around the mean estimates of **oxygen_consumption**, you can produce confidence intervals around the means.

- A 95% confidence interval for the mean says that you are 95% confident your interval contains the population mean of Y for a particular X.
- Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.



Prediction Intervals for a Single Observation

Suppose the mean **oxygen_consumption** at a fixed value of **performance** is not the focus. If you are interested in establishing an inference on a future single observation, you need a prediction interval.

- A 95% prediction interval is one that you are 95% confident will contain a new observation.
- Prediction intervals are wider than confidence intervals because single observations have more variability than sample means.

Producing Confidence and Prediction Intervals

m3demo16.sas

Example: Invoke PROC REG and produce confidence intervals for the mean and individual values of **performance**.

```
proc reg data=predoxy;
   model oxygen_consumption=performance / clm cli
                                       alpha=.05;
   id name performance;
   plot oxygen_consumption*performance / conf pred;
   symbol1 c=red v=dot;
   symbol2 c=red;
   symbol3 c=blue;
   symbol4 c=blue;
   symbol5 c=green;
   symbol6 c=green;
   title;
run;
quit;
```

Selected REG procedure statement:

PLOT prints scatter plots with y-variables on the vertical axis and x-variables on the horizontal axis.

Selected PROC REG statement option (not shown above):

LINEPRINTER creates plots requested as line printer plots. If you do **not** specify this option, requested plots are created on a high resolution graphics device. This option is required if plots are requested and you do not have SAS/GRAPH software.

Selected MODEL statement options:

CLM produces all P option output, plus standard errors of the predicted values, and upper and lower 95% confidence bounds for the mean at each value of the predictor variable.

CLI produces all P option output, plus standard errors of the predicted values, and upper and lower 95% prediction intervals at each value of the predictor variable.

ALPHA= sets the significance level used for the construction of confidence intervals.

Selected PLOT statement options:

CONF requests overlaid plots of confidence intervals.

PRED requests overlaid plots of predicted values.

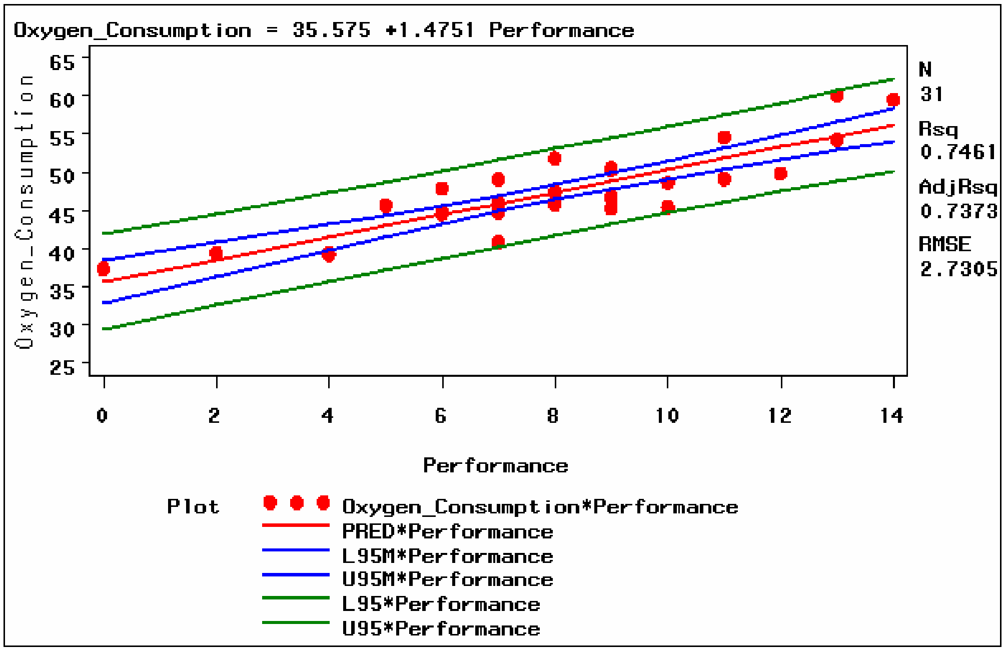
Partial PROC REG Output

```
                                              Std Error
      Obs   Name                Performance   Mean Predict        95% CL Mean

       32                                 0         1.3692    32.7750     38.3755
       33                                 3         0.9375    38.0831     41.9178
       34                                 6         0.5854    43.2285     45.6228
       35                                 9         0.5158    47.7960     49.9058
       36                                12         0.8056    51.6284     54.9238
```

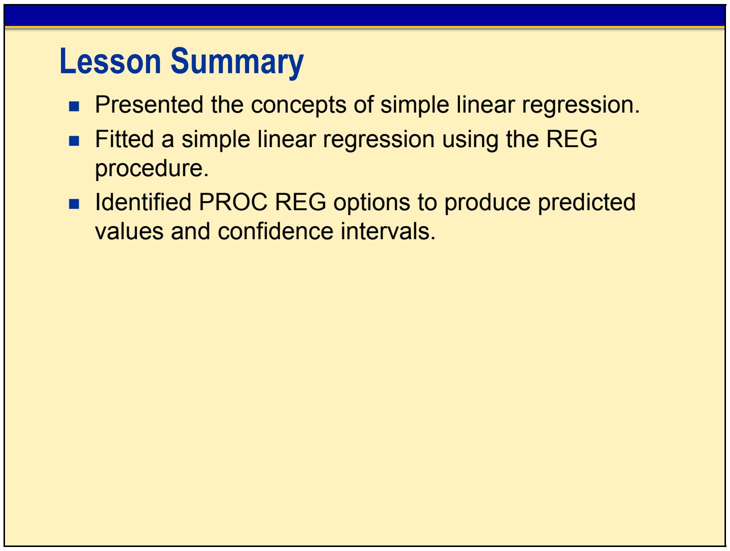
```
       Obs   Name                Performance      95% CL Predict       Residual

        32                                 0      29.3280     41.8225           .
        33                                 3      34.0960     45.9049           .
        34                                 6      38.7143     50.1370           .
        35                                 9      43.1676     54.5341           .
        36                                12      47.4536     59.0986           .
```

When **performance** is 6,

- the confidence interval for the mean of **oxygen_consumption** is (43.23, 45.62)
- the prediction interval for **oxygen_consumption** is (38.71, 50.14).



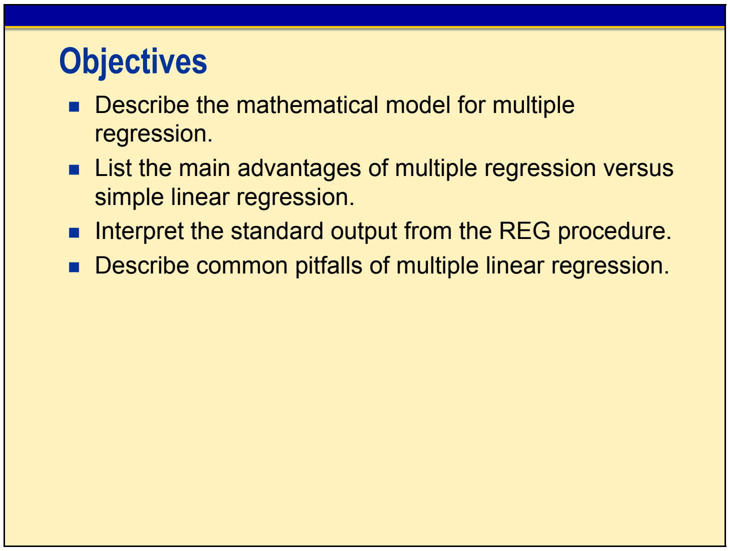
The data, regression line, confidence intervals, and predictions intervals are plotted in the graph above.



Lesson Summary

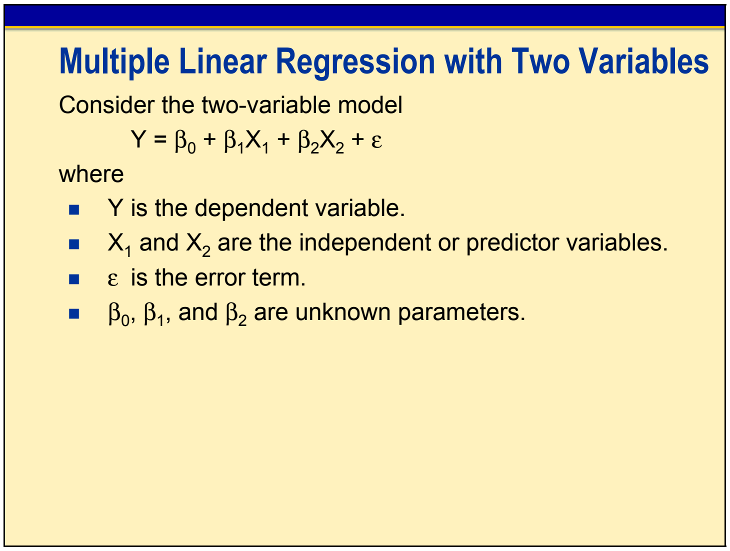
- Presented the concepts of simple linear regression.
- Fitted a simple linear regression using the REG procedure.
- Identified PROC REG options to produce predicted values and confidence intervals.

3.5 Concepts of Multiple Regression



Objectives

- Describe the mathematical model for multiple regression.
- List the main advantages of multiple regression versus simple linear regression.
- Interpret the standard output from the REG procedure.
- Describe common pitfalls of multiple linear regression.



Multiple Linear Regression with Two Variables

Consider the two-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

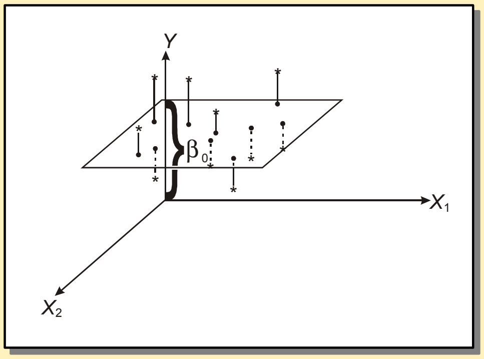
where

- Y is the dependent variable.
- X_1 and X_2 are the independent or predictor variables.
- ε is the error term.
- β_0, β_1, and β_2 are unknown parameters.

In simple linear regression, you can model the relationship between the two variables (two dimensions) with a line (one dimension).

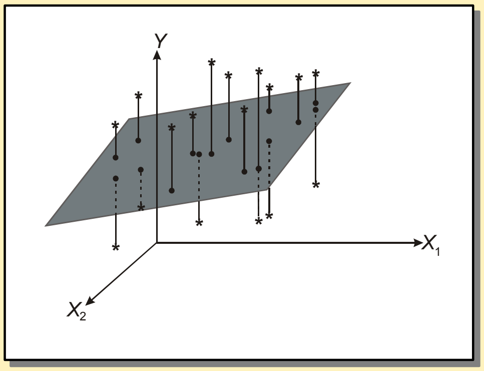
For the two-variable model, you can model the relationship of three variables (three dimensions) with a plane (two dimensions).

Picturing the Model: No Relationship



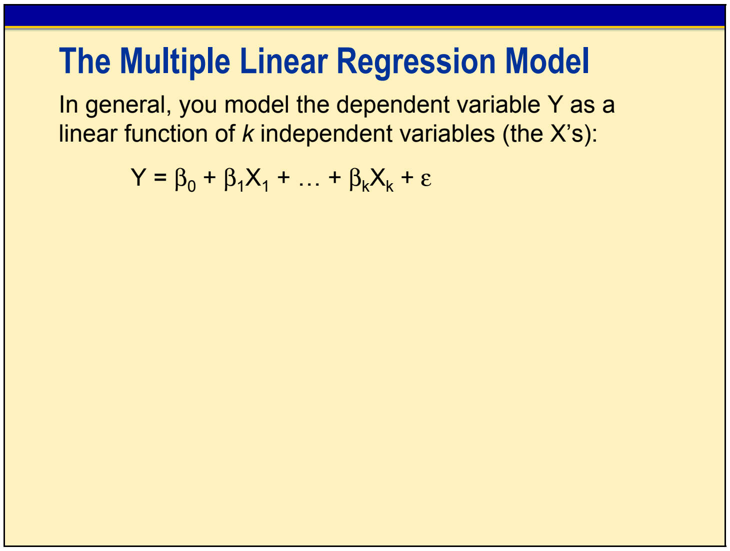
If there is no relationship among Y and X_1 and X_2, the model is a horizontal plane passing through the point ($Y = \beta_{0,}$ $X_1 = 0$, $X_2 = 0$).

Picturing the Model: A Relationship



If there is a relationship among Y and X_1 and X_2, the model is a sloping plane passing through three points:

- ($Y = \beta_0$, $X_1 = 0$, $X_2 = 0$)
- ($Y = \beta_0 + \beta_1$, $X_1 = 1$, $X_2 = 0$)
- ($Y = \beta_0 + \beta_2$, $X_1 = 0$, $X_2 = 1$).



The Multiple Linear Regression Model

In general, you model the dependent variable Y as a linear function of *k* independent variables (the X's):

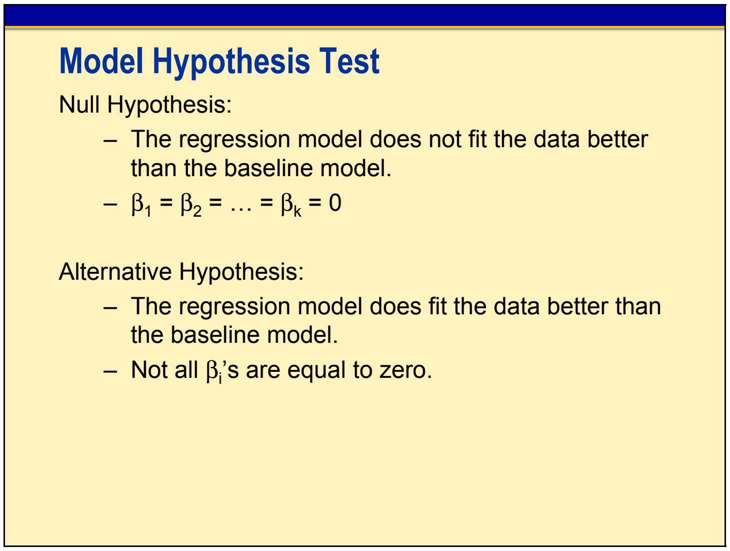
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

You investigate the relationship of $k + 1$ variables ($k + 1$ dimensions) with a k-dimensional surface.

The multiple general linear model is not restricted to modeling only planes. By using higher order terms, such as quadratic or cubic powers of the X's or cross products of one X with another, more complex surfaces than planes can be modeled.

In the examples, the models are limited to relatively simple surfaces, such as planes.

The model has $p = k + 1$ parameters (the β's) because of the intercept, $\beta_{0.}$



Model Hypothesis Test

Null Hypothesis:

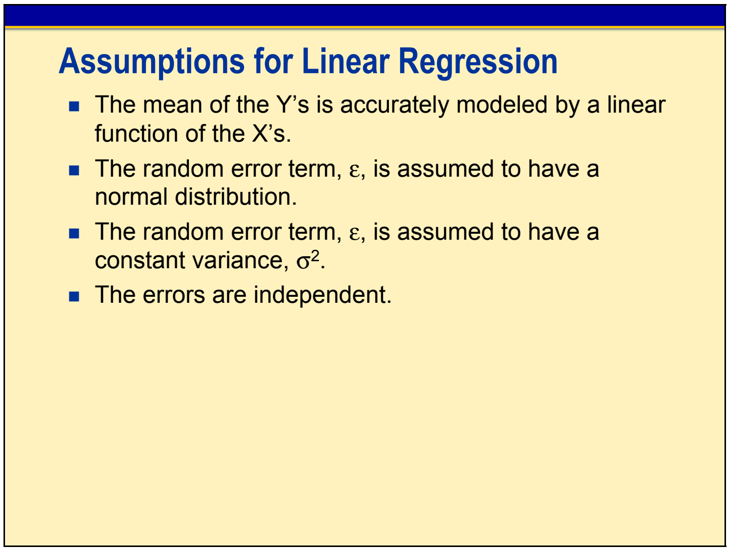
- The regression model does not fit the data better than the baseline model.
- $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

Alternative Hypothesis:

- The regression model does fit the data better than the baseline model.
- Not all β_i's are equal to zero.

If the estimated linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you do **not** have enough evidence to say that all of the slopes of the regression in the population are **not** 0 and that the predictor variables explain a significant amount of variability in the response variable.

If the estimated linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that at least one slope of the regression in the population is **not** 0 and that at least one predictor variable explains a significant amount of variability in the response variable.



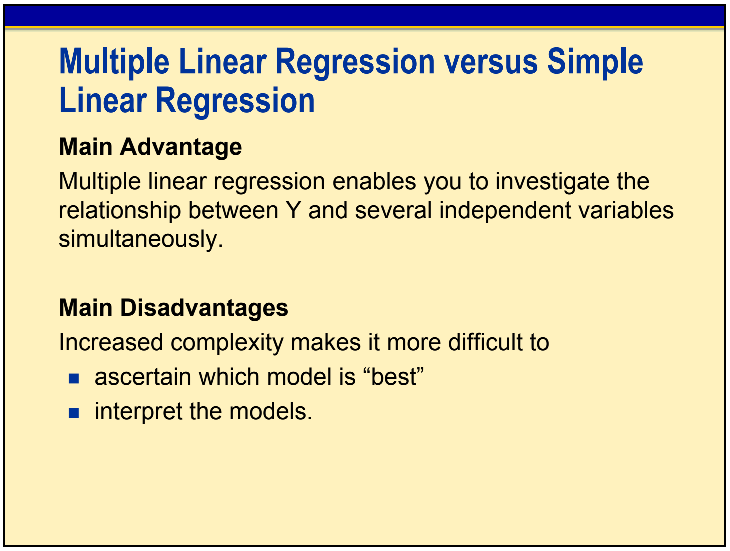
Assumptions for Linear Regression

- The mean of the Y's is accurately modeled by a linear function of the X's.
- The random error term, ε, is assumed to have a normal distribution.
- The random error term, ε, is assumed to have a constant variance, σ^2.
- The errors are independent.

Techniques to evaluate the validity of these assumptions are discussed in Module 4.

Because of the central limit theorem, the assumption that the errors are normally distributed is not as restrictive as you may think.

You also estimate σ^2 from the data.



Multiple Linear Regression versus Simple Linear Regression

Main Advantage

Multiple linear regression enables you to investigate the relationship between Y and several independent variables simultaneously.

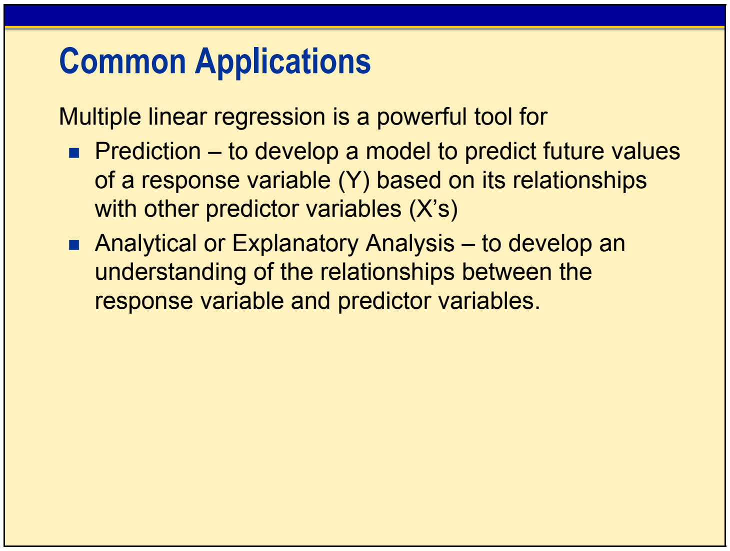
Main Disadvantages

Increased complexity makes it more difficult to

- ascertain which model is "best"
- interpret the models.

The advantages far outweigh the disadvantages. In practice, many responses depend on multiple factors that could interact in some way.

SAS tools help you decide upon a "best" model, a choice that can depend upon the purposes of the analysis as well as subject matter expertise



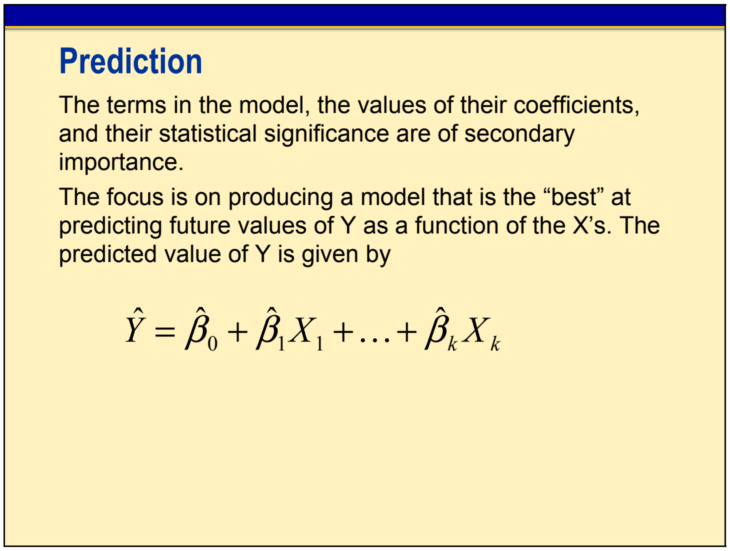
Common Applications

Multiple linear regression is a powerful tool for

- Prediction – to develop a model to predict future values of a response variable (Y) based on its relationships with other predictor variables (X's)
- Analytical or Explanatory Analysis – to develop an understanding of the relationships between the response variable and predictor variables.

The distinction between using multiple regression for an analytic analysis and prediction modeling is somewhat artificial. A model developed for an analytic study may be a good prediction model, and the reverse might also be true.

Myers (1986) actually refers to four applications of regression: prediction, variable screening, model specifications, and parameter estimation. The term *analytical analysis* is similar to Myers' parameter estimation application and variable screening.



Prediction

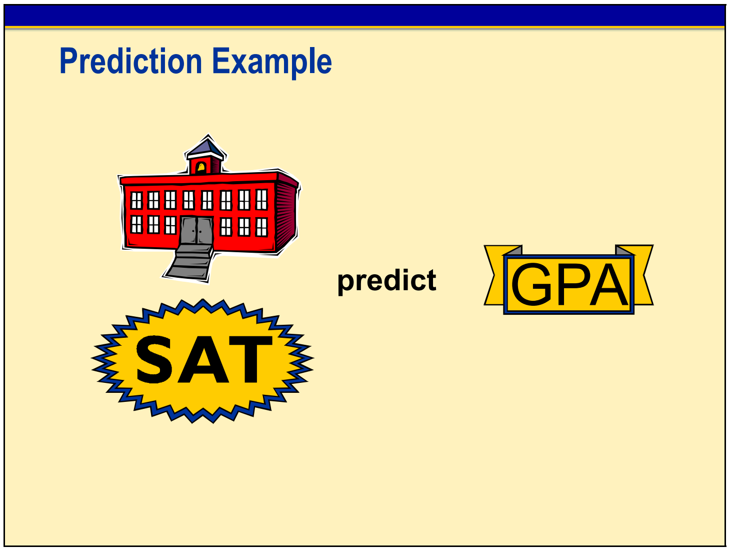
The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.

The focus is on producing a model that is the "best" at predicting future values of Y as a function of the X's. The predicted value of Y is given by

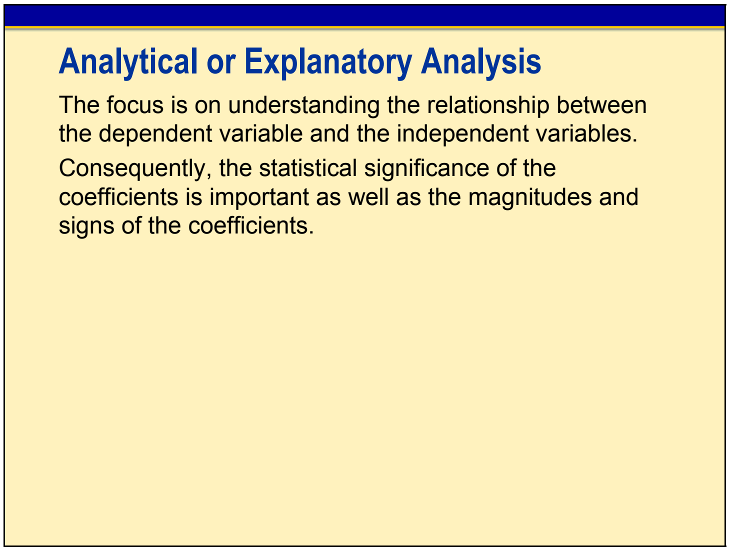
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

Most investigators do not ignore the terms in the model (the X's), the values of their coefficients (the β's), or their statistical significance (the *p*-values). They use these statistics to help choose among models with different numbers of terms and predictive capability.

R.H. Myers, *Classical and Modern Regression with Applications* (Boston: Duxbury Press, 1986).



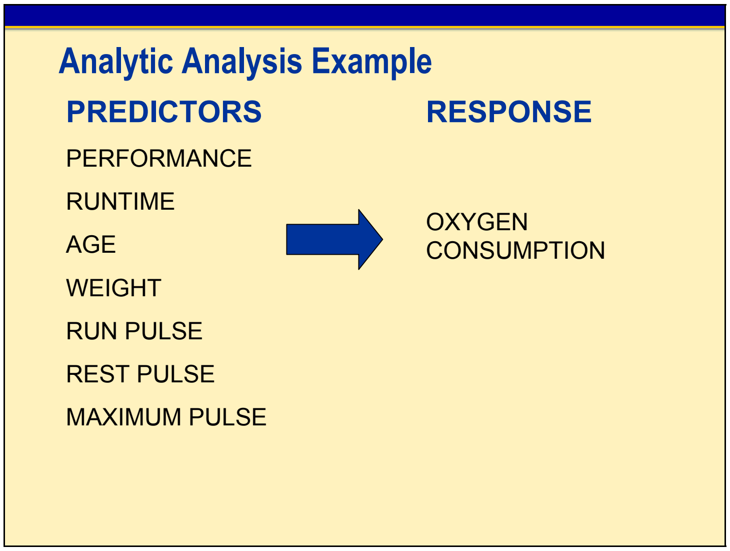
Prediction Example



Analytical or Explanatory Analysis

The focus is on understanding the relationship between the dependent variable and the independent variables.

Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.



Analytic Analysis Example

PREDICTORS

PERFORMANCE
RUNTIME
AGE
WEIGHT
RUN PULSE
REST PULSE
MAXIMUM PULSE

RESPONSE

OXYGEN
CONSUMPTION

An analyst knows from doing a simple linear regression that the measure of performance is an important variable in explaining the oxygen consumption capability of a club member. The analyst is interested in investigating other information to ascertain whether other variables are important in explaining the oxygen consumption capability.

Recall that you did a simple linear regression on **`oxygen_consumption`** with **`performance`** as the independent variable.

The R square for this model was 0.7461, which suggests that more of the variation in the oxygen consumption is still unexplained. Consequently, adding other variables to the model, such as **`runtime`** or **`age`**, could provide a significantly better model.

Fitting a Multiple Linear Regression Model

m3demo17.sas

Example: Invoke PROC REG and perform multiple linear regression analysis of **oxygen_consumption** on **performance** and **runtime**. Interpret the output for the two-variable model.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=performance runtime;
   title 'Multiple Linear Regression for b_fitness Data';
run;
quit;
```

The only required statement for PROC REG is the MODEL statement.

General form of the MODEL statement:

MODEL $Y = X_1\ X_2\ \ldots\ X_k$;

where

Y the dependent variable

$X_1\ X_2\ \ldots\ X_k$ a list of the independent variables that will be included in the model.

PROC REG Output

```
              Multiple Linear Regression for b_fitness Data

                            The REG Procedure
                              Model: MODEL1
                 Dependent Variable: Oxygen_Consumption

                          Analysis of Variance

                                  Sum of          Mean
Source                  DF       Squares        Square   F Value   Pr > F

Model                    2     637.96565     318.98283     41.82   <.0001
Error                   28     213.58890       7.62818
Corrected Total         30     851.55455


          Root MSE              2.76192    R-Square     0.7492
          Dependent Mean       47.37581    Adj R-Sq     0.7313
          Coeff Var             5.82980


                          Parameter Estimates

                            Parameter      Standard
Variable             DF      Estimate         Error   t Value   Pr > |t|

Intercept             1      55.37940      33.79380      1.64     0.1125
Performance           1       0.85780       1.06475      0.81     0.4272
Runtime               1      -1.40429       2.39427     -0.59     0.5622
```

Examine the sections of the output separately.

```
                           Analysis of Variance

                                   Sum of          Mean
Source                   DF       Squares        Square   F Value   Pr > F

Model                     2     637.96565     318.98283     41.82   <.0001
Error                    28     213.58890       7.62818
Corrected Total          30     851.55455
```

Model DF is 2, the number of parameters minus 1.

Error DF is 28, the total numbers of observations (31) minus the number of parameters in the model (3).

Corrected Total DF is 30, the number of observations minus 1.

Model Sum of Squares is the total variation in the Y explained by the model.

Error Sum of Squares is the variation in the Y **not** explained by the model.

Corrected Total Sum of Squares is the total variation in the Y.

Model Mean Square is the Model Sum of Squares divided by the Model DF.

Mean Square Error is the Error Sum of Squares divided by the Error DF and is an estimate of σ^2, the variance of the random error term.

F Value is the (Mean Square Model)/(Mean Square Error).

Pr > F is small; therefore, you reject H_0: $\beta_1 = \beta_2 = 0$ and conclude that at least one $\beta_i \neq 0$.

```
            Root MSE             2.76192    R-Square     0.7492
            Dependent Mean      47.37581    Adj R-Sq     0.7313
            Coeff Var            5.82980
```

The R square for this model, 0.7492, is only slightly larger than the R square for the model in which **performance** is the only predictor variable, 0.7461.

The R square always increases as you include more terms in the model. However, choosing the “best” model is not as simple as just making the R square as large as possible.

The adjusted R square is a measure similar to R square, but it takes into account the number of terms in the model. The adjusted R^2 for this model is 0.7313, smaller than the adjusted R^2 of 0.7373 for the **performance** only model. This strongly suggests that the variable **runtime** does not explain the oxygen consumption capacity if you know **performance**.

```
                            Parameter Estimates

                                 Parameter      Standard
 Variable             DF          Estimate         Error   t Value   Pr > |t|

 Intercept             1          55.37940      33.79380      1.64     0.1125
 Performance           1           0.85780       1.06475      0.81     0.4272
 Runtime               1          -1.40429       2.39427     -0.59     0.5622
```

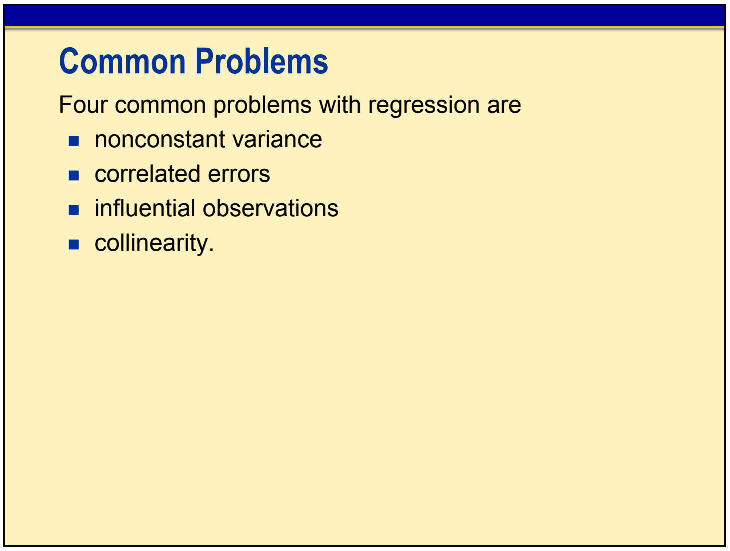
Using the estimates for β_0, β_1, and β_2 above, this model can be written as

oxygen_consumption = 55.3794 + 0.8578***performance** – 1.40429***runtime**

Both the *p*-values for **performance** and **runtime** are large, which suggests that neither slope is significantly different from 0. The reason is that the test for β_i=0 is conditioned on the other terms in the model. So the test for β_1=0 is conditional on or adjusted for X_2 (**runtime**). Similarly, the test for β_2=0 is conditional on X_1 (**performance**).

The variable **Performance** was significant when it was the only term in the model, but it is not significant when **runtime** is included. This implies that the variables are correlated with each other.

The significance level of the test does **not** depend on the order that you list the independent variables in the MODEL statement, but it does depend upon the variables included in the MODEL statement.



Common Problems

Four common problems with regression are

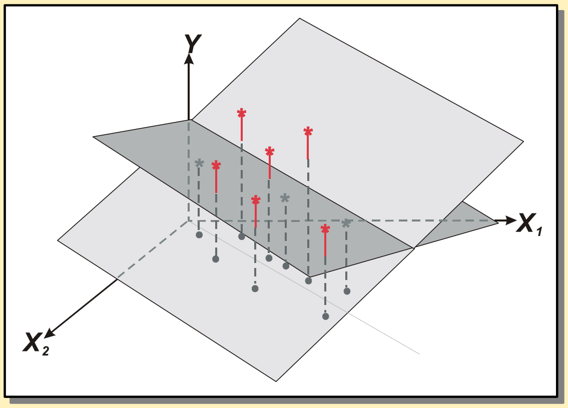
- nonconstant variance
- correlated errors
- influential observations
- collinearity.

The first three problems can arise in simple linear regression or multiple regression. The first two problems are always violations of the assumptions. The third can be a violation of the assumptions, but not always.

The fourth problem, however, is unique to multiple linear regression. *Collinearity* is redundant information among the independent variables. Collinearity is **not** a violation of assumptions of multiple regression.

When the number of potential X's is large, the likelihood of collinearity becoming a problem increases.

Generic Illustration of Collinearity



X_1 and X_2 almost follow a straight line $X_1 = X_2$ in the $(X_{1,}\ X_2)$ plane. Consequently, one variable provides nearly as much information as the other does. They are redundant.

Why is this a problem? Two reasons exist:

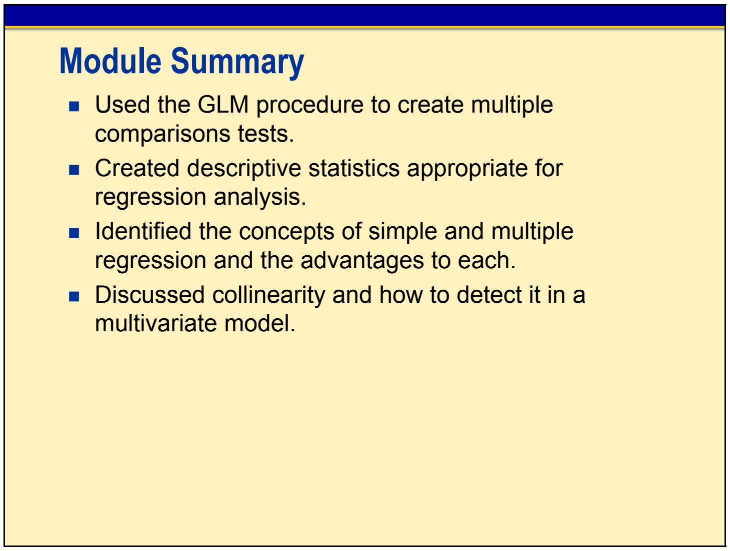
1. Neither variable can appear to be significant when both are in the model; however, both can be significant when only one is in the model. Thus, collinearity can hide significant variables.

2. Collinearity also increases the variance of the parameter estimates and consequently increases prediction error.

When collinearity is a problem, the estimates of the coefficients are unstable. This means they have a large variance. Consequently, the true relationship between Y and the X's can be quite different from that suggested by the magnitude and sign of the coefficients.

Lesson Summary

- Described the mathematical model for multiple regression.
- Listed the main advantages of multiple regression versus simple linear regression.
- Described issues of multiple linear regression.

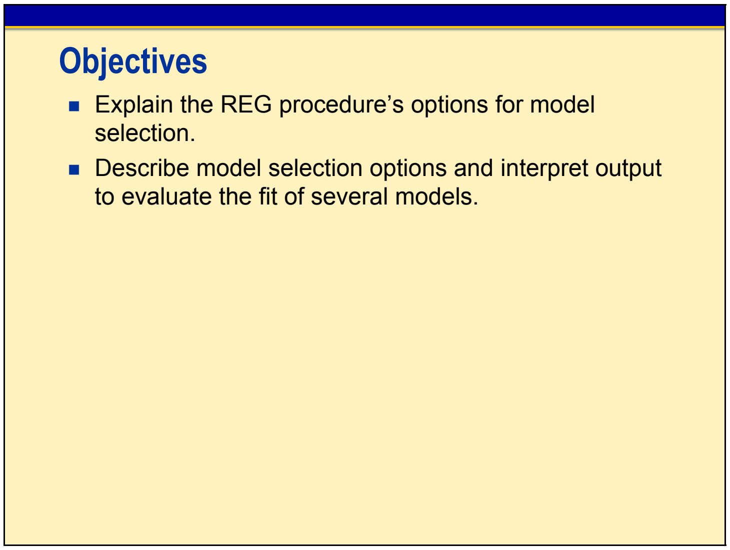


Module Summary

- Used the GLM procedure to create multiple comparisons tests.
- Created descriptive statistics appropriate for regression analysis.
- Identified the concepts of simple and multiple regression and the advantages to each.
- Discussed collinearity and how to detect it in a multivariate model.

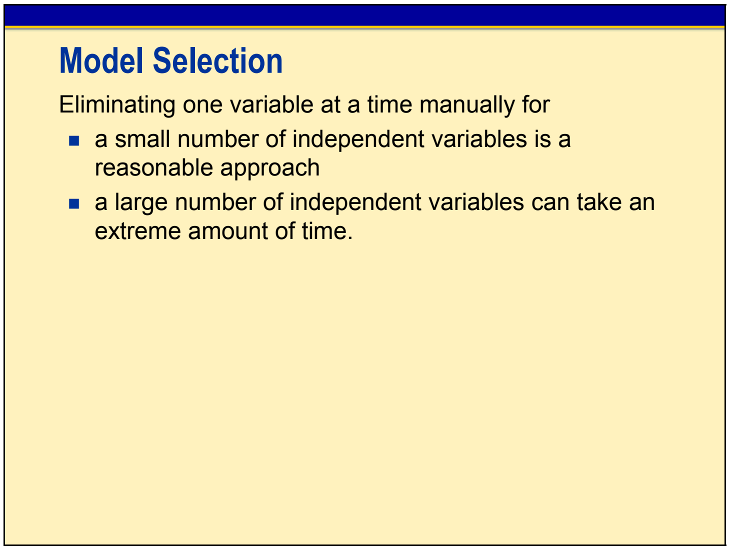
Module 4 Model Building and Assumption Verification

4.1 Model Building and Interpretation



Objectives

- Explain the REG procedure’s options for model selection.
- Describe model selection options and interpret output to evaluate the fit of several models.



Model Selection

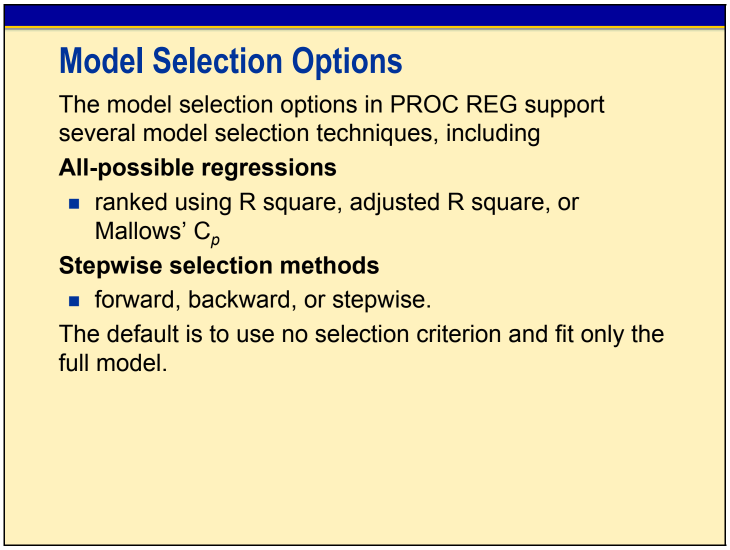
Eliminating one variable at a time manually for

- a small number of independent variables is a reasonable approach
- a large number of independent variables can take an extreme amount of time.

The exercises are designed to walk you through a model selection process. You start with all the variables in the **`b_fitness`** data set and eliminate the least significant terms.

For this small example, a model can be developed in a reasonable amount of time. However, if you start with a large model, eliminating one variable at a time can take an extreme amount of time.

You continue this process until only terms with a *p*-value less than a specified value, such as 0.10 or 0.05, remain.



Model Selection Options

The model selection options in PROC REG support several model selection techniques, including

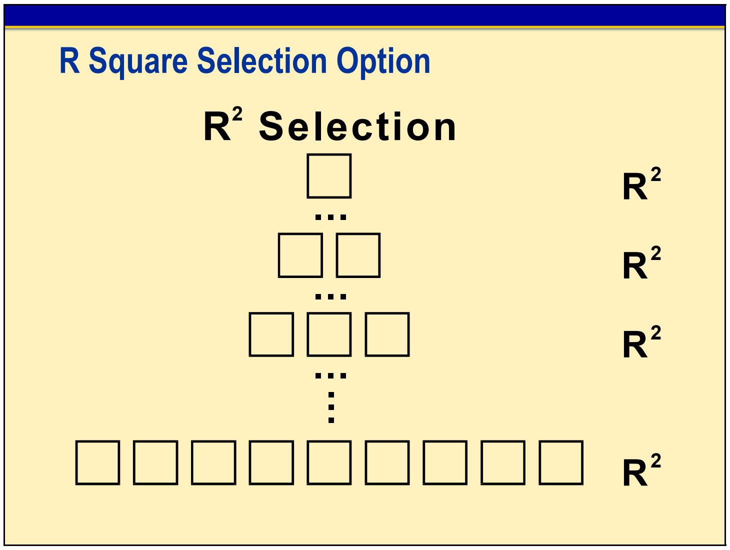
All-possible regressions

- ranked using R square, adjusted R square, or Mallows' C_p

Stepwise selection methods

- forward, backward, or stepwise.

The default is to use no selection criterion and fit only the full model.

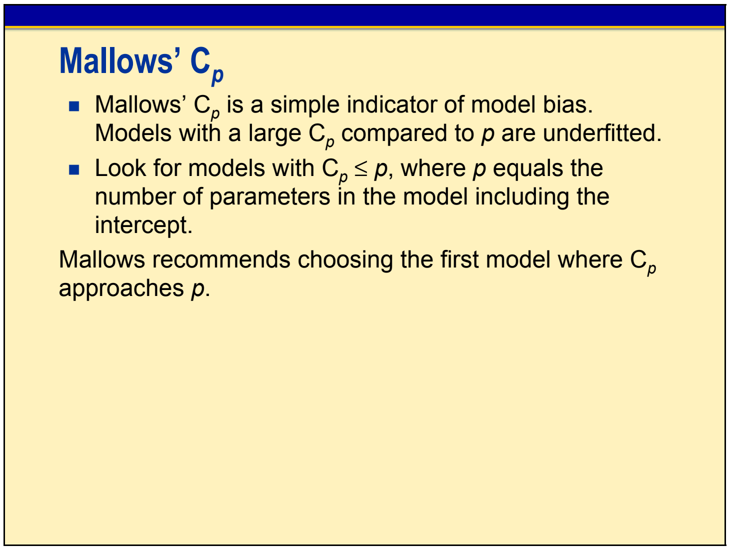


R Square Selection Option

In the **`b_fitness`** data set, there are 7 possible independent variables. Therefore, there are $2^7-1=127$ possible regression models. There are 7 possible one-variable models, 21 possible two-variable models, 35 possible three-variable models, and so on.

You will only look at the best four models as measured by the model R square for k=1, 2, 3, …, 7. This option only reduces the output. All regressions are still calculated.

If there were 20 possible independent variables, there would be over 1,000,000 models. In a later demonstration, you see another technique that does not have to examine all the models to help you choose a set of candidate models.



Mallows' C_p

- Mallows' C_p is a simple indicator of model bias. Models with a large C_p compared to *p* are underfitted.
- Look for models with $C_p \leq p$, where *p* equals the number of parameters in the model including the intercept.

Mallows recommends choosing the first model where C_p approaches *p*.

Mallows' C_p is estimated by

$$C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$$

where

MSE_p the mean square error for the model with *p* parameters

MSE_{full} the mean square error for the full model used to estimate the true residual variance

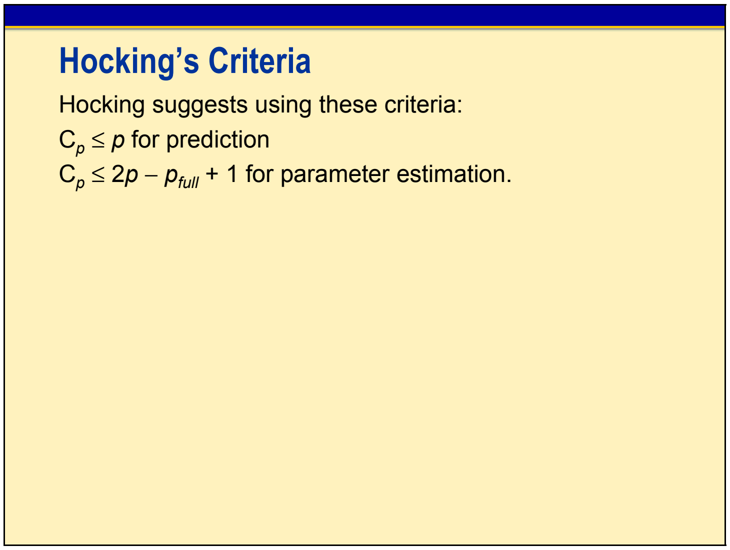
n the number of observations.

Bias in this context refers to the model overfitting the sample. In other words, variables are selected that appear to be important predictors in the sample but would not be important predictors in the population.

Notes about the Mallows' C_p selection method.

- Mallows' C_p consists of a variance component plus a bias component.
- If an important variable has been left out of the model, then the Mallows' statistic is less than *p*.
- If all the important variables are in the model, then the Mallows' statistic is approximately equal to *p*.
- For the full model, $C_p = p$.

C. L. Mallows, "Some Comments on C_p," *Technometrics* 15 (1973): 661-675.



Hocking's Criteria

Hocking suggests using these criteria:

$C_p \le p$ for prediction

$C_p \le 2p - p_{full} + 1$ for parameter estimation.

R. R. Hocking, "The Analysis and Selection of Variables in Linear Regression," *Biometrics* 32 (1976): 1-49.

Automatic Model Selection

m04demo01.sas

Example: Invoke PROC REG to produce a regression of **oxygen_consumption** on all the other variables in the **sasuser.b_fitness** data set.

```
proc reg data=sasuser.b_fitness;
   ALL_REG: model oxygen_consumption= performance runtime
                  age weight run_pulse rest_pulse
                  maximum_pulse / selection=rsquare adjrsq
                                  cp best=4;
   plot cp.*np. / vaxis=0 to 30 by 5
                  haxis=0 to 7 by 1
                  chocking=blue
                  cmallows=red;
   symbol v=plus h=1;
   title 'Best 4 Models Using All Regression Option';
run;
quit;
```

Selected MODEL statement options:

SELECTION= enables you to choose the different regression methods.

Selected SELECTION= option methods:

RSQUARE tells PROC REG to use the model R square to rank the model from best to worst for a given number of variables.

ADJRSQ prints the adjusted R square for each model.

CP prints Mallows' C_p statistic for each model.

BEST=*n* limits the output to only the best *n* models for a fixed number of variables.

The PLOT statement specifies that the values of the Mallows' C_p statistic (CP.) be plotted using the vertical axis and that the number of terms in the model (NP.) be plotted using the horizontal axis.

Selected PLOT statement options:

VAXIS= specifies the range for the vertical axis.

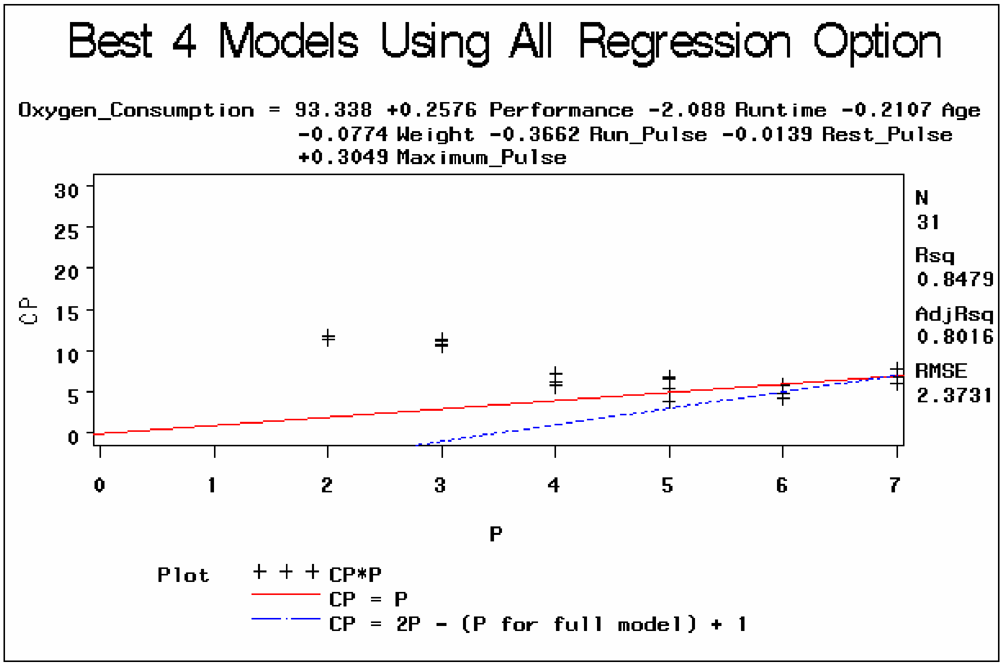
HAXIS= specifies the range for the horizontal axis. The default is the range of the data.

CHOCKING= requests a $2p - p_{full} + 1$ reference line in addition to the CMALLOWS reference line and specifies a color.

CMALLOWS= requests a $C_p = p$ reference line and specifies a color.

The models are ranked by their R-square values.

Partial PROC REG Output



Best 4 Models Using All Regression Option

```
Oxygen_Consumption = 93.338 +0.2576 Performance -2.088 Runtime -0.2107 Age
                     -0.0774 Weight -0.3662 Run_Pulse -0.0139 Rest_Pulse
                     +0.3049 Maximum_Pulse
```

The line $C_p = p$ is plotted to help you identify models that satisfy the criterion $C_p \le p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \le 2p - p_{full} + 1$ for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. The first model to fall below the line for Mallows' criterion has five parameters. The first model to fall below Hocking's criterion has six parameters.

PROC REG Output (continued)

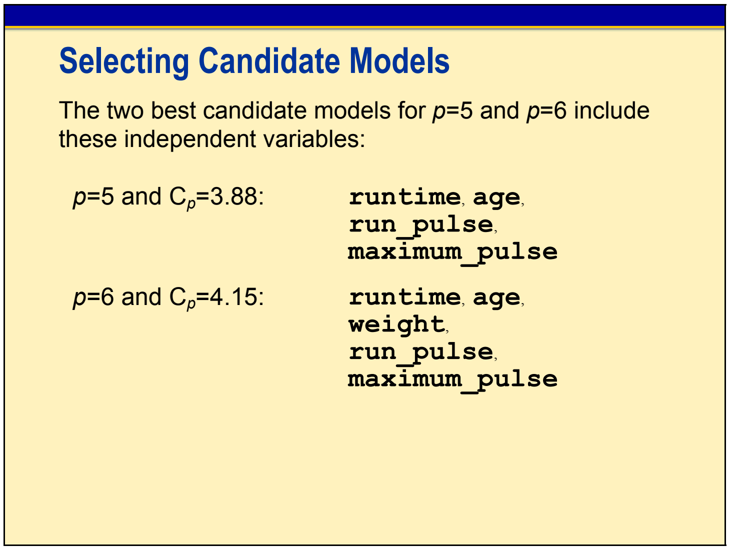
```
                  Best 4 Models Using All Regression Option

                              The REG Procedure
                               Model: ALL_REG
                   Dependent Variable: Oxygen_Consumption

                          R-Square Selection Method

Number in            Adjusted
  Model     R-Square  R-Square       C(p)  Variables in Model

       1      0.7461    0.7373    11.3942  Performance
       1      0.7434    0.7345    11.8074  Runtime
       1      0.1595    0.1305   100.1000  Rest_Pulse
       1      0.1585    0.1294   100.2529  Run_Pulse
---------------------------------------------------------------------------
```

PROC REG Output (continued)

```
       2      0.7647     0.7479    10.5794  Runtime Age
       2      0.7640     0.7472    10.6839  Performance Run_Pulse
       2      0.7614     0.7444    11.0743  Runtime Run_Pulse
       2      0.7597     0.7425    11.3400  Performance Age
---------------------------------------------------------------------------
       3      0.8101     0.7890     5.7169  Runtime Run_Pulse Maximum_Pulse
       3      0.8096     0.7884     5.7963  Runtime Age Run_Pulse
       3      0.8072     0.7858     6.1523  Performance Run_Pulse
                                            Maximum_Pulse
       3      0.8003     0.7781     7.2046  Performance Age Run_Pulse
---------------------------------------------------------------------------
       4      0.8355     0.8102     3.8790  Runtime Age Run_Pulse
                                            Maximum_Pulse
       4      0.8253     0.7984     5.4191  Performance Age Run_Pulse
                                            Maximum_Pulse
       4      0.8181     0.7901     6.5036  Performance Weight Run_Pulse
                                            Maximum_Pulse
       4      0.8160     0.7877     6.8265  Runtime Weight Run_Pulse
                                            Maximum_Pulse
---------------------------------------------------------------------------
       5      0.8469     0.8163     4.1469  Runtime Age Weight Run_Pulse
                                            Maximum_Pulse
       5      0.8421     0.8105     4.8787  Performance Age Weight Run_Pulse
                                            Maximum_Pulse
       5      0.8356     0.8027     5.8571  Runtime Age Run_Pulse Rest_Pulse
                                            Maximum_Pulse
       5      0.8355     0.8026     5.8738  Performance Runtime Age Run_Pulse
                                            Maximum_Pulse
---------------------------------------------------------------------------
       6      0.8476     0.8096     6.0381  Performance Runtime Age Weight
                                            Run_Pulse Maximum_Pulse
       6      0.8475     0.8094     6.0633  Runtime Age Weight Run_Pulse
                                            Rest_Pulse Maximum_Pulse
       6      0.8421     0.8026     6.8779  Performance Age Weight Run_Pulse
                                            Rest_Pulse Maximum_Pulse
       6      0.8356     0.7945     7.8565  Performance Runtime Age Run_Pulse
                                            Rest_Pulse Maximum_Pulse
---------------------------------------------------------------------------
       7      0.8479     0.8016     8.0000  Performance Runtime Age Weight
                                            Run_Pulse Rest_Pulse Maximum_Pulse
```

In this example, p_{full} equals 8—that is, 7 variables plus the intercept.

For p=5 (k=4), the "best" model has a C_p=3.879, satisfying Mallows' criterion. For p=6 (k=5), four models satisfy Mallows' criterion, but only two models also satisfy Hocking's criterion.



Selecting Candidate Models

The two best candidate models for *p*=5 and *p*=6 include these independent variables:

p=5 and C_p=3.88:
`runtime`, **`age`**,
`run_pulse`,
`maximum_pulse`

p=6 and C_p=4.15:
`runtime`, **`age`**,
`weight`,
`run_pulse`,
`maximum_pulse`

In practice, you might not want to limit your subsequent investigation to only the best model for a given number of terms. Some models may be essentially equivalent based on their R square or other measures.

A limitation of the evaluation you have done thus far is that you do not know the magnitudes or signs of the coefficients of the candidate models or their statistical significance.

Estimating and Testing the Coefficients for the Selected Models

m04demo02.sas

Example: Invoke PROC REG to compare the ANOVA tables and parameter estimates for the two candidate models in the **sasuser.b_fitness** data set.

```
proc reg data=sasuser.b_fitness;
   best4: model oxygen_consumption=runtime age run_pulse
                maximum_pulse;
   best5: model oxygen_consumption=runtime age weight
                run_pulse maximum_pulse;
   title 'Check "Best" Two Candidate Models';
run;
quit;
```

PROC REG can have more than one MODEL statement. You can assign a label to each MODEL statement to identify the output generated for each model.

Output for the BEST4 Model

```
                      Check "Best" Two Candidate Models

                              The REG Procedure
                                Model: BEST4
                   Dependent Variable: Oxygen_Consumption

                             Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     4       711.45087      177.86272      33.01    <.0001
Error                    26       140.10368        5.38860
Corrected Total          30       851.55455


          Root MSE              2.32134    R-Square     0.8355
          Dependent Mean       47.37581    Adj R-Sq     0.8102
          Coeff Var             4.89984


                            Parameter Estimates

                              Parameter      Standard
Variable             DF        Estimate         Error    t Value    Pr > |t|

Intercept             1        97.16952      11.65703       8.34      <.0001
Runtime               1        -2.77576       0.34159      -8.13      <.0001
Age                   1        -0.18903       0.09439      -2.00      0.0557
Run_Pulse             1        -0.34568       0.11820      -2.92      0.0071
Maximum_Pulse         1         0.27188       0.13438       2.02      0.0534
```

The R square and adjusted R square are the same as calculated during the model selection program. However, if there are missing values in the data set, this might not be true.

The model *F* is large and highly significant. The variables **`age`** and **`maximum_pulse`** are not significant at the 0.05 level of significance. However, all terms are significant at alpha=0.10.

The adjusted R square is close to the R square, which suggests that there are not too many variables in the model.

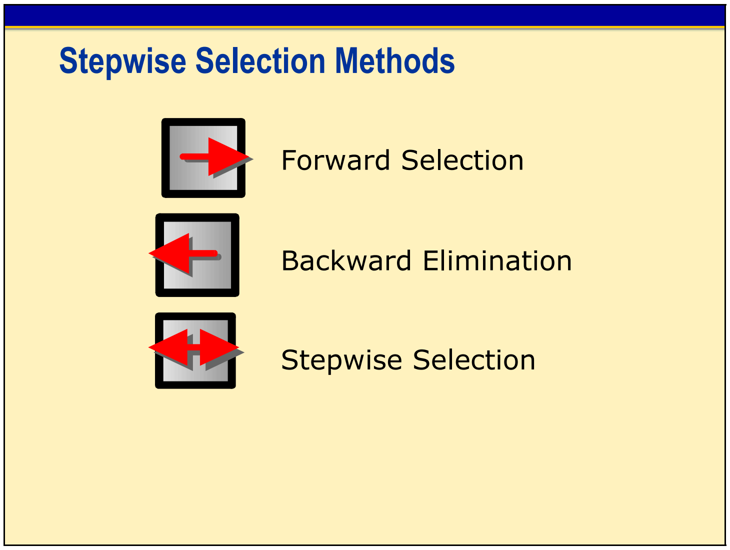
Output for the BEST5 Model

```
                      Check "Best" Two Candidate Models

                              The REG Procedure
                                Model: BEST5
                   Dependent Variable: Oxygen_Consumption

                            Analysis of Variance

                                    Sum of           Mean
Source                  DF         Squares         Square    F Value    Pr > F

Model                    5       721.20532      144.24106      27.66    <.0001
Error                   25       130.34923        5.21397
Corrected Total         30       851.55455


          Root MSE              2.28341    R-Square     0.8469
          Dependent Mean       47.37581    Adj R-Sq     0.8163
          Coeff Var             4.81978


                            Parameter Estimates

                              Parameter      Standard
Variable             DF        Estimate         Error    t Value    Pr > |t|

Intercept             1       101.33835      11.86474       8.54      <.0001
Runtime               1        -2.68846       0.34202      -7.86      <.0001
Age                   1        -0.21217       0.09437      -2.25      0.0336
Weight                1        -0.07332       0.05360      -1.37      0.1836
Run_Pulse             1        -0.37071       0.11770      -3.15      0.0042
Maximum_Pulse         1         0.30603       0.13452       2.28      0.0317
```

The adjusted R square is slightly larger than in the BEST4 model and very close to the R square.

The model *F* is large, but it is smaller than in the BEST4 model. However, it is still highly significant. All terms included in the model are significant except **`weight`**. Note that the *p*-values for **`age`**, **`run_pulse`**, and **`maximum_pulse`** are smaller in this model than they were in the BEST4 model.

Including the additional variable in the model changes the coefficients of the other terms and changes the *t* statistics for all.



Stepwise Selection Methods

Forward Selection

Backward Elimination

Stepwise Selection

The all-possible regressions technique that was discussed can be computer-intensive, especially if there are a large number of potential independent variables.

The REG procedure also offers these stepwise selection options:

FORWARD first selects the best one-variable model, based on the smallest *p*-value for all independent variables. Then it selects the next smallest *p*-value of the remaining variables, producing a two-variable model. FORWARD continues this process and stops when it reaches the point where no additional variables have a *p*-value < 0.50.

BACKWARD begins with the full model. Next, the variable that is least significant, given the other variables, is removed from the model based on the largest *p*-value for all independent variables. BACKWARD continues this process until all of the remaining variables have a *p*-value < 0.10.

STEPWISE starts like FORWARD but allows the possibility of a variable being removed once it is in the model. The default entry *p*-value is 0.15 and the default stay *p*-value is also 0.15.

The SLENTRY= and SLSTAY= options can be used to change the default values.

Stepwise Regression

m4demo03.sas

Example: Select a model for predicting **oxygen_consumption** in the **sasuser.b_fitness** data set by using the FORWARD stepwise selection method.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption= performance runtime age weight
         run_pulse rest_pulse maximum_pulse
         / selection=forward;
   title 'Stepwise Regression Using the FORWARD Option';
run;
quit;
```

```
                Stepwise Regression Using the FORWARD Option

                              The REG Procedure
                                Model: MODEL1
                   Dependent Variable: Oxygen_Consumption

                          Forward Selection: Step 1

    Variable Performance Entered: R-Square = 0.7461 and C(p) = 11.3942


                            Analysis of Variance

                                  Sum of          Mean
 Source                  DF       Squares        Square   F Value   Pr > F

 Model                    1     635.34150     635.34150     85.22   <.0001
 Error                   29     216.21305       7.45562
 Corrected Total         30     851.55455


                     Parameter     Standard
Variable              Estimate        Error   Type II SS  F Value  Pr > F

Intercept             35.57526      1.36917   5033.48080   675.13  <.0001
Performance            1.47507      0.15979    635.34150    85.22  <.0001

                     Bounds on condition number: 1, 1
---------------------------------------------------------------------------
```

PROC REG Output (continued)

```
                         Forward Selection: Step 2


     Variable Run_Pulse Entered: R-Square = 0.7640 and C(p) = 10.6839


                           Analysis of Variance

                                   Sum of           Mean
 Source                   DF      Squares         Square   F Value   Pr > F

 Model                     2    650.60420      325.30210     45.33   <.0001
 Error                    28    200.95035        7.17680
 Corrected Total          30    851.55455

               Stepwise Regression Using the FORWARD Option

                             The REG Procedure
                               Model: MODEL1
                  Dependent Variable: Oxygen_Consumption

                         Forward Selection: Step 2

                      Parameter     Standard
Variable               Estimate        Error   Type II SS  F Value  Pr > F

Intercept              48.60983      9.03851    207.58002    28.92  <.0001
Performance             1.39954      0.16511    515.66060    71.85  <.0001
Run_Pulse              -0.07327      0.05024     15.26270     2.13  0.1559

                Bounds on condition number: 1.1091, 4.4366
---------------------------------------------------------------------------
```

PROC REG Output (continued)

```
                         Forward Selection: Step 3


    Variable Maximum_Pulse Entered: R-Square = 0.8072 and C(p) = 6.1523


                           Analysis of Variance

                                   Sum of           Mean
 Source                   DF      Squares         Square   F Value   Pr > F

 Model                     3    687.38657      229.12886     37.68   <.0001
 Error                    27    164.16798        6.08030
 Corrected Total          30    851.55455


                      Parameter     Standard
Variable               Estimate        Error   Type II SS  F Value  Pr > F

Intercept              39.50427      9.10596    114.43553    18.82  0.0002
Performance             1.32166      0.15524    440.73994    72.49  <.0001
Run_Pulse              -0.35931      0.12515     50.11542     8.24  0.0079
Maximum_Pulse           0.33522      0.13629     36.78237     6.05  0.0206

                Bounds on condition number: 8.1227, 50.931
---------------------------------------------------------------------------

                         Forward Selection: Step 4


         Variable Age Entered: R-Square = 0.8253 and C(p) = 5.4191


                           Analysis of Variance

                                   Sum of           Mean
 Source                   DF      Squares         Square   F Value   Pr > F

 Model                     4    702.77828      175.69457     30.70   <.0001
 Error                    26    148.77627        5.72216
 Corrected Total          30    851.55455


                      Parameter     Standard
Variable               Estimate        Error   Type II SS  F Value  Pr > F

Intercept              55.88849     13.33542    100.50593    17.56  0.0003
Performance             1.23818      0.15897    347.15423    60.67  <.0001
Age                    -0.16144      0.09844     15.39171     2.69  0.1130
Run_Pulse              -0.33710      0.12216     43.56925     7.61  0.0105
Maximum_Pulse           0.26739      0.13854     21.31755     3.73  0.0646

                Bounds on condition number: 8.4502, 77.481
---------------------------------------------------------------------------
```

PROC REG Output (continued)

```
                          Forward Selection: Step 5


       Variable Weight Entered: R-Square = 0.8421 and C(p) = 4.8787


                            Analysis of Variance

                                    Sum of           Mean
 Source                  DF        Squares         Square    F Value    Pr > F

 Model                    5      717.08415      143.41683      26.66    <.0001
 Error                   25      134.47041        5.37882
 Corrected Total         30      851.55455


                      Parameter     Standard
Variable               Estimate        Error   Type II SS  F Value  Pr > F

Intercept              62.17928     13.49230    114.23682    21.24  0.0001
Performance             1.19926      0.15596    318.04934    59.13  <.0001
Age                    -0.18877      0.09690     20.41315     3.80  0.0627
Weight                 -0.08827      0.05412     14.30587     2.66  0.1155
Run_Pulse              -0.36603      0.11976     50.24137     9.34  0.0053
Maximum_Pulse           0.30806      0.13661     27.35207     5.09  0.0331

                Bounds on condition number: 8.7415, 105.27
---------------------------------------------------------------------------
```

PROC REG Output (continued)

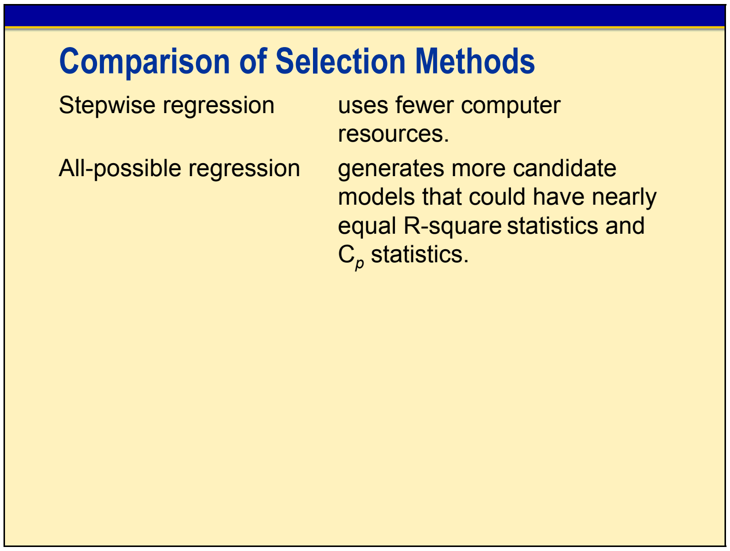
```
                          Forward Selection: Step 6


        Variable Runtime Entered: R-Square = 0.8476 and C(p) = 6.0381


                            Analysis of Variance

                                    Sum of          Mean
 Source                    DF       Squares        Square   F Value   Pr > F

 Model                      6     721.81791     120.30298     22.25   <.0001
 Error                     24     129.73665       5.40569
 Corrected Total           30     851.55455


                       Parameter     Standard
Variable                Estimate        Error   Type II SS  F Value  Pr > F

Intercept               90.83022     33.47159     39.80699     7.36  0.0121
Performance              0.32048      0.95201      0.61258     0.11  0.7393
Runtime                 -1.98433      2.12049      4.73376     0.88  0.3587
Age                     -0.20470      0.09862     23.28867     4.31  0.0488
Weight                  -0.07689      0.05560     10.33766     1.91  0.1794
Run_Pulse               -0.36818      0.12008     50.81482     9.40  0.0053
Maximum_Pulse            0.30593      0.13697     26.96687     4.99  0.0351

                Bounds on condition number: 48.957, 700.99
---------------------------------------------------------------------------


  No other variable met the 0.5000 significance level for entry into the
                                  model.


                        Summary of Forward Selection

      Variable            Number  Partial   Model
 Step Entered             Vars In R-Square R-Square  C(p)   F Value Pr > F

   1  Performance            1     0.7461   0.7461  11.3942   85.22 <.0001
   2  Run_Pulse              2     0.0179   0.7640  10.6839    2.13 0.1559
   3  Maximum_Pulse          3     0.0432   0.8072   6.1523    6.05 0.0206
   4  Age                    4     0.0181   0.8253   5.4191    2.69 0.1130
   5  Weight                 5     0.0168   0.8421   4.8787    2.66 0.1155
   6  Runtime                6     0.0056   0.8476   6.0381    0.88 0.3587
```

The model selected at each step is printed and a summary of the sequence of steps is given at the end of the output. In the summary, the variables are listed in the order in which they were selected. The partial R square shows the increase in the model R square as each term was added.

The model that STEPWISE selected has more variables than the models chosen using the all-regressions techniques.

In this example, no variables were deleted. Remember that this is not always the case.

Exercise: Refer to your course workbook.

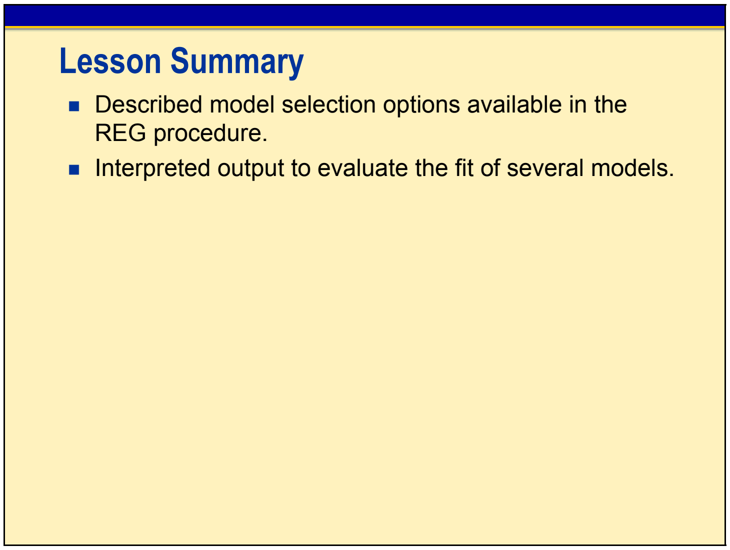


Comparison of Selection Methods

Stepwise regression	uses fewer computer resources.
All-possible regression	generates more candidate models that could have nearly equal R-square statistics and C_p statistics.

The stepwise regression methods have an advantage when there is a large number of independent variables.

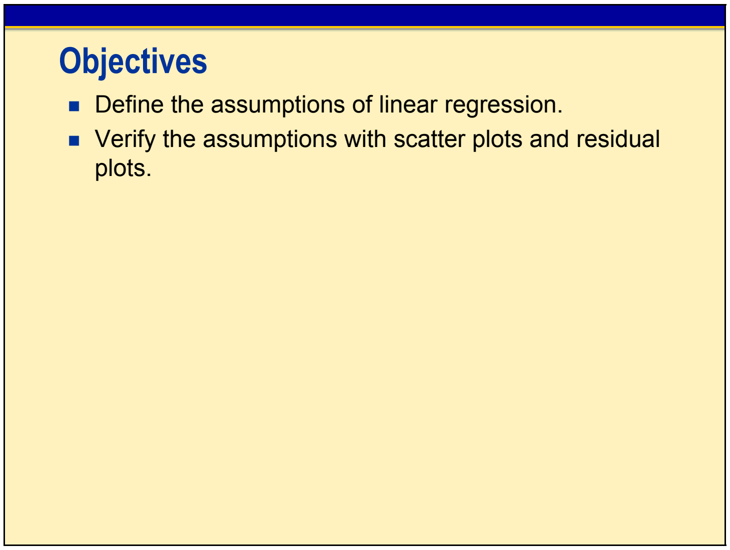
With the all-possible regressions techniques, you can compare essentially equivalent models and use your knowledge of the data set and subject area to select a model that is more easily interpreted. The adjusted R square is close to the R square, which suggests that there are not too many variables in the model.



Lesson Summary

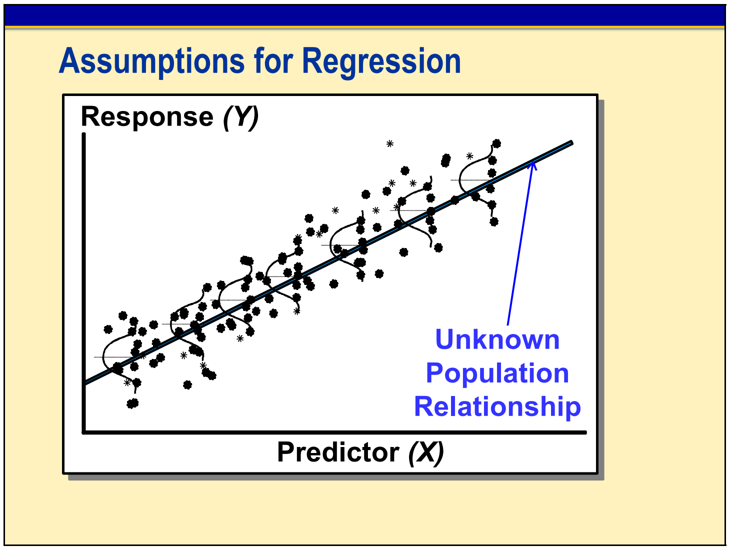
- Described model selection options available in the REG procedure.
- Interpreted output to evaluate the fit of several models.

4.2 Examining Residuals



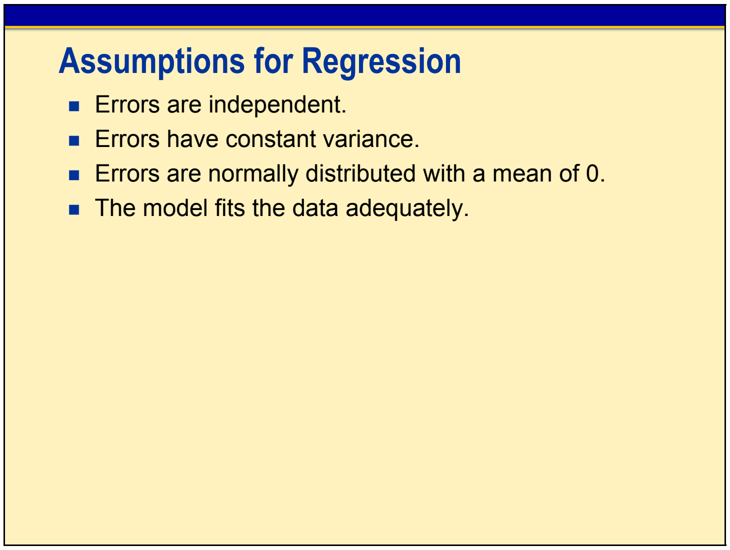
Objectives

- Define the assumptions of linear regression.
- Verify the assumptions with scatter plots and residual plots.



Assumptions for Regression

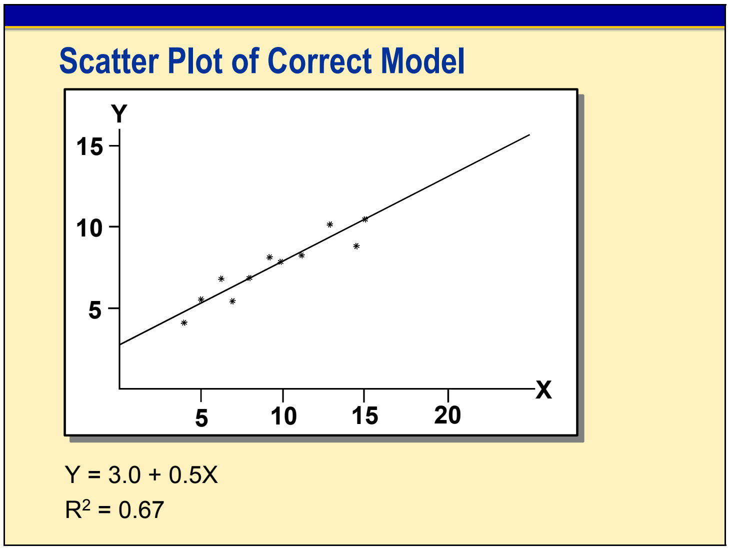
Recall that the model for linear regression has the form $Y=\beta_0 + \beta_1 X + \varepsilon$. When you perform a regression analysis, several assumptions about the error terms must be met to provide valid tests of hypothesis and confidence intervals.



Assumptions for Regression

- Errors are independent.
- Errors have constant variance.
- Errors are normally distributed with a mean of 0.
- The model fits the data adequately.

You can use scatter plots and residual plots to help verify these assumptions.



Scatter Plot of Correct Model

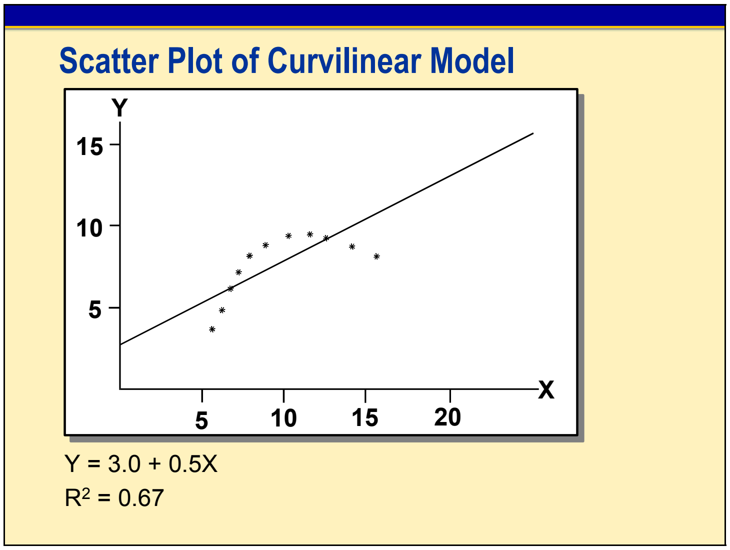
Y = 3.0 + 0.5X

$R^2 = 0.67$

To illustrate the importance of plotting data, four examples were developed by Anscombe. In each example, the scatter plot of the data values is different. However, the regression equation and the R-square statistic are the same.

In the first plot, a regression line adequately describes the data.

F. Anscombe, "Graphs in Statistical Analysis," *The American Statistician* 27 (1973): 17-21.

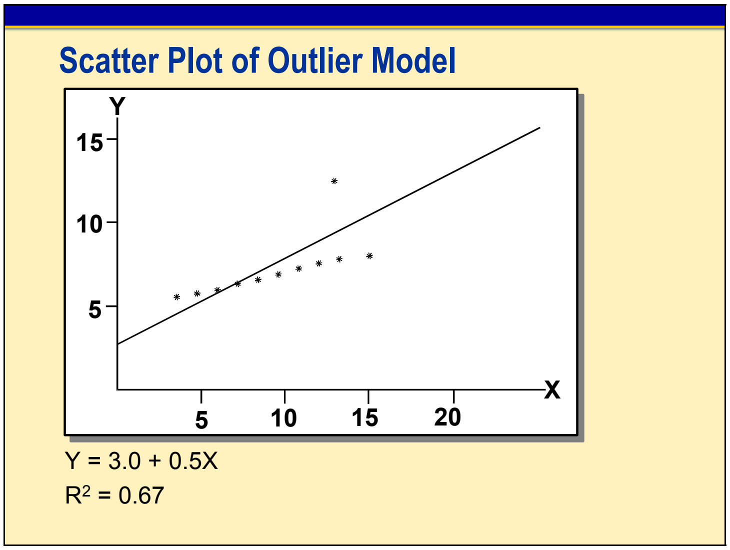


Scatter Plot of Curvilinear Model

$Y = 3.0 + 0.5X$

$R^2 = 0.67$

In the second plot, a simple linear regression model is not appropriate because you are fitting a straight line through a curvilinear relationship.

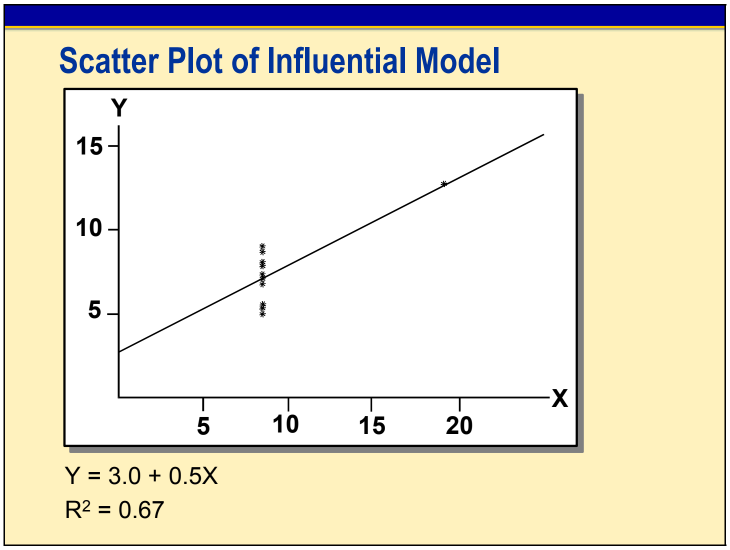


Scatter Plot of Outlier Model

$Y = 3.0 + 0.5X$

$R^2 = 0.67$

In the third plot, there seems to be an outlying data value that is affecting the regression line. This outlier is an influential data value in that it is substantially changing the fit of the regression line.



Scatter Plot of Influential Model

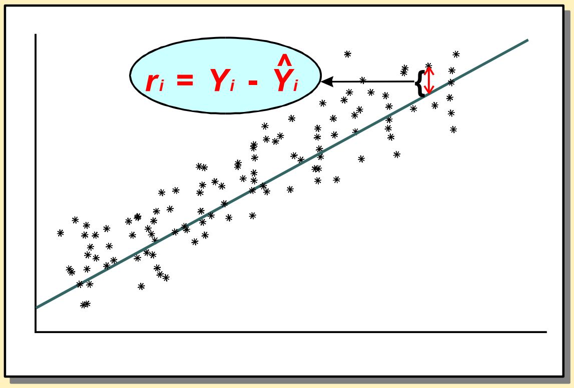
$Y = 3.0 + 0.5X$

$R^2 = 0.67$

In the fourth plot, the outlying data point dramatically changes the fit of the regression line. In fact, the slope would be undefined without the outlier.

The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the R-square statistics are the same even though the relationships between the two variables are different. Always produce a scatter plot before you conduct a regression analysis.

Verifying Assumptions



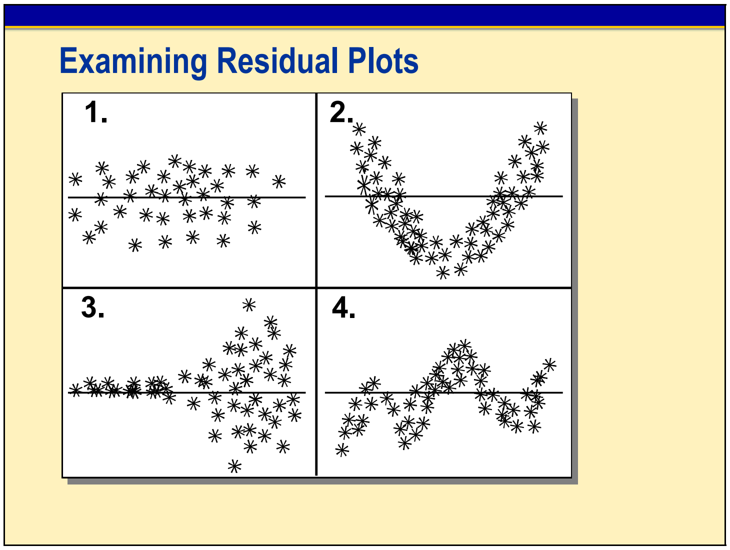
To verify the assumptions for regression, you can use the residual values from the regression analysis. Residuals are defined as

$$r_i = Y_i - \hat{Y}_i$$

where $\hat{Y}_i$ is the predicted value for the i^{th} value of the dependent variable.

You can examine two types of plots when verifying assumptions:

- the residuals versus the predicted values
- the residuals versus the values of the independent variable.



Examining Residual Plots

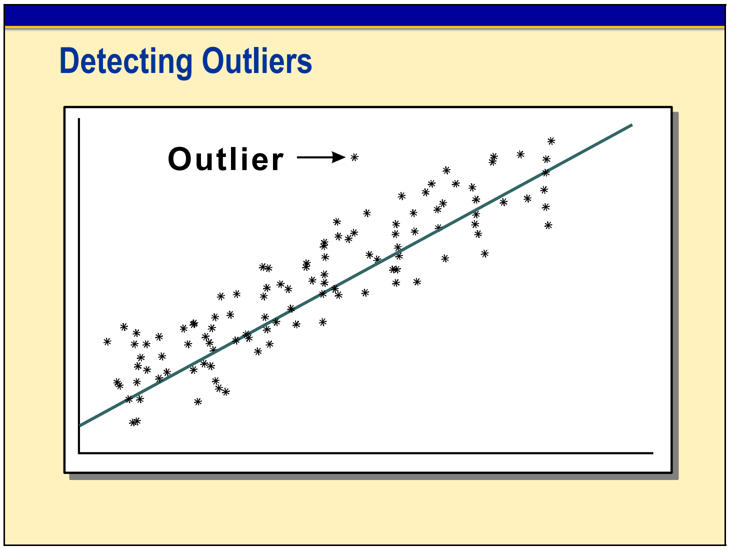
The graphs above are plots of residual values versus predicted values or predictor variable values for four models fit to different sets of data. If model assumptions are valid, the residual values should be randomly scattered about a reference line at 0. Any patterns or trends in the residuals can indicate problems in the model.

1. The model form appears to be adequate because the residuals are randomly scattered about a reference line at 0 and no patterns appear in the residual values.

2. The model form is incorrect. The plot indicates that the model should take into account curvature in the data. One possible solution is to add a quadratic term as one of the predictor variables.

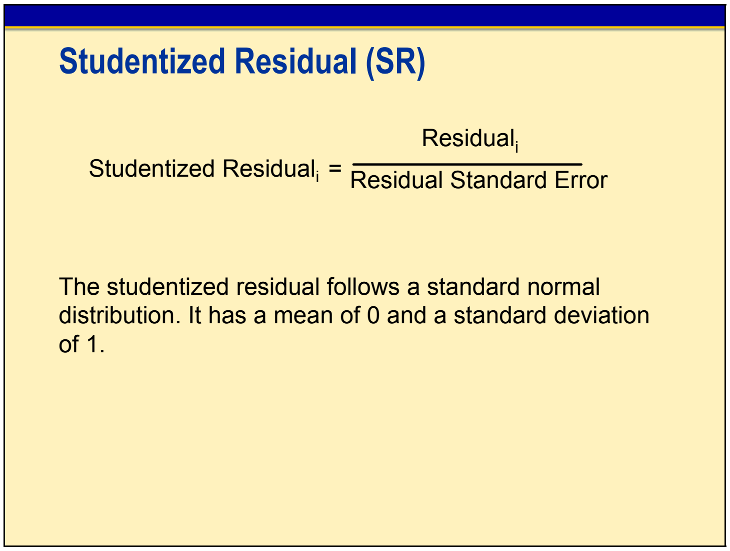
3. The variance is not constant; this is called *heteroskedasticity*. As you move from left to right, the variance increases. The variance also might decrease from left to right, or it might be smaller at the middle than at the ends (resulting in a bow-tie shape). One possible solution is to transform your dependent variable.

4. The observations are not independent. For this graph, the residuals tend to be followed by residuals with the same sign, which is called *autocorrelation*. This problem can occur when you have observations that have been collected over time. A possible solution is to use the AUTOREG procedure in SAS/ETS software.



Detecting Outliers

In addition to verifying assumptions, it is also important to check for outliers. Observations that are outliers are far away from the bulk of your data. These observations are often data errors or they reflect unusual circumstances. In either case, it is good statistical practice to detect these outliers and find out why they occurred.

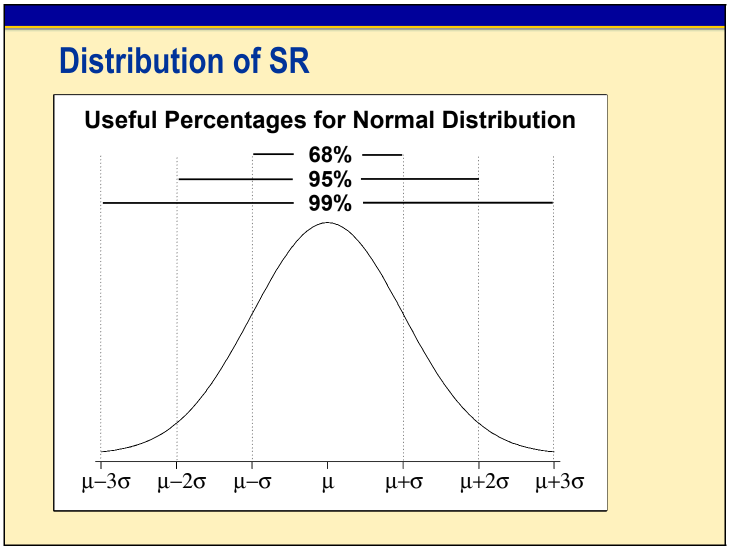


Studentized Residual (SR)

$$\text{Studentized Residual}_i = \frac{\text{Residual}_i}{\text{Residual Standard Error}}$$

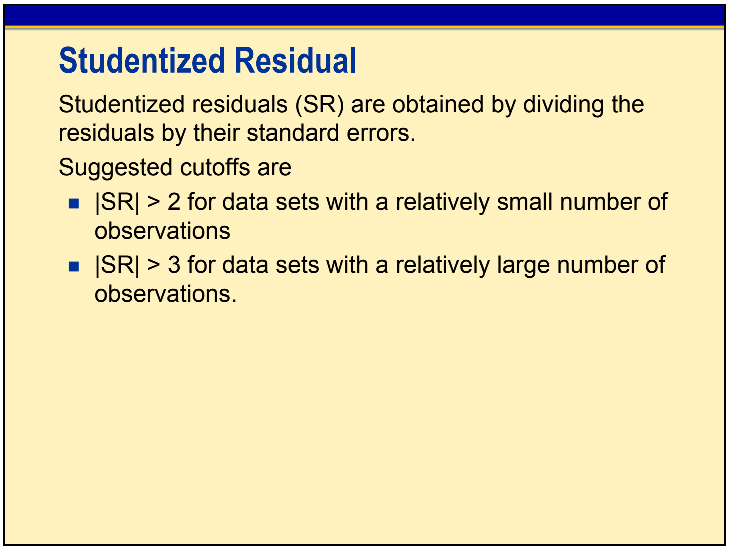
The studentized residual follows a standard normal distribution. It has a mean of 0 and a standard deviation of 1.

One way to check for outliers is to use the studentized residuals. These are calculated by dividing the residual values by their standard errors.



Distribution of SR

Useful Percentages for Normal Distribution



Studentized Residual

Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.

Suggested cutoffs are

- |SR| > 2 for data sets with a relatively small number of observations
- |SR| > 3 for data sets with a relatively large number of observations.

For a model that fits the data well and has no outliers, most of the studentized residuals should be close to 0. In general, studentized residuals that have an absolute value less than 2.0 could have easily occurred by chance. Studentized residuals that are between an absolute value of 2.0 to 3.0 occur infrequently and could be outliers. Studentized residuals that are larger than an absolute value of 3.0 occur rarely by chance alone and should be investigated.

Residual and Outlier Plots

m4demo04.sas

Example: Invoke the REG procedure and use a PLOT statement to produce high-resolution residual plots and diagnostic plots of the simple linear regression for the **sasuser.b_fitness** data set.

```
goptions reset=all;

proc reg data=sasuser.b_fitness;
   model oxygen_consumption=performance;
   plot r.*(p. performance);
   plot student.*obs. / vref=3 2 -2 -3
                        haxis=0 to 32 by 1;
   plot nqq.*student.;
   symbol v=dot;
   title 'Plots of Diagnostic Statistics';
run;
quit;
```

Selected REG procedure statement:

PLOT produces plots of variables from the input data set and statistics from the analysis. The statistics you plot can be any that are available in the OUTPUT data set. To plot a statistic from the analysis, follow the keyword with a period to indicate that it is not a variable from the input data set.

Selected PLOT statement options:

VREF specifies where reference lines perpendicular to the vertical axis are to appear.

VAXIS specifies range and tick marks for the vertical axis.

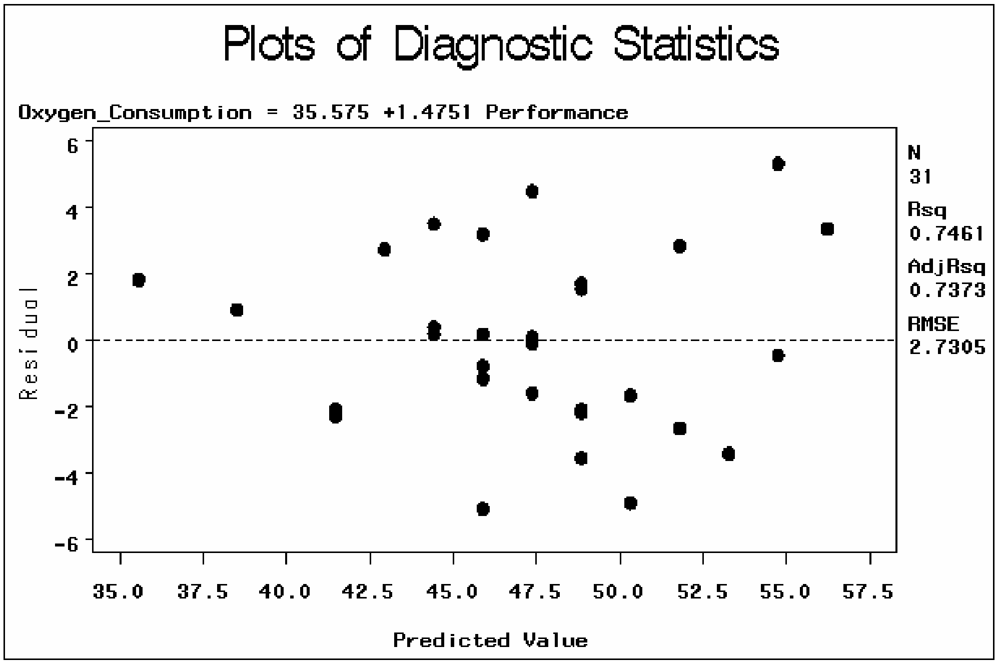
HAXIS specifies range and tick marks for the horizontal axis.

Selected keywords for the PLOT statement:

R.	residuals
P.	predicted values
STUDENT.	student residuals
NQQ.	normal quantile values
OBS.	observation number in the data set.

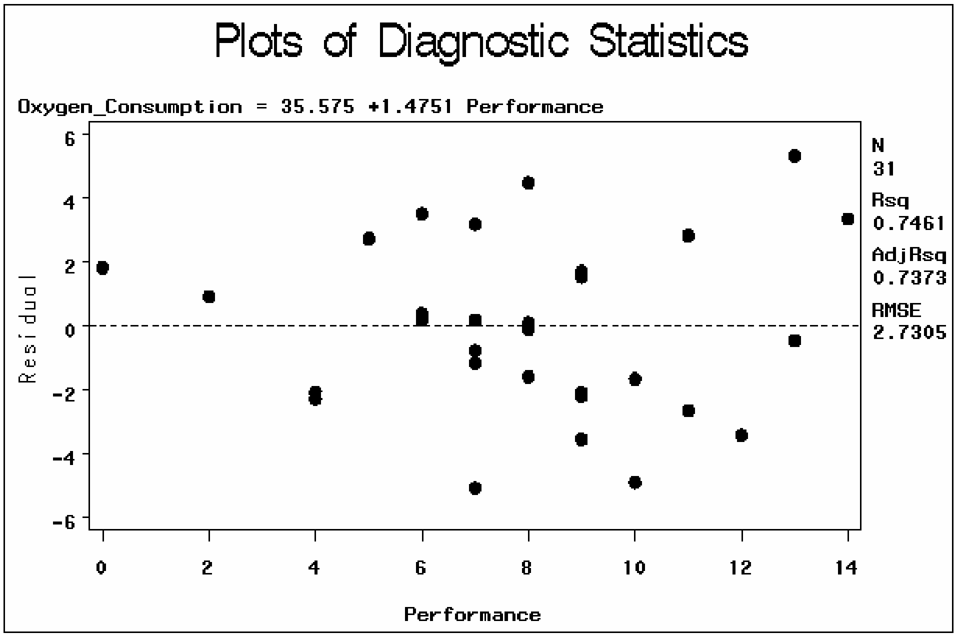
✎ The normal quantile-quantile plot helps to indicate whether the residuals are normally distributed. The assumption of normality should be verified, but it is not as important as the other regression assumptions.

The plot of the residuals by predicted values of **oxygen_consumption** is shown below. The residual values appear to be randomly scattered about the reference line at 0. There are no apparent trends or patterns in the residuals.

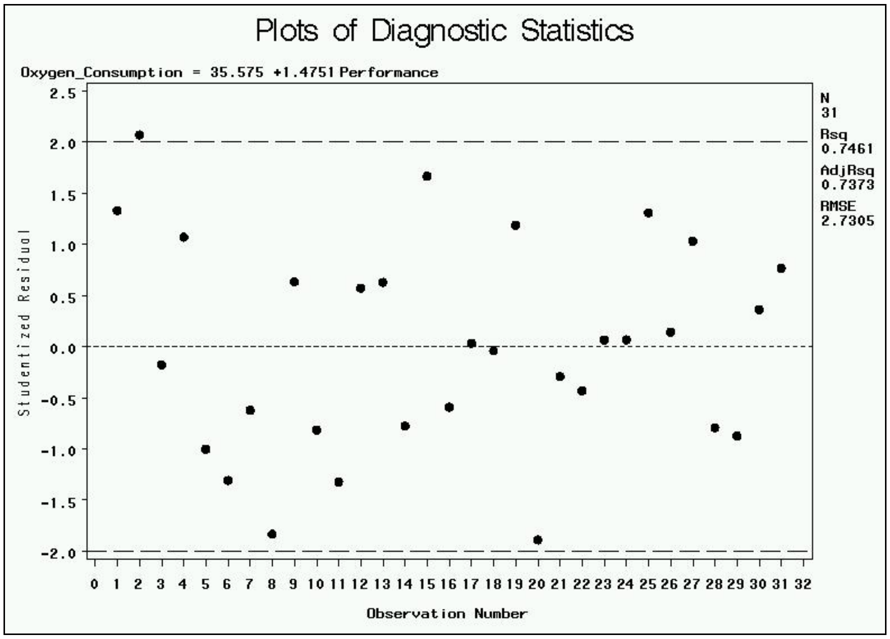


✐ The statistics printed on the side are the same as those found in the PROC REG output.

The plot of the residuals versus the values of the independent variable, **performance**, is shown below. There is also no apparent trend or pattern in the residuals.



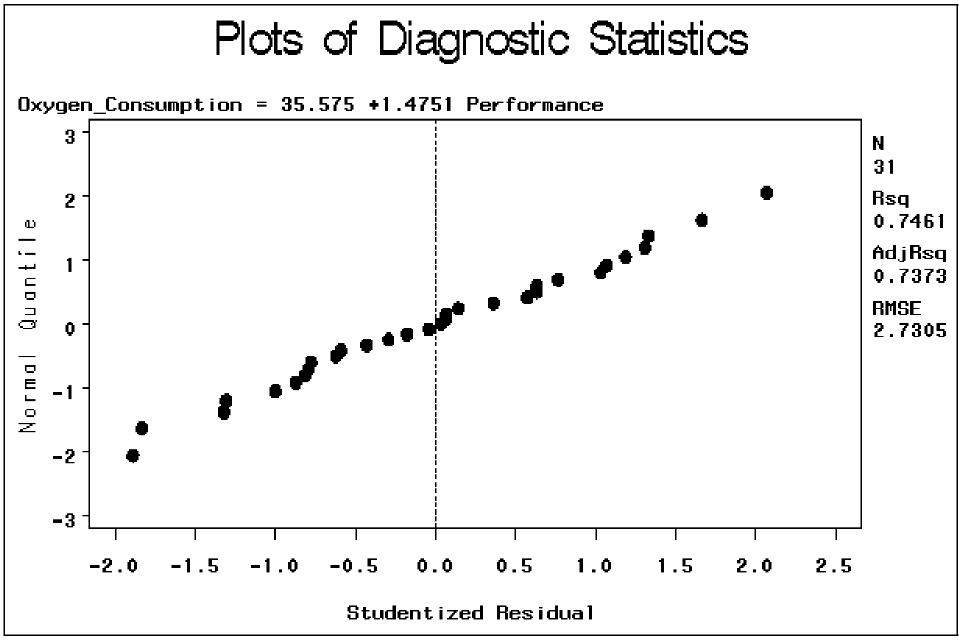
The plot of the student residuals by observation number is shown below. Reference lines are drawn on the student residual axis at 3.0, 2.0, –2.0, and –3.0. There do not appear to be any unusually large residuals.



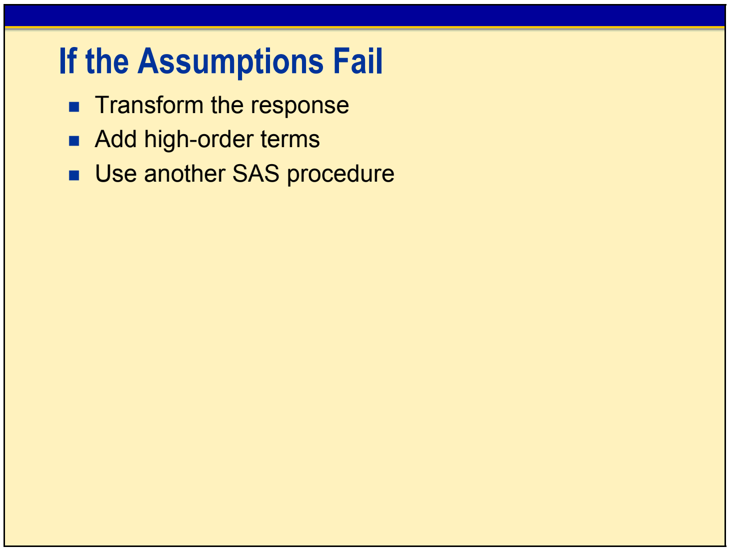
✎ You can also use the R option in the MODEL statement of PROC REG to obtain residual diagnostics. Output from the R option includes the values of the response variable, the predicted values of the response variable, the standard error of the predicted values, the residuals, the standard error of the residuals, the student residuals, and a plot of the student residuals in tabular rather than graphic form.

The plot of the normal quantiles versus the student residuals is shown below. The plot is obtained by plotting the student residuals to their expected quantiles if the residuals come from a normal distribution. If the residuals are normally distributed, the plot should appear to be a straight line with a slope of about 1. If the plot deviates substantially from the ideal, there is evidence against normality.

The plot below shows no deviation from the expected pattern. Thus, you can conclude that the residuals do not significantly violate the normality assumption. If the residuals did violate the normality assumption, a transformation of the response variable might be warranted.

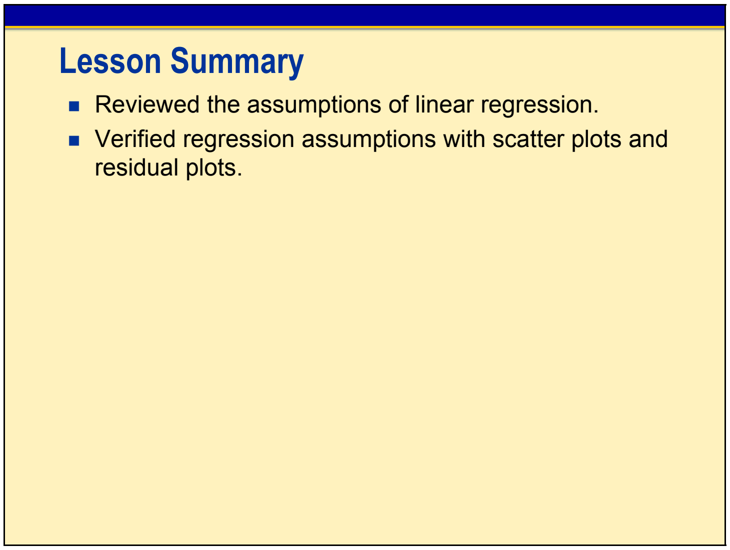


You can use the NORMAL option in the UNIVARIATE procedure to obtain normal statistics. This might be necessary if you feel the plot above shows a violation of the normality assumption. First, you must create an output data set with the residuals in PROC REG using an OUTPUT statement or the Output Delivery System. Then use that data set as the input data set in PROC UNIVARIATE.



If the Assumptions Fail

- Transform the response
- Add high-order terms
- Use another SAS procedure

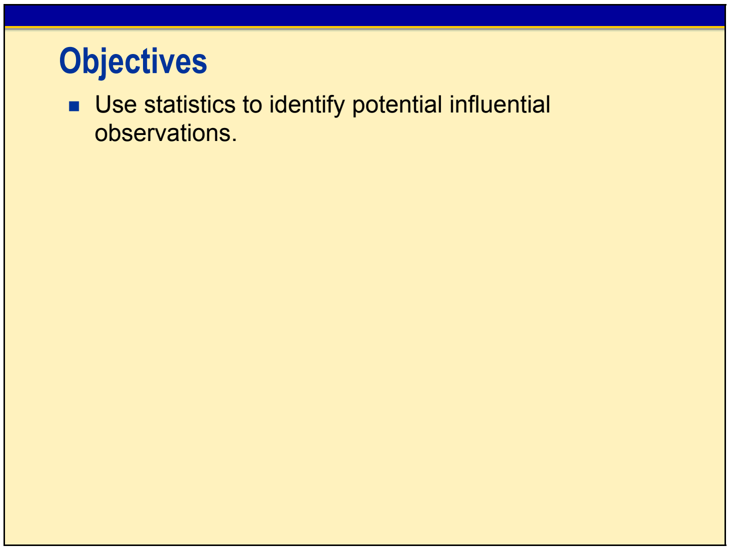


Lesson Summary

- Reviewed the assumptions of linear regression.
- Verified regression assumptions with scatter plots and residual plots.

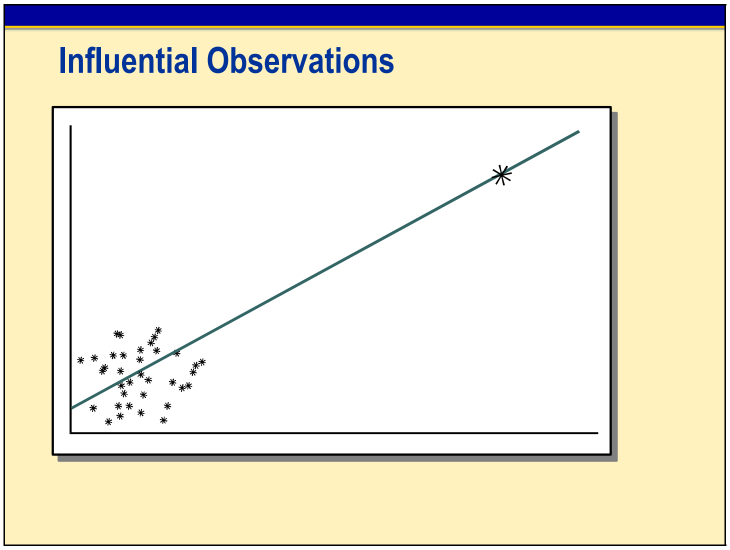
Exercise: Refer to your course workbook.

4.3 Influential Observations



Objectives

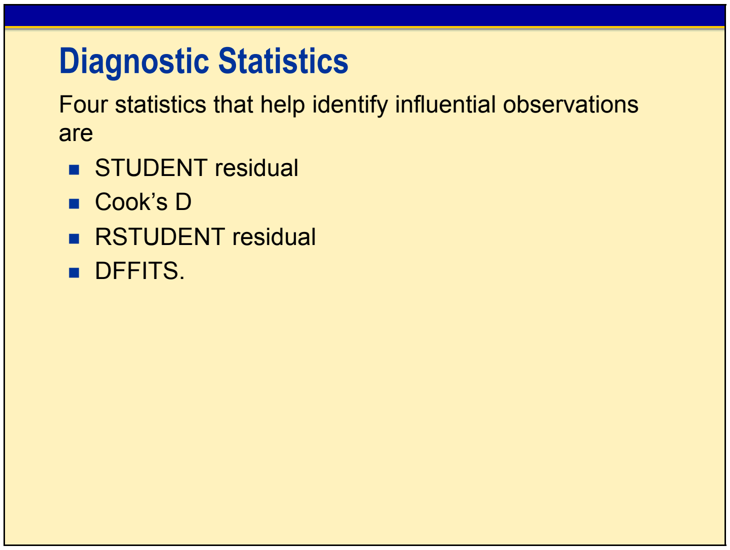
- Use statistics to identify potential influential observations.



Influential Observations

Recall in the previous lesson that you saw examples of data sets where the simple linear regression model fits were essentially the same. However, plotting the data revealed that the model fits were different. One of the examples showed a highly influential observation like the example above.

Identifying influential observations in a multiple linear regression is more complex because you have more predictors to consider. The REG procedure has options to calculate statistics that help identify influential observations.

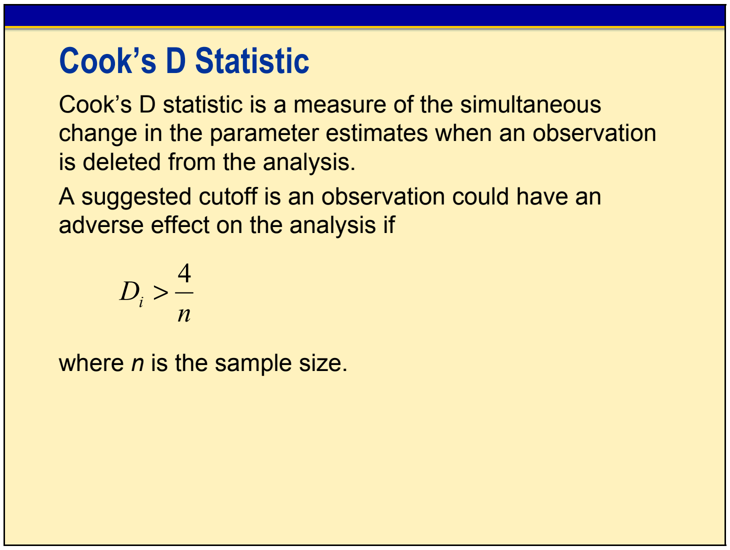


Diagnostic Statistics

Four statistics that help identify influential observations are

- STUDENT residual
- Cook's D
- RSTUDENT residual
- DFFITS.

You learned earlier how to save residual and predicted values into an output data set. You can use options to produce the Cook's D and DFFITS statistics shown above.



Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis.

A suggested cutoff is an observation could have an adverse effect on the analysis if

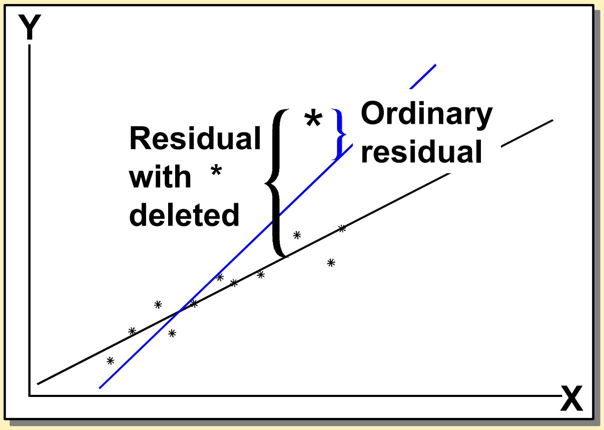
$$D_i > \frac{4}{n}$$

where *n* is the sample size.

To detect influential observations, you can use Cook's D statistic. This statistic measures the change in the parameter estimates that results from deleting each observation.

Identify observations above the cutoff and investigate the reasons they occurred.

RSTUDENT



Recall that STUDENT residuals are the ordinary residuals divided by their standard errors. The RSTUDENT residuals are similar to the STUDENT residuals except they are calculated after deleting the *i*th observation.

If the RSTUDENT residual is different from the STUDENT residual for a specific observation, that observation is likely to be influential.

DFFITS

The DFFITS(i) measures the impact that the *i*th observation has on the predicted value.

$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

$\hat{Y}_i$ is the *i*th predicted value.

$\hat{Y}_{(i)}$ is the *i*th predicted value when the *i*th observation is deleted.

$s(\hat{Y}_i)$ is the standard error of the *i*th predicted value.

$$|Diffits| > 2\sqrt{\frac{p}{n}}$$

The suggested cutoff is |DFFITS$_i$| > $2\sqrt{\frac{p}{n}}$.

See D. A. Belsey, E., Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (New York: Wiley, 1980).

Looking for Influential Observations

m04demo05.sas

Example: Generate the RSTUDENT and DFFITS influence statistics for the BEST4 variable model. Save the statistics to an output data set and create a data set with only observations that exceed the suggested cutoffs of the influence statistics.

```
proc reg data=sasuser.b_fitness;
   best4: model oxygen_consumption=runtime age run_pulse
                maximum_pulse / r influence;
   id name;
   output out=fitnessout rstudent=rstud dffits=dfits
          cookd=cooksd;
   title;
run;
quit;
```

Selected REG procedure statement:

OUTPUT requests that the statistics associated with the variables be saved in a WORK data set.

Selected MODEL statement option:

INFLUENCE requests that the diagnostics be printed.

Partial PROC REG Output

```
                              The REG Procedure
                                Model: BEST4
                   Dependent Variable: Oxygen_Consumption

                            Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     4      711.45087      177.86272      33.01    <.0001
Error                    26      140.10368        5.38860
Corrected Total          30      851.55455


          Root MSE              2.32134    R-Square     0.8355
          Dependent Mean       47.37581    Adj R-Sq     0.8102
          Coeff Var             4.89984


                            Parameter Estimates

                              Parameter      Standard
Variable              DF       Estimate         Error    t Value    Pr > |t|

Intercept              1       97.16952      11.65703       8.34      <.0001
Runtime                1       -2.77576       0.34159      -8.13      <.0001
Age                    1       -0.18903       0.09439      -2.00      0.0557
Run_Pulse              1       -0.34568       0.11820      -2.92      0.0071
Maximum_Pulse          1        0.27188       0.13438       2.02      0.0534
```

The ANOVA table and the Parameter Estimates table are identical to the previous example.

```
                              The REG Procedure
                                Model: BEST4
                   Dependent Variable: Oxygen_Consumption


                              Output Statistics

                                        Dep Var   Predicted      Std Error
      Obs Name             Oxygen_Consumption        Value   Mean Predict

        1 Donna                       59.5700     55.9333         0.9104
        2 Gracie                      60.0600     57.8362         1.6123
        3 Luanne                      54.3000     56.7812         1.0775
        4 Mimi                        54.6300     54.6309         1.0870
        5 Chris                       49.1600     51.1400         1.0944
```

Partial PROC REG Output (continued)

```
                              Output Statistics

                                       Std Error   Student
       Obs Name              Residual   Residual  Residual    -2-1 0 1 2

         1 Donna               3.6367      2.135     1.703  |      |***   |
         2 Gracie              2.2238      1.670     1.332  |      |**    |
         3 Luanne             -2.4812      2.056    -1.207  |    **|      |
         4 Mimi             -0.000855      2.051   -0.0004  |      |      |
         5 Chris              -1.9800      2.047    -0.967  |     *|      |
```

```
                              Output Statistics

                               Cook's          Hat Diag      Cov
        Obs Name                    D  RStudent         H    Ratio   DFFITS

          1 Donna               0.105    1.7718    0.1538   0.7959   0.7554
          2 Gracie              0.331    1.3526    0.4824   1.6512   1.3059
          3 Luanne              0.080   -1.2179    0.2155   1.1625  -0.6383
          4 Mimi                0.000 -0.000409    0.2193   1.5584  -0.0002
          5 Chris               0.053   -0.9659    0.2223   1.3025  -0.5164
```

```
                              Output Statistics

                             -------------------DFBETAS------------------
                                                                 Maximum_
       Obs Name              Intercept  Runtime      Age Run_Pulse    Pulse

         1 Donna                0.3224  -0.4897  -0.2658    0.0429  -0.0645
         2 Gracie              -0.2501  -0.2278  -0.1814   -0.9617   1.0269
         3 Luanne              -0.2127   0.1280   0.1711    0.4084  -0.3017
         4 Mimi                -0.0001   0.0000   0.0000    0.0001  -0.0000
         5 Chris                0.3170   0.3586  -0.2798    0.0185  -0.1792
```

```
               Sum of Residuals                          0
               Sum of Squared Residuals          140.10368
               Predicted Residual SS (PRESS)     190.90531
```

Because the INFLUENCE option is used, the statistics are calculated and printed. The PRESS statistic is the sum of the PRESS residuals. The residuals measure the deviation of the i^{th} observation about the regression line formed when that observation is deleted from the analysis. In other words, it measures how well the regression model predicts the i^{th} observation as though it were a new observation.

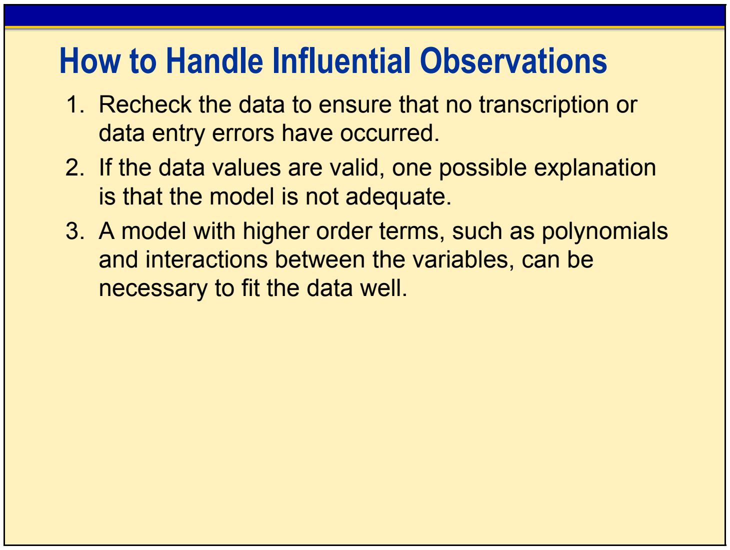
When the PRESS statistic is large compared to the Sum of the Squared Residuals, it indicates the presence of influential observations. The PRESS statistic is most useful when comparing several candidate models, such as comparing the BEST4 and BEST5 models that were examined earlier.

```
data influential;
   set fitnessout;
   p=5;
   n=31;
   cutdfits=2*((p/n)**0.5);
   cutcookd=4/n;
   rstud_i=(abs(rstud)>3);
   dfits_i=(abs(dfits)>cutdfits);
   cookd_i=(cooksd>cutcookd);
   sum_i=rstud_i+dfits_i+cookd_i;
   if sum_i>0;
run;
```

The DATA step sets 0/1 indicator variables (**rstud_i**, **dfits_i**, and **cookd_i**) for the diagnostic statistics using the suggested cutoffs. The **sum_i** variable is the total number of diagnostic statistics that exceed the cutoffs for the observation. The last line subsets the file so the data set **influential** includes only those observations that have at least one statistic that exceeds the cutoff. If the number of influential observations is large, you might not have the proper model.

```
proc print data=influential;
   var name cooksd rstud dfits cutcookd cutdfits cookd_i
       rstud_i dfits_i sum_i;
   title 'Observations that Exceed Suggested Cutoffs';
run;
```

```
                  Observations that Exceed Suggested Cutoffs

                                             c        c
                                             u        u     c  r  d
                  c                          t        t     o  s  f
                  o        r        d        c        d     o  t  i  s
         N        o        s        f        o        f     k  u  t  u
    O    a        k        t        i        o        i     d  d  s  m
    b    m        s        u        t        k        t     _  _  _  _
    s    e        d        d        s        d        s     i  i  i  i

    1  Gracie  0.33051  1.35265  1.30587  0.12903  0.80322  1  0  1  2
```



How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors have occurred.
2. If the data values are valid, one possible explanation is that the model is not adequate.
3. A model with higher order terms, such as polynomials and interactions between the variables, can be necessary to fit the data well.

If the unusual data is erroneous, correct the errors and reanalyze the data.

In this course, time does not permit discussion of higher order models in any depth.

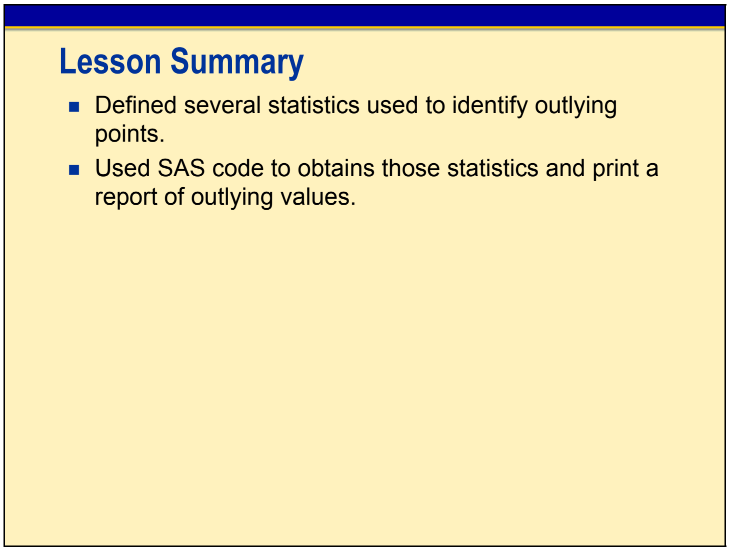
Another possibility is that the observation, though valid, may be unusual. If you have a larger sample size, there could be more observations like the unusual ones.

You might have to collect more data to confirm the relationship suggested by the influential observation.

In general, do not exclude data. In many circumstances, some of the unusual observations contain important information.

If you do choose to exclude some observations, include a description of the types of observations you exclude and provide an explanation. Also discuss the limitation of your conclusions, given the exclusions, as part of your report or presentation.

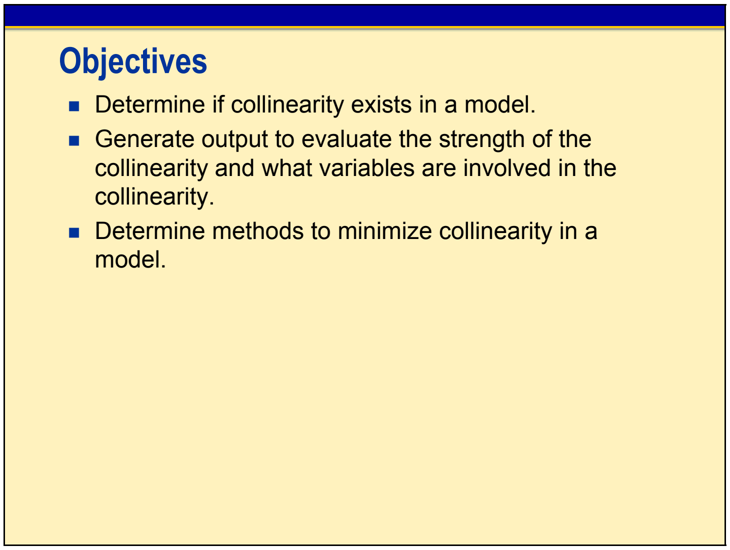
Exercise: Refer to your course workbook.



Lesson Summary

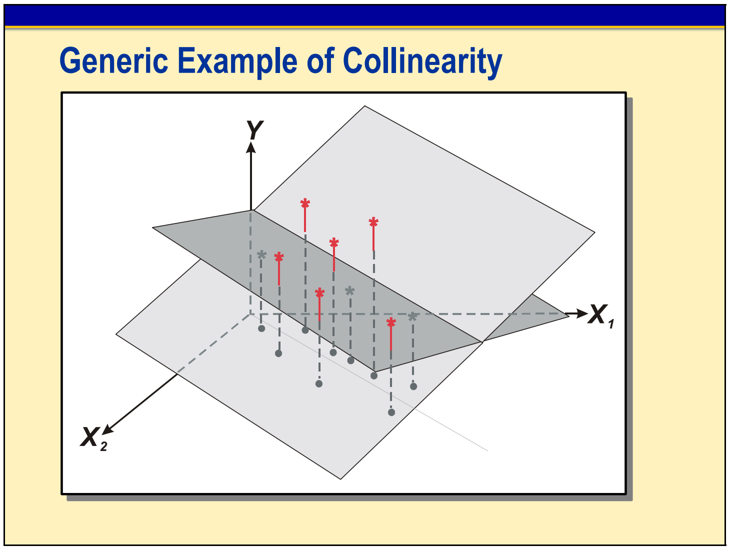
- Defined several statistics used to identify outlying points.
- Used SAS code to obtains those statistics and print a report of outlying values.

4.4 Collinearity



Objectives

- Determine if collinearity exists in a model.
- Generate output to evaluate the strength of the collinearity and what variables are involved in the collinearity.
- Determine methods to minimize collinearity in a model.



Generic Example of Collinearity

Recall the “picket fence” graph shown in Module 3 to illustrate the idea of collinearity.

Collinearity can cause these problems in your model:

- Truly significant terms might be hidden.
- The variances of the coefficients are increased, which results in less precise estimates of the parameters and the predicted values.

Collinearity is **not** a violation of the assumptions

Recall that collinearity arises when the X’s contain redundant information; for example, **`performance`** and **`runtime`** are highly correlated with each other.

Example of Collinearity

m04demo06.sas

Example: Generate a regression with **oxygen_consumption** as the dependent variable and **performance**, **runtime**, **age**, **weight**, **run_pulse**, **rest_pulse**, and **maximum_pulse** as the independent variables. Compare this model with the BEST4 model from the previous lesson.

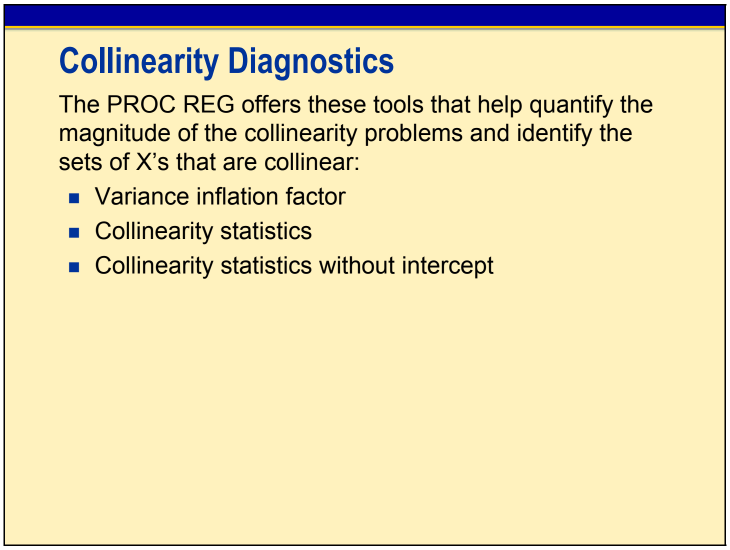
```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=performance runtime age weight
         run_pulse rest_pulse maximum_pulse;
   title 'Collinearity -- Full Model';
run;
quit;
```

```
                         Collinearity -- Full Model

                              The REG Procedure
                                Model: MODEL1
                   Dependent Variable: Oxygen_Consumption

                            Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     7      722.03251      103.14750      18.32    <.0001
Error                    23      129.52204        5.63139
Corrected Total          30      851.55455


          Root MSE              2.37306    R-Square     0.8479
          Dependent Mean       47.37581    Adj R-Sq     0.8016
          Coeff Var             5.00900


                            Parameter Estimates

                              Parameter      Standard
Variable              DF       Estimate         Error    t Value    Pr > |t|

Intercept              1       93.33753      36.49782       2.56      0.0176
Performance            1        0.25756       1.02373       0.25      0.8036
Runtime                1       -2.08804       2.22856      -0.94      0.3585
Age                    1       -0.21066       0.10519      -2.00      0.0571
Weight                 1       -0.07741       0.05681      -1.36      0.1862
Run_Pulse              1       -0.36618       0.12299      -2.98      0.0067
Rest_Pulse             1       -0.01389       0.07114      -0.20      0.8469
Maximum_Pulse          1        0.30490       0.13990       2.18      0.0398
```

The Model *F* is highly significant and the R square is large. These statistics suggest that the model fits the data well. However, when you examine the *p*-values of the parameters, only **`run_pulse`** and **`maximum_pulse`** are statistically significant.

Recall that the BEST4 model included **`runtime`**; however, in the full model, this same variable is not statistically significant (*p*-value=0.3585). Including all the terms in the model hid at least one significant term.

When you have a significant Model *F*, but no highly significant terms, collinearity is a likely problem.



Collinearity Diagnostics

The PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the sets of X’s that are collinear:

- Variance inflation factor
- Collinearity statistics
- Collinearity statistics without intercept

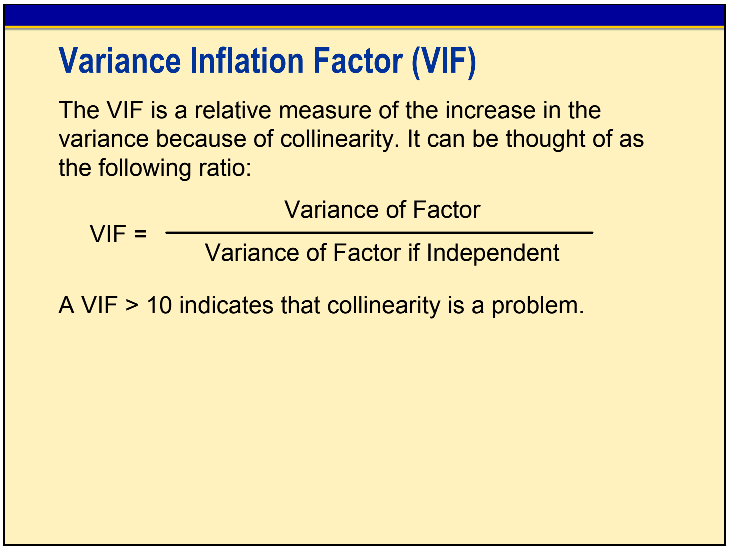
Selected collinearity tools:

The variance inflation factor provides a measure of the magnitude of the collinearity.

Collinearity statistics include the intercept vector when analyzing the X′X matrix for collinearity. These statistics are requested using the COLLIN option.

Collinearity statistics without the intercept exclude the intercept vector. These statistics are requested using the COLLINOINT option.

Statistics are also generated to provide a measure of the magnitude of the collinearity as well as give information that can be used to identify the sets of X’s that are the source of the collinearity.



Variance Inflation Factor (VIF)

The VIF is a relative measure of the increase in the variance because of collinearity. It can be thought of as the following ratio:

$$\text{VIF} = \frac{\text{Variance of Factor}}{\text{Variance of Factor if Independent}}$$

A VIF > 10 indicates that collinearity is a problem.

You can calculate a VIF for each term in the model.

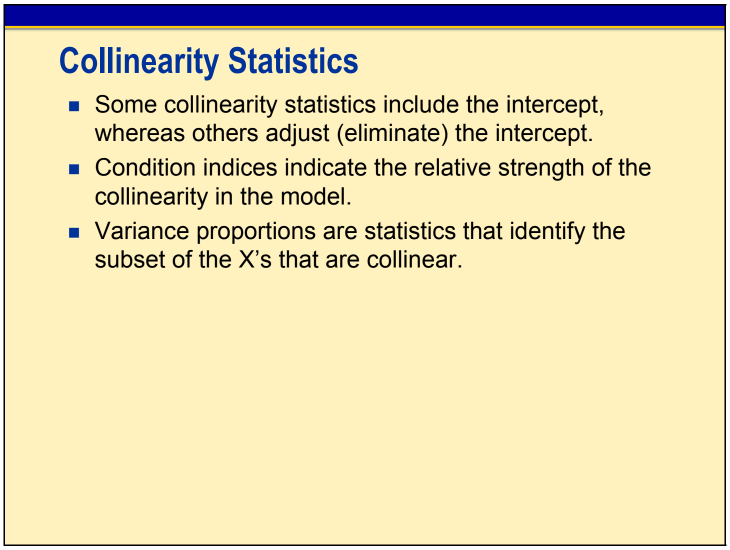
Marquardt suggests that a VIF > 10 indicates the presence of strong collinearity in the model.

$\text{VIF}_i = 1/(1 - R_i^2)$, where R_i^2 is the R square of X_i, regressed on all the other X's in the model.

For example, if the model is Y=X1 X2 X3 X4, i=1 to 4.

To calculate the R square for X3, fit the model X=X1 X2 X4. Take the R square from the model with X_3 as the dependent variable and replace it in the formula $\text{VIF}_3=1/(1 - R_3^2)$. If VIF_3 is greater than 10, X3 is possibly involved in collinearity.

D. W. Marquardt, "You Should Standardize the Predictor Variables in Your Regression Models," *Journal of the American Statistical Association* 75 (1980): 74-103.



Collinearity Statistics

- Some collinearity statistics include the intercept, whereas others adjust (eliminate) the intercept.
- Condition indices indicate the relative strength of the collinearity in the model.
- Variance proportions are statistics that identify the subset of the X's that are collinear.

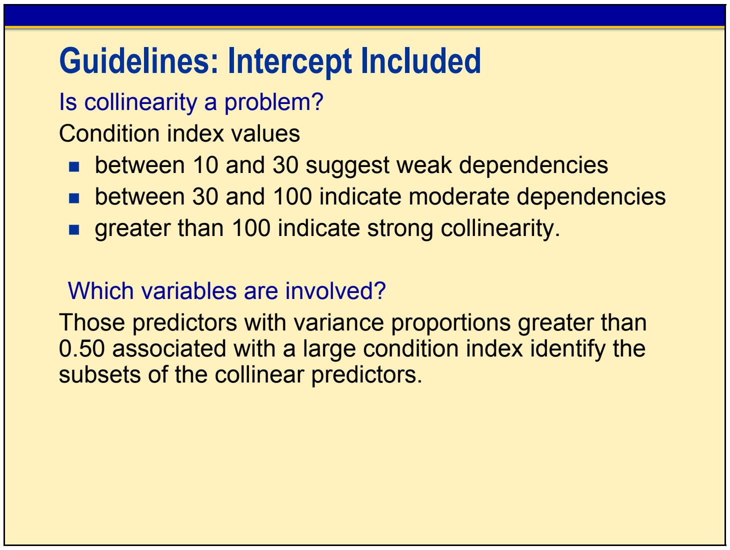
Collinearity statistics include

- eigenvalues
- condition indices
- variance proportions.

Eigenvalues are also called characteristic roots. Eigenvalues near zero indicate strong collinearity. A value λ is called an eigenvalue if there exists a nonzero vector z such that $\mathbf{(X'X)z} = \lambda\mathbf{z}$. The *condition index*, η_i, is the square root of the largest eigenvalue divided by λ_i.

The *variance proportions* used in combination with the condition index can be used to identify the sets of X's that are collinear. Variance proportions greater than 0.50 indicate which terms are correlated. The variance proportions are calculated for each term in the model.

The variance proportions for each term sum to 1.



Guidelines: Intercept Included

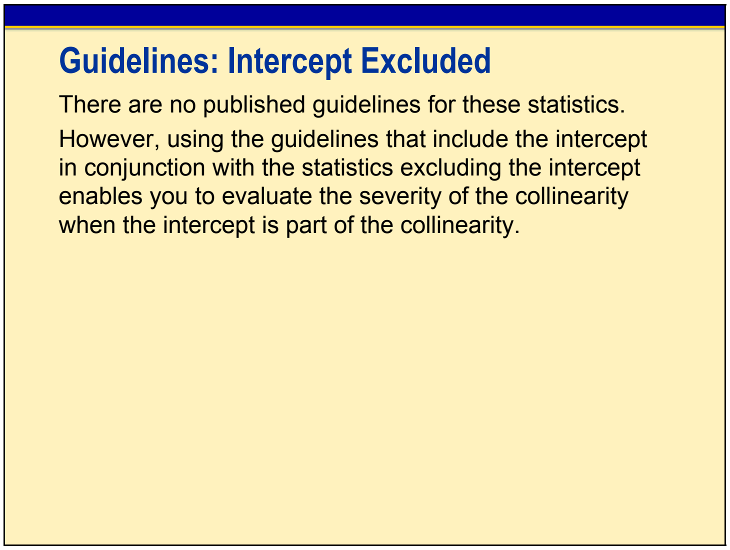
Is collinearity a problem?

Condition index values

- between 10 and 30 suggest weak dependencies
- between 30 and 100 indicate moderate dependencies
- greater than 100 indicate strong collinearity.

Which variables are involved?

Those predictors with variance proportions greater than 0.50 associated with a large condition index identify the subsets of the collinear predictors.



Guidelines: Intercept Excluded

There are no published guidelines for these statistics.

However, using the guidelines that include the intercept in conjunction with the statistics excluding the intercept enables you to evaluate the severity of the collinearity when the intercept is part of the collinearity.

Collinearity Diagnostics

m04demo07.sas, m04demo08.sas, m04demo09.sas

Example: Invoke PROC REG and use the VIF, COLLIN, and COLLINOINT options to assess the magnitude of the collinearity problem and identify the terms involved in the problem.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=performance runtime age weight
         run_pulse rest_pulse maximum_pulse
         / vif collin collinoint;
run;
quit;
```

Partial PROC REG Output

```
                            Parameter Estimates

                              Parameter      Standard
Variable             DF        Estimate         Error    t Value    Pr > |t|

Intercept             1        93.33753      36.49782       2.56      0.0176
Performance           1         0.25756       1.02373       0.25      0.8036
Runtime               1        -2.08804       2.22856      -0.94      0.3585
Age                   1        -0.21066       0.10519      -2.00      0.0571
Weight                1        -0.07741       0.05681      -1.36      0.1862
Run_Pulse             1        -0.36618       0.12299      -2.98      0.0067
Rest_Pulse            1        -0.01389       0.07114      -0.20      0.8469
Maximum_Pulse         1         0.30490       0.13990       2.18      0.0398

                             Parameter Estimates

                                           Variance
                  Variable             DF     Inflation

                  Intercept             1             0
                  Performance           1      54.34236
                  Runtime               1      50.92913
                  Age                   1       1.63228
                  Weight                1       1.19280
                  Run_Pulse             1       8.46965
                  Rest_Pulse            1       1.56516
                  Maximum_Pulse         1       8.75615
```

Some of the VIFs are much larger than 10. A severe collinearity problem is present.

COLLIN Option Output

```
                     Collinearity -- Full Model

                          The REG Procedure
                            Model: MODEL1
               Dependent Variable: Oxygen_Consumption

                       Collinearity Diagnostics

                       Condition  -------Proportion of Variation-------
  Number   Eigenvalue      Index    Intercept  Performance      Runtime

       1      7.81224    1.00000   0.00000223   0.00003396   0.00000516
       2      0.14978    7.22204  4.610439E-7      0.01283   0.00026016
       3      0.01739   21.19723   0.00006157   0.00023609   0.00028745
       4      0.01246   25.03710   0.00000120      0.00120   0.00016004
       5      0.00606   35.90012   0.00027949   0.00007171      0.00149
       6      0.00179   66.03652      0.01276      0.03405      0.07620
       7   0.00018592  204.98810      0.00326      0.03584      0.02721
       8   0.00009415  288.05165      0.98363      0.91573      0.89439

                       Collinearity Diagnostics

          --------------------Proportion of Variation--------------------
                                                                 Maximum_
  Number          Age       Weight    Run_Pulse   Rest_Pulse        Pulse

       1   0.00011543   0.00015063   0.00000679   0.00019829   0.00000501
       2   0.00032355   0.00018997   0.00001537      0.00374   0.00000627
       3      0.24299      0.00908   0.00032301      0.24059   0.00022961
       4      0.05498      0.39864   0.00016217      0.33791   0.00022890
       5      0.09288      0.45536      0.01695      0.29325      0.00969
       6      0.38685      0.10219      0.04272      0.01670      0.01335
       7      0.01651      0.01929      0.92679   0.00001297      0.92625
       8      0.20535      0.01510      0.01303      0.10759      0.05024
```

Two condition indices are well above 100. For the largest, the variance proportions for the intercept and the variables **`performance`** and **`runtime`** are greater than 0.50.

COLLINOINT Option Output

```
              Collinearity Diagnostics(intercept adjusted)

                         Condition  -------Proportion of Variation-------
  Number   Eigenvalue        Index  Performance      Runtime          Age

       1      2.92687      1.00000      0.00124      0.00133      0.00328
       2      1.87356      1.24988      0.00196      0.00194      0.10087
       3      0.94035      1.76424   0.00014220   0.00035679      0.00167
       4      0.74998      1.97550   0.00001910   0.00003187      0.20986
       5      0.43947      2.58069      0.00329      0.00519      0.57367
       6      0.06022      6.97181   0.00019461   0.00012410      0.03802
       7      0.00955     17.50829      0.99315      0.99103      0.07263


             ----------------Proportion of Variation----------------
                                                          Maximum_
     Number       Weight      Run_Pulse     Rest_Pulse        Pulse

          1      0.00953        0.00870        0.03205      0.00750
          2      0.01834        0.00620        0.00309      0.00967
          3      0.74750        0.00695        0.03473      0.00343
          4   0.00001480        0.02020        0.43182      0.01612
          5      0.16190        0.00433        0.41363      0.00220
          6      0.02856        0.95340        0.00431      0.96071
          7      0.03416     0.00023243        0.08038   0.00036791
```

A similar pattern of collinearity appears when using the COLLINOINT option. Examining the last row of the above table reveals that **`performance`** (0.99315) and **`runtime`** (0.99103) possess variance proportions greater than 0.50. You can conclude that these two variables are involved in the collinearity.

Begin the process of eliminating collinear terms by returning to the Parameter Estimates table and recording the *p*-values of the identified subset of the independent variables:

`performance` *p*-value=0.8036

`runtime` *p*-value=0.3585

With this subset of variables, eliminate **`performance`** from the model. Note that this variable also has a high VIF.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=runtime age weight run_pulse
         rest_pulse maximum_pulse / vif collin collinoint;
   title 'Collinearity - PERFORMANCE Removed';
run;
quit;
```

Partial PROC REG Output

```
                           Parameter Estimates

                             Parameter      Standard
 Variable            DF       Estimate         Error   t Value   Pr > |t|

 Intercept            1      101.96313      12.27174      8.31     <.0001
 Runtime              1       -2.63994       0.38532     -6.85     <.0001
 Age                  1       -0.21848       0.09850     -2.22     0.0363
 Weight               1       -0.07503       0.05492     -1.37     0.1845
 Run_Pulse            1       -0.36721       0.12050     -3.05     0.0055
 Rest_Pulse           1       -0.01952       0.06619     -0.29     0.7706
 Maximum_Pulse        1        0.30457       0.13714      2.22     0.0360

                           Parameter Estimates

                                              Variance
                  Variable            DF     Inflation

                  Intercept            1             0
                  Runtime              1       1.58432
                  Age                  1       1.48953
                  Weight               1       1.15973
                  Run_Pulse            1       8.46034
                  Rest_Pulse           1       1.41004
                  Maximum_Pulse        1       8.75535
```

The variables **run_pulse** and **maximum_pulse** are significant in this model as they were in the previous model, but now both **runtime** and **age** are significant in this model.

Note that the VIFs are now all less than 10.

Partial PROC REG Output (continued)

```
                  Collinearity - PERFORMANCE Removed

                           The REG Procedure
                             Model: MODEL1
                Dependent Variable: Oxygen_Consumption

                        Collinearity Diagnostics

                        Condition  -------Proportion of Variation-------
   Number   Eigenvalue      Index    Intercept      Runtime          Age

      1      6.94983      1.00000   0.00002395   0.00021174   0.00015997
      2      0.01856     19.35297      0.00224      0.02439      0.15550
      3      0.01521     21.37532   0.00069190      0.12332      0.15174
      4      0.00914     27.57505      0.00635      0.61945      0.03075
      5      0.00603     33.94799      0.00139      0.12581      0.11951
      6      0.00105     81.17086      0.79602      0.09233      0.47800
      7   0.00017900    197.04044      0.19329      0.01449      0.06435

                       Collinearity Diagnostics

           -----------------Proportion of Variation----------------
                                                            Maximum_
     Number       Weight      Run_Pulse     Rest_Pulse         Pulse

        1   0.00019576     0.00000860     0.00027961    0.00000633
        2      0.00878     0.00000185        0.39351    0.00000723
        3      0.23637        0.00113        0.03259       0.00121
        4      0.17375        0.00152        0.19195       0.00125
        5      0.45090        0.01510        0.35859       0.00840
        6      0.10834        0.06682        0.01756       0.00556
        7      0.02167        0.91542        0.00552       0.98356
```

The largest condition index is still greater than 100, indicating that there is still collinearity in this model. For the largest condition index, the variance proportions for **`run_pulse`** (0.91542) and **`maximum_pulse`** (0.98356) are greater than 0.5. Note that the intercept is not involved in collinearity, so there is no need to examine the COLLINOINT output.

Because the variable **`maximum_pulse`** (0.0360) has a higher *p*-value than **`run_pulse`** (0.0055), generate another model and eliminate the variable **`maximum_pulse`** from the MODEL statement.

```
proc reg data=sasuser.b_fitness;
   model oxygen_consumption=runtime age weight run_pulse
         rest_pulse / vif collin collinoint;
   title 'Collinearity - MAXIMUM_PULSE and PERFORMANCE Removed';
run;
quit;
```

PROC REG Output

```
          Collinearity - MAXIMUM_PULSE and PERFORMANCE Removed

                            The REG Procedure
                              Model: MODEL1
                 Dependent Variable: Oxygen_Consumption

                          Analysis of Variance

                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F

Model                    5     694.98323      138.99665      22.19    <.0001
Error                   25     156.57132        6.26285
Corrected Total         30     851.55455


          Root MSE              2.50257    R-Square     0.8161
          Dependent Mean       47.37581    Adj R-Sq     0.7794
          Coeff Var             5.28238


                          Parameter Estimates

                            Parameter      Standard
Variable             DF      Estimate         Error    t Value    Pr > |t|

Intercept             1     115.46115      11.46893      10.07      <.0001
Runtime               1      -2.71594       0.41288      -6.58      <.0001
Age                   1      -0.27650       0.10217      -2.71      0.0121
Weight                1      -0.05300       0.05811      -0.91      0.3704
Run_Pulse             1      -0.12213       0.05207      -2.35      0.0272
Rest_Pulse            1      -0.02485       0.07116      -0.35      0.7298

                           Parameter Estimates

                                                Variance
                  Variable             DF      Inflation

                  Intercept             1              0
                  Runtime               1        1.57183
                  Age                   1        1.38477
                  Weight                1        1.12190
                  Run_Pulse             1        1.36493
                  Rest_Pulse            1        1.40819
```

The variables **`weight`** and **`rest_pulse`** are not statistically significant, indicating that they could be removed from the model. All VIFs are relatively small.

PROC REG Output (continued)

```
          Collinearity - MAXIMUM_PULSE and PERFORMANCE Removed

                            The REG Procedure
                              Model: MODEL1
                 Dependent Variable: Oxygen_Consumption

                          Collinearity Diagnostics

                          Condition  -------Proportion of Variation-------
   Number   Eigenvalue        Index    Intercept      Runtime          Age

        1      5.95261      1.00000   0.00004324   0.00029113   0.00023471
        2      0.01855     17.91390      0.00296      0.02190      0.17447
        3      0.01434     20.37297      0.00139      0.09587      0.14694
        4      0.00882     25.97155      0.01086      0.75407      0.04148
        5      0.00465     35.78017      0.02723      0.02828      0.18069
        6      0.00102     76.21454      0.95752      0.09958      0.45619

                          Collinearity Diagnostics

                      ---------Proportion of Variation---------
              Number       Weight      Run_Pulse     Rest_Pulse

                   1   0.00027579     0.00007258     0.00038178
                   2      0.00826     0.00002193        0.38754
                   3      0.36846        0.00674        0.02990
                   4      0.06095        0.00710        0.27246
                   5      0.46144        0.26977        0.29881
                   6      0.10061        0.71629        0.01090
```

The largest condition index is now approximately 76. This indicates that there are some moderate dependencies between the predictor variables in this model. Examination of the variance proportions indicates that the intercept and **`run_pulse`** are involved in collinearity.

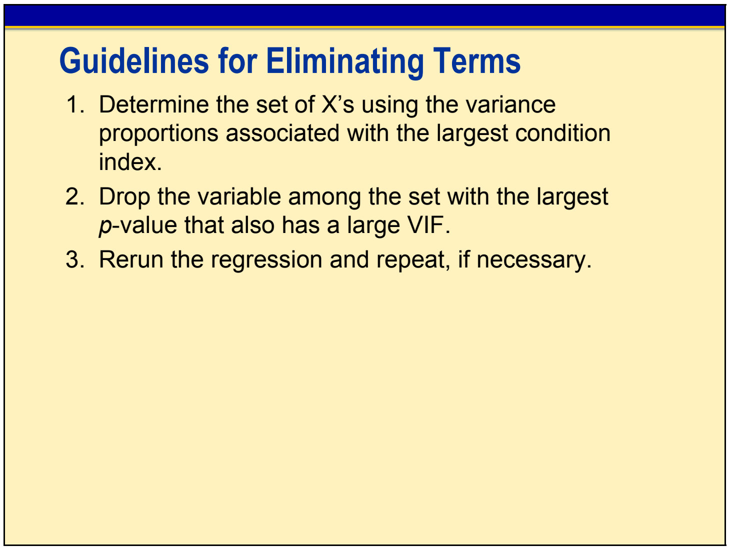
```
                Collinearity Diagnostics(intercept adjusted)

                                              Condition
                      Number   Eigenvalue         Index

                           1      1.86111       1.00000
                           2      1.28404       1.20392
                           3      0.89216       1.44433
                           4      0.59808       1.76404
                           5      0.36462       2.25927

                 Collinearity Diagnostics(intercept adjusted)

            --------------------Proportion of Variation--------------------
  Number      Runtime          Age       Weight    Run_Pulse   Rest_Pulse

       1      0.07701      0.02981      0.04373      0.12341      0.11184
       2      0.14039      0.27964      0.09841      0.01037      0.03290
       3      0.04970      0.07934      0.68711      0.03614      0.08003
       4      0.00449      0.05979      0.03567      0.66283      0.45266
       5      0.72841      0.55142      0.13508      0.16726      0.32257
```

Using the COLLINOINT option output, you can determine that **runtime** (variance proportion=0.72841) and **age** (variance proportion=0.55142) are involved in collinearity.

Return to the Parameter Estimates table, and record the *p*-values of **runtime** (<0.0001) and **age** (0.0121).

You can accept the current model without deleting any more variables because **runtime** and **age** are both statistically significant. Furthermore, remember that the COLLIN Condition Index is approximately 76 for this model and that falls into the moderate range of collinearity.

As noted earlier, the variables **weight** (*p*-value=0.3704) and **rest_pulse** (*p*-value=0.7298) are not statistically significant; you might want to eliminate **rest_pulse** from the model and re-execute the reduced model.



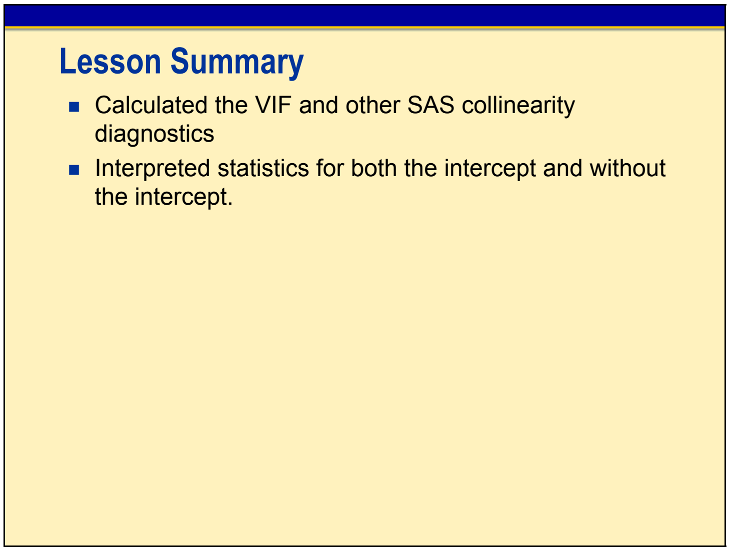
Guidelines for Eliminating Terms

1. Determine the set of X's using the variance proportions associated with the largest condition index.
2. Drop the variable among the set with the largest *p*-value that also has a large VIF.
3. Rerun the regression and repeat, if necessary.

In the previous demonstration you saw how to identify the sets of X's that were collinear.

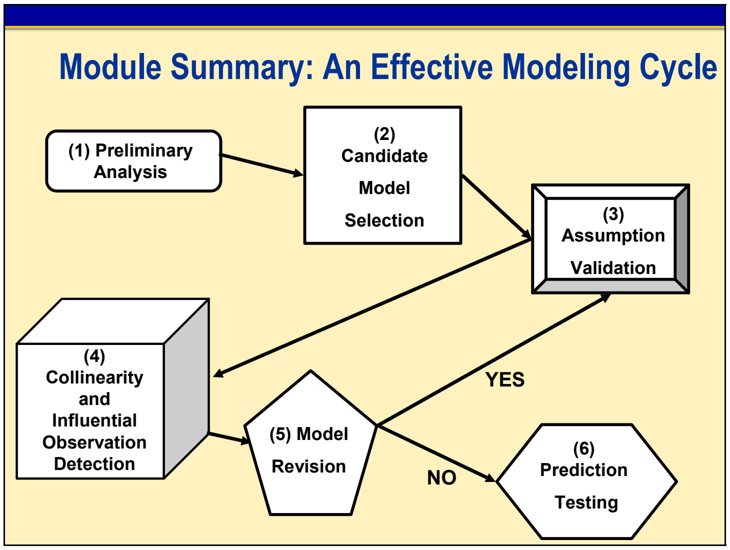
The natural question is, "Which terms should be dropped?" Subject matter expertise should be used, as well as the suggested guidelines above.

There are other approaches to dealing with collinearity. Two techniques are ridge regression and principle components regression. In addition, recentering the predictor variables can sometimes eliminate collinearity problems, especially in a polynomial regression.



Lesson Summary

- Calculated the VIF and other SAS collinearity diagnostics
- Interpreted statistics for both the intercept and without the intercept.


Module Summary: An Effective Modeling Cycle
(1) Preliminary Analysis
(2) Candidate Model Selection
(3) Assumption Validation
(4) Collinearity and Influential Observation Detection
(5) Model Revision
YES
NO
(6) Prediction Testing

(1) Preliminary Analysis

This step includes the use of descriptive statistics, graphs, and correlation analysis.

(2) Candidate Model Selection

This step uses the numerous selection options in linear regression to identify one or more candidate models.

(3) Assumption Validation

This step includes the plots of residuals and graphs of the residuals versus the predicted values. It also includes a test for equal variances.

(4) Collinearity and Influential Observation Detection

The former includes the use of the VIF statistic, condition indices, and variance proportions; the latter includes the examination of Rstudent residuals, Cook's D statistic, and DFFITS statistics and their respective cutoffs.

(5) Model Revision

If steps (3) and (4) indicate the need for model revision, generate a new model by returning to these two steps.

(6) Prediction Testing

If possible, validate the model with data not used to build the model.

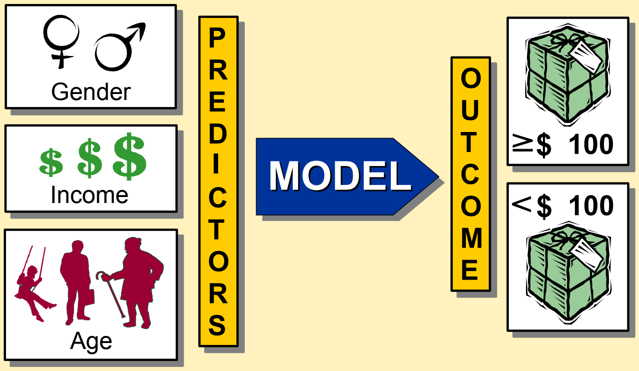
Module 5 Tests of Association and Logistic Regression

5.1 Tests of Association

Objectives

- Perform a chi-square test for association.
- Calculate the strength of the association.
- Produce exact *p*-values for the chi-square test for association.
- Perform a Mantel-Haenszel chi-square test.

Sample Data Set



Gender

Income

Age

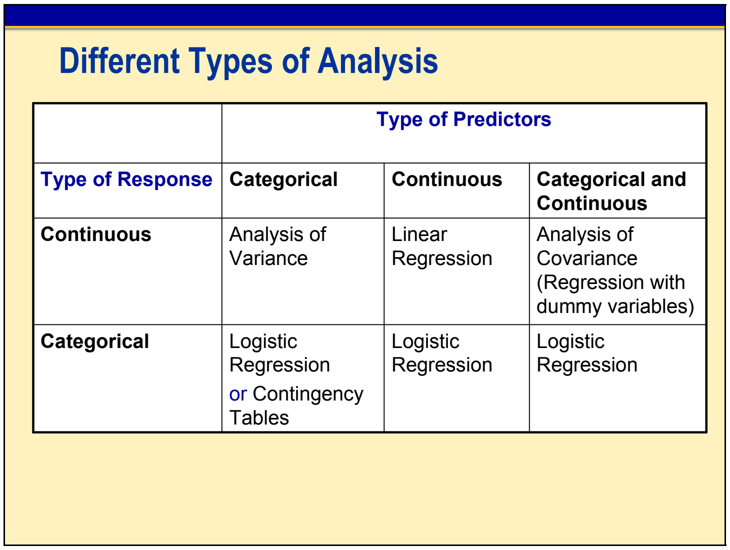
PREDICTORS

MODEL

OUTCOME

≥$ 100

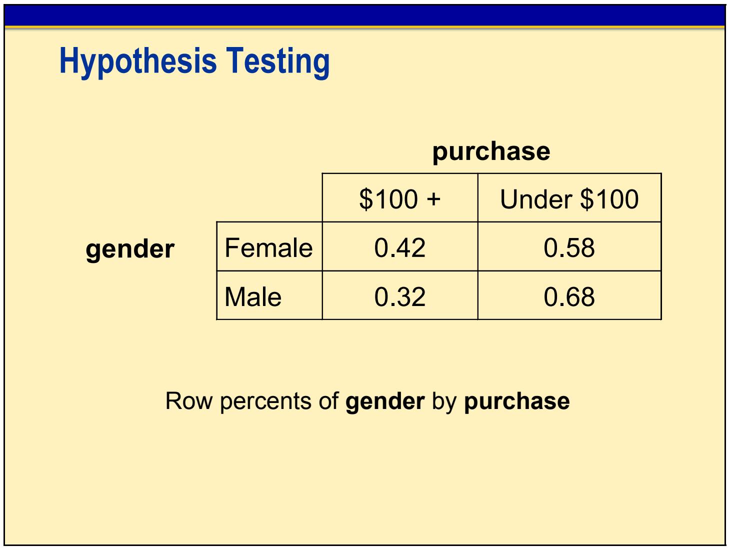
<$ 100



Different Types of Analysis

	Type of Predictors		
Type of Response	**Categorical**	**Continuous**	**Categorical and Continuous**
Continuous	Analysis of Variance	Linear Regression	Analysis of Covariance (Regression with dummy variables)
Categorical	Logistic Regression or Contingency Tables	Logistic Regression	Logistic Regression

The mathematical model for a model is mostly determined by whether the outcome is continuous or categorical.

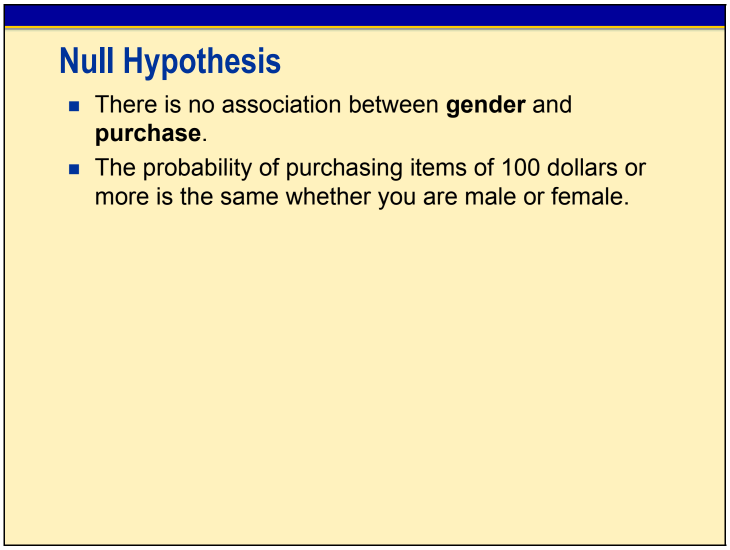


Hypothesis Testing

		purchase	
		$100 +	Under $100
gender	Female	0.42	0.58
	Male	0.32	0.68

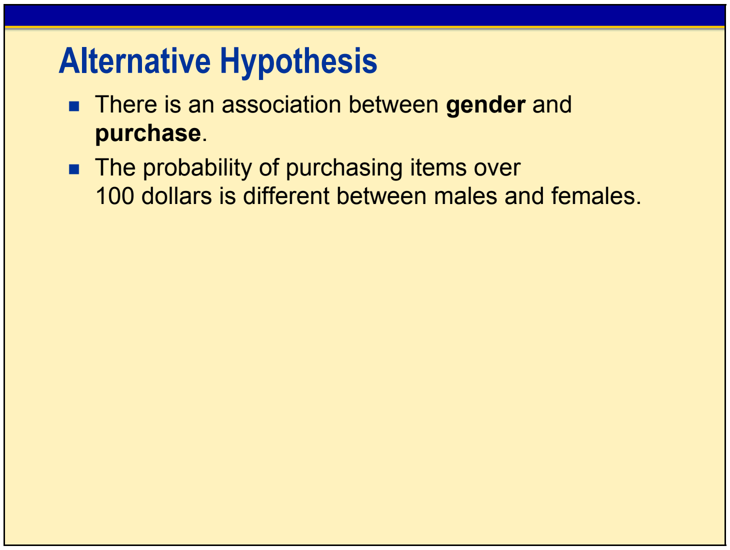
Row percents of **gender** by **purchase**

There appears to be an association between `gender` and `purchase` because the row percentages are different in each column. To test for this association, you are assessing whether the probability of females purchasing items of 100 dollars or more (0.42) is significantly different from the probability of males purchasing items of 100 dollars or more (0.32).



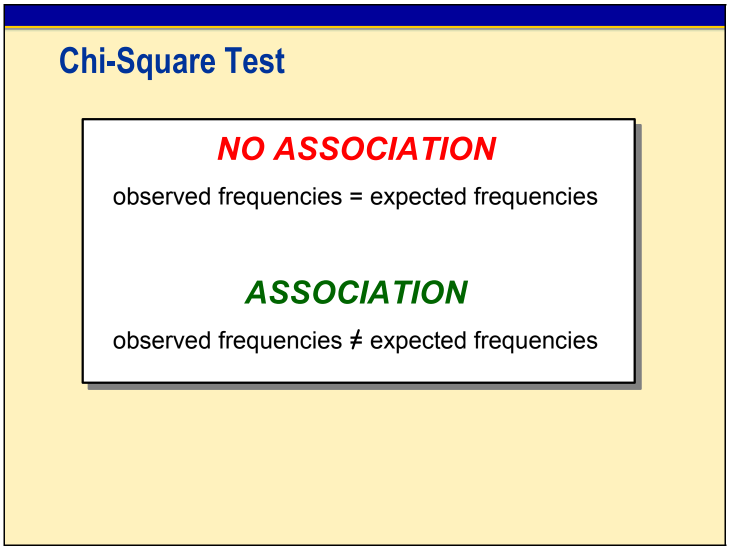
Null Hypothesis

- There is no association between **gender** and **purchase**.
- The probability of purchasing items of 100 dollars or more is the same whether you are male or female.



Alternative Hypothesis

- There is an association between **gender** and **purchase**.
- The probability of purchasing items over 100 dollars is different between males and females.



Chi-Square Test

NO ASSOCIATION

observed frequencies = expected frequencies

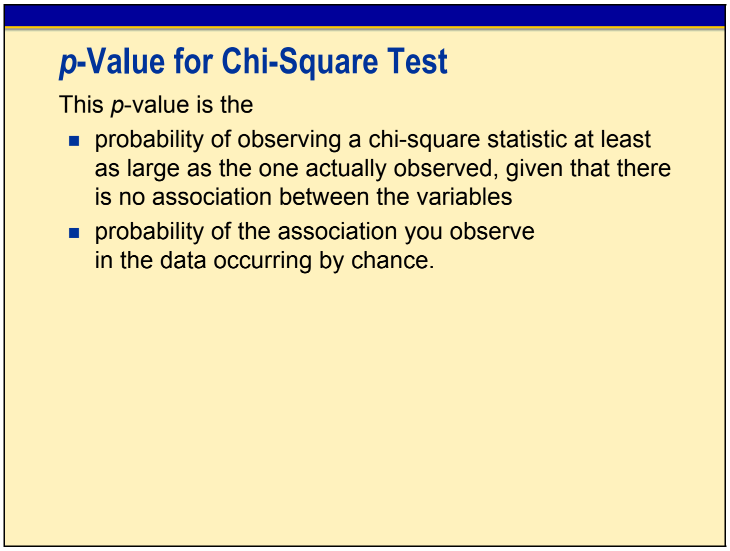
ASSOCIATION

observed frequencies ≠ expected frequencies

A commonly used test that examines whether there is an association between two categorical variables is the Pearson chi-square test. The chi-square test measures the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If you have a significant chi-square statistic, there is strong evidence that an association exists between the variables.

The expected frequencies are calculated using the formula

(row total * column total) / sample size.

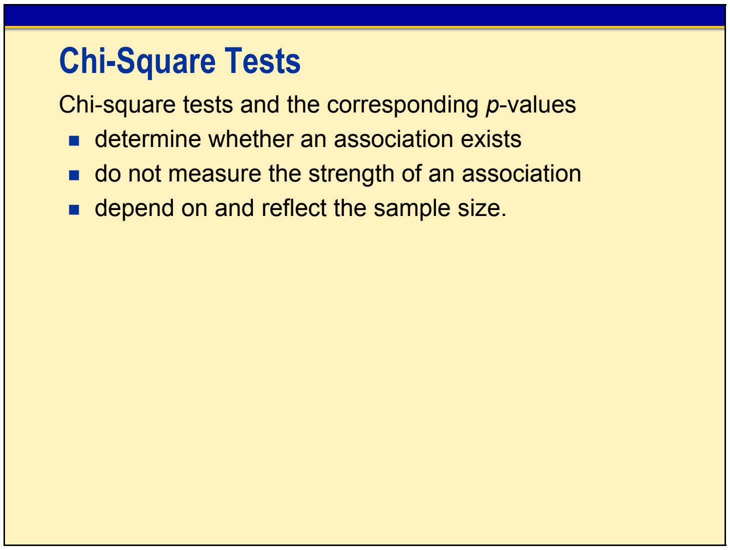


p-Value for Chi-Square Test

This *p*-value is the

- probability of observing a chi-square statistic at least as large as the one actually observed, given that there is no association between the variables
- probability of the association you observe in the data occurring by chance.

In general, the larger the chi-square values, the smaller the *p*-value, which means you have more evidence against the null hypothesis.

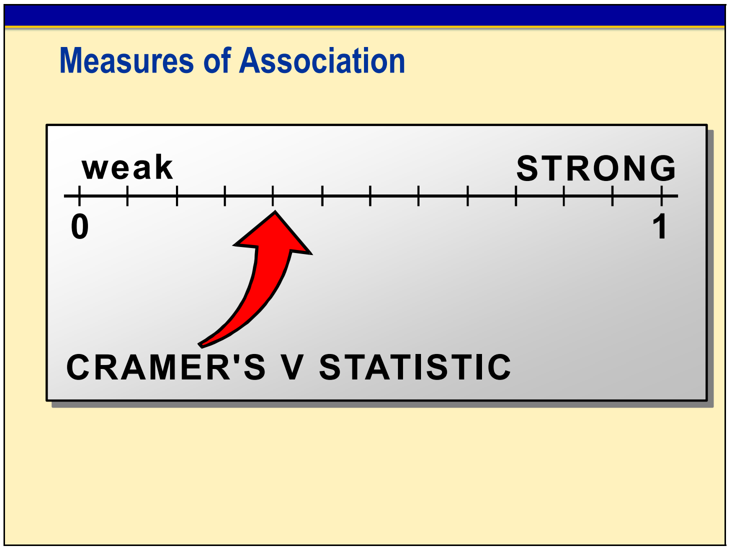


Chi-Square Tests

Chi-square tests and the corresponding *p*-values

- determine whether an association exists
- do not measure the strength of an association
- depend on and reflect the sample size.

If you double the size of your sample by copying each observation, you double the chi-square statistic even though the strength of the association does not change.



Measures of Association

weak

STRONG

0

1

CRAMER'S V STATISTIC

One measure of the strength of the association between two nominal variables is Cramer’s V statistic. It is in the range of -1 to 1 for 2-by-2 tables and 0 to 1 for larger tables. Values further away from 0 indicate the presence of a relatively strong association.

Cramer’s V statistic is derived from the Pearson chi-square statistic.

Chi-Square Test

m5demo02.sas

Example: Use the FREQ procedure to test for an association between the variables **gender** and **purchase**. Also generate the expected cell frequencies and the cell's contribution to the total chi-square statistic.

```
proc freq data=sasuser.b_sales_inc;
   tables gender*purchase
          / chisq expected cellchi2 nocol nopercent;
   format purchase purfmt.;
   title1 'Association between GENDER and PURCHASE';
run;
```

Selected TABLES statement options:

CHISQ produces the chi-square test of association and the measures of association based upon the chi-square statistic.

EXPECTED prints the expected cell frequencies under the hypothesis of no association.

CELLCHI2 prints each cell's contribution to the total chi-square statistic.

NOCOL suppresses printing the column percentages.

NOPERCENT suppresses printing the cell percentages.

The frequency table is shown below.

```
          Association between GENDER and PURCHASE

                    The FREQ Procedure

                Table of gender by purchase

         gender          purchase

         Frequency      |
         Expected       |
         Cell Chi-Square|
         Row Pct        |< $100  |$100 +  |  Total
         ---------------+--------+--------+
         Female         |    139 |    101 |    240
                        | 149.79 | 90.209 |
                        | 0.7774 | 1.2909 |
                        |  57.92 |  42.08 |
         ---------------+--------+--------+
         Male           |    130 |     61 |    191
                        | 119.21 | 71.791 |
                        | 0.9769 | 1.6221 |
                        |  68.06 |  31.94 |
         ---------------+--------+--------+
         Total               269      162      431
```

It appears that the cell for **purchase**=1 ($100 dollars or more) and **gender**=Male contributes the most to the chi-square statistic.

The cell chi-square is calculated using the formula

(observed frequency – expected frequency)^2 / expected frequency

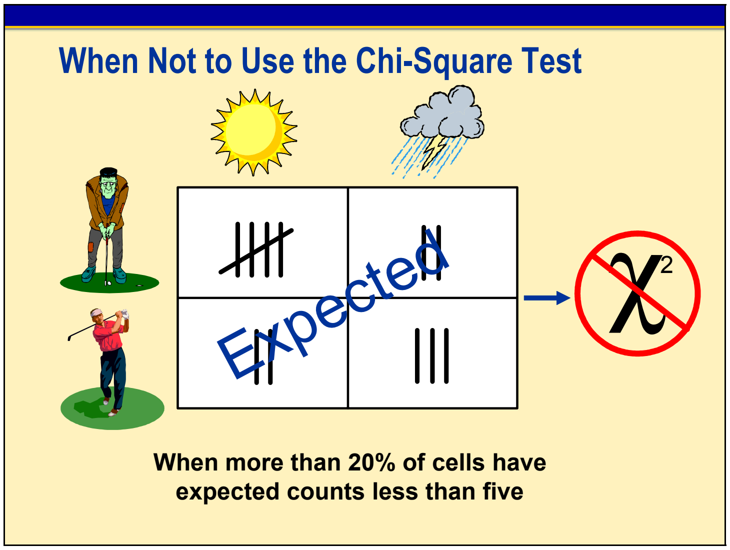
The table showing the chi-square test and Cramer's V is shown below.

```
            Statistics for Table of gender by purchase

           Statistic                     DF       Value      Prob
           ------------------------------------------------------
           Chi-Square                     1      4.6672    0.0307
           Likelihood Ratio Chi-Square    1      4.6978    0.0302
           Continuity Adj. Chi-Square     1      4.2447    0.0394
           Mantel-Haenszel Chi-Square     1      4.6564    0.0309
           Phi Coefficient                      -0.1041
           Contingency Coefficient               0.1035
           Cramer's V                           -0.1041


                           Fisher's Exact Test
                     ----------------------------------
                     Cell (1,1) Frequency (F)       139
                     Left-sided Pr <= F          0.0195
                     Right-sided Pr >= F         0.9883

                     Table Probability (P)       0.0078
                     Two-sided Pr <= P           0.0355

                              Sample Size = 431
```

Because the *p*-value for the chi-square statistic is 0.0307 (and, thus, is below .05), you reject the null hypothesis at the 0.05 level and conclude that there is evidence of an association between **`gender`** and **`purchase`**. However, Cramer's V indicates that the association detected with the chi-square test is relatively weak. This means that the association was detected because of the large sample size, not because of its strength.

The chi-square statistic is calculated by summing the cell chi-square values. It exploits the property that the frequency distributions tend toward a normal distribution in very large samples. The formula is

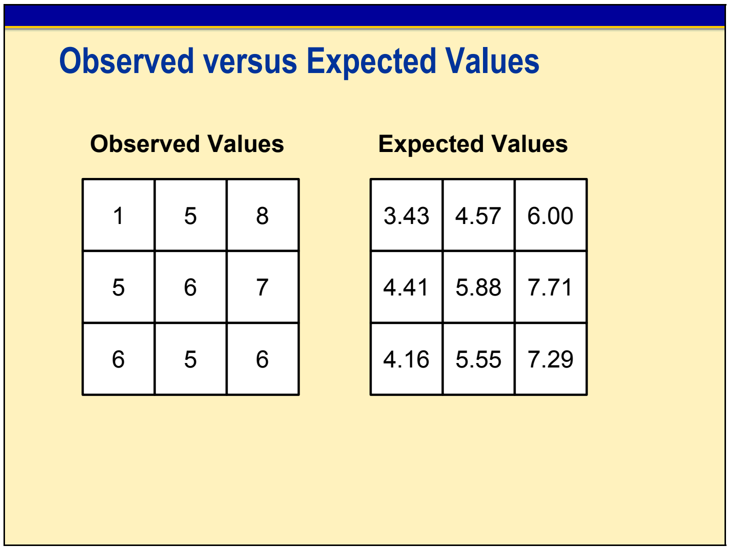
$$\Sigma(\text{observed} - \text{expected})^2 / \text{expected}$$



When Not to Use the Chi-Square Test

When more than 20% of cells have expected counts less than five

There are times when the chi-square test might not be appropriate. In fact, when more than 20% of the cells have an expected cell frequency of less than 5, the chi-square test might not be valid. This is because the *p*-values are based on the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. Therefore, when the sample sizes are small, the asymptotic (large sample) *p*-values might not be valid.



Observed versus Expected Values

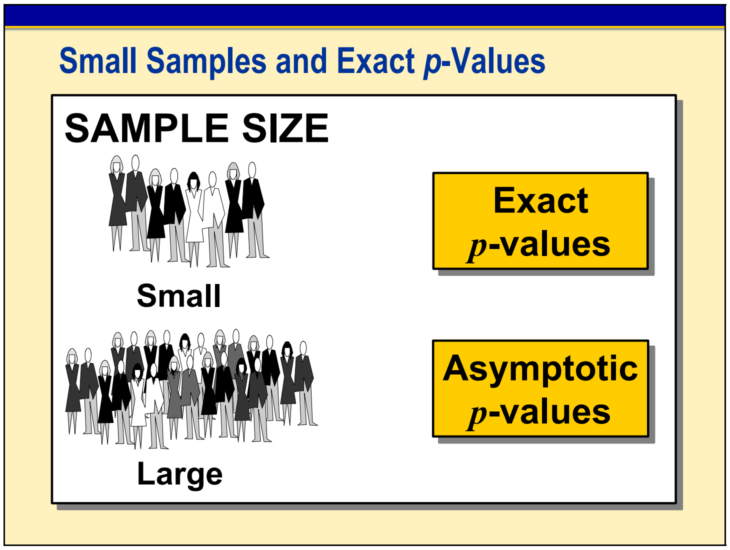
Observed Values

1	5	8
5	6	7
6	5	6

Expected Values

3.43	4.57	6.00
4.41	5.88	7.71
4.16	5.55	7.29

The criterion for the chi-square test is based on the expected values, not the observed values. In the slide above, 8 out of 9, or 89%, have observed values of 5 or more. However, 5 out of 9, or 56%, have expected values of 5 or more. Therefore, the chi-square test might not be valid.



Small Samples and Exact *p*-Values

SAMPLE SIZE

Small

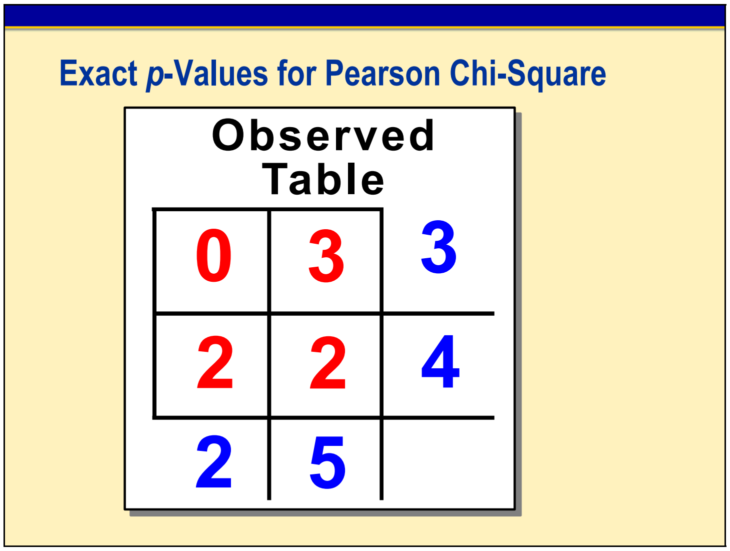
Large

Exact *p*-values

Asymptotic *p*-values

You can obtain exact *p*-values for many tests in the FREQ procedure. Exact *p*-values are useful when the sample size is small, in which case the asymptotic *p*-values might not be useful.

However, large data sets (in terms of sample size, number of rows, and number of columns) can require a prohibitive amount of time and memory for computing exact *p*-values. For large data sets, consider whether exact *p*-values are needed or whether asymptotic *p*-values might be quite close to the exact *p*-values.



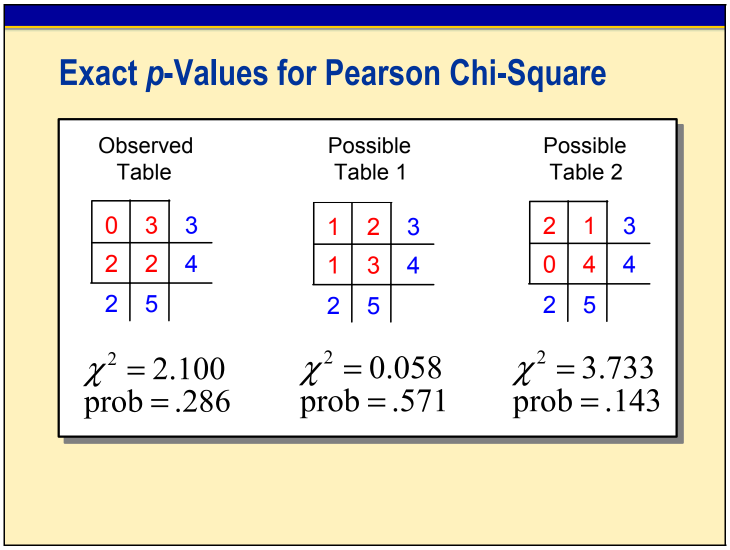
Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	

Exact *p*-values reflect the probability of observing a table with at least as much evidence of an association as the one actually observed, given there is no association between the variables. If your significance level is .05, exact *p*-values below .05 reflect significant associations.

For example, consider the table above. With such a small sample size, the asymptotic *p*-values would not be valid.



Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	

$\chi^2 = 2.100$
prob = .286

Possible Table 1

1	2	3
1	3	4
2	5	

$\chi^2 = 0.058$
prob = .571

Possible Table 2

2	1	3
0	4	4
2	5	

$\chi^2 = 3.733$
prob = .143

A key assumption behind the computation of exact *p*-values is that the column totals and row totals are fixed. Thus, there are only three possible tables.

To compute an exact *p*-value for this example, examine the chi-square value for each table and the probability that the table occurs given the three tables (the probabilities add up to 1). The Observed Table has a chi-square value of 2.100, so any table with a chi-square value of 2.100 or higher would be used to compute the exact *p*-value. Thus, the exact *p*-value would be 0.286 (Observed Table)+0.143 (Possible Table 2=.429. This means you have a 43% chance of obtaining, simply by random chance, a table with at least as much of an association as the observed table.

Exact *p*-Values for Pearson Chi-Square Test

m5demo03.sas

Example: Invoke PROC FREQ and produce exact *p*-values for the Pearson chi-square test. Use the data set **sasuser.b_exact**, which has the data from the previous example.

```
proc freq data=sasuser.b_exact;
   tables a*b;
   exact pchi;
   title 'Association using EXACT PCHI statement';
run;
```

Selected FREQ procedure statements:

EXACT produces exact *p*-values for the statistics listed as keywords. If you use only one TABLES statement, you do not need to specify options in the TABLES statement to perform the analyses the EXACT statement requests.

Selected EXACT statement options:

PCHI requests exact *p*-values for the chi-square statistics. Also produces Cramer's V and other related statistics.

If you use multiple TABLES statements and want exact computations, you must specify options in the TABLES statement to compute the desired statistics.

The frequency table is shown below.

```
                  Association using EXACT PCHI statement

                            The FREQ Procedure

                              Table of a by b

                    a         b

                    Frequency│
                    Percent  │
                    Row Pct  │
                    Col Pct  │       1│       2│  Total
                    ─────────┼────────┼────────┼
                           1 │      0 │      3 │      3
                             │   0.00 │  42.86 │  42.86
                             │   0.00 │ 100.00 │
                             │   0.00 │  60.00 │
                    ─────────┼────────┼────────┼
                           2 │      2 │      2 │      4
                             │  28.57 │  28.57 │  57.14
                             │  50.00 │  50.00 │
                             │ 100.00 │  40.00 │
                    ─────────┼────────┼────────┼
                    Total           2        5        7
                                28.57    71.43   100.00
```

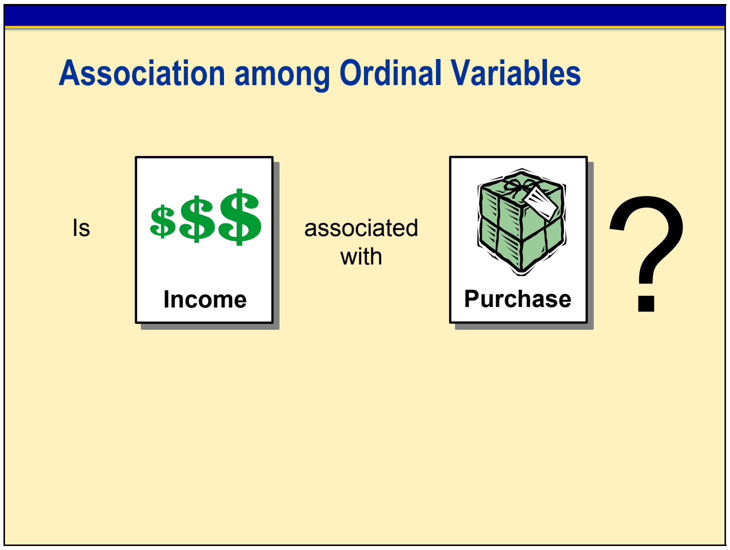
This is the observed table from the previous example.

The Pearson Chi-Square Test table contains the Exact Pr >= ChiSq value of 0.4286, and is shown below.

```
                   Statistics for Table of a by b

          Statistic                     DF       Value      Prob
          ------------------------------------------------------
          Chi-Square                     1      2.1000    0.1473
          Likelihood Ratio Chi-Square    1      2.8306    0.0925
          Continuity Adj. Chi-Square     1      0.3646    0.5460
          Mantel-Haenszel Chi-Square     1      1.8000    0.1797
          Phi Coefficient                      -0.5477
          Contingency Coefficient               0.4804
          Cramer's V                           -0.5477

       WARNING: 100% of the cells have expected counts less than 5.
                (Asymptotic) Chi-Square may not be a valid test.


                      Pearson Chi-Square Test
                    ----------------------------------
                    Chi-Square                  2.1000
                    DF                               1
                    Asymptotic Pr >  ChiSq      0.1473
                    Exact      Pr >= ChiSq      0.4286
```

Notice the difference between the exact *p*-value (0.4286) and the asymptotic *p*-value (0.1473) in the Pearson Chi-Square Test table. Exact *p*-values tend to be larger than asymptotic *p*-values because the exact tests are more conservative.

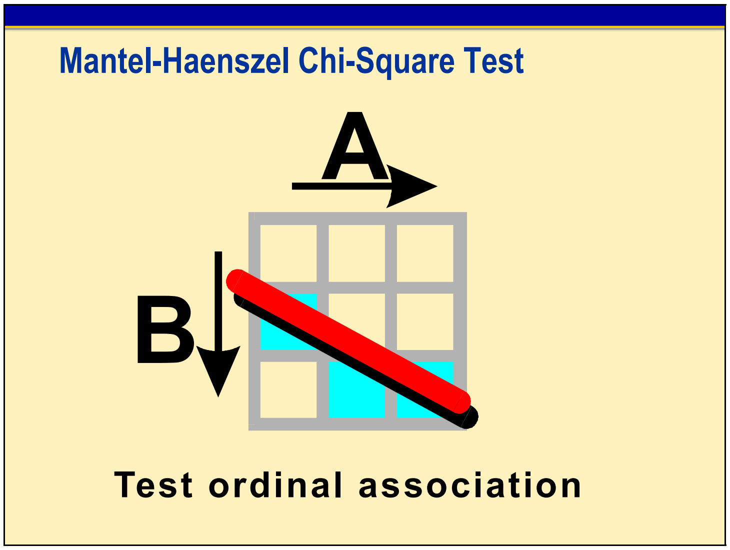
The warning message informs you that because of the small sample size, the asymptotic chi-square may not be a valid test.



Association among Ordinal Variables

Is $$$ Income associated with Purchase ?

You have already seen that **`purchase`** and **`gender`** have a significant association. Another question you can ask is whether **`purchase`** and **`income`** have a significant association. You can use the chi-square test. However, because **`income`** is ordinal and **`purchase`** can be considered ordinal, you might want to test for an ordinal association. The appropriate test for ordinal associations is the Mantel-Haenszel chi-square test.



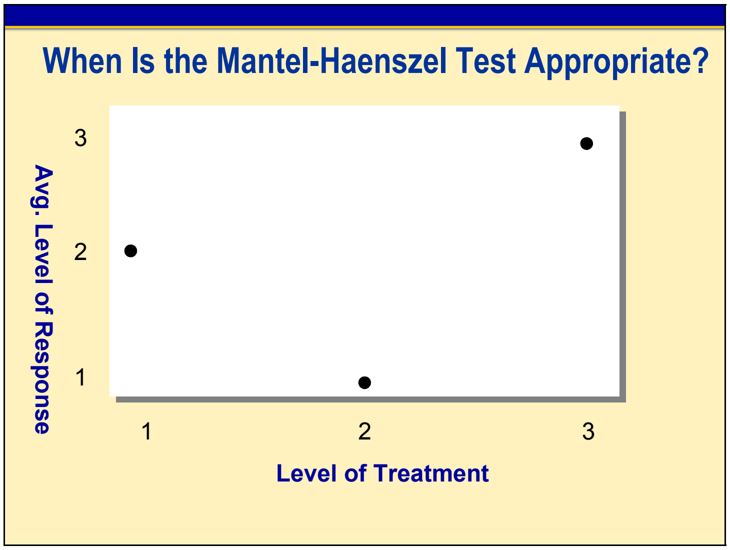
Mantel-Haenszel Chi-Square Test

Test ordinal association

The Mantel-Haenszel chi-square test is particularly sensitive to ordinal associations. An *ordinal association* implies that as one variable increases, the other variable tends to increase or decrease. For the test results to be meaningful when there are variables with more than two levels, the levels must be in a logical order.

Null hypothesis: There is no ordinal association between the row and column variables.

Alternative hypothesis: There is an ordinal association between the row and column variables.



When Is the Mantel-Haenszel Test Appropriate?

The Mantel-Haenszel test expects the effect of the predictor to be constant. That is, if the average value of the response decreases from the first level to the second level, it expects the average level of the response to decrease from the second level to the third. The above relationship would not be statistically significant under Mantel-Haenszel.

Mantel-Haenszel Chi-Square Test

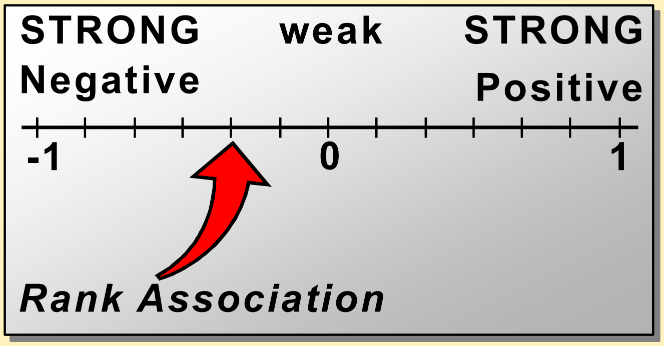
The Mantel-Haenszel chi-square test

- determines whether an ordinal association exists
- does not measure the strength of the ordinal association because it depends upon and reflects the sample size.

The Mantel-Haenszel chi-square statistic is more powerful than the general association chi-square statistic for detecting an ordinal association. The reasons are

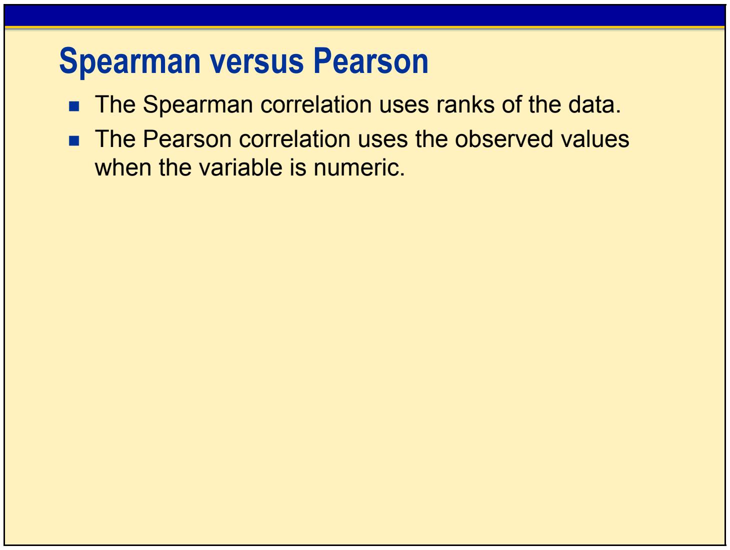
- all of the Mantel-Haenszel statistic's power is concentrated toward that objective
- the power of the general association statistic is dispersed over a greater number of alternatives.

Spearman Correlation Statistic



To measure the strength of the ordinal association, you can use the Spearman correlation statistic. This statistic

- has a range between -1 and 1
- has values close to 1, if there is a relatively high degree of positive correlation
- has values close to -1, if there is a relatively high degree of negative correlation
- is appropriate only if both variables are ordinally scaled and the values are in a logical order.



Spearman versus Pearson

- The Spearman correlation uses ranks of the data.
- The Pearson correlation uses the observed values when the variable is numeric.

The Spearman statistic can be interpreted as the Pearson correlation between the ranks on variable X and the ranks on variable Y.

Detecting Ordinal Associations

m5demo04.sas

Example: Use PROC FREQ to test whether an ordinal association exists between **purchase** and **income**. Use the variable **inclevel** and the appropriate format to ensure the income levels are in a logical order.

```
proc freq data=sasuser.b_sales_inc;
   tables inclevel*purchase / chisq measures cl;
   format inclevel incfmt. purchase purfmt.;
   title1 'Ordinal Association between INCLEVEL and PURCHASE?';
run;
```

Selected TABLES statement options:

CHISQ produces the Pearson chi-square, the likelihood-ratio chi-square, and the Mantel-Haenszel chi-square. It also produces measures of association based on chi-square such as the phi coefficient, the contingency coefficient, and Cramer's V.

MEASURES produces the Spearman correlation statistic along with other measures of association.

CL produces confidence bounds for the MEASURES statistics.

The crosstabulation is shown below.

```
                Ordinal Association between INCLEVEL and PURCHASE?

                             The FREQ Procedure

                        Table of inclevel by purchase

                  inclevel       purchase

                  Frequency     |
                  Percent       |
                  Row Pct       |
                  Col Pct       |< $100   |$100 +   |  Total
                  --------------+---------+---------+
                  Low Income    |     90  |     42  |    132
                                |  20.88  |   9.74  |  30.63
                                |  68.18  |  31.82  |
                                |  33.46  |  25.93  |
                  --------------+---------+---------+
                  Medium Income |     98  |     46  |    144
                                |  22.74  |  10.67  |  33.41
                                |  68.06  |  31.94  |
                                |  36.43  |  28.40  |
                  --------------+---------+---------+
                  High Income   |     81  |     74  |    155
                                |  18.79  |  17.17  |  35.96
                                |  52.26  |  47.74  |
                                |  30.11  |  45.68  |
                  --------------+---------+---------+
                  Total              269       162       431
                                   62.41     37.59    100.00
```

The results of the Mantel-Haenszel chi-square test are shown below.

```
Statistics for Table of inclevel by purchase

          Statistic                     DF       Value      Prob
          ------------------------------------------------------
          Chi-Square                     2     10.6404    0.0049
          Likelihood Ratio Chi-Square    2     10.5425    0.0051
          Mantel-Haenszel Chi-Square     1      8.1174    0.0044
          Phi Coefficient                       0.1571
          Contingency Coefficient               0.1552
          Cramer's V                            0.1571
```

Because the *p*-value of the Mantel-Haenszel chi-square is 0.0044, you can conclude at the 0.05 significance level that there is evidence of an ordinal association between **`purchase`** and **`income`**.

The Spearman correlation statistic is shown below.

```
           Ordinal Association between INCLEVEL and PURCHASE?

                          The FREQ Procedure

              Statistics for Table of inclevel by purchase

         Statistic                               Value       ASE
         ------------------------------------------------------
         Gamma                                  0.2324    0.0789
         Kendall's Tau-b                        0.1312    0.0454
         Stuart's Tau-c                         0.1466    0.0508

         Somers' D C|R                          0.1102    0.0382
         Somers' D R|C                          0.1562    0.0540

         Pearson Correlation                    0.1374    0.0480
         Spearman Correlation                   0.1391    0.0481

         Lambda Asymmetric C|R                  0.0000    0.0000
         Lambda Asymmetric R|C                  0.0616    0.0470
         Lambda Symmetric                       0.0388    0.0300

         Uncertainty Coefficient C|R            0.0185    0.0114
         Uncertainty Coefficient R|C            0.0112    0.0069
         Uncertainty Coefficient Symmetric      0.0139    0.0086
```

The Spearman correlation statistic indicates that there is a relatively small positive ordinal relationship between **`income`** and **`purchase`** (as **`income`** levels increase, **`purchase`** levels increase).

The ASE is the asymptotic standard error, which is what the standard error approaches as your sample size increases to infinity.

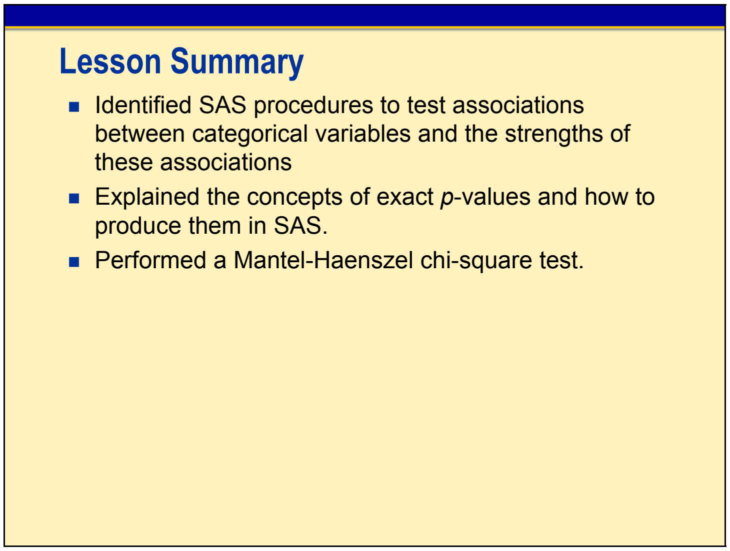
The 95% confidence bounds for the statistics are shown below.

```
                                                       95%
         Statistic                             Confidence Limits
         ───────────────────────────────────────────────────────
         Gamma                                 0.0777     0.3871
         Kendall's Tau-b                       0.0423     0.2201
         Stuart's Tau-c                        0.0471     0.2461

         Somers' D C|R                         0.0353     0.1850
         Somers' D R|C                         0.0505     0.2620

         Pearson Correlation                   0.0433     0.2315
         Spearman Correlation                  0.0449     0.2334

         Lambda Asymmetric C|R                 0.0000     0.0000
         Lambda Asymmetric R|C                 0.0000     0.1536
         Lambda Symmetric                      0.0000     0.0976

         Uncertainty Coefficient C|R           0.0000     0.0408
         Uncertainty Coefficient R|C           0.0000     0.0246
         Uncertainty Coefficient Symmetric     0.0000     0.0307

                             Sample Size = 431
```

Because the 95% confidence interval for the Spearman correlation statistic does not contain 0, the relationship is significant at the 0.05 significance level.

The confidence bounds are valid only if your sample size is large. A general guideline is to have a sample size of at least 25 for each degree of freedom in the Pearson chi-square statistic.

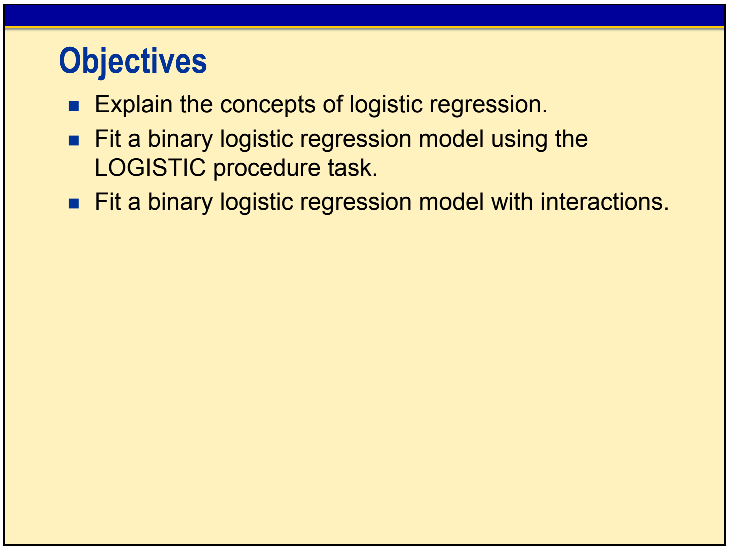
Exercise: Refer to your course workbook.



Lesson Summary

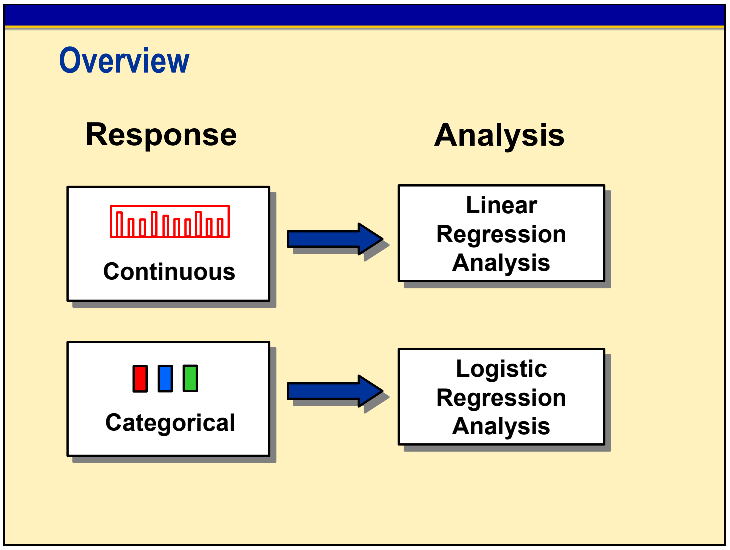
- Identified SAS procedures to test associations between categorical variables and the strengths of these associations
- Explained the concepts of exact *p*-values and how to produce them in SAS.
- Performed a Mantel-Haenszel chi-square test.

5.2 Introduction to Logistic Regression



Objectives

- Explain the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure task.
- Fit a binary logistic regression model with interactions.



Overview

Response

Analysis

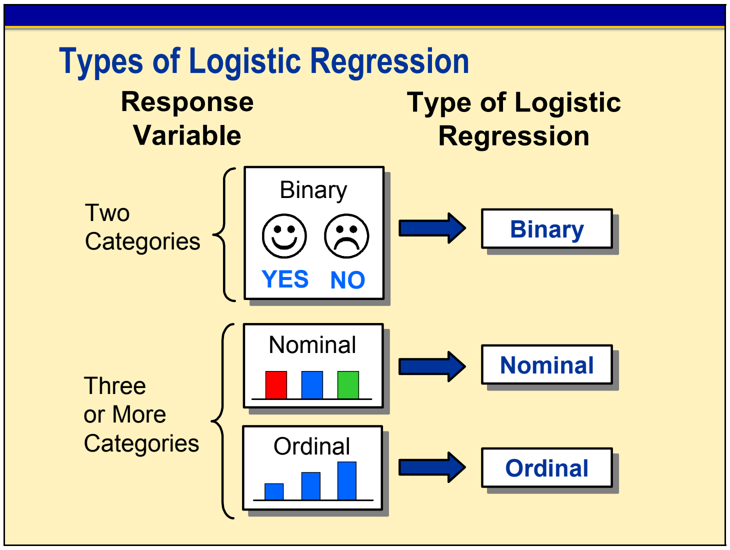
Continuous

Linear Regression Analysis

Categorical

Logistic Regression Analysis

Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical.



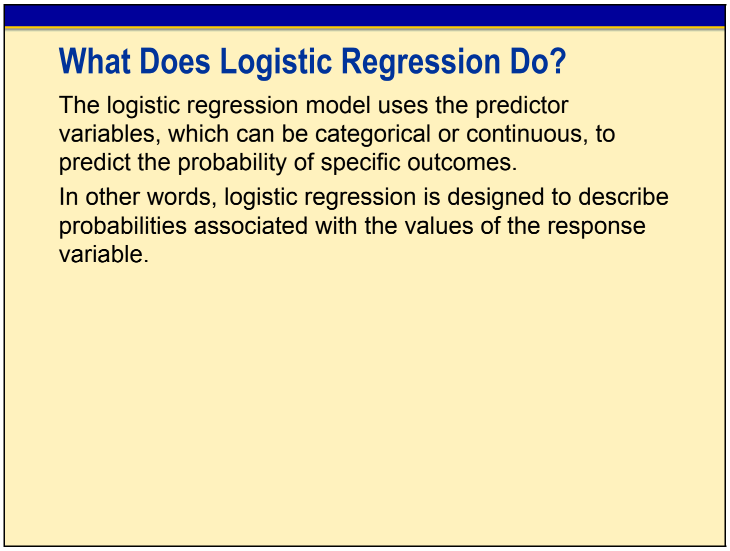
Types of Logistic Regression

If the response variable is dichotomous (two categories), the appropriate logistic regression model is binary logistic regression.

If you have more than two categories (levels) within the response variable, there are two possible logistic regression models:

1. If the response variable is nominal, you fit a nominal logistic regression.

2. If the response variable is ordinal, you fit an ordinal logistic regression



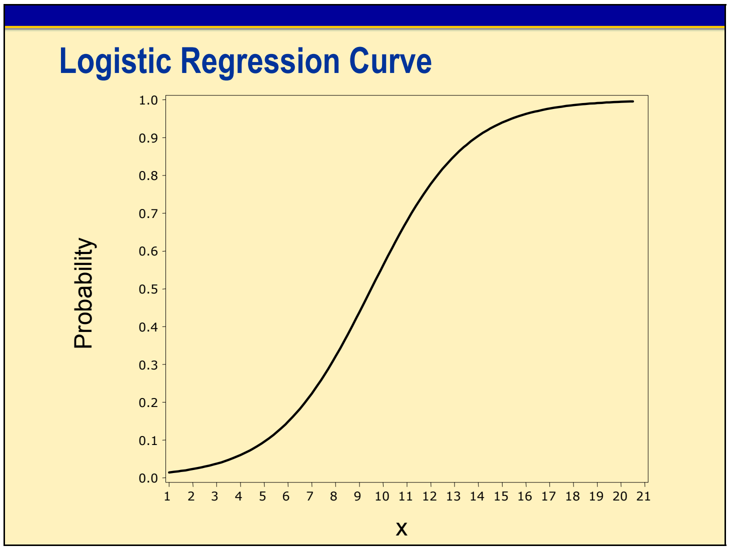
What Does Logistic Regression Do?

The logistic regression model uses the predictor variables, which can be categorical or continuous, to predict the probability of specific outcomes.

In other words, logistic regression is designed to describe probabilities associated with the values of the response variable.

Because you are modeling probabilities, a continuous linear regression model is not appropriate. One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1. Logistic regression models ensure that the estimated probabilities are between 0 and 1.

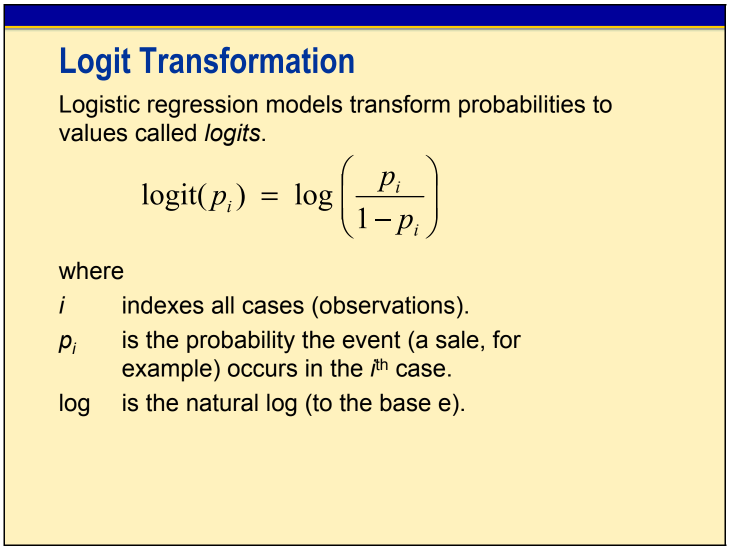
Another problem is that the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an S-shaped curve.



Logistic Regression Curve

The nonlinear relationship between the probability of the outcome and the predictor variables is solely due to the constrained scale of the probabilities. Furthermore, the relationship is fairly linear in the middle of the range of the probabilities (.20 to .80) and fairly nonlinear at the end of the range (0 to .20 and .80 to 1).

The parameter estimate of this curve determines the rate of increase or decrease of the estimated curve. When the parameter estimate is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter estimate is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter estimate increases, the curve has a steeper rate of change. When the parameter estimate is equal to 0, the curve resembles a straight line.



Logit Transformation

Logistic regression models transform probabilities to values called *logits*.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where

i indexes all cases (observations).

p_i is the probability the event (a sale, for example) occurs in the i^{th} case.

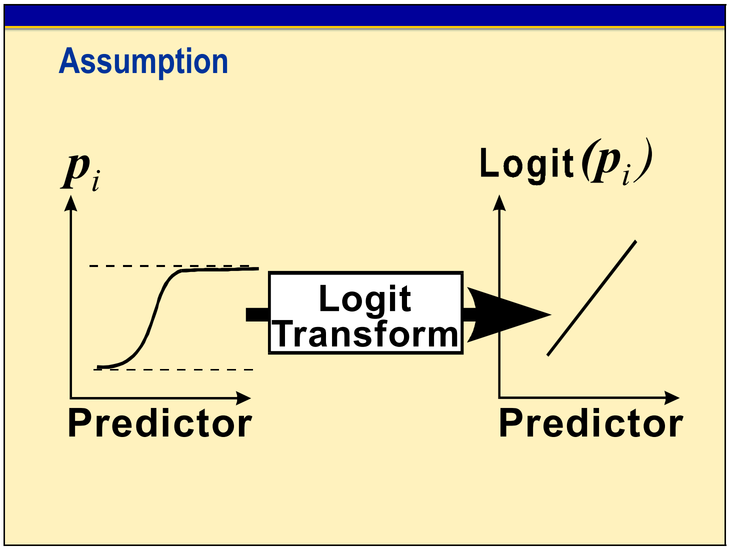
log is the natural log (to the base e).

A logistic regression model applies a transformation to the probabilities. The probabilities are transformed because the relationship between the probabilities and the predictor variable is nonlinear.

The logit transformation ensures the model generates estimated probabilities between 0 and 1.

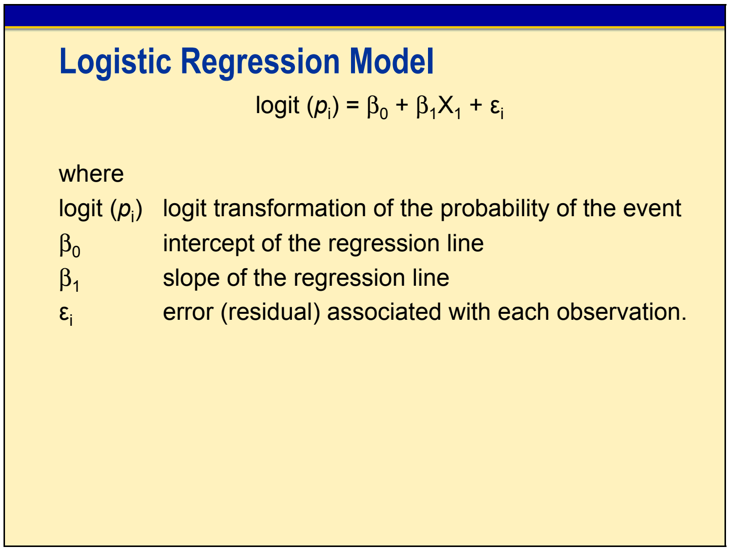
The ratio [p / (1 – p)] is also known as *odds*, and it is discussed later in this module. In the current example using **`b_sales`**, the probability of interest is whether the customer purchased $100 or more.

Assumption



Assumption in logistic regression: The logit transformation of the probabilities results in a linear relationship with the predictor variables.

To verify this assumption, it would be useful to plot the logits by the predictor variable.



Logistic Regression Model

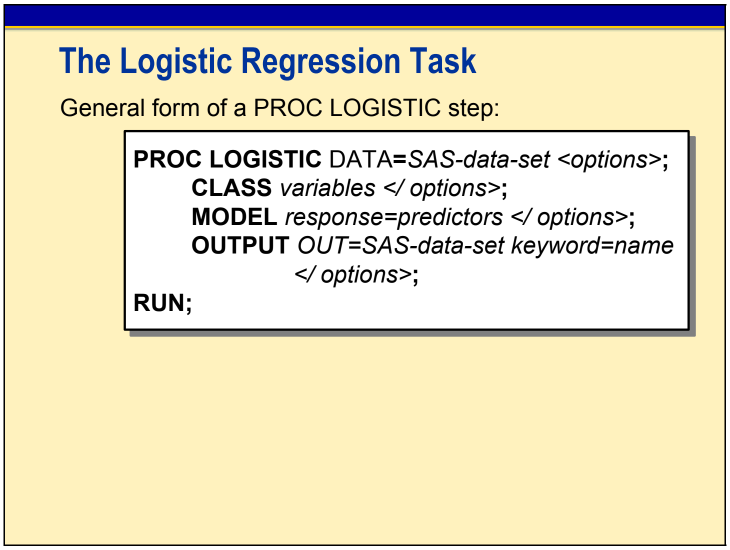
$$\text{logit}\ (p_i) = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

where

logit (p_i)	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line
ε_i	error (residual) associated with each observation.

For a binary outcome variable, the linear logistic model with one predictor variable has the form above.

Unlike linear regression, the categorical response is not normally distributed and the variances are not the same. Also, logistic regression usually requires a more complex iterative estimation method to estimate the parameters than linear regression does.



The Logistic Regression Task

General form of a PROC LOGISTIC step:

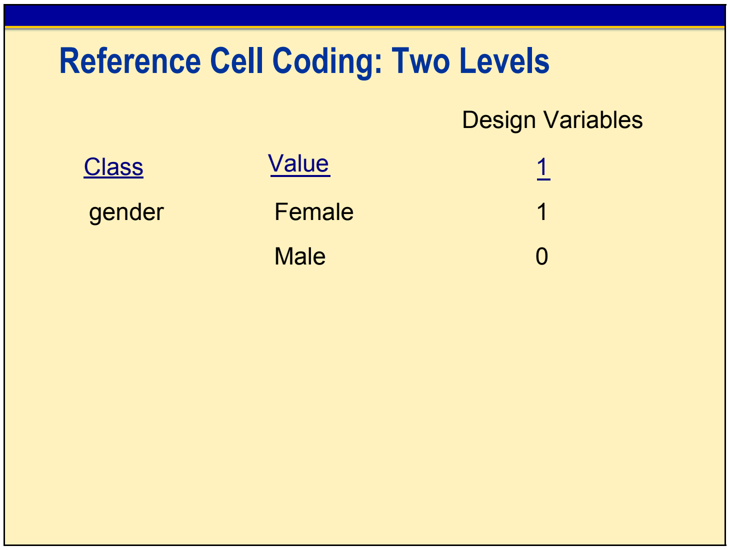
```
PROC LOGISTIC DATA=SAS-data-set <options>;
     CLASS variables </ options>;
     MODEL response=predictors </ options>;
     OUTPUT OUT=SAS-data-set keyword=name
              </ options>;
RUN;
```

Selected LOGISTIC procedure statements:

CLASS names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement.

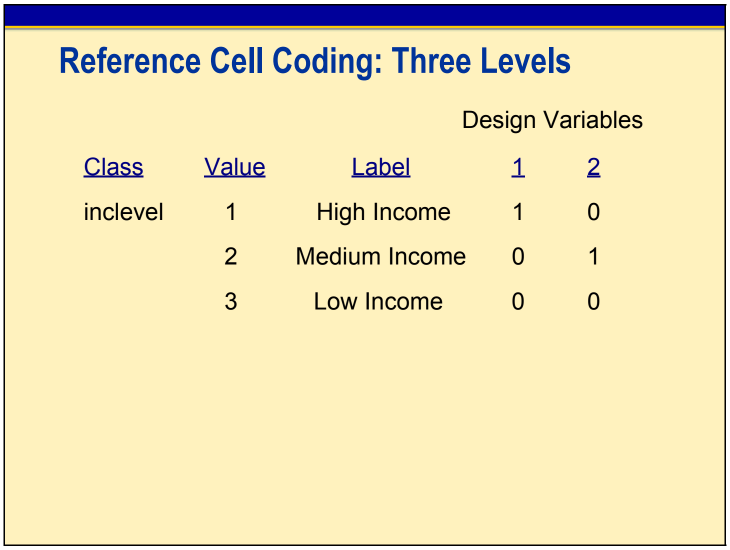
MODEL specifies the response variable and the predictor variables.

OUTPUT creates an output data set containing all the variables from the input data set and the requested statistics.



Reference Cell Coding: Two Levels

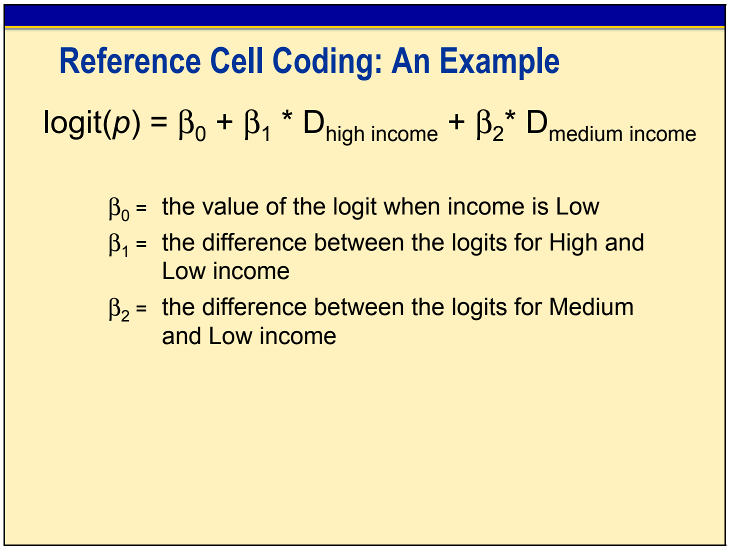
		Design Variables
Class	Value	1
gender	Female	1
	Male	0



Reference Cell Coding: Three Levels

			Design Variables	
Class	Value	Label	1	2
inclevel	1	High Income	1	0
	2	Medium Income	0	1
	3	Low Income	0	0

For *reference cell coding*, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level, called the *reference level*. For example, the effect for the level Low estimates the difference between Low and High. You can choose the reference level in the CLASS statement.



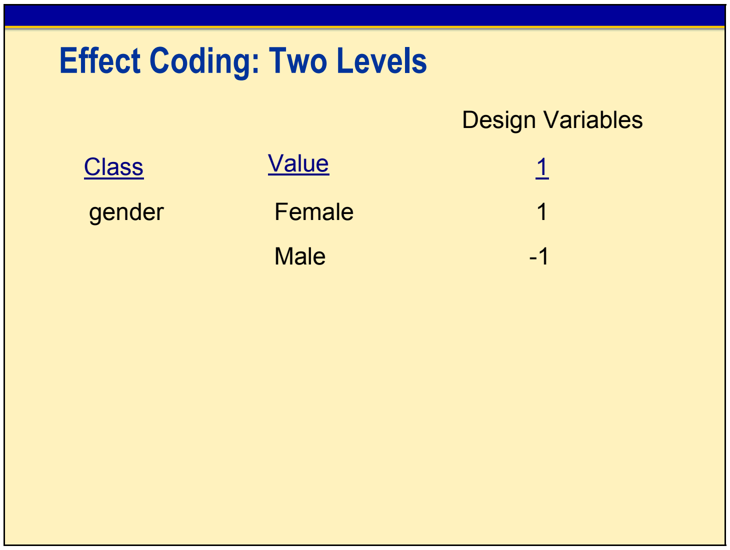
Reference Cell Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{high income}} + \beta_2 * D_{\text{medium income}}$$

β_0 = the value of the logit when income is Low

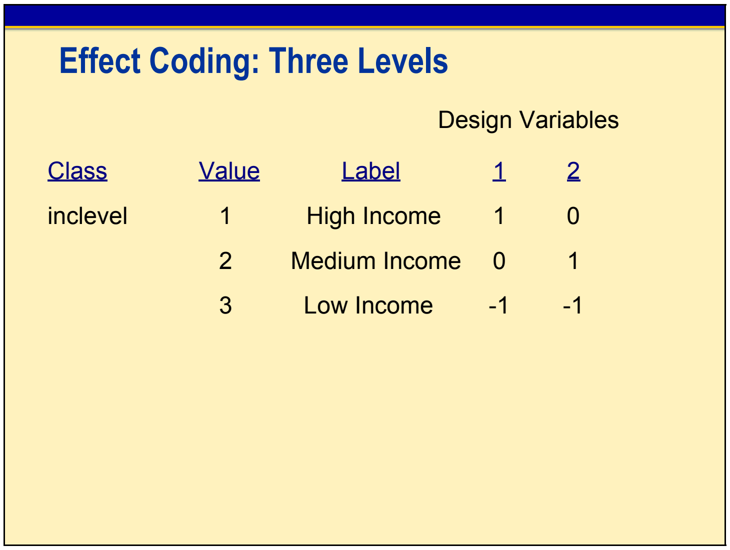
β_1 = the difference between the logits for High and Low income

β_2 = the difference between the logits for Medium and Low income



Effect Coding: Two Levels

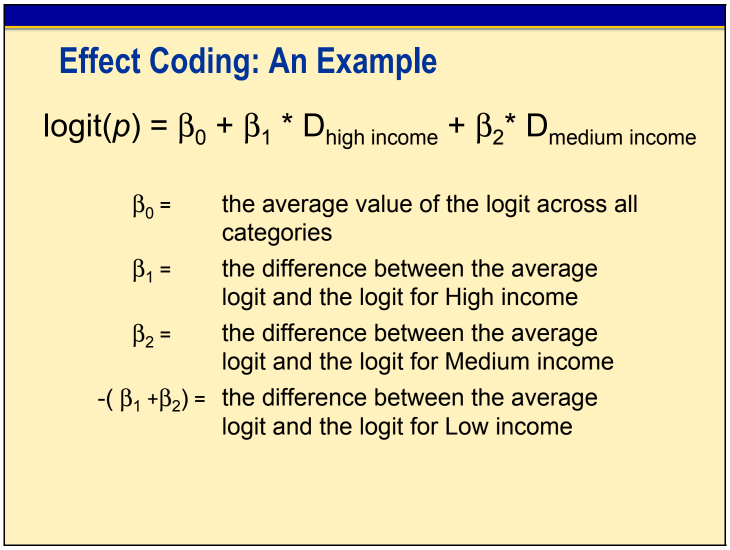
		Design Variables
Class	Value	1
gender	Female	1
	Male	-1



Effect Coding: Three Levels

			Design Variables	
Class	Value	Label	1	2
inclevel	1	High Income	1	0
	2	Medium Income	0	1
	3	Low Income	-1	-1

For *effect coding* (also called *deviation from the mean coding*), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, because the variable **`income`** has three levels, two design variables were created. For the last level of the CLASS variable (`Low`), all the design variables have a value of –1. Parameter estimates of the CLASS main effects using this coding scheme estimate the difference between the effect of each level and the average effect over all levels.



Effect Coding: An Example

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{high income}} + \beta_2 * D_{\text{medium income}}$$

β_0 = the average value of the logit across all categories

β_1 = the difference between the average logit and the logit for High income

β_2 = the difference between the average logit and the logit for Medium income

$-(\beta_1 + \beta_2)$ = the difference between the average logit and the logit for Low income

Binary Logistic Regression

m5demo05.sas

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **purchase** as the outcome variable and **gender** as the predictor variable. Specify reference cell coding and select `Male` as the reference group. Also use the DESCENDING option to model the probability of spending 100 dollars or more and request confidence intervals around the estimated odds ratios.

```
proc logistic data=sasuser.b_sales_inc descending;
   class gender (param=ref ref='Male');
   model purchase = gender / clodds=wald;
   title1 'LOGISTIC MODEL (1):  purchase = gender';
run;
```

Selected PROC LOGISTIC statement option:

DESCENDING reverses the sorting order for the levels of the response variable.

Selected CLASS statement options:

PARAM= specifies the parameterization method for the classification variable or variables. Design matrix columns are created from CLASS variables according to the following coding schemes. There are at least five codes that can be used, but two are listed below:

EFFECT specifies effect coding (default).

REFERENCE | REF specifies reference cell coding.

REF= specifies the reference level for PARAM=EFFECT or PARAM=REFERENCE.

Selected MODEL statement option:

CLODDS=WALD requests confidence intervals for the odds ratios of all predictor variables, based on the individual Wald tests.

✏ If there are numerous levels in the CLASS variable, you might want to reduce the number of levels using subject matter knowledge. This is especially important when the levels have few or no observations. This must be done in a DATA step.

Partial PROC LOGISTIC Output

```
     LOGISTIC MODEL (1):  purchase = gender

             The LOGISTIC Procedure

                Model Information

Data Set                      SASUSER.B_SALES_INC
Response Variable             purchase
Number of Response Levels     2
Number of Observations        431
Link Function                 Logit
Optimization Technique        Fisher's scoring

                 Response Profile

        Ordered                      Total
          Value     purchase     Frequency

              1            1           162
              2            0           269

             Class Level Information

                                 Design
                               Variables

         Class      Value              1

         gender     Female             1
                    Male               0
```

The Model Information table describes the data set, the response variable, the number of observations, and the link function. The *link function* is the term used to describe the transformation applied to the probabilities. For this example, the logit transformation is used. Other link functions in PROC LOGISTIC include PROBIT or NORMIT (inverse standard normal distribution function) and CLOGLOG (complementary log-log function).

The Response Profile table shows the response variable values listed according to their ordered values. By default, PROC LOGISTIC orders the response variable alphanumerically so that it bases the logistic regression model on the probability of the smallest value. Because you used the DESCENDING option, in this example the model is based on the probability of purchasing items of 100 dollars or more (PURCHASE=1).

The Response Profile table also shows the value of the response variable and the frequency.

The Class Level Information table includes the predictor variable in the CLASS statement. Because you used the PARAM=REF and REF='Male' options, this table reflects your choice of **gender**=Male as the reference level. The design variable is 1 when the value of **gender** is Female and 0 when the value is Male.

Partial PROC LOGISTIC Output (continued)

```
                          Model Convergence Status

               Convergence criterion (GCONV=1E-8) satisfied.


                            Model Fit Statistics

                                                Intercept
                                 Intercept         and
                   Criterion        Only        Covariates

                   AIC            572.649         569.951
                   SC             576.715         578.084
                   -2 Log L       570.649         565.951
```

The Model Convergence Status simply informs you that the convergence criterion was met. There are a number of options to control the convergence criterion, but the default is the gradient convergence criterion with a default value of 1E-8 (0.00000001).

The Model Fit Statistics table provides three tests: AIC is Akaike's 'A' information criterion, SC is the Schwarz criterion, and –2Log L is the –2 log likelihood. These are goodness-of-fit measures you can use to compare one model to another. Lower values indicate a more desirable model. AIC and SC adjust for the number of predictor variables and the number of observations.

```
                   Testing Global Null Hypothesis: BETA=0

           Test                 Chi-Square       DF     Pr > ChiSq

           Likelihood Ratio         4.6978        1         0.0302
           Score                    4.6672        1         0.0307
           Wald                     4.6436        1         0.0312


                        Type III Analysis of Effects

                                      Wald
                 Effect      DF    Chi-Square    Pr > ChiSq
                 gender       1        4.6436        0.0312
```

The Testing Global Null Hypothesis: BETA=0 table provides three statistics to test the null hypothesis that all regression coefficients of the model are 0.

A significant *p*-value for the Likelihood Ratio test provides evidence that at least one of the regression coefficients for an explanatory variable is nonzero. (In this example the *p*-value is 0.0302, which is significant at the .05 level.) This statistic is similar to the overall *F* test in linear regression. The Score and Wald tests are also used to test whether all the regression coefficients are 0.

The Type III Analysis of Effects table is generated when a predictor variable is used in the CLASS statement. The listed effect (variable) is tested using the Wald chi-square statistic (in this example, 4.6436 with a *p*-value of 0.0312). This analysis is similar to the individual *t*-test in the REG procedure. Because **`gender`** is the only variable in the CLASS statement, the value listed in the table will be identical to the Wald test in the Testing Global Null Hypothesis table.

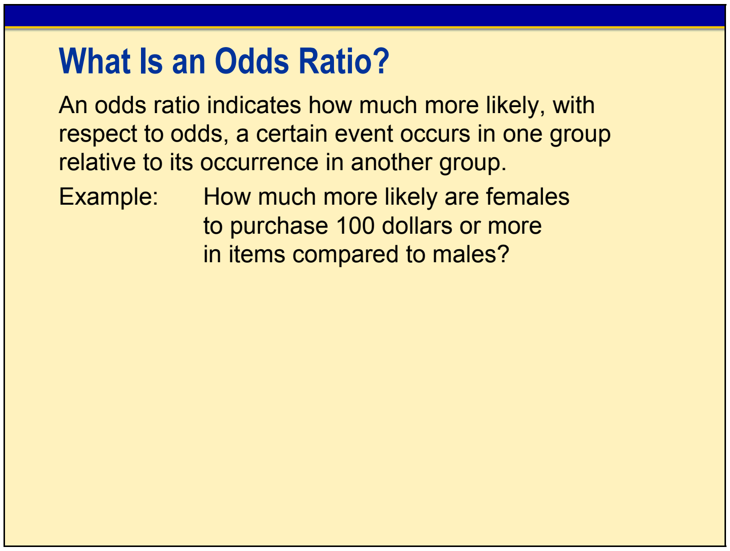
A reference for AIC can be found in D. F. Findley and E. Parzen, "A Conversation with Hirotugu Akaike," *Statistical Science,* Vol. 10, No. 1 (1995): 104-117.

The Analysis of Maximum Likelihood Estimates table lists the estimated model parameters, their standard errors, Wald tests, and odds ratios.

```
                  Analysis of Maximum Likelihood Estimates

                                            Standard
Parameter           DF     Estimate        Error     Chi-Square    Pr > ChiSq

Intercept            1      -0.7566       0.1552        23.7700        <.0001
gender     Female    1       0.4373       0.2029         4.6436        0.0312
```

The parameter estimates are the estimated coefficients of the fitted logistic regression model. The logistic regression equation is logit($\hat{p}$)=–0.7566 + 0.4373***`gender`**, for this example.

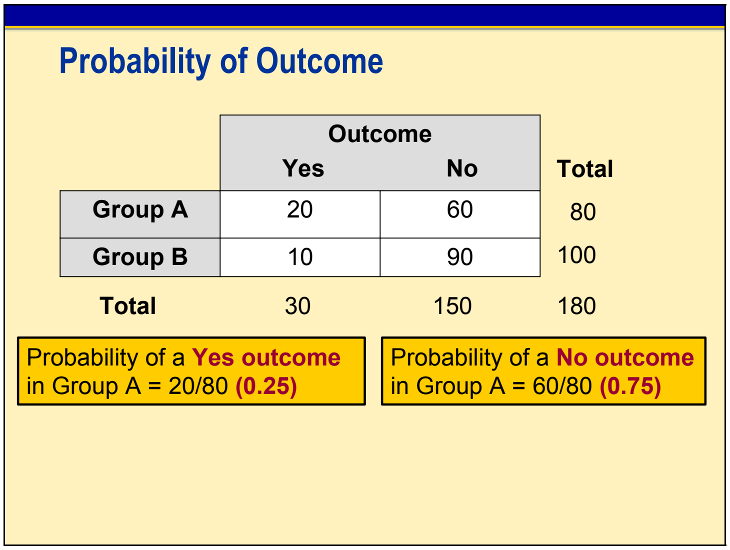
The Wald chi-square, and its associated *p*-value, tests whether the parameter estimate is significantly different from 0. For this example, both the *p*-values for the intercept and the variable **`gender`** are significant at the 0.05 significance level.



What Is an Odds Ratio?

An odds ratio indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Example: How much more likely are females to purchase 100 dollars or more in items compared to males?



Probability of Outcome

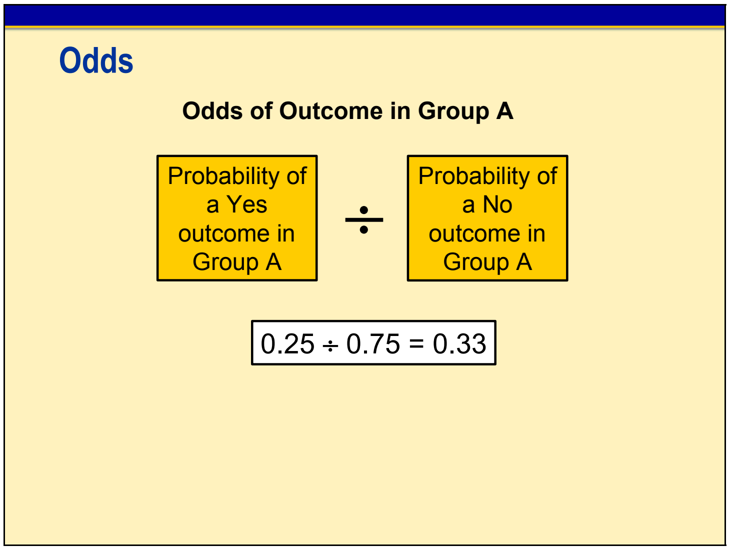
	Outcome		
	Yes	No	Total
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Probability of a **Yes outcome** in Group A = 20/80 **(0.25)**

Probability of a **No outcome** in Group A = 60/80 **(0.75)**

You have a probability of 0.25 of getting the outcome in group A.

What is the probability of getting the outcome in group B?



Odds

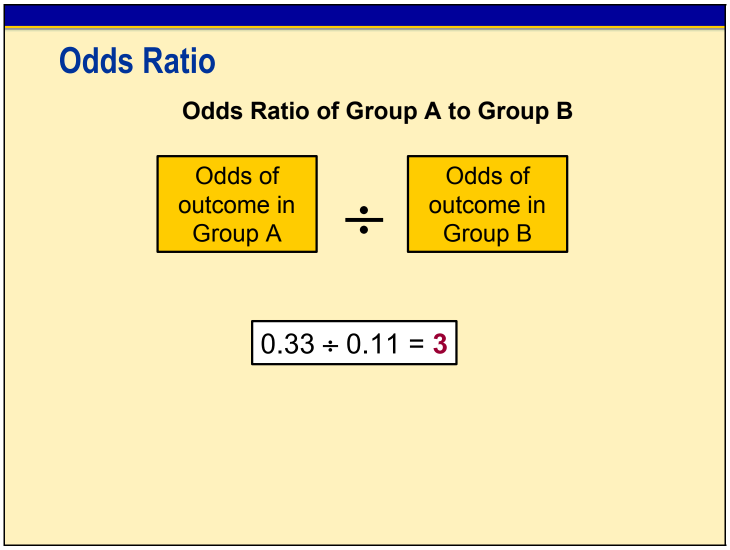
Odds of Outcome in Group A

Probability of a Yes outcome in Group A ÷ Probability of a No outcome in Group A

0.25 ÷ 0.75 = 0.33

The odds of an outcome is the ratio of the expected number of times that the outcome will occur to the expected number of times the outcome will **not** occur. In other words, the odds is simply the ratio of the probability of the outcome to the probability of no outcome. The odds for group A equals 0.33 indicating that you expect only 1/3 as many occurrences as non-occurrences in group A.

What are the odds of getting the outcome in group B?



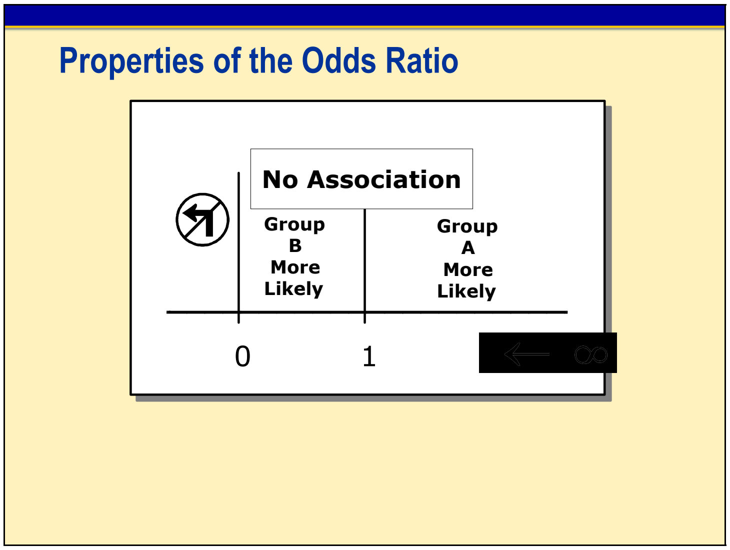
Odds Ratio

Odds Ratio of Group A to Group B

Odds of outcome in Group A ÷ Odds of outcome in Group B

0.33 ÷ 0.11 = **3**

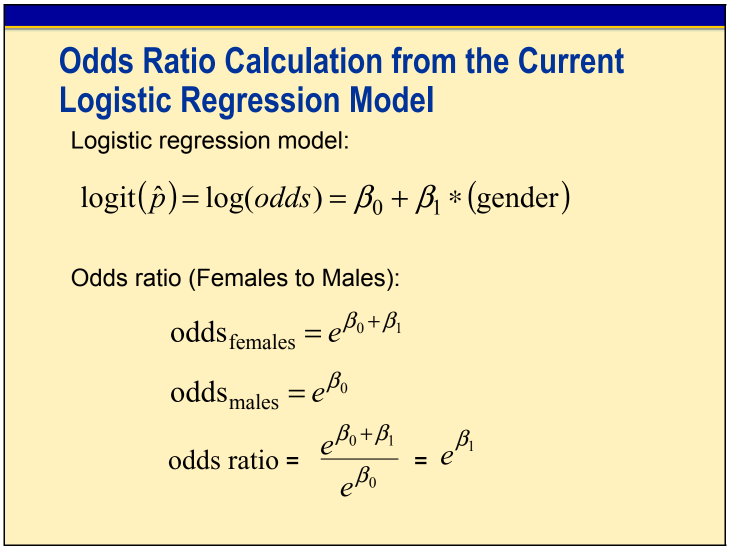
The odds ratio of group A to B equals 3, indicating that the odds of getting the outcome in group A is 3 times the odds in group B.



Properties of the Odds Ratio

No Association

The odds ratio shows the strength of the association between the predictor variable and the outcome variable. If the odds ratio is 1, then there is no association between the predictor variable and the outcome. If the odds ratio is greater than 1, then group A is more likely to have the outcome. If the odds ratio is less than 1, then group B is more likely to have the outcome. For example, an odds ratio of 3 indicates that the odds of getting the outcome in group A is 3 times that in group B.



Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\operatorname{logit}(\hat{p}) = \log(odds) = \beta_0 + \beta_1 * (\text{gender})$$

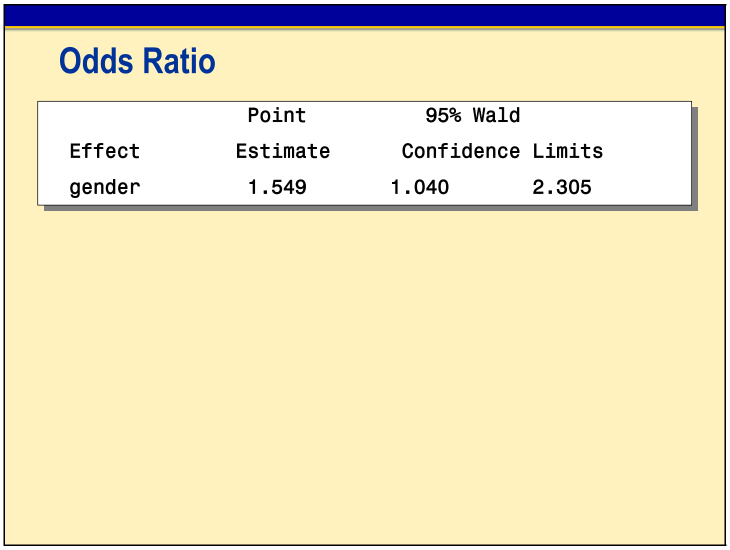
Odds ratio (Females to Males):

$$\text{odds}_{\text{females}} = e^{\beta_0 + \beta_1}$$

$$\text{odds}_{\text{males}} = e^{\beta_0}$$

$$\text{odds ratio} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

The odds ratio is computed by exponentiating the parameter estimate for the predictor variable.



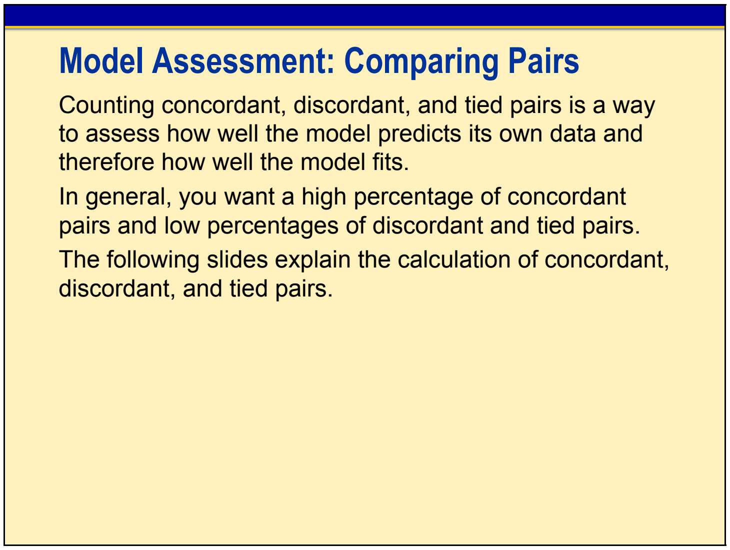
Odds Ratio

```
                  Point           95% Wald
Effect          Estimate      Confidence Limits
gender            1.549       1.040        2.305
```

The odds ratio indicates that females are 1.55 times more likely to purchase 100 dollars or more in items than males.

The 95% confidence limits indicate that you are 95% confident that the true odds ratio is between 1.04 and 2.31. Because the 95% confidence interval does not include 1.00, the odds ratio is significant at the .05 significance level.

If you want a different significance level for the confidence intervals, you can use the ALPHA= option in the MODEL statement. The value must be between 0 and 1. The default value of .05 results in the calculation of a 95% confidence interval.

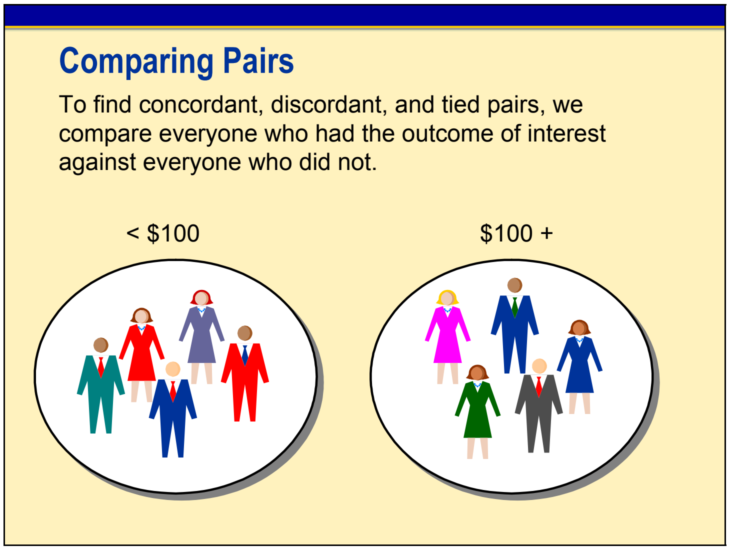


Model Assessment: Comparing Pairs

Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits.

In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.

The following slides explain the calculation of concordant, discordant, and tied pairs.

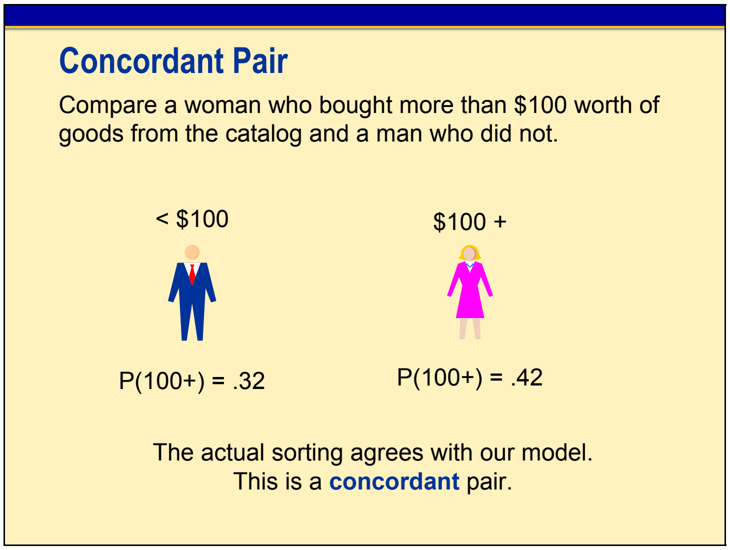


Comparing Pairs

To find concordant, discordant, and tied pairs, we compare everyone who had the outcome of interest against everyone who did not.

< $100

$100 +



Concordant Pair

Compare a woman who bought more than $100 worth of goods from the catalog and a man who did not.

< $100

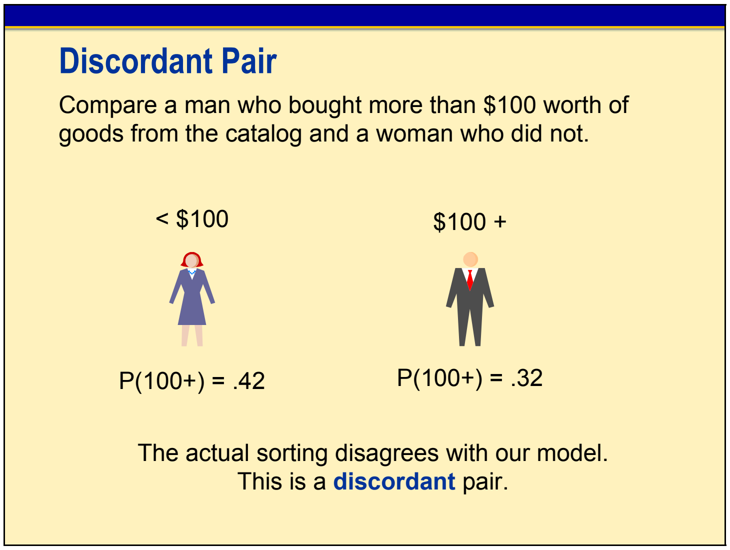
$100 +

P(100+) = .32

P(100+) = .42

The actual sorting agrees with our model.
This is a **concordant** pair.

For all pairs of observations with different values of the response variable, a pair is *concordant* if the observation with the outcome has a **higher** predicted outcome probability (based on the model) than the observation without the outcome.



Discordant Pair

Compare a man who bought more than $100 worth of goods from the catalog and a woman who did not.

< $100

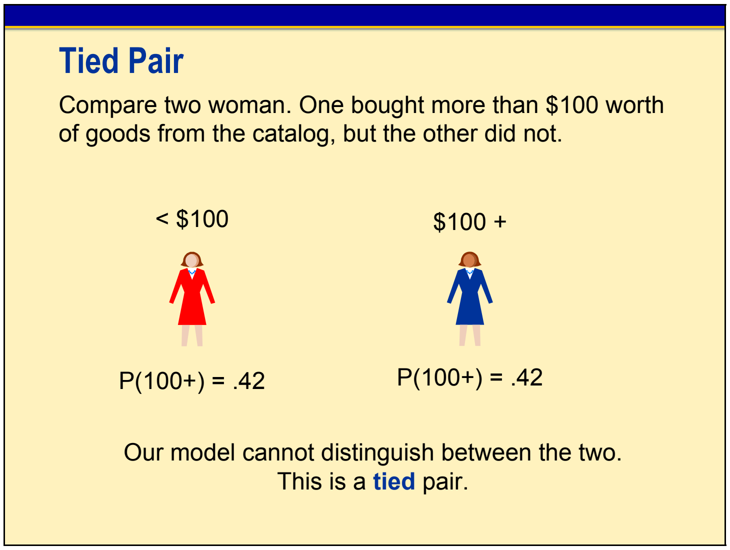
$100 +

P(100+) = .42

P(100+) = .32

The actual sorting disagrees with our model.
This is a **discordant** pair.

A pair is *discordant* if the observation with the outcome has a **lower** predicted outcome probability than the observation without the outcome.

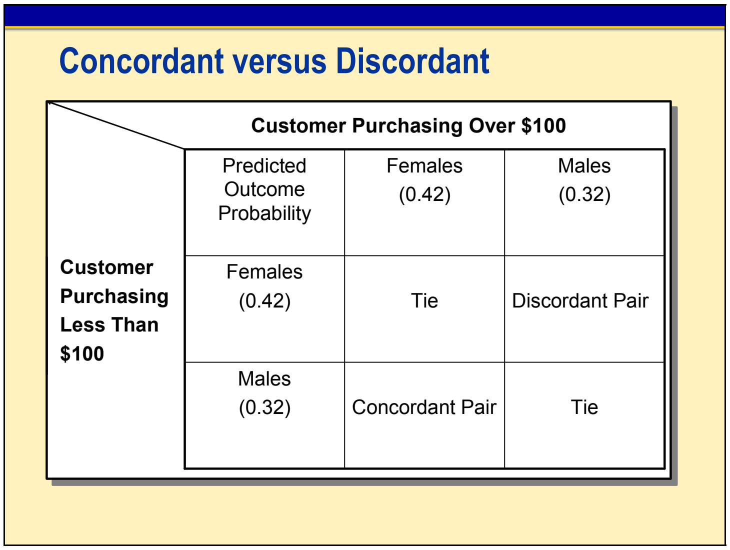


Tied Pair

Compare two woman. One bought more than $100 worth of goods from the catalog, but the other did not.

Our model cannot distinguish between the two.
This is a **tied** pair.

A pair is *tied* if it is neither concordant nor discordant.

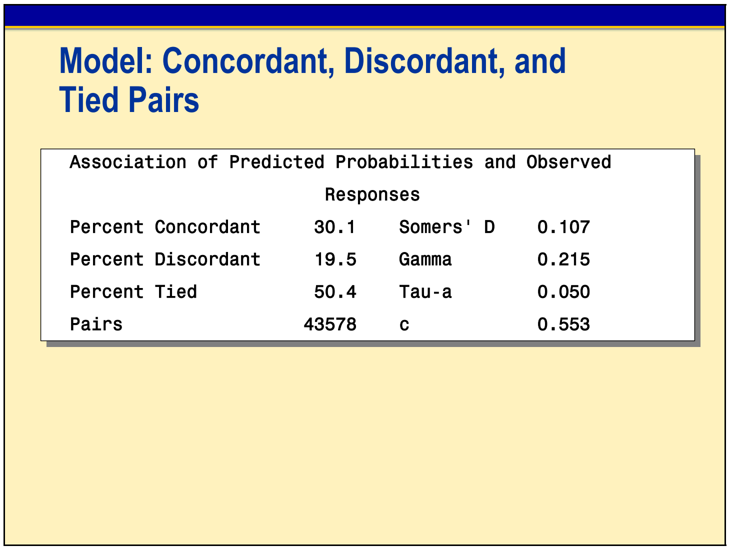


Concordant versus Discordant

		Customer Purchasing Over $100	
	Predicted Outcome Probability	Females (0.42)	Males (0.32)
Customer Purchasing Less Than $100	Females (0.42)	Tie	Discordant Pair
	Males (0.32)	Concordant Pair	Tie

This table shows the difference between discordant and concordant pairs. Because the predictor variable (**`gender`**) has only two levels, there are only two predicted outcome probabilities for purchasing items of 100 dollars or more (`female`=.42 and `male`=.32). For all pairs of observations with different outcomes (making purchases of 100 dollars or more versus making purchases of less than 100 dollars), a comparison is made of the predicted outcome probabilities. If the observation with the outcome (in this case making purchases of 100 dollars or more) has a higher predicted outcome probability compared to an observation without the outcome, the pair is concordant. However, if the observation with the outcome has a lower predicted outcome probability compared to the predicted outcome probability of an observation without the outcome, the pair is discordant. If the predicted outcome probabilities are tied, the pair is tied.

In more complex models, there are more than two predicted outcome probabilities. However, the same comparisons are made across all pairs of observations with different outcomes.



Model: Concordant, Discordant, and Tied Pairs

```
Association of Predicted Probabilities and Observed
                      Responses
Percent Concordant     30.1    Somers' D    0.107
Percent Discordant     19.5    Gamma        0.215
Percent Tied           50.4    Tau-a        0.050
Pairs                 43578    c            0.553
```

Goodness of Fit Statistics

The Association of Predicted Probabilities and Observed Responses table lists several measures of association to help you assess the predictive ability of the logistic model.

Concordant represents the percentage of concordant pairs of observations.

Discordant represents the percentage of discordant pairs of observations.

Tied represents the percentage of tied pairs of observations.

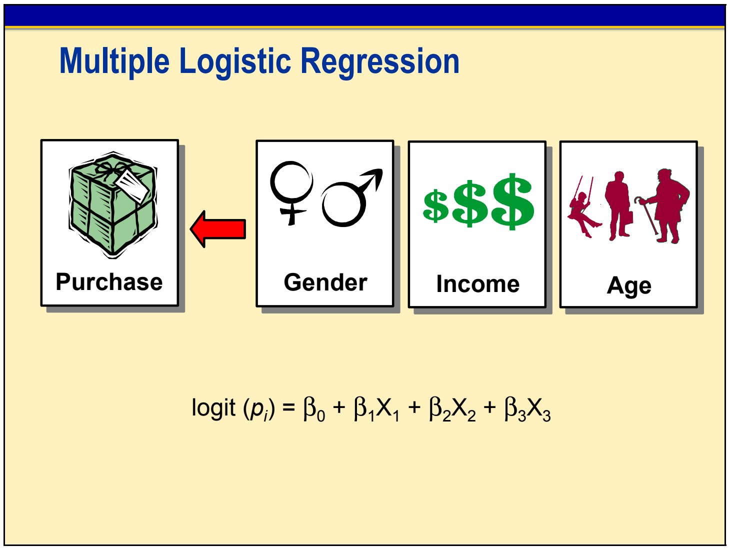
You can use these percentages as goodness-of-fit measures to compare one model to another. In general, higher percentages of concordant pairs and lower percentages of discordant pairs indicate a more desirable model.

The Association of Predicted Probabilities and Observed Responses table also shows the number of observation pairs upon which the percentages are based. For this example, there are 162 observations with an outcome of `100 dollars or more` and 269 observations with an outcome of `Under 100 dollars`. This creates 162*269 = 43578 pairs of observations with different outcome values.

Measures of Prediction

The four rank correlation indexes (Somer's D, Gamma, Tau-a, and *c*) are computed from the numbers of concordant, discordant, and tied pairs of observations. The difference between them is how they treat the tied pairs. In general, a model with higher values for these indexes has better predictive ability than a model with lower values for these indexes.

Exercise: Refer to your course workbook.



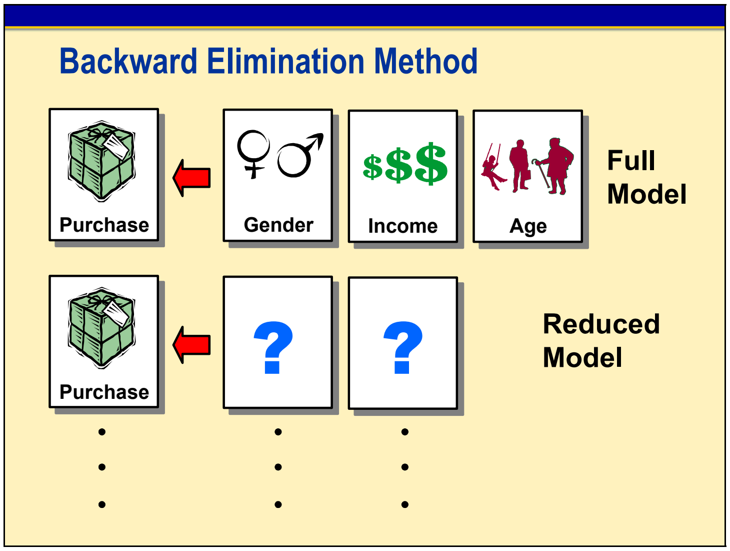
Multiple Logistic Regression

Purchase ← Gender, Income, Age

$$\text{logit}\,(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

In multiple logistic regression models, several continuous or categorical predictor variables are trying to explain the variability of the response variable. The goal in multiple logistic regression is similar to that in linear multiple regression. Find the best subset of variables by eliminating unnecessary ones. Models that are parsimonious, or simple, are more likely to be numerically stable and easier to generalize.

If you have a large number of variables, you might need to try a variable reduction method such as variable clustering before modeling with logistic regression.



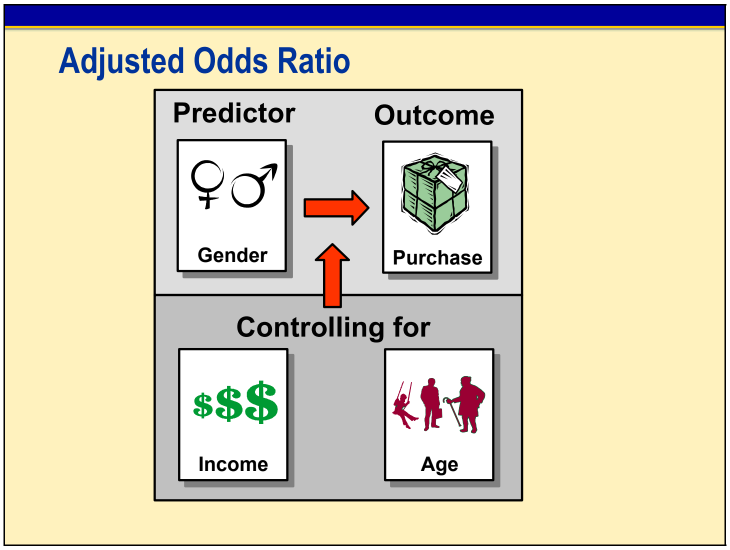
Backward Elimination Method

Purchase ← Gender, Income, Age — Full Model

Purchase ← ?, ? — Reduced Model

One way to eliminate unnecessary terms in a model is the *backward elimination method*. Backward logistic regression in SAS begins by fitting the full model with all the main effects. It then eliminates the nonsignificant parameter estimates one at a time, starting with the least significant term (the one with the largest *p*-value). The final model should only have significant main effects.

The significance level you choose depends on how much evidence you need in the significance of the predictor variables. The smaller your significance level, the more evidence you need to keep the predictor variable (in other words, the smaller the *p*-value has to be to keep the predictor variable).



Adjusted Odds Ratio

Predictor

Outcome

Gender

Purchase

Controlling for

Income

Age

One major difference between a model with one predictor variable and a model with more than one predictor variable is that the reported odds ratios are now adjusted odds ratios.

Adjusted odds ratios measure the effect between a predictor variable and a response variable, while holding all the other predictor variables constant. In other words, the levels of the predictor variables would remain the same across the observations.

For example, the odds ratio for the variable **`gender`** would measure the effect of **`gender`** on **`purchase`** while holding **`income`** and **`age`** constant (all observations are held at the same income and at the same age).

The assumption is that the odds ratio for **`gender`** is the same regardless of the level of **`income`** or **`age`**. If that assumption is not true, you have an interaction. This is discussed later in the module.

Multiple Logistic Regression

m5demo06.sas

Example: Fit a multiple logistic regression model using the backward elimination method. The full model should include all the main effects.

```
proc logistic data=sasuser.b_sales_inc descending;
   class gender (param=ref ref='Male')
         income (param=ref ref='Low');
   model purchase = gender age income / selection=backward;
   title1 'LOGISTIC MODEL (2):  purchase = gender age income';
run;
```

Because **income** is a character variable, it has been added to the CLASS statement using the PARAM=REF and REF='Low' options to choose Low as the reference group.

Selected MODEL statement option:

SELECTION= specifies the method to select the variables in the model. BACKWARD requests backward elimination, FORWARD requests forward selection, NONE fits the complete model specified in the MODEL statement, STEPWISE requests stepwise selection, and SCORE requests best subset selection. The default is NONE.

✎ The default significance level for the backward elimination method is .05. If you want to change the significance level, you can use the SLSTAY= option in the MODEL statement. Values must be between 0 and 1.

The Model Information and Response Profile of the PROC LOGISTIC output is the same as the first model, but the title has been changed to reflect the new model.

```
              LOGISTIC MODEL (2):  purchase = gender age income

                           The LOGISTIC Procedure

                              Model Information

              Data Set                      SASUSER.B_SALES_INC
              Response Variable             purchase
              Number of Response Levels     2
              Number of Observations        431
              Link Function                 Logit
              Optimization Technique        Fisher's scoring


                               Response Profile

                      Ordered                      Total
                        Value     purchase     Frequency

                            1            1           162
                            2            0           269
```

PROC LOGISTIC identifies the chosen BACKWARD selection method, and then provides a Class Level Information table. The variable **`income`** has been added to this table, and because there are three levels, two design variable columns are displayed. You have chosen `Low` as the reference value using the PARAM=REF and REF=’Low’ options in the CLASS statement. PROC LOGISTIC has generated two Design Variables for the three levels of **`income`**. Design Variable 1 will be 1 when the value of **`income`** is `High` and will be 0 when **`income`** is `Low` or `Medium`. Design variable 2 will be 1 when the value of **`income`** is `Medium` and 0 when **`income`** is `High` or `Low`.

```
                        Backward Elimination Procedure


                           Class Level Information

                                       Design
                                      Variables

                        Class      Value       1      2

                        gender     Female      1
                                   Male        0

                        income     High        1      0
                                   Low         0      0
                                   Medium      0      1
```

The next part of the output shows the backward elimination process in PROC LOGISTIC. At Step 0, the intercept and three predictor variables are entered into the model. The Model Fit Statistics and Testing Global Null Hypothesis tables are presented for this step.

```
Step  0. The following effects were entered:

Intercept  gender  age  income


                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.


                          The LOGISTIC Procedure

                           Model Fit Statistics

                                              Intercept
                               Intercept         and
                  Criterion       Only        Covariates

                  AIC             572.649        562.208
                  SC              576.715        582.539
                  -2 Log L        570.649        552.208


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        18.4410        4         0.0010
          Score                   18.2729        4         0.0011
          Wald                    17.6172        4         0.0015
```

At Step 1, the variable **age** was removed from the model and the Model Fit Statistics and Testing Global Null Hypothesis tables are updated.

```
Step  1. Effect age is removed:


                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.

                           Model Fit Statistics

                                                Intercept
                                 Intercept         and
                  Criterion        Only        Covariates

                  AIC              572.649       562.190
                  SC               576.715       578.454
                  -2 Log L         570.649       554.190


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        16.4592        3         0.0009
          Score                   16.3718        3         0.0010
          Wald                    15.8824        3         0.0012


                         Residual Chi-Square Test

                    Chi-Square       DF     Pr > ChiSq

                        1.9836        1         0.1590
```

The Residual Chi-Square Test table displays the joint significance of the variables not in the model (in this case, **age**). This score chi-squared statistic has an asymptotic chi-squared distribution with the degrees of freedom being the difference between the full and reduced models.

When the selection process is complete, a note states that no additional variables met the specified significance level for removal from the model, and a Summary of Backward Elimination table is generated.

```
NOTE: No (additional) effects met the 0.05 significance level for removal
      from the model.


                      Summary of Backward Elimination

                Effect               Number          Wald
        Step    Removed      DF          In    Chi-Square    Pr > ChiSq

           1    age           1           2        1.9729        0.1601


                       Type III Analysis of Effects

                                      Wald
                Effect      DF    Chi-Square    Pr > ChiSq

                gender       1        5.8211        0.0158
                income       2       11.6669        0.0029
```

In the next part of the output, the Summary of Backward Elimination table lists the step number, the name of each predictor variable (effect) that is removed from the model at each step, degrees of freedom, the number of the predictor variable in the MODEL statement, the Wald Chi-Square statistic for each variable, and the corresponding *p*-value upon which each variable's removal from the model is based.

The Type III Analysis of Effects table for this model indicates that the coefficients for **gender** and **income** are statistically different from 0 at the 0.05 level of significance. Note that **income** has two degrees of freedom, because it has three levels.

```
                Analysis of Maximum Likelihood Estimates

                                          Standard
Parameter           DF    Estimate        Error    Chi-Square    Pr > ChiSq

Intercept            1     -1.1125       0.2403       21.4255        <.0001
gender    Female     1      0.5040       0.2089        5.8211        0.0158
income    High       1      0.7605       0.2515        9.1447        0.0025
income    Medium     1      0.0963       0.2628        0.1342        0.7141
```

The Analysis of Maximum Likelihood Estimates table is now examined. The *p*-value for **gender**=Female (0.0158) indicates that its coefficient is statistically different from 0 at the 0.05 level of significance. In addition, you can also state that females and males are statistically different from one another in terms of purchasing 100 dollars or more.

The coefficient for **income**=High is also statistically different from 0, based on its *p*-value (0.0025). Because **income**=Low is the reference group, you can state that high- and low-income people are statistically different from one another with respect to purchasing 100 dollars or more. When examining **income**=Medium, the *p*-value of 0.7141 indicates that this coefficient is not different from 0. Again, because Low is the reference group, you can state that medium- and low-income people are not statistically different and have similar purchasing trends.

What action can you take at this point? If your analysis goal is building predictive models, you can write a DATA step to, in essence, collapse the Low and Medium observations into a single group. The new variable (**highinc**) would be equal to High when **income**=High, or Low/Medium otherwise. You would then replace **income** in the MODEL statement with **highinc** and execute PROC LOGISTIC again. Remember to correctly interpret the coefficient of **highinc**.

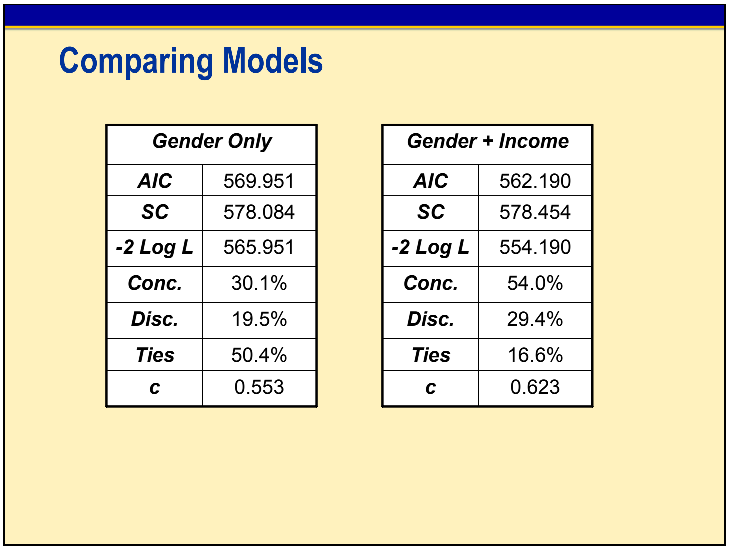
```
                    Odds Ratio Estimates

                              Point          95% Wald
 Effect                    Estimate      Confidence Limits

 gender Female vs Male        1.655       1.099       2.493
 income High vs Low           2.139       1.307       3.502
 income Medium vs Low         1.101       0.658       1.843


Association of Predicted Probabilities and Observed Responses

      Percent Concordant     54.0    Somers' D    0.246
      Percent Discordant     29.4    Gamma        0.295
      Percent Tied           16.6    Tau-a        0.116
      Pairs                 43578    c            0.623
```

The last part of the output provides the Odds Ratio Estimates table as well as the Association of Predicted Probabilities and Observed Responses table.

The effects for **`gender Female vs Male`** and **`income High vs Low`** both indicate that they are statistically significant at the 0.05 level because their 95% Wald Confidence Intervals do not include 1.000. Note that the 95% confidence interval for **`income Medium vs Low`** is not significant. The interval (0.658, 1.843) includes 1.000.

When you compare the percentages of this model with the previous model where **`gender`** was the only predictor variable, the concordant percentage increased (from 30.1 to 54.0), but the discordant percentage also increased (from 19.5 to 29.4). The tied percentage showed the most change, decreasing from 50.4 to 16.6.

The *c* statistic increased (0.553 to 0.623) from the simple **`gender`** model, which is desirable.

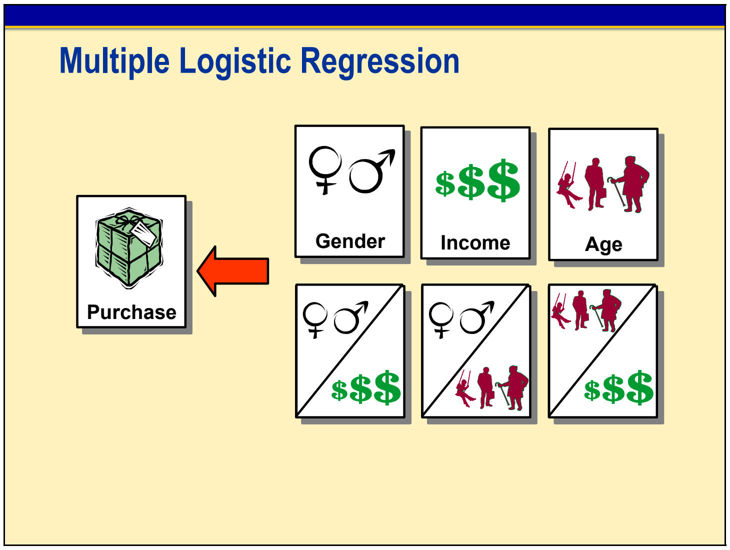


Comparing Models

Gender Only	
AIC	569.951
SC	578.084
-2 Log L	565.951
Conc.	30.1%
Disc.	19.5%
Ties	50.4%
c	0.553

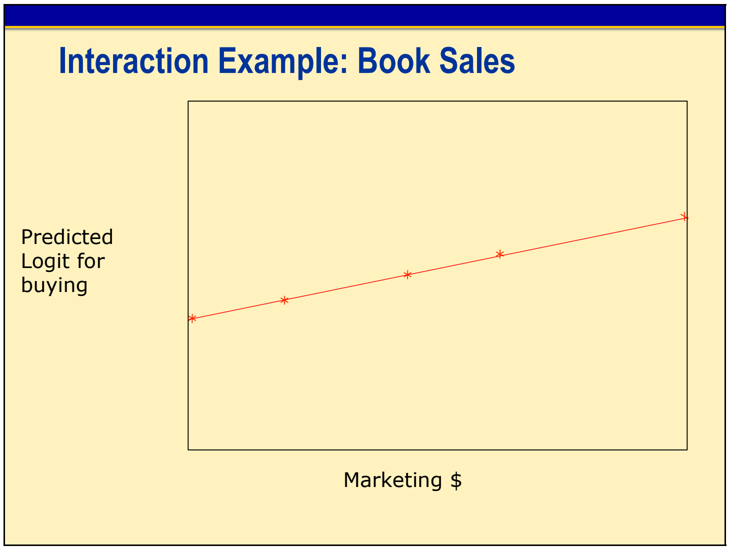
Gender + Income	
AIC	562.190
SC	578.454
-2 Log L	554.190
Conc.	54.0%
Disc.	29.4%
Ties	16.6%
c	0.623

Adding income to the model decreases the AIC and the SC, and it increases the number of concordant pairs. Although discordant pairs increased, tied pairs decreased. Adding `income` improves the model.



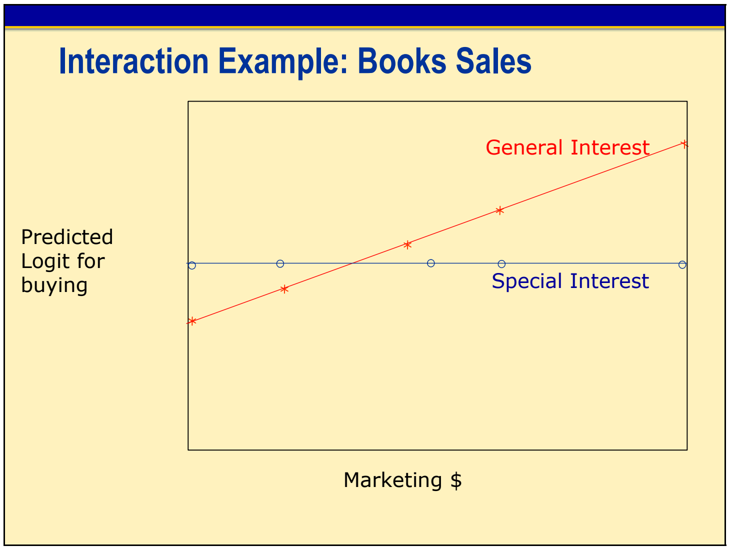
Multiple Logistic Regression

In the last example, a multiple logistic regression model was fitted with only the main effects (just predictor variables are in the model). Thus, you are assuming that the effect of each variable on the outcome is the same regardless of the levels of the other variables. For example, you are assuming that the effect of `gender` (Female to Male) on the probability of making purchases of 100 dollars or more is the same regardless of `income` level. If this assumption is not correct, you might want to fit a more complex model that has interactions.



Interaction Example: Book Sales

The above example assumes that one dollar of marketing money has the same effect for all books.

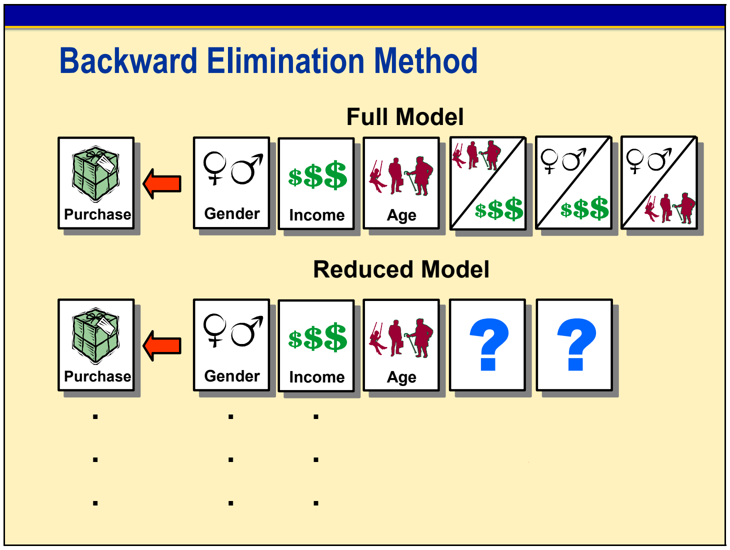


Interaction Example: Books Sales

However, if you consider the type of book to be sold, there seems to be a difference in the effect of marketing dollars on general interest books versus special interest. This is called an interaction. An *interaction* between two variables A and B is said to occur when the effect of A on the outcome depends on the observed level of B, or when the effect of B on the outcome depends on the observed level of A.

In the example above, the effect of **`marketing`** depends on the level of **`booktype`**. For **`booktype`**=`General Interest`, as **`marketing`** increases, the probability of buying increases. However, for **`booktype`**=`Special Interest`, as **`marketing`** increases, the probability of buying does not change.

Therefore, there is a **`marketing`** by **`booktype`** interaction.



Backward Elimination Method

Full Model

Reduced Model

When you use the backward elimination method with interactions in the model, you begin by fitting the full model with all the main effects and interactions. You then eliminate the nonsignificant interactions one at a time, starting with the least significant interaction (the one with the largest *p*-value). Next, you eliminate the nonsignificant main effects not involved in any significant interactions. The final model should only have significant interactions, the main effects involved in the interactions, and the significant main effects.

The requirement that for any interaction in the model, all effects it contains must also be in the model is called *model hierarchy*. For example, if the interaction **`gender*income`** is in the model, then the main effects **`gender`** and **`income`** must also be in the model. This ensures that you have a hierarchically well-formulated model.

For a more customized analysis, the HIERARCHY= option specifies whether hierarchy is maintained and whether a single effect or multiple effects are allowed to enter or leave the model in one step for forward, backward, and stepwise selection. The default is HIERARCHY=SINGLE. You can change this option by inserting the HIERARCHY = option in the MODEL statement. See the SAS/STAT User's Guide in the SAS OnlineDoc for more on using this option. In the LOGISTIC procedure, HIERARCH=SINGLE is the default, meaning SAS will not drop a main effect before dropping all its interactions.

Multiple Logistic Regression with Interactions

m5demo07.sas

Example: Fit a multiple logistic regression model using the backward elimination method. In the MODEL statement, specify all the main effects and the two-factor interactions.

```
proc logistic data=sasuser.b_sales_inc descending;
   class gender (param=ref ref='Male')
         income (param=ref ref='Low');
   model purchase=gender|age|income @2/ selection=backward;
   title1 'LOGISTIC MODEL (3): purchase = gender age '
          'income gender*age';
   title2 'gender*income  age*income / sel=backward';
run;
```

Selected MODEL statement option:

SELECTION= specifies the method used to select the variables in the model.

✎ The bar notation with the @2 constructs a model with all the main effects and the two-factor interactions. If you increased it to @3, then you would construct a model with all of the main effects, the two-factor interactions, and the three-factor interaction. However, the three-factor interaction can be more difficult to interpret.

The Model Information, Response Profile, and Class Level Information tables have not changed.

```
            LOGISTIC MODEL (3): purchase = gender age income gender*age
                        gender*income  age*income / sel=backward


                          The LOGISTIC Procedure

                             Model Information

             Data Set                      SASUSER.B_SALES_INC
             Response Variable             purchase
             Number of Response Levels     2
             Number of Observations        431
             Link Function                 Logit
             Optimization Technique        Fisher's scoring


                              Response Profile

                     Ordered                      Total
                       Value     purchase     Frequency

                           1            1           162
                           2            0           269


                      Backward Elimination Procedure


                          Class Level Information

                                              Design
                                            Variables

                      Class      Value       1      2

                      gender     Female      1
                                 Male        0

                      income     High        1      0
                                 Low         0      0
                                 Medium      0      1


Step  0. The following effects were entered:

Intercept  gender  age  age*gender  income  gender*income  age*income


                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.
```

PROC LOGISTIC Output (continued)

```
                           Model Fit Statistics

                                                 Intercept
                                   Intercept        and
                  Criterion          Only       Covariates

                  AIC                572.649       560.330
                  SC                 576.715       600.991
                  -2 Log L           570.649       540.330


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        30.3195        9         0.0004
          Score                   28.9614        9         0.0007
          Wald                    26.7755        9         0.0015


Step  1. Effect age*income is removed:

                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                                 Intercept
                                   Intercept        and
                  Criterion          Only       Covariates

                  AIC                572.649       557.936
                  SC                 576.715       590.465
                  -2 Log L           570.649       541.936


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        28.7135        7         0.0002
          Score                   26.8148        7         0.0004
          Wald                    24.7124        7         0.0009


                         Residual Chi-Square Test

                    Chi-Square       DF     Pr > ChiSq

                        1.5966        2         0.4501


Step  2. Effect age*gender is removed:
```

PROC LOGISTIC Output (continued)

```
                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.


                           Model Fit Statistics

                                                Intercept
                                 Intercept         and
                  Criterion        Only        Covariates

                  AIC              572.649        557.592
                  SC               576.715        586.054
                  -2 Log L         570.649        543.592


                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        27.0577        6         0.0001
          Score                   25.6386        6         0.0003
          Wald                    24.1104        6         0.0005


                         Residual Chi-Square Test

                    Chi-Square       DF     Pr > ChiSq

                        3.2232        3         0.3585


Step  3. Effect age is removed:

                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.

                           Model Fit Statistics

                                                Intercept
                                 Intercept         and
                  Criterion        Only        Covariates

                  AIC              572.649        557.194
                  SC               576.715        581.591
                  -2 Log L         570.649        545.194

                  Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        25.4552        5         0.0001
          Score                   24.1139        5         0.0002
          Wald                    22.7265        5         0.0004
```

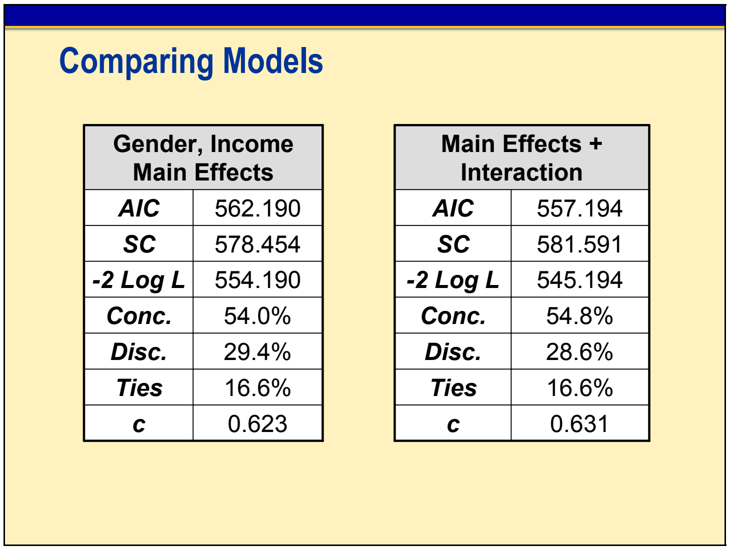
PROC LOGISTIC Output (continued)

```
                         Residual Chi-Square Test

                    Chi-Square       DF     Pr > ChiSq

                        4.7980        4         0.3087


NOTE: No (additional) effects met the 0.05 significance level for removal
      from the model.

                      Summary of Backward Elimination

             Effect                   Number          Wald
     Step    Removed            DF        In    Chi-Square    Pr > ChiSq

        1    age*income          2         5        1.5891        0.4518
        2    age*gender          1         4        1.6408        0.2002
        3    age                 1         3        1.5965        0.2064


                       Type III Analysis of Effects

                                            Wald
             Effect             DF    Chi-Square    Pr > ChiSq

             gender              1        4.9207        0.0265
             income              2       18.8745        <.0001
             gender*income       2        8.8363        0.0121


                  Analysis of Maximum Likelihood Estimates

                                           Standard
Parameter                    DF  Estimate     Error  Chi-Square  Pr > ChiSq

Intercept                     1   -1.4759    0.3919     14.1841      0.0002
gender        Female          1    0.9949    0.4485      4.9207      0.0265
income        High            1    1.5026    0.4549     10.9113      0.0010
income        Medium          1    0.1235    0.4873      0.0642      0.7999
gender*income Female High     1   -1.2223    0.5523      4.8979      0.0269
gender*income Female Medium   1    0.1026    0.5851      0.0307      0.8608

       Association of Predicted Probabilities and Observed Responses

             Percent Concordant     54.8    Somers' D    0.261
             Percent Discordant     28.6    Gamma        0.314
             Percent Tied           16.6    Tau-a        0.123
             Pairs                 43578    c            0.631
```

The interactions between **`age*income`** and **`age*gender`** are eliminated from the model because their *p*-values are greater than the default value of 0.05, as reported in the Summary of Backward Elimination table. However, because the interaction of **`gender`** and **`income`** is significant, the main effects **`gender`** and **`income`** must remain in the model. Because the main effect of **`age`** is not significant and not involved in a significant interaction, the term is dropped from the model.

Comparing the goodness-of-fit statistics and the statistics that assess the predictive ability of the full model and the final model shows that the full model has better predictive ability (because of the higher *c* statistic), whereas the final model has better goodness-of-fit statistics (because of the lower AIC and SBC statistics).

Statistic	**Full Model** **purchase=gender age income gender*age gender*income age*income**	**Final Model** **purchase=gender income gender*income**
AIC	560.330	557.194
SBC	600.991	581.591
% Concordant	64.3	54.8
% Discordant	34.5	28.6
% Tied	1.1	16.6
c	0.649	0.631

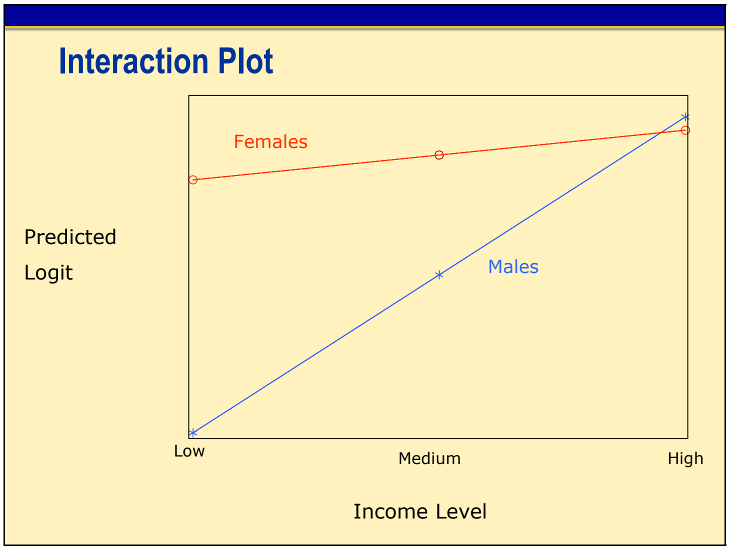


Comparing Models

Gender, Income Main Effects	
AIC	562.190
SC	578.454
-2 Log L	554.190
Conc.	54.0%
Disc.	29.4%
Ties	16.6%
c	0.623

Main Effects + Interaction	
AIC	557.194
SC	581.591
-2 Log L	545.194
Conc.	54.8%
Disc.	28.6%
Ties	16.6%
c	0.631

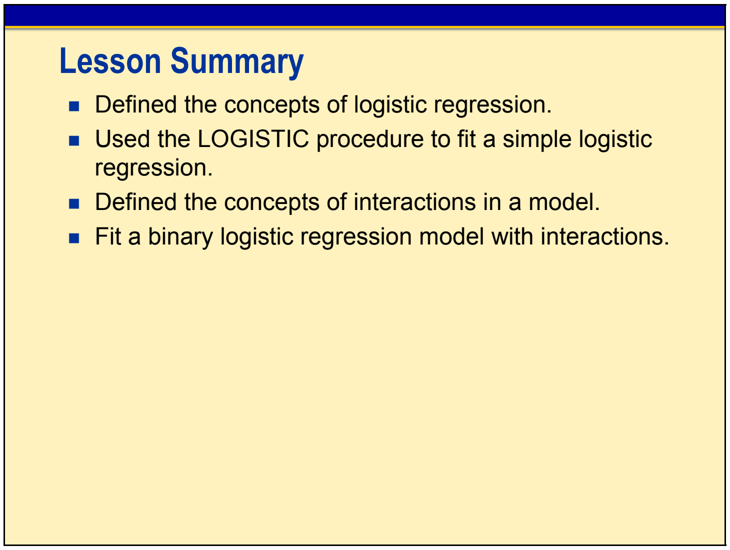
AIC decreased (improved) for this model, but SC increased. This indicates that adding the interaction term might have improved the model's inference, but it also might have worsened its ability to predict. Overall, a model should be chosen based on the researcher's intent.



Interaction Plot

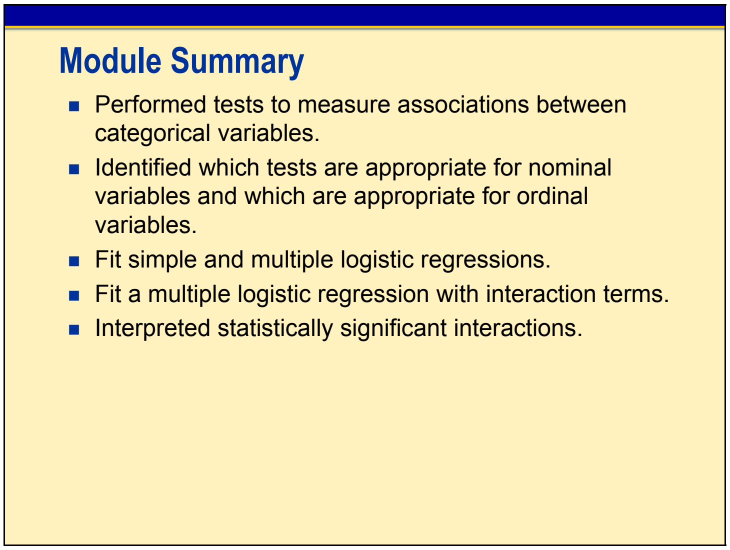
To visualize the interaction between **`gender`** and **`income`**, you could do an interaction plot. The plot would show two slopes for **`income`**, one for males and one for females. If there is no interaction between **`gender`** and **`income`**, then the slopes should be relatively parallel. However, the graph above shows that the slopes are not parallel. The reason for the interaction is that the probability of making purchases of 100 dollars or more is highly related to income for men but is weakly related to income for women.

✎ The code for the interaction plot is shown in Appendix D, "Advanced Programs."



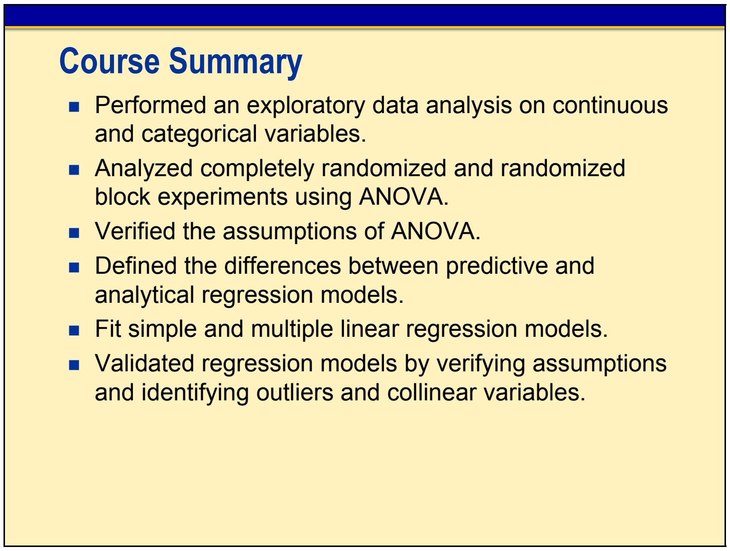
Lesson Summary

- Defined the concepts of logistic regression.
- Used the LOGISTIC procedure to fit a simple logistic regression.
- Defined the concepts of interactions in a model.
- Fit a binary logistic regression model with interactions.



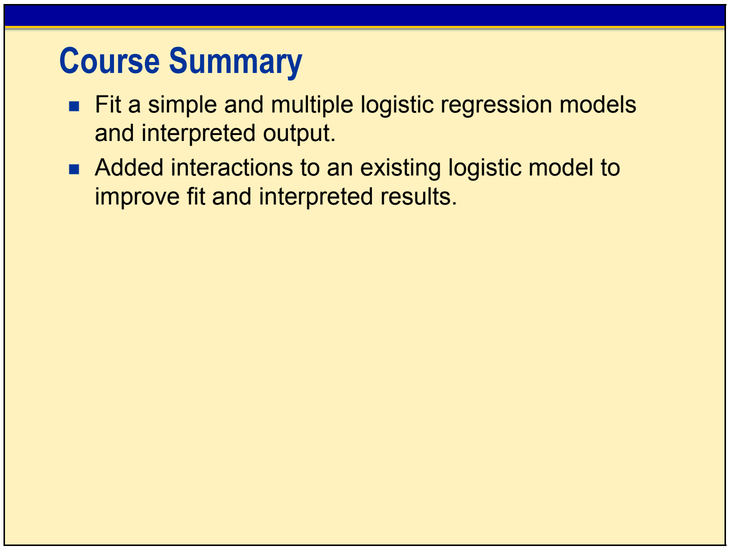
Module Summary

- Performed tests to measure associations between categorical variables.
- Identified which tests are appropriate for nominal variables and which are appropriate for ordinal variables.
- Fit simple and multiple logistic regressions.
- Fit a multiple logistic regression with interaction terms.
- Interpreted statistically significant interactions.



Course Summary

- Performed an exploratory data analysis on continuous and categorical variables.
- Analyzed completely randomized and randomized block experiments using ANOVA.
- Verified the assumptions of ANOVA.
- Defined the differences between predictive and analytical regression models.
- Fit simple and multiple linear regression models.
- Validated regression models by verifying assumptions and identifying outliers and collinear variables.

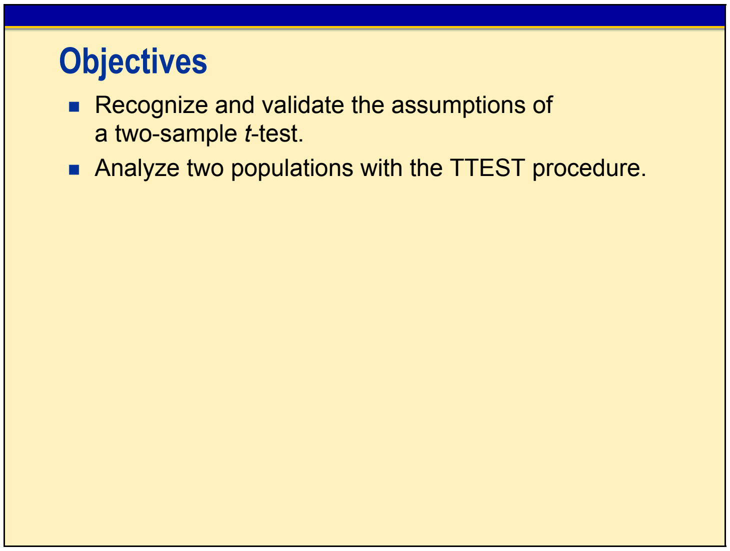


Course Summary

- Fit a simple and multiple logistic regression models and interpreted output.
- Added interactions to an existing logistic model to improve fit and interpreted results.

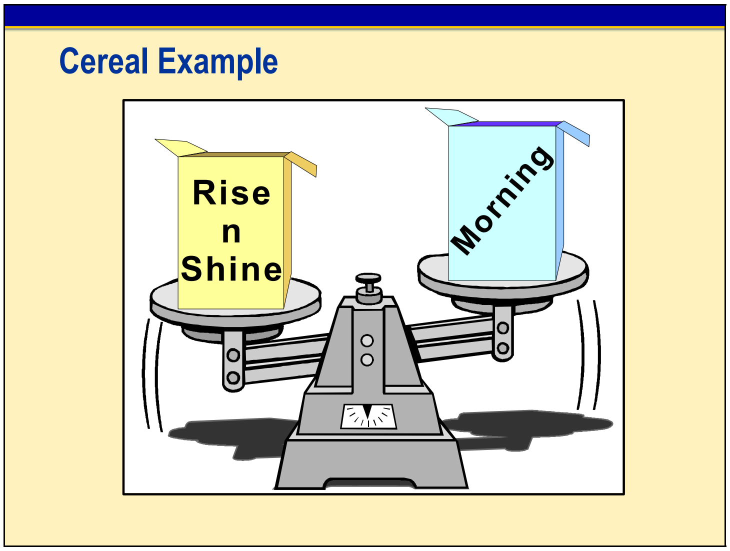
Appendix A Self-Study

A.1 Two-Sample *t*-Tests



Objectives

- Recognize and validate the assumptions of a two-sample *t*-test.
- Analyze two populations with the TTEST procedure.



Cereal Example

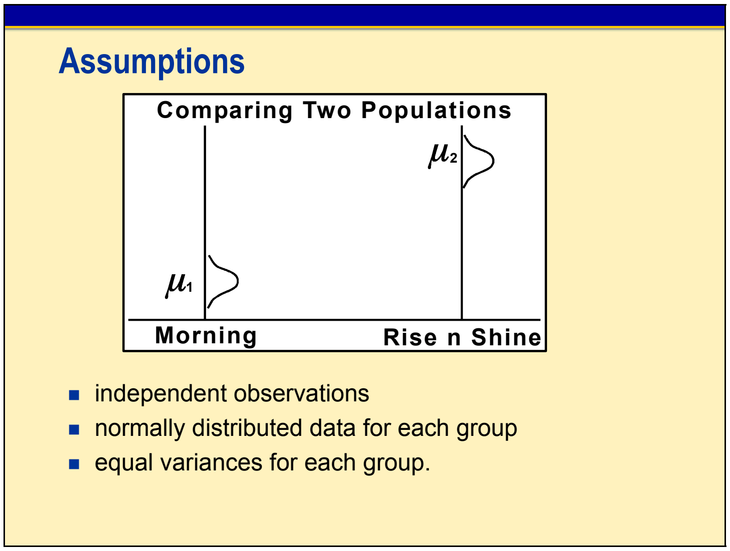
Example: A consumer advocacy group wants to determine whether two popular cereal brands, Rise n Shine and Morning, have the same amount of cereal. Both brands advertise that they have 15 ounces of cereal per box. A random sample of both brands is selected and the number of ounces of cereal is recorded. The data is stored in a data set called **`sasuser.b_cereal`**.

The variables in the data set are

`brand` two groups, `Rise n Shine` and `Morning`, corresponding to the two brands

`weight` weight of the cereal in ounces

`idnumber` the identification number for each cereal box.



Assumptions

- independent observations
- normally distributed data for each group
- equal variances for each group.

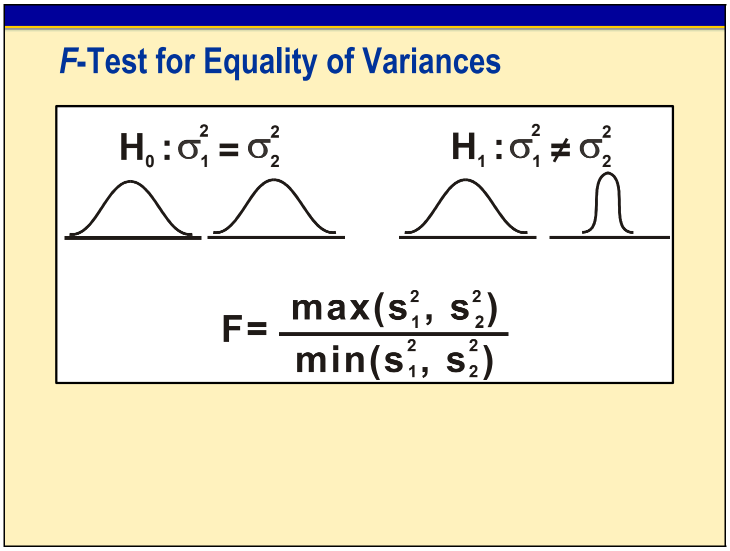
Before you start the analysis, examine the data to verify that the assumptions are valid.

The assumption of independent observations means that no observations provide any information about any other observation you collect. For example, measurements are not repeated on the same subject. This assumption can be verified during the design stage.

The assumption of normality can be relaxed if the data is approximately normally distributed or if enough data is collected. This assumption can be verified by examining plots of the data.

There are several tests for equal variances. If this assumption is not valid, an approximate *t*-test can be performed.

If these assumptions are **not** valid, the probability of drawing incorrect conclusions from the analysis could be increased.



F-Test for Equality of Variances

$H_0: \sigma_1^2 = \sigma_2^2$

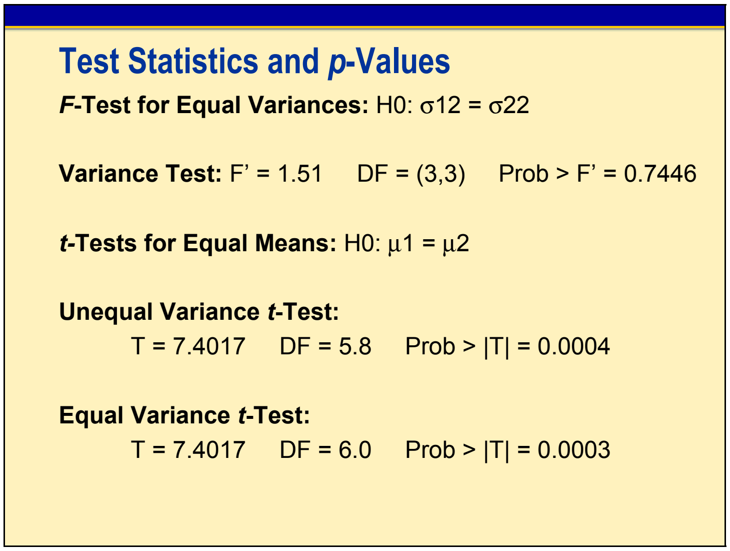
$H_1: \sigma_1^2 \neq \sigma_2^2$

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

When performing this test, note that if the null hypothesis is true, *F* tends to be close to 1.

If you reject the null hypothesis, it is recommended that you use the unequal variance *t*-test in the TTEST procedure for testing the equality of group means.

This test is valid **only** for independent samples from normal distributions. Normality is required even for large sample sizes.



Test Statistics and *p*-Values

***F*-Test for Equal Variances:** H0: σ12 = σ22

Variance Test: F' = 1.51 DF = (3,3) Prob > F' = 0.7446

***t*-Tests for Equal Means:** H0: µ1 = µ2

Unequal Variance *t*-Test:

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

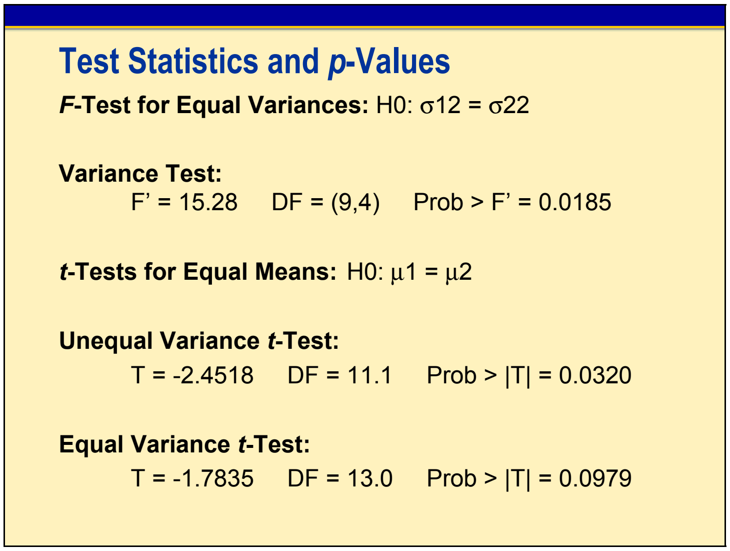
Equal Variance *t*-Test:

T = 7.4017 DF = 6.0 Prob > |T| = 0.0003

First, check the assumption for equal variances and then use the appropriate test for equal means. Because the *p*-value of the test *F*-statistic is 0.7446, there is not enough evidence to reject the null hypothesis of equal variances. Use the Equal Variance *t*-test line in the output to test whether the means of the two populations are equal.

The null hypothesis that the group means are equal is rejected at the 0.05 level. You conclude that there is a difference between the means of the groups.

The equal variance *F*-test is found at the bottom of the PROC TTEST output.



Test Statistics and *p*-Values

***F*-Test for Equal Variances:** H0: σ12 = σ22

Variance Test:

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

***t*-Tests for Equal Means:** H0: μ1 = μ2

Unequal Variance *t*-Test:

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

Equal Variance *t*-Test:

T = -1.7835 DF = 13.0 Prob > |T| = 0.0979

Again, first check the assumption for equal variances and use the appropriate test for equal means. Because the *p*-value of the test *F*-statistic is less than alpha=0.05, there is enough evidence to reject the null hypothesis of equal variances. Use the unequal variance *t*-test line in the output to test whether the means of the two populations are equal.

The null hypothesis that the group means are equal is rejected at the .05 level.

However, notice that if you choose the equal variance *t*-test, you would not reject the null hypothesis at the .05 level. This shows the importance of choosing the appropriate *t*-test.

Testing for Equality of Means

Example: Print the data in the **`sasuser.b_cereal`** data set and do an initial check of the assumptions of the *t*-test and the *F*-test using the UNIVARIATE procedure. Then invoke PROC TTEST to test the hypothesis that the means are equal for the two groups.

```
proc print data=sasuser.b_cereal (obs=15);
   title 'Partial Listing of Cereal Data';
run;
```

Part of the data is shown below.

```
                     Partial Listing of Cereal Data

              OBS    BRAND             WEIGHT       ID

                1    Morning          14.9982    61469897
                2    Rise n Shine     15.0136    33081197
                3    Morning          15.0100    68137597
                4    Rise n Shine     14.9982    37070397
                5    Morning          15.0052    64608797
                6    Rise n Shine     14.9930    60714297
                7    Morning          14.9733    16907997
                8    Rise n Shine     15.0812     9589297
                9    Morning          15.0037    93891897
               10    Rise n Shine     15.0418    85859397
               11    Morning          14.9957    38152597
               12    Rise n Shine     15.0639    99108497
               13    Morning          15.0099    59666697
               14    Rise n Shine     15.0613    70847197
               15    Morning          14.9943    47613397
```

```
proc sort data=sasuser.b_cereal out=sorted_cereal;
   by brand;
run;

proc univariate data=sorted_cereal normal plot;
   var weight;
   by brand;
   probplot weight / normal (mu=est sigma=est
                            color=blue w=1);
   title 'Univariate Analysis of the Cereal Data';
run;
```

In order to generate the analysis for each cereal brand, the data must be sorted by the variable **brand**. The SORT procedure step is needed before PROC UNIVARIATE, and the same BY variable used in PROC SORT is needed in PROC UNIVARIATE.

Selected PROC UNIVARIATE statement options:

NORMAL produces four test statistics and their corresponding *p*-values for testing

H_0: Normal Distribution
versus
H_1: Nonnormal Distribution.

PLOT produces a stem-and-leaf plot, a box-and-whisker plot, and a normal probability plot. When a BY statement is used in combination with the PLOT option, side-by-side box-and-whisker plots are produced.

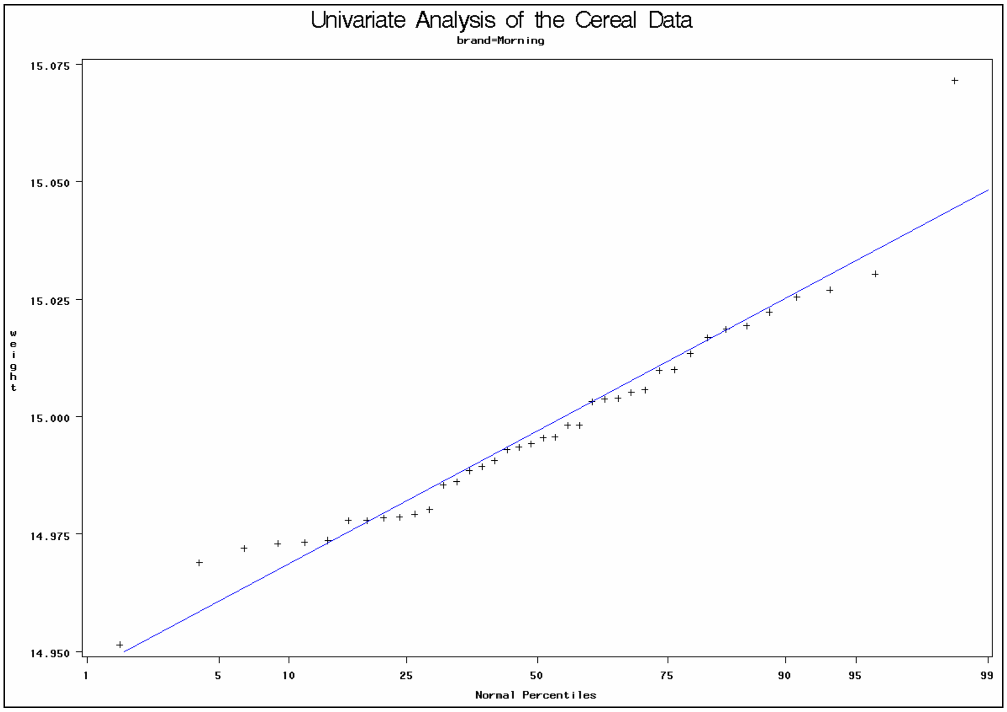
Partial PROC UNIVARIATE Output

```
                 Univariate Analysis of the Cereal Data

------------------------------ brand=Morning ------------------------------

                         The UNIVARIATE Procedure
                             Variable:  weight

                                  Moments

      N                          40    Sum Weights                 40
      Mean               14.9970125    Sum Observations      599.8805
      Std Deviation      0.02201048    Variance            0.00048446
      Skewness           0.87481049    Kurtosis            2.07993397
      Uncorrected SS     8996.43425    Corrected SS        0.01889398
      Coeff Variation    0.14676575    Std Error Mean      0.00348016


                        Basic Statistical Measures

              Location                    Variability

          Mean     14.99701     Std Deviation           0.02201
          Median   14.99490     Variance              0.0004845
          Mode     14.97790     Range                   0.12010
                                Interquartile Range     0.03095

  NOTE: The mode displayed is the smallest of 2 modes with a count of 2.


                        Tests for Location: Mu0=0

             Test           -Statistic-    -----p Value------

             Student's t    t  4309.286    Pr > |t|    <.0001
             Sign           M        20    Pr >= |M|   <.0001
             Signed Rank    S       410    Pr >= |S|   <.0001


                           Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W      0.95094    Pr < W      0.0817
        Kolmogorov-Smirnov    D     0.078487    Pr > D     >0.1500
        Cramer-von Mises      W-Sq  0.049936    Pr > W-Sq  >0.2500
        Anderson-Darling      A-Sq  0.414338    Pr > A-Sq  >0.2500
```

Examine the Tests for Normality table above. The null hypothesis is that the data is normally distributed. Because all the observed *p*-values are greater than 0.05, there is insufficient evidence to reject the null hypothesis.

```
Stem Leaf                     #              Boxplot
1507 2                        1                 0
1506
1505
1504
1503 0                        1                 |
1502 257                      3                 |
1501 003799                   6              +-----+
1500 34456                    5              |     |
1499 13446688                 8              *--+--*
1498 06689                    5              |     |
1497 233488899                9              +-----+
1496 9                        1                 |
1495 2                        1                 |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-2
```

The stem-and-leaf plot and the box-and-whisker plot show one extreme value. Otherwise, the data for `Morning` appears to be symmetric.



The normal probability plot shows no serious departures from normality, allowing for the one extreme point previously noted. There appears to be no pattern for the data that reflects skewness or kurtosis.

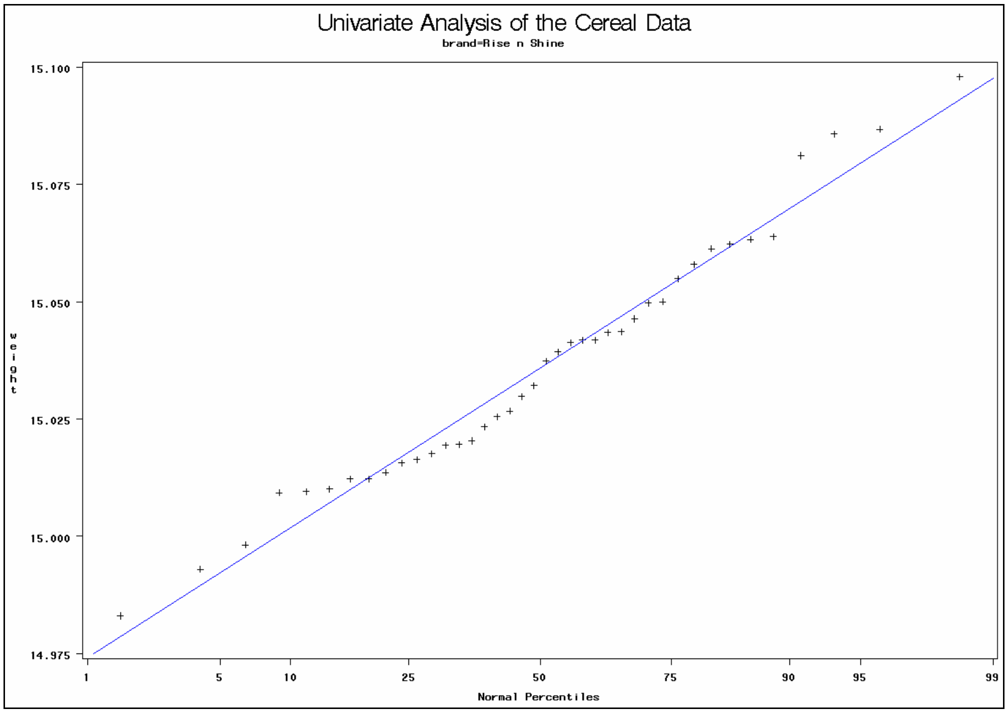
PROC UNIVARIATE Output (continued)

```
                   Univariate Analysis of the Cereal Data

-------------------------- brand=Rise n Shine ---------------------------

                          The UNIVARIATE Procedure
                              Variable:  weight

                                   Moments

      N                           40    Sum Weights                 40
      Mean                  15.03596    Sum Observations      601.4384
      Std Deviation       0.02654963    Variance            0.00070488
      Skewness            0.39889232    Kurtosis            -0.1975717
      Uncorrected SS      9043.23122    Corrected SS        0.02749044
      Coeff Variation     0.17657424    Std Error Mean      0.00419787


                         Basic Statistical Measures

              Location                    Variability

          Mean     15.03596     Std Deviation            0.02655
          Median   15.03480     Variance               0.0007049
          Mode     15.01220     Range                    0.11490
                                Interquartile Range      0.03650

  NOTE: The mode displayed is the smallest of 2 modes with a count of 2.


                            Tests for Normality

        Test                  --Statistic---    -----p Value------

        Shapiro-Wilk          W     0.974477    Pr < W      0.4926
        Kolmogorov-Smirnov    D     0.096086    Pr > D     >0.1500
        Cramer-von Mises      W-Sq  0.059304    Pr > W-Sq  >0.2500
        Anderson-Darling      A-Sq  0.387763    Pr > A-Sq  >0.2500
```

The tests for normality for the brand `Rise n Shine` are not significant. Therefore, there is insufficient evidence to conclude that the data is not normally distributed.

The stem-and-leaf plot and the box-and-whisker plot illustrate that the data is fairly symmetric. There are also no extreme values. The normal probability plot shows no serious departures from normality.

```
Stem Leaf                      #            Boxplot
1509 8                         1                |
1508 167                       3                |
1507                                            |
1506 1234                      4                |
1505 0058                      4            +-----+
1504 122446                    6            |     |
1503 0279                      4            *--+--*
1502 00367                     5            |     |
1501 002246689                 9            +-----+
1500 9                         1                |
1499 38                        2                |
1498 3                         1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-2
```



```
 Univariate Analysis of the Cereal Data

        The UNIVARIATE Procedure
             Schematic Plots

      |
 15.1 +                        |
      |                        |
      |                        |
      |                        |
15.08 +                        |
      |                        |
      |            0           |
      |                        |
15.06 +                        |
      |                     +-----+
      |                     |     |
      |                     |     |
15.04 +                     |     |
      |                     *--+--*
      |            |        |     |
      |            |        |     |
15.02 +            |        |     |
      |            |        +-----+
      |         +-----+        |
      |         |     |        |
   15 +         |     |        |
      |         *--+--*        |
      |         |     |        |
      |         |     |        |
14.98 +         +-----+
      |            |
      |            |
      |            |
14.96 +            |
      |            |
      |            |
      |
14.94 +
       ------------+----------+-----------
brand           Morning    Rise n S
```

The comparative box-and-whisker plots show that the weights of the brand `Rise n Shine` have a larger mean and more variability than `Morning` cereal weights.

Because both brands have weights that are normally distributed, the assumptions of the *F*-test for equal variances are verified. The assumption of the *t*-test regarding the normality of the distribution of sample means is also satisfied. You could have used the central limit theorem to validate the assumption for the *t*-test because both brands have 40 observations.

Invoke the TTEST procedure and interpret the output.

```
proc ttest data=sasuser.b_cereal;
   class brand;
   var weight;
   title 'Testing the Equality of Means for Two Cereal '
         'Brands';
run;
```

```
             Testing the Equality of Means for Two Cereal Brands

                             The TTEST Procedure

❶                                     Statistics

                              Lower CL            Upper CL  Lower CL
Variable  brand            N      Mean    Mean      Mean   Std Dev  Std Dev

weight    Morning         40     14.99  14.997    15.004     0.018    0.022
weight    Rise n          40    15.027  15.036    15.044    0.0217   0.0265
          Shine
weight    Diff (1-2)             -0.05  -0.039    -0.028    0.0211   0.0244

                                    Statistics

                          Upper CL
     Variable  brand       Std Dev    Std Err    Minimum    Maximum

     weight    Morning      0.0283     0.0035     14.952     15.072
     weight    Rise n       0.0341     0.0042     14.983     15.098
               Shine
     weight    Diff (1-2)   0.0289     0.0055

❸                                      T-Tests

   Variable    Method           Variances      DF    t Value    Pr > |t|

   weight      Pooled           Equal          78      -7.14      <.0001
   weight      Satterthwaite    Unequal      75.4      -7.14      <.0001

                            Equality of Variances

       Variable    Method      Num DF    Den DF    F Value    Pr > F

❷       weight      Folded F        39        39       1.45    0.2460
```

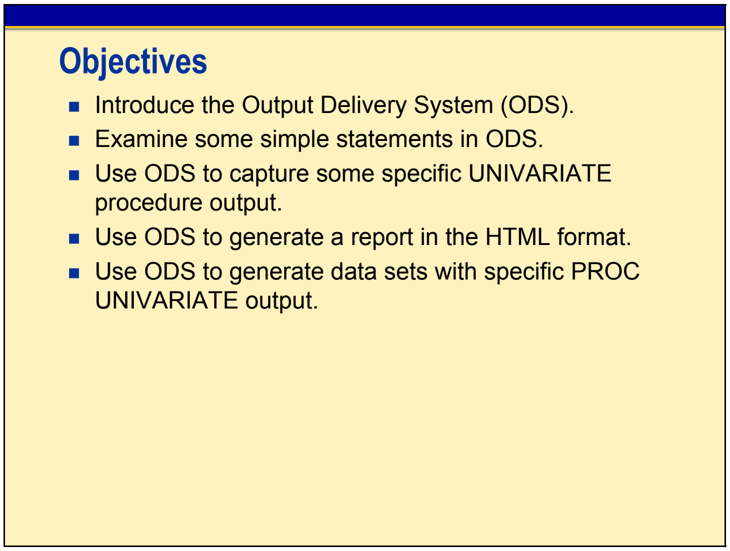
❶ In the Statistics table, examine the descriptive statistics for each group and their differences. The confidence limits for the sample mean and sample standard deviation are also shown.

❷ Look at the Equality of Variances table that appears at the bottom of the output. The *F*-test for equal variances has a *p*-value of 0.2460. In this case, do not reject the null hypothesis. Conclude that there is insufficient evidence to indicate that the variances are not equal.

❸ Based on the *F*-test for equal variances, you then look in the T-Tests table at the *t*-test for the hypothesis of equal means. Using the equal variance *t*-test, you reject the null hypothesis that the group means are equal. Conclude that there is a difference in the average weight of the cereal between the Rise n Shine brand and the Morning brand.

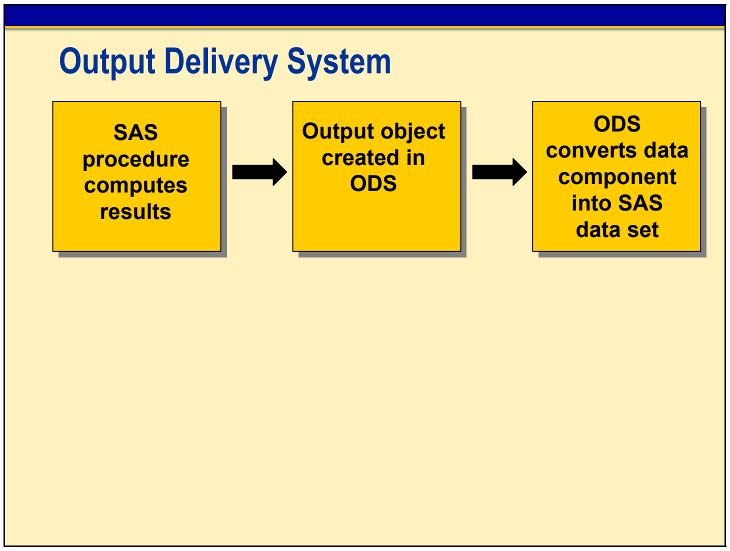
Return your attention to the Statistics table. Because the confidence interval for the mean (-0.05, -0.028) does not include 0, you can conclude that there is a significant difference between the two cereal means.

A.2 Output Delivery System



Objectives

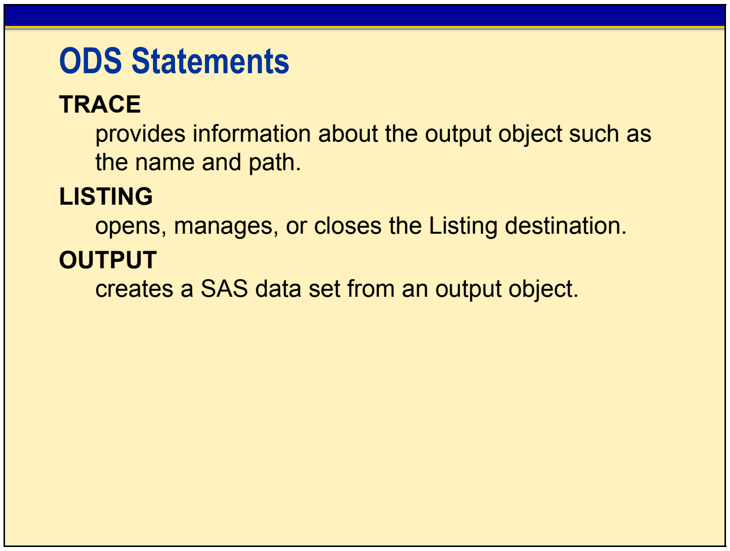
- Introduce the Output Delivery System (ODS).
- Examine some simple statements in ODS.
- Use ODS to capture some specific UNIVARIATE procedure output.
- Use ODS to generate a report in the HTML format.
- Use ODS to generate data sets with specific PROC UNIVARIATE output.



Output Delivery System

SAS procedure computes results → Output object created in ODS → ODS converts data component into SAS data set

The Output Delivery System (ODS) enables you to take output from a SAS procedure and convert it to a SAS data set. Instead of writing to the listing file directly, SAS procedures can now create an output object for each piece of output that is displayed. For example, each table produced in the UNIVARIATE procedure is now a separate entity in ODS. You can then take the data component of the output object and convert it to a SAS data set. This means that every number in every table of every procedure can be accessed via a data set.



ODS Statements

TRACE
provides information about the output object such as the name and path.

LISTING
opens, manages, or closes the Listing destination.

OUTPUT
creates a SAS data set from an output object.

The TRACE statement is used to obtain the name of the output object. The LISTING statement is used to manage the Output window, and the OUTPUT statement is used to create SAS data sets.

Output Delivery System

Example: Examine some basic functionality of the Output Delivery System.

The ODS TRACE ON statement produces a trace record in the SAS Log window, including the name and label of each output object. The ODS LISTING CLOSE statement instructs ODS not to produce any results in the Output window.

```
ods trace on;

/*--- --- --- --- --- --- --- --- --- --- --- --- --- ---*/
/*  -do not generate any results in the output window    */
/*-generate and examine table definitions for UNIVARIATE */
/*--- --- --- --- --- --- --- --- --- --- --- --- --- ---*/

ods listing close;

proc univariate data=sasuser.b_rise normal plot;
   var weight;
   id idnumber;
run;
ods trace off;
```

SAS Log

```
13   ods trace on;
14   /*--- --- --- --- --- --- --- --- --- --- --- --- --- --- ---*/
15   /*  -do not generate any results in the output window       */
16   /*  -generate and examine table definitions for UNIVARIATE   */
17   /*--- --- --- --- --- --- --- --- --- --- --- --- --- --- ---*/
18   ods listing close;
19   proc univariate data=sasuser.b_rise normal plot;
20      var weight;
21      id idnumber;
22   run;

WARNING: No output destinations active.

Output Added:
-------------
Name:       Moments
Label:      Moments
Template:   base.univariate.Moments
Path:       Univariate.weight.Moments
-------------
Output Added:
-------------
```

SAS Log (continued)

```
Name:       BasicMeasures
Label:      Basic Measures of Location and Variability
Template:   base.univariate.Measures
Path:       Univariate.weight.BasicMeasures
-------------

Output Added:
-------------
Name:       TestsForLocation
Label:      Tests For Location
Template:   base.univariate.Location
Path:       Univariate.weight.TestsForLocation
-------------

Output Added:
-------------
Name:       TestsForNormality
Label:      Tests For Normality
Template:   base.univariate.Normal
Path:       Univariate.weight.TestsForNormality
-------------

Output Added:
-------------
Name:       Quantiles
Label:      Quantiles
Template:   base.univariate.Quantiles
Path:       Univariate.weight.Quantiles
-------------

Output Added:
-------------
Name:       ExtremeObs
Label:      Extreme Observations
Template:   base.univariate.ExtObs
Path:       Univariate.weight.ExtremeObs
-------------

Output Added:
-------------
Name:       Plots
Label:      Plots
Data Name:  BatchOutput
Path:       Univariate.weight.Plots
-------------
NOTE: There were 40 observations read from the data set SASUSER.B_RISE.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.10 seconds
      cpu time            0.10 seconds
```

For each table, Name, Label, Template or Data Name, and Path are listed. Please note the warning that you have not generated any output:

```
WARNING: No output destinations active.
```

You can now select only those tables of interest. The tables of interest for a management presentation might only be the following: Moments, BasicMeasures, and Plots.

```
ods select
    Moments
    BasicMeasures
    Plots
    ;
ods listing;
proc univariate data=sasuser.b_rise normal plot;
   var weight;
   id idnumber;
   title1 'Selected Results using ODS';
run;
```

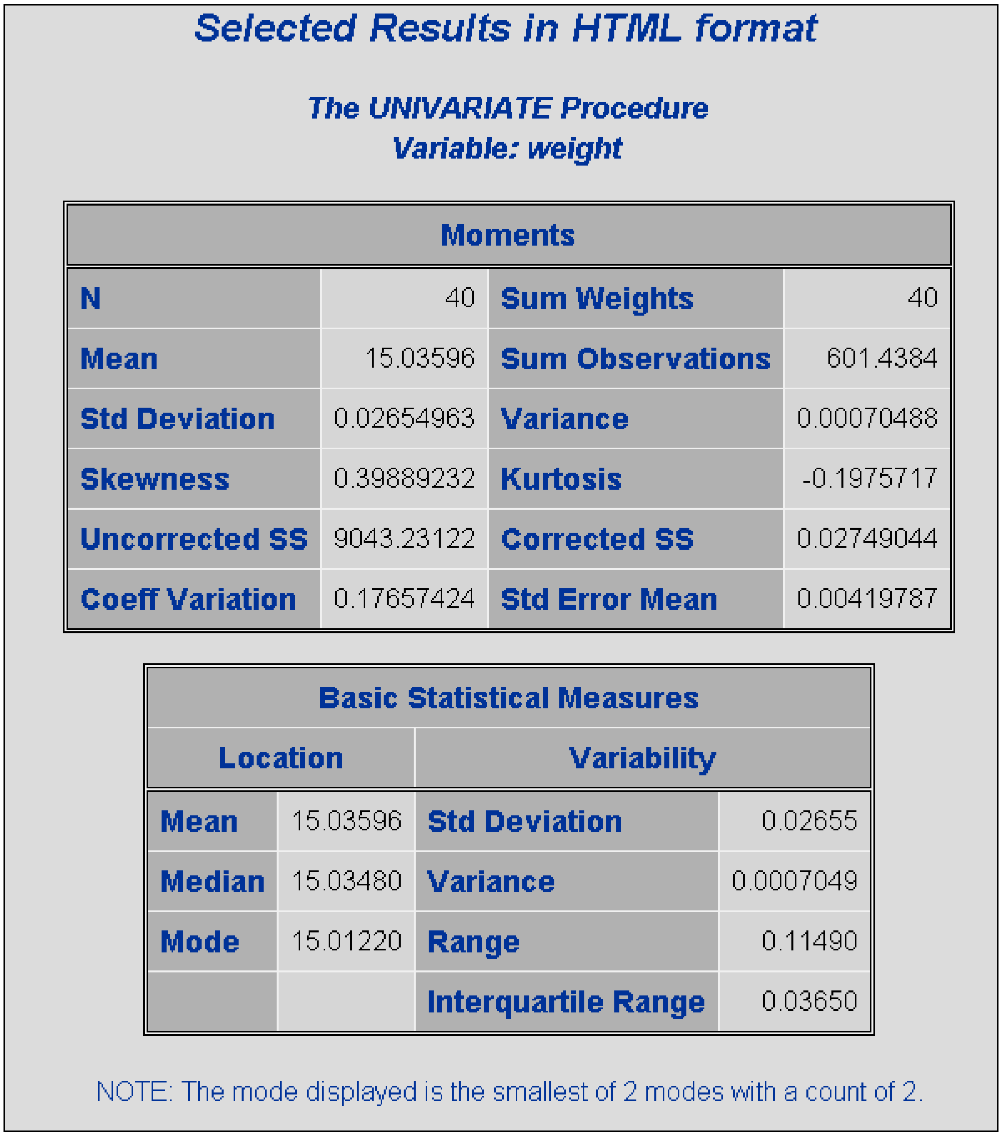
```
                        Selected Results using ODS

                         The UNIVARIATE Procedure
                             Variable:  weight

                                  Moments

      N                           40    Sum Weights                 40
      Mean                  15.03596    Sum Observations      601.4384
      Std Deviation       0.02654963    Variance            0.00070488
      Skewness            0.39889232    Kurtosis            -0.1975717
      Uncorrected SS      9043.23122    Corrected SS        0.02749044
      Coeff Variation     0.17657424    Std Error Mean      0.00419787

                        Basic Statistical Measures

              Location                    Variability

          Mean     15.03596     Std Deviation            0.02655
          Median   15.03480     Variance               0.0007049
          Mode     15.01220     Range                    0.11490
                                Interquartile Range      0.03650

  NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

             Stem Leaf                     #             Boxplot
             1509 8                        1                |
             1508 167                      3                |
             1507                                           |
             1506 1234                     4                |
             1505 0058                     4             +-----+
             1504 122446                   6             |     |
             1503 0279                     4             *--+--*
             1502 00367                    5             |     |
             1501 002246689                9             +-----+
             1500 9                        1                |
             1499 38                       2                |
             1498 3                        1                |
                  ----+----+----+----+
              Multiply Stem.Leaf by 10**-2
```

SAS Output (continued)

```
                        Selected Results using ODS

                         The UNIVARIATE Procedure
                             Variable:  weight

                              Normal Probability Plot
         15.095+                                              +***
               |                                      * * *+++
               |                                      ++++
               |                                  ****
               |                               ***+
               |                           *****
               |                       ++**
               |                   +++***
               |              **+*****
               |            **++
               |        *+*+
         14.985+   *++++
                +----+----+----+----+----+----+----+----+----+----+
                    -2        -1         0        +1        +2
```

Although these reports are effective, in order to make them easier to distribute, use ODS to generate them in HTML format.

```
ods listing close;

ods html
    body='sel_u.htm';
ods select
    Moments
    BasicMeasures
    Plots
    ;

proc univariate data=sasuser.b_rise normal plot;
   var weight;
   id idnumber;
   title1 'Selected Results in HTML format';
run;

ods html close;
```



Selected Results in HTML format

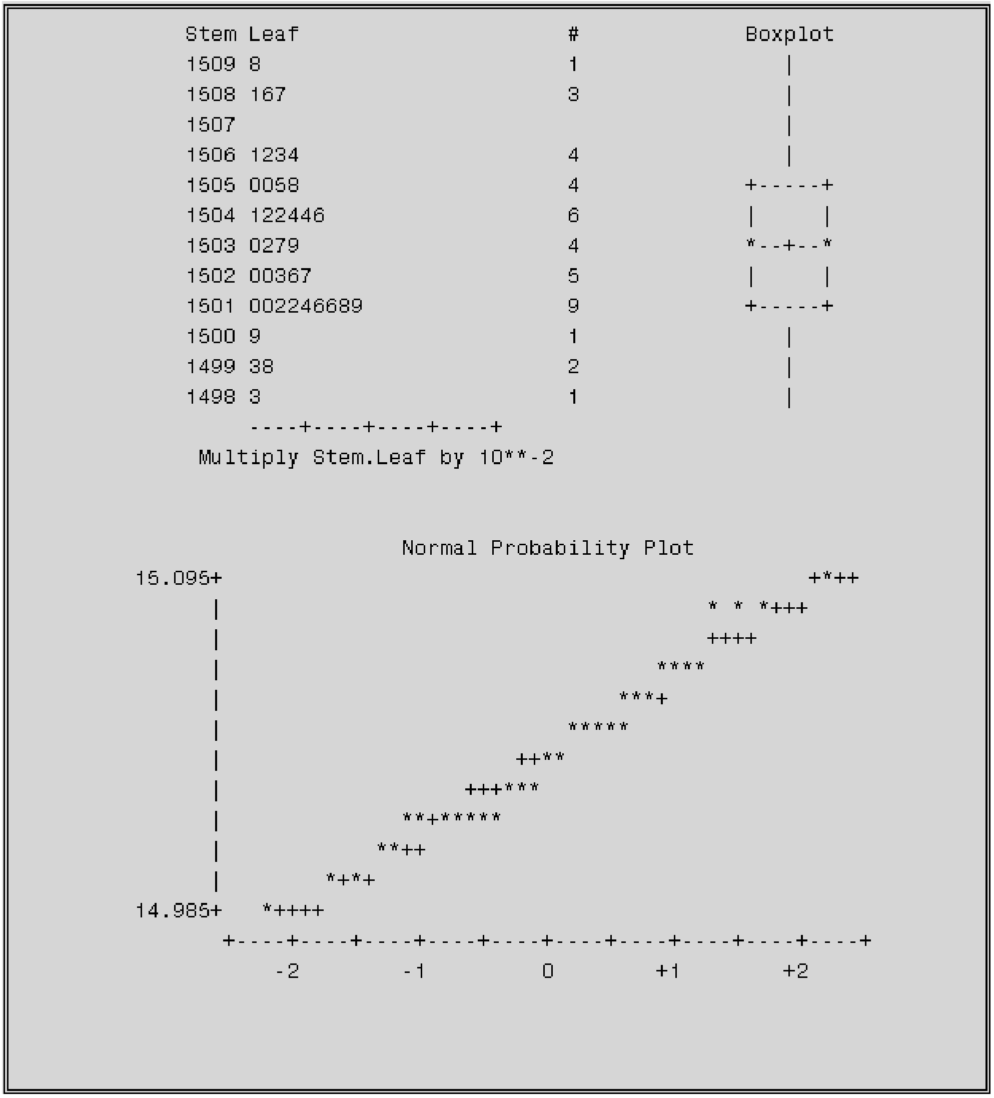
The UNIVARIATE Procedure
Variable: weight

Moments			
N	40	Sum Weights	40
Mean	15.03596	Sum Observations	601.4384
Std Deviation	0.02654963	Variance	0.00070488
Skewness	0.39889232	Kurtosis	-0.1975717
Uncorrected SS	9043.23122	Corrected SS	0.02749044
Coeff Variation	0.17657424	Std Error Mean	0.00419787

Basic Statistical Measures			
Location		Variability	
Mean	15.03596	Std Deviation	0.02655
Median	15.03480	Variance	0.0007049
Mode	15.01220	Range	0.11490
		Interquartile Range	0.03650

NOTE: The mode displayed is the smallest of 2 modes with a count of 2.

HTML Output (continued)



```
    Stem Leaf                     #             Boxplot
    1509 8                        1                |
    1508 167                      3                |
    1507                                           |
    1506 1234                     4                |
    1505 0058                     4             +-----+
    1504 122446                   6             |     |
    1503 0279                     4             *--+--*
    1502 00367                    5             |     |
    1501 002246689                9             +-----+
    1500 9                        1                |
    1499 38                       2                |
    1498 3                        1                |
         ----+----+----+----+
     Multiply Stem.Leaf by 10**-2


                     Normal Probability Plot
15.095+                                              +*++
      |                                      * * *+++
      |                                      ++++
      |                                  ****
      |                               ***+
      |                           *****
      |                       ++**
      |                   +++***
      |              **+*****
      |            **++
      |        *+*+
14.985+   *++++
       +----+----+----+----+----+----+----+----+----+----+
           -2        -1         0        +1        +2
```

The file containing this HTML, sel_u.htm, is located in the root directory of where SAS resides in your environment.

If you are in the Windows environment, this HTML output is displayed immediately and is also available in the Results window.

You can also generate SAS data sets to extract specific values in later programming steps or for future analyses.

```
ods listing close;

ods output
    Moments=o_moments
    BasicMeasures=o_basic
    TestsforNormality=o_tnormal
    Quantiles=o_quant
    ;

proc univariate data=sasuser.b_rise normal plot;
   var weight;
   id idnumber;
run;

ods listing;
```

```
72   ods listing close;
73   ods output
74       Moments=o_moments
75       BasicMeasures=o_basic
76       TestsforNormality=o_tnormal
77       Quantiles=o_quant
78       ;

79   proc univariate data=sasuser.b_rise normal plot;
80      var weight;
81      id idnumber;
82   run;

NOTE: The data set WORK.O_QUANT has 11 observations and 3 variables.
NOTE: The data set WORK.O_TNORMAL has 4 observations and 7 variables.
NOTE: The data set WORK.O_BASIC has 4 observations and 5 variables.
NOTE: The data set WORK.O_MOMENTS has 6 observations and 7 variables.
NOTE: There were 40 observations read from the data set SASUSER.B_RISE.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.09 seconds
      cpu time            0.09 seconds


83   ods listing;
```

The SAS data sets generated with the OUTPUT statement are stored in the **work** library. To store them in a SAS data set, use a two-level SAS name.

Appendix B Sampling Macros

B.1 Random Samples

Selecting Random Samples

The SURVERYSELECT procedure (Version 8 and higher) selects a random sample from a SAS data set.

```
PROC SURVEYSELECT DATA=name-of-SAS-data-set
                  OUT=name-of-output-data-set
                  METHOD = method-of-random-sampling
                  SEED=seed-value
                  SAMPSIZE=number of observations desired in
                           sample
                  ;
<STRATA stratification- variable(s)>;
RUN;
```

Selected PROC SURVEYSELECT statement options:

DATA= identifies the data set to be selected from.

OUT= indicates the name of the output data set.

METHOD= specifies the random sampling method to be used. For simple random sampling without replacement, use METHOD=SRS. For simple random sampling with replacement, use METHOD=URS. For other selection methods and details on sampling algorithms, see the SAS documentation for PROC SURVEYSELECT.

SEED= specifies the initial seed for random number generation. If no SEED option is specified, SAS uses the system time as its seed value. This creates a different random sample every time the procedure is run.

SAMPSIZE= indicates the number of observations to be included in the sample. To select a certain fraction of the original data set rather than a given number of observations, use the SAMPRATE= option.

Selected PROC SURVEYSELECT statements:

STRATA enables the user to specify one or more stratification variables. If no STRATA statement is specified, no stratification takes place.

Other statements and options for the SURVERYSELECT procedure can be found in SAS OnlineDoc.

Appendix C Percentile Definitions

C.1 Calculating Percentiles

Using the UNIVARIATE Procedure

Example: Calculate the 25th percentile for the following data using the five definitions available in PROC UNIVARIATE:

1 3 7 11 14

For all of these calculations (except definition 4), you use the value *np*=(5)(0.25)=1.25. This can be viewed as an observation number. However, there is obviously no observation 1.25.

Definition 1 returns a weighted average. The value returned is 25% (25% is the fractional part of 1.25 expressed as a percentage) of the distance between observations 1 and 2:

$$\text{percentile} = 1 + (0.25)(3 - 1) = 1.5$$

Definition 2 rounds to the nearest observation number. Thus, the value 1.25 is rounded to 1 and the first observation, 1, is taken as the 25th percentile. If *np* were 1.5, then the second observation is selected as the 25th percentile.

Definition 3 always rounds up. Thus, 1.25 rounds up to 2 and the second data value, 3, is taken as the 25th percentile.

Definition 4 is a weighted average similar to definition 1, except instead of using *np*, definition 4 uses (*n*+1)*p*=1.5.

$$\text{percentile} = 1 + (0.5)(3 - 1) = 2$$

Definition 5 rounds up to the next observation number unless *np* is an integer, in which case an average of the observations represented by *np* and (*np* + 1) is calculated. In this example, definition 5 rounds up, and the 25th percentile is 3.

Appendix D Advanced Programs

D.1 Interaction Plot

To visualize the interaction, output the final parameter estimates to a data set using the OUTEST= option in the LOGISTIC procedure. It is a good idea to examine the data set **betas** to see what the variable names are.

```
proc logistic data=sasuser.b_sales_inc des outest=betas;
   class gender (param=ref ref='Male');
   model purchase=gender inclevel gender*inclevel;
run;
```

A DATA step with two DO loops is used to create a data set with plotting points. The data points include all possible combinations of **gender** and **income** and the interaction of **gender*income**.

```
data plot;
   do genderfemale=0,1;
      do inclevel=1,2,3;
         genderfemaleinclevel=genderfemale*inclevel;
         output;
      end;
   end;
run;
```

The SCORE procedure multiplies values from two SAS data sets, one containing the coefficients and the other containing data to be scored using the coefficients from the first data set.

Selected PROC SCORE statement options:

OUT= names the SAS data set created by PROC SCORE.

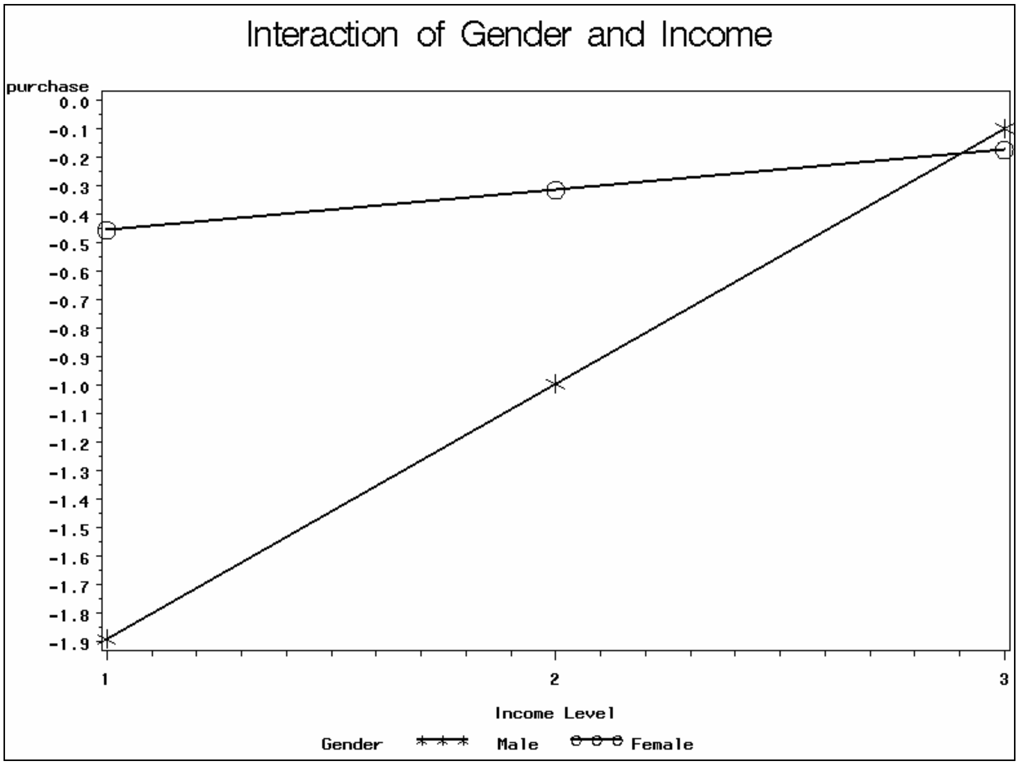
SCORE= names the data set that contains the coefficients.

TYPE= specifies the observations in the SCORE= data set that contain scoring coefficients.

```
proc score data=plot out=scored score=betas type=parms;
   var genderfemale inclevel genderfemaleinclevel;
run;
```

The GPLOT procedure is used to create the interaction plot. The variable **genderfemale** (produced in the **betas** data set) is formatted, and labels are written in the horizontal axis and the legend.

```
proc format;
   value genfmt 1='Female'
                0='Male';
run;

proc gplot data=scored;
   plot purchase*inclevel=genderfemale / haxis=axis1
   legend=legend1;
   format genderfemale genfmt.;
   axis1 label=("Income Level");
   legend1 label=("Gender");
   symbol1 c=black w=2 h=2 i=join v=star;
   symbol2 c=black w=2 h=2 i=join v=circle;
   title "Interaction of Gender and Income";
run;
quit;
```



Appendix E Randomization Technique

E.1 Randomize Paints

A DATA step is used to generate the 28 observations for the completely randomized experiment. Each of the seven roads is given four stripe identification numbers. The variable **`random`** has been generated using a seed of 47, yet any positive integer would suffice. Selected variables of the data set **`stripes`** are printed for verification of the data.

```
options ls=75 ps=55  nodate nonumber;

/* associate a road with a number */
proc format;
   value roadid  1='Center  '
                 2='Broadway'
                 3='Main    '
                 4='Elm     '
                 5='Station '
                 6='Park    '
                 7='Beech   '
                 ;
run;

data stripes;

   stripe_id = 0;
   do r = 1 to 7; /* # of roads */

      road = put(r,$roadid.);
      do s = 1 to 4; /* # of paints      */
                     /* 7 * 4 = 28 obs.  */
         stripe_id = stripe_id + 1;
         random = ranuni(47);
         output;
      end; /* s */
   end; /* r */

   drop
      r s;
run;

proc print data=stripes;
   id road;
   var stripe_id;
   title 'Stripe-ID for each Road';
run;

proc sort data=stripes;
   by random;
run;
```

The data set **`stripes`** is now sorted by the variable **`random`**. The four paints, identified with values `Paint-1`, `Paint-2`, `Paint-3`, and `Paint-4`, are assigned to each of the 28 stripes.

```
/* generate values for paint based on the MOD function, */
/* described below.                                     */
proc format;
   value paintid  0='Paint-4'
                  1='Paint-2'
                  2='Paint-1'
                  3='Paint-3'
                  ;
run;

/* associate the modular of 4 with a paint via the */
/* format PAINTID                                  */
data paints;
   set stripes;
   by random;  /* NOTE: data is sorted by this variable */

   break = mod(_n_,4);/* _n_ is observation number.     */
                      /* MOD computes the remainder of  */
                      /*  the first argument divided by */
                      /*  the second argument.          */

   select (break); /* use select instead of if-then-else */
     when (0) assigned_paint = put(break,$paintid.);
     when (1) assigned_paint = put(break,$paintid.);
     when (2) assigned_paint = put(break,$paintid.);
     when (3) assigned_paint = put(break,$paintid.);
     otherwise;
     end;

   drop
      break random;
run;

proc datasets library=work nolist;
   delete stripes;
run;
```

The data set **paints** is now sorted in two ways: by the paint that was assigned to each stripe and by the road/stripe combination. The latter is best used in the field.

```
proc sort data=paints out=grpd_paints;
   by assigned_paint;
run;

proc print data=grpd_paints;
   by assigned_paint;
   id assigned_paint;
   var road stripe_id;
   title 'Paint #(1,2,3 or 4) ... on Road/Stripe-ID';
run;

proc sort data=paints out=grpd_paints;
   by road stripe_id;
run;

proc print data=grpd_paints;
   by road;
   id road;
   var stripe_id assigned_paint;
   title 'On Road/Stripe-ID, Assign Paint #(1,2,3, or 4)';
run;
```

```
Stripe-ID for each Road

              stripe_
  road           id

  Center          1
  Center          2
  Center          3
  Center          4
  Broadway        5
  Broadway        6
  Broadway        7
  Broadway        8
  Main            9
  Main           10
  Main           11
  Main           12
  Elm            13
  Elm            14
  Elm            15
  Elm            16
  Station        17
  Station        18
  Station        19
  Station        20
  Park           21
  Park           22
  Park           23
  Park           24
  Beech          25
  Beech          26
  Beech          27
  Beech          28
```

```
Paint #(1,2,3 or 4) ... on Road/Stripe-ID

    assigned_                stripe_
      paint      road           id

     Paint-1     Main           10
                 Broadway        5
                 Park           22
                 Broadway        7
                 Station        20
                 Center          3
                 Elm            16

     Paint-2     Elm            13
                 Park           23
                 Beech          25
                 Main           11
                 Main           12
                 Beech          28
                 Station        19

     Paint-3     Elm            14
                 Main            9
                 Station        18
                 Broadway        6
                 Center          1
                 Station        17
                 Elm            15

     Paint-4     Center          4
                 Park           21
                 Park           24
                 Center          2
                 Beech          26
                 Beech          27
                 Broadway        8
```

```
               On Road/Stripe-ID, Assign Paint #(1,2,3, or 4)

                                stripe_    assigned_
                      road         id        paint

                      Beech        25       Paint-2
                                   26       Paint-4
                                   27       Paint-4
                                   28       Paint-2

                      Broadway      5       Paint-1
                                    6       Paint-3
                                    7       Paint-1
                                    8       Paint-4

                      Center        1       Paint-3
                                    2       Paint-4
                                    3       Paint-1
                                    4       Paint-4

                      Elm          13       Paint-2
                                   14       Paint-3
                                   15       Paint-3
                                   16       Paint-1

                      Main          9       Paint-3
                                   10       Paint-1
                                   11       Paint-2
                                   12       Paint-2

                      Park         21       Paint-4
                                   22       Paint-1
                                   23       Paint-2
                                   24       Paint-4

                      Station      17       Paint-3
                                   18       Paint-3
                                   19       Paint-2
                                   20       Paint-1
```