

Chapter 1 Introduction to Statistics

1.1 Fundamental Statistical Concepts

Objectives

- Decide what tasks to complete before analyzing the data.
- Use the MEANS procedure to produce descriptive statistics.

3


Defining the Problem

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




4


As a project, students in Ms. Chao's statistics course must assess whether the students at magnet schools (schools with special curricula) in their district accomplished a goal of the Board of Education. The board wants the graduating class to attain a combined score of 1200 on the Math and Verbal portions of the SAT (the Scholastic Aptitude Test, a college admissions exam). Each section of the SAT has a maximum score of 800. Eighty students are selected at random from among magnet school students in the district. The total scores are recorded and each sample member is assigned an identification number.




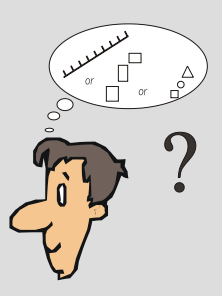
Variable Type and Level of Measurement

VARIABLE

 AGREE

 NO OPINION


 DISAGREE



Before analyzing, identify the variable type (continuous or categorical) and level of measurement (nominal or ordinal).



5

There are a variety of statistical methods for analyzing data. To choose the appropriate method, you must determine the type and level of measurement for your variables.



Continuous versus Categorical Variables

Variable: Temperature of Beverage



Variable: Gender

6

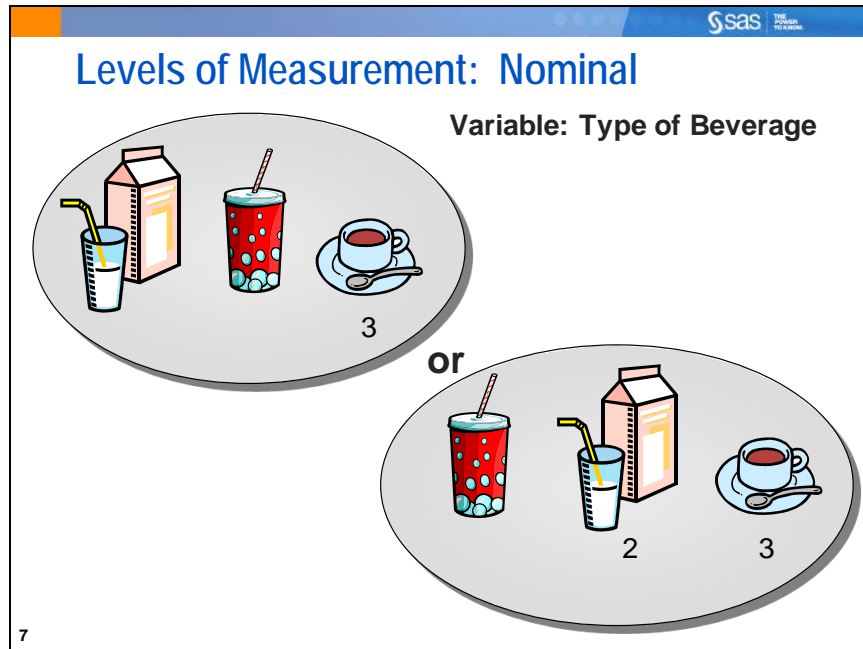
Continuous variables can, in theory, take on any of an infinite number of possible values between two numbers. An example of a continuous variable would be temperature of a beverage. There is no restriction to the number of values between 22° and 23°. You could have a temperature of 22.45° or 22.98° Fahrenheit. Notice that a variable can be continuous even if your measurement system has finite intervals.

Some numeric variables are not continuous. These include variables such as counts that can only take on specific values (for example, integers). In many statistical applications, methods for continuous data can be applied to these variables as well.


Categorical variables are variables that represent groupings. Categorical variables can be stored as numeric or non-numeric values in SAS. Examples of categorical variables include **gender** (**male** or **female**) and **size** of a product (**Small**, **Medium**, **Large**).



It should be noted that continuous variables, through the process of binning, could be made into categorical variables.




Nominal variables have values with no logical ordering. In this example, drink **type** is a nominal variable, even though numeric values are assigned to the categories.





Levels of Measurement: Ordinal


Variable: Size of Beverage



8

Ordinal variables have values with a logical order. However, the relative distances between the values are not clear. In this example, drink **size** is categorical. The number assignments to the categories convey information of the relative size of the drinks, but not precise information about the quantitative differences.

-  For mathematical completion, there are two other levels of measurement (interval and ratio). Interval and ratio variables have a shared property that we are able to calculate the distance between two ranked values. Ratio variables differ from that of interval in the existence of a “true” zero point. This zero point allows ratios to be calculated.
-  The SAT score variable used in this chapter is not truly continuous. It is argued that educational test scores have a linear relationship with measures that truly are continuous and therefore can be analyzed as if they were continuous. An SAT score is treated as a continuous, interval-level measure in this course.

<div>  </div>			
Overview of Statistical Models			
<div> <div>Type of Predictors</div> <div>Type of Response</div> </div>	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression


This course deals with statistical modeling. The type of modeling depends on the level of measurement of two types of variables.

The first type of variable is called *Response Variables*. These are the variables that generally are the focus of business or research. They are also known as *outcome variables* or *target variables* or (in designed experiments) *dependent variables*.

The second type of variable is referred to as *Predictor Variables*. These are the measures that are theoretically associated with the response variables. They can therefore be used to “predict” the value of the response variables. They are also known as *independent variables* in analysis of data from designed experiments.

Categorical data analysis is concerned with categorical responses, regardless of whether the predictor variables are categorical or continuous. Categorical responses have a measurement scale consisting of a set of categories.

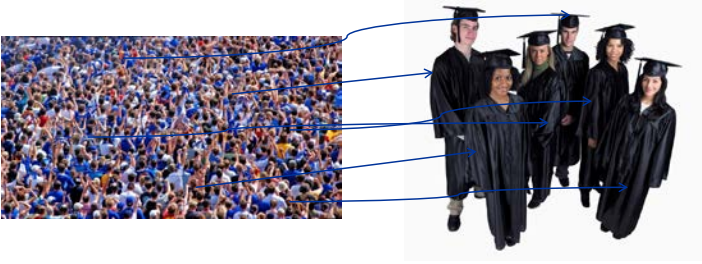
Continuous data analysis is concerned with the analysis of continuous responses, regardless of whether the predictor variables are categorical or continuous



THE POWER OF DATA

Populations and Samples

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

Sample – a subset of a population drawn to enable inferences to the population.



 **Assumption for This Course – The sample that is drawn is *representative* of the population.**

10

A *population* is a collection of all objects about which information is desired, for example:


- all potential customers of a bank
- all copper wires of 1/8" diameter and 36" length
- all students in Carver schools magnet programs

A *sample* is a subset of the population. The sample should be ***representative*** of the population, meaning that the sample's characteristics are similar to the population's characteristics. Examples of samples are as follows:

- 500 bank customers responding to a survey
- 50 randomly selected copper wires of 1/8" diameter and 36" length
- 80 students in Carver schools magnet programs

Simple random sampling, a technique in which each member of the population has an equal probability of being selected, is used by Ms. Chao's students. Random sampling can help ensure that the sample is representative of the population.

In a simple random sample, every member of the population has an equal chance of being included. In the test scores example, each student has an equal chance of being selected for the study.

 See the appendix for information about how to generate random samples without replacement and with replacement.

Why not select only the students from Ms. Chao's class?

When you only select students that are easily available to you, you are using *convenience sampling*. Convenience sampling can lead to biased samples. A *biased* sample is one that is not representative of the population from which it is drawn.

In the example, the average test scores of only Ms. Chao's students might not be close to the true average of the population. This can cause the students to reach incorrect conclusions about the true average score and the variability of scores in the school district. This would not impress Ms. Chao.

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

11

Parameters are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

Suppose you have x_1, x_2, \dots, x_n , a sample from some population.

$$\bar{x} = \frac{1}{n} \sum x_i$$

The mean is an average, a typical value in the distribution.

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

The variance measures the sample variability.

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

The standard deviation measures variability. It is reported in the same units as the mean.

Describing Your Data

When you describe data, your goals are as follows:

- characterize the central tendency
- inspect the spread and shape of continuous variables
- screen for unusual data values

12

After you select a random sample of the population, 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, must be identified.
- An extreme value of a variable, if not detected, could cause gross errors in the interpretation of the statistics.

1.01 Multiple Answer Poll

A sample from a population should be which of the following?

- a. Random
- b. Representative
- c. Normal

14



Test Score Data Set

<u>Gender</u>	<u>SATScore</u>	<u>IDNumber</u>
Male	1170	61469897
Female	1090	33081197
Male	1240	68137597
Female	1000	37070397
Male	1210	64608797
Female	970	60714297
Male	1020	16907997
Female	1490	9589297
Male	1200	93891897
Female	1260	5859397
...

16


Example: The identification number of each student (**IDNumber**) and the total score on the SAT (**SATScore**) are recorded. The data are stored in the **sasuser.testscores** data set.



You might be curious as to whether the girls in the schools have a different average score than the boys. This possibility is discussed later in the chapter.



The SAT is not a truly continuous measure. Scores are functions of counts of correct, incorrect, and unanswered questions. Measures of this type exist in many areas of statistical analysis. While the measure is not truly continuous, scores on the SAT behave similar to continuous measures in analysis. Therefore, statistical techniques created for continuous measures on discrete, multi-level measures are often used.



Distributions

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

- the range of possible data values
- the frequency of data values
- whether the data values accumulate in the middle of the distribution or at one end

17

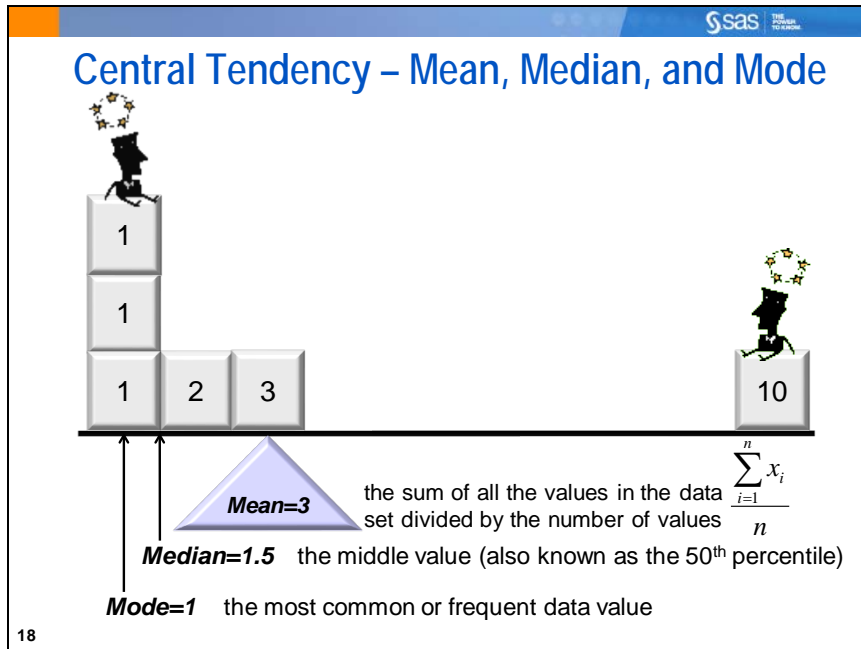
A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any type of data, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.

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

- Are the values of **SATScore** symmetrically distributed?
- Are any values of **SATScore** unusual?

You can answer these questions using descriptive statistics.

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



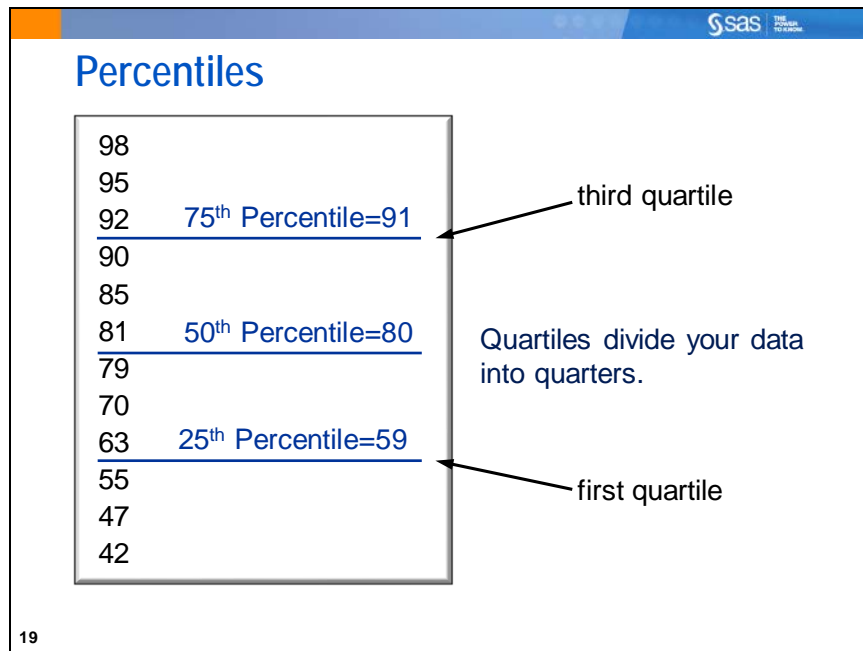
Descriptive statistics that locate the center of your data are called *measures of central tendency*. The most commonly reported measure of central tendency is the sample mean. It is most appropriate for variables measured on an interval or ratio scale with an approximately symmetrical distribution.

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, $\sum (x_i - \bar{x}) = 0$.

The *mean* is the arithmetic balancing point of your data.

The *median* is the data point in the middle of a sorted sequence. It is appropriate for either rank scores (variables measured on an ordinal scale) or variables measured on an interval or ratio scale with a skewed distribution.

The *mode* is the data point that occurs most frequently. It is most appropriate for variables measured on a nominal scale. There might be several modes in a distribution.



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

Commonly reported percentile values are the following:

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

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

20

Measures of dispersion enable you to characterize the dispersion, or spread, of the data.

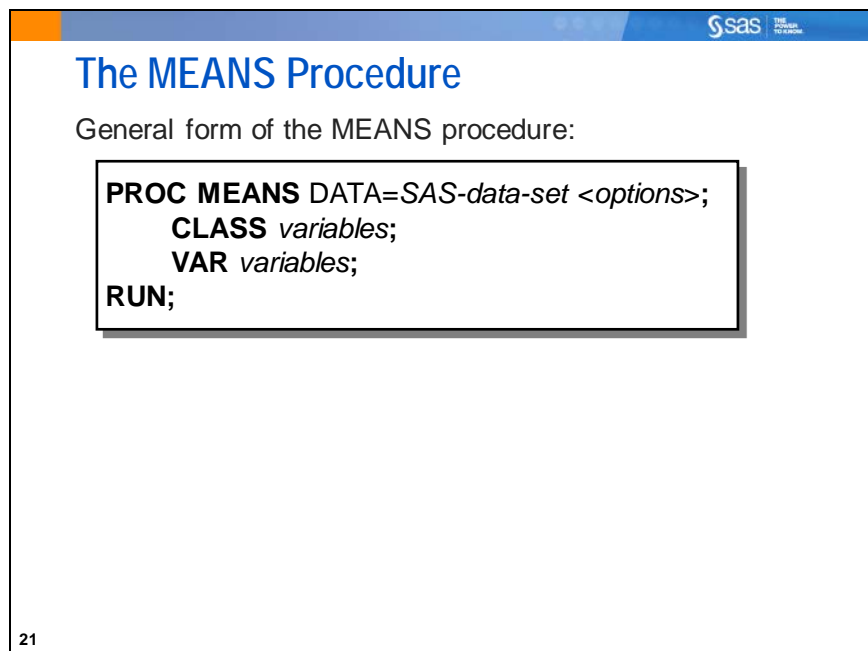
Formula for sample variance: $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$



Another measure of variation is the coefficient of variation (C.V.), which is the standard deviation as a percentage of the mean.

It is defined as $\frac{s}{\bar{x}} \times 100$.

The variance and standard deviation are typically reported where the measure of central tendency is the mean. Where the distribution is skewed, the data contains several extreme outliers, or the variable is measured on an ordinal scale, a value better suited to reflect dispersion is the *interquartile range*. The interquartile range shows the range of the middle 50% of data values.



The slide features a blue header with the SAS logo and the text 'THE POWER OF DATA'. The main title is 'The MEANS Procedure' in blue. Below it, the text 'General form of the MEANS procedure:' is followed by a code block containing the SAS syntax for the MEANS procedure. The slide number '21' is in the bottom left corner.

The MEANS Procedure

General form of the MEANS procedure:

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

21

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

Selected MEANS procedure statements:

CLASS specifies the variables whose values define the subgroup combinations for the analysis. Class variables are numeric or character. Class variables can have continuous values, but they typically have a few discrete values that define levels of the variable. ***You do not have to sort the data by CLASS variables.***

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** in the command box. This opens the Help window, which accesses SAS documentation. After you locate the appropriate procedure, select **syntax** to see all options available for that procedure.



Descriptive Statistics

Example: Use the PRINT procedure to list the first 10 observations in the **sasuser.testscores** data set. Then use PROC MEANS to generate descriptive statistics for **SATScore**.



Submit the program **st100d01.sas** before running the programs in this course. Formats for the data sets are written there.

Partial Code

```
options nodate nonumber ls=95 ps=80 formdlim='- ';
ods noproctitle;

proc format;
...

data sasuser.testscores;
  input Gender $ 1-6 SATScore 8-11 IDNumber 13-20;
  datalines;
...
```

Selected SAS system options:

NODATE	specifies that the date and the time are not printed.
NONUMBER	specifies that SAS not print the page number on the first title line of each page of SAS output.
LINESIZE= (LS=) <i>n</i>	specifies the line size (printer line width) in characters for the SAS log and the SAS output that are used by the DATA step and procedures.
PAGESIZE= (PS=) <i>n</i>	specifies the number of lines that compose a page.
FORMDLIM=	specifies in quotation marks a character written to delimit pages. Normally, the delimit character is null.

Selected ODS statement options.

NOPROCTITLE	suppresses the writing of the title of the procedure that produces the results.
-------------	---

Code for the demonstration starts here:

```
/*st101d01.sas*/ /*Part A*/
proc print data=sasuser.testscores (obs=10);
  title 'Listing of the SAT Data Set';
run;
```

Obs	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397
5	Male	1210	64608797
6	Female	970	60714297
7	Male	1020	16907997
8	Female	1490	9589297
9	Male	1200	93891897
10	Female	1260	85859397

```

/*st101d01.sas*/  /*Part B*/
proc means data=sasuser.testscores;
  var SATScore;
  title 'Descriptive Statistics Using PROC MEANS';
run;

```

Analysis Variable : SATScore				
N	Mean	Std Dev	Minimum	Maximum
80	1190.63	147.0584466	890.0000000	1600.00

By default, PROC MEANS prints the number of nonmissing observations (N), the mean, the standard deviation, the minimum value, and the maximum value. You can add options to the MEANS statement to request additional or alternate statistics. When you add options to request specific statistics, only the requested statistics appear in the output. In addition, you can control the number of decimal places that are displayed.

```

/*st101d01.sas*/  /*Part C*/
proc means data=sasuser.testscores
  maxdec=2
  n mean median std q1 q3 qrange;
  var SATScore;
  title 'Selected Descriptive Statistics for SAT Scores';
run;

```

Selected PROC MEANS statement options:

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

Analysis Variable : SATScore						
N	Mean	Median	Std Dev	Lower Quartile	Upper Quartile	Quartile Range
80	1190.63	1170.00	147.06	1085.00	1280.00	195.00



Exercises

1. Calculating Basic Statistics in PROC MEANS

The data in **sasuser.NormTemp** comes from an article in the *Journal of Statistics Education* by Dr. Allen L. Shoemaker from the Psychology Department at Calvin College. The data are based on an article in a 1992 edition of *JAMA (Journal of the American Medical Association)*, which questions the notion that the true mean body temperature is 98.6. There are 65 males and 65 females. There is also some question about whether mean body temperatures for women are the same as for men. The variables in the data set are as follows:

ID	Identification number
BodyTemp	Body temperature (degrees Fahrenheit)
Gender	Coded (Male , Female)
HeartRate	Heart rate (beats per minute)

Use PROC MEANS to answer these questions:

- What is the overall mean and standard deviation of body temperature in the sample?
- What is the interquartile range of body temperature?
- Do the mean values seem to differ between men and women?

Hint: Use the **CLASS** statement in PROC MEANS, with **Gender** as the class variable.

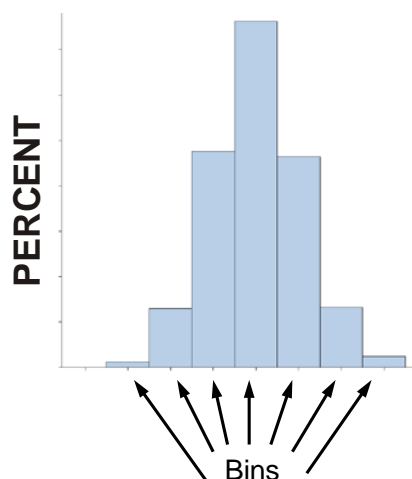
1.2 Picturing Distributions

Objectives

- Look at distributions of continuous variables.
- Describe the normal distribution.
- Use the UNIVARIATE procedure to generate histograms and normal probability plots and to produce descriptive statistics.
- Use the SGPLOT procedure to generate box plots.

26

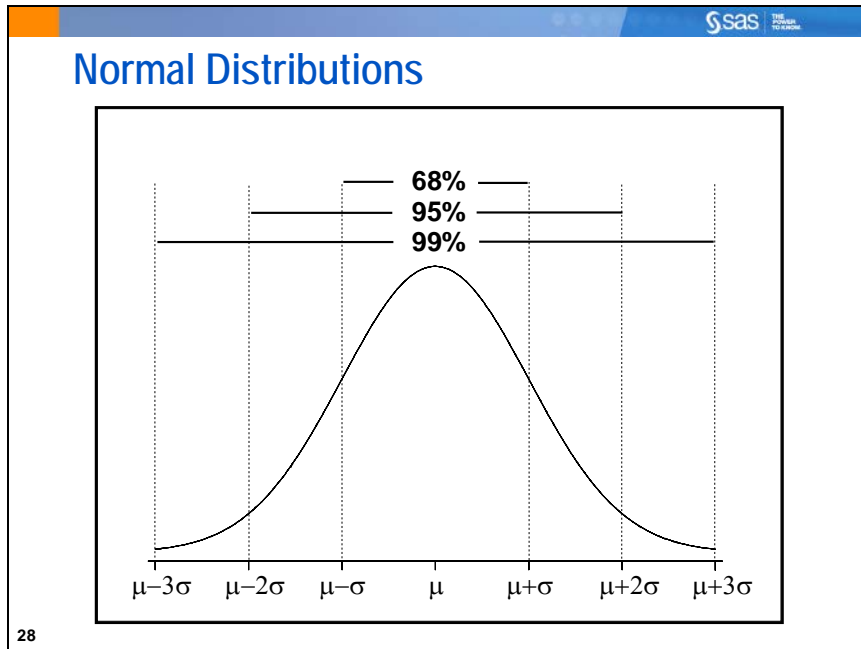
Picturing Distributions: Histogram



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

27

Most elementary statistical procedures assume some underlying population probability distribution. It is a good idea to look at your data to see whether the distribution of your sample data can reasonably be assumed to come from a population with the assumed distribution. A histogram is a good way to determine how the probability distribution is shaped.



Quite often in analysis, although not always, a normal distribution is assumed.

The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the “probability density” at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle. The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ . A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its center point) and the standard deviation scales it.

An observation value is considered unusual if it is far away from the mean. How far is far? You can use the mathematical properties of the normal probability density function (PDF) to determine that. If a population follows a normal distribution, then approximately the following is true:

- 68% of the data fall within 1 standard deviation of the mean.
- 95% of the data fall within 2 standard deviations of the mean.
- 99.7% of the data fall within 3 standard deviations of the mean.

Often, values that are more than two standard deviations from the mean are regarded as unusual. Now you can see why. Only about 5% of all values are at least that far away from the mean.

You use this information later when you discuss the concepts of confidence intervals and hypothesis tests.

Normal Distributions

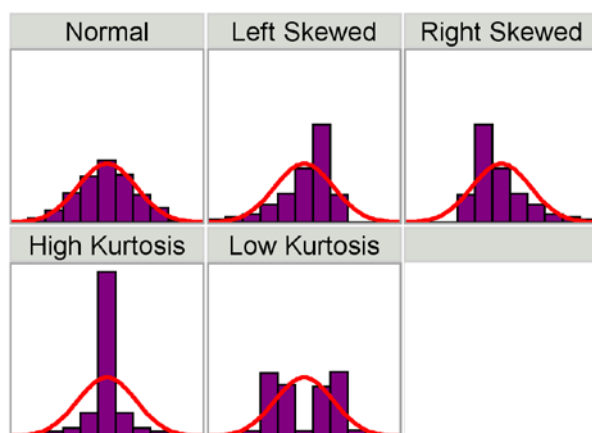
A normal distribution

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

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

29

Data Distributions Compared to Normal

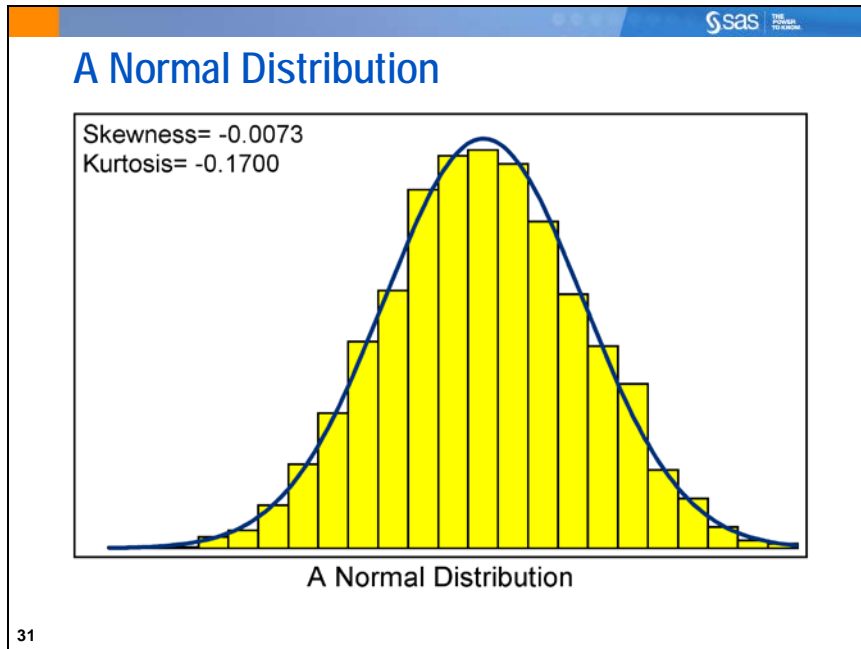


30

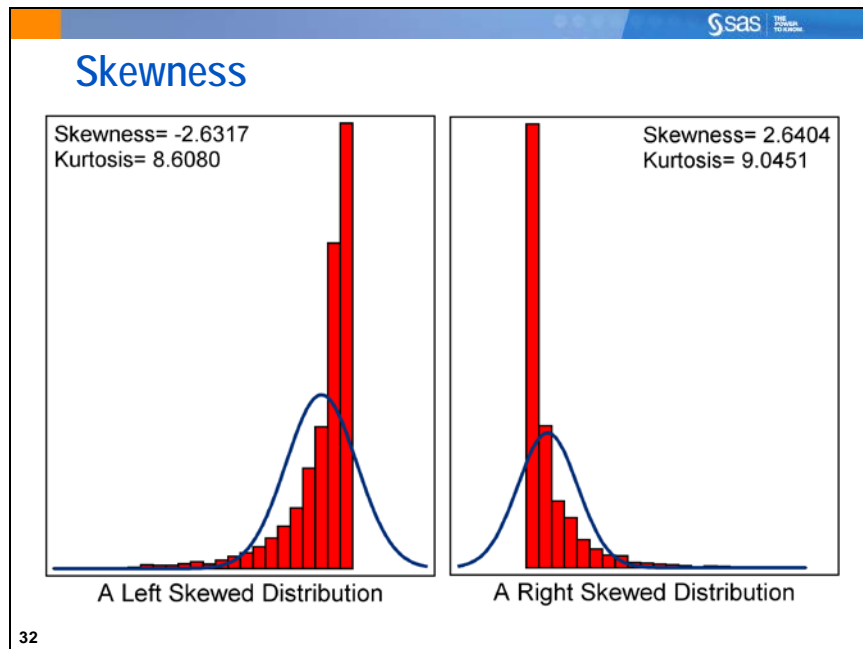
The distribution of your data might not look normal. There are an infinite number of different ways that a population can be distributed. When you look at your data, you might notice the features of the distribution that indicate similarity or difference from the normal distribution.

In evaluating distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.

Two such measures are skewness and kurtosis.



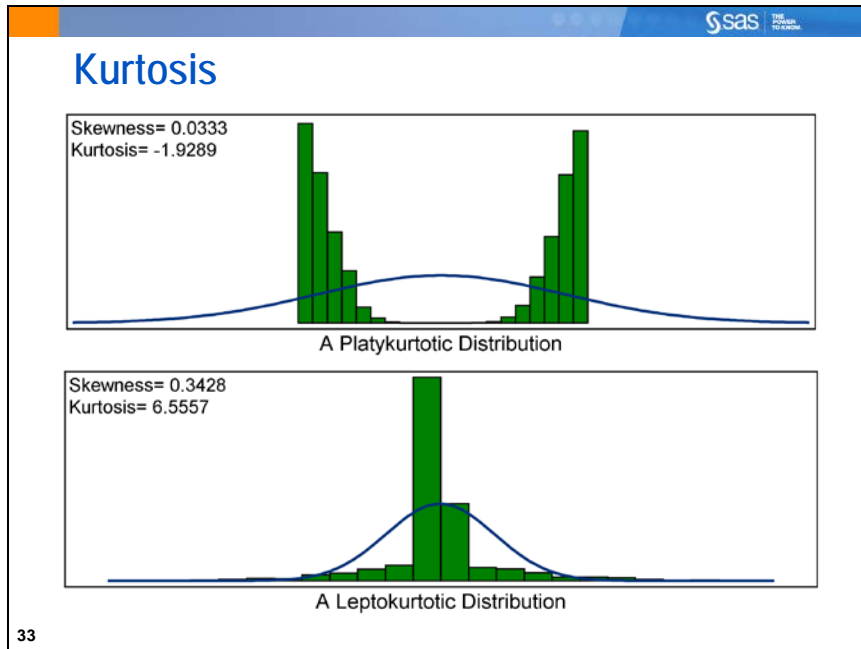
A histogram of data from a sample drawn from a normal population generally shows values of skewness and kurtosis near 0 in SAS output.



One measure of the shape of a distribution is skewness. 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, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- **right** side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.



Kurtosis measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to 0 in SAS. Kurtosis is often very difficult to assess visually.

If the value of your kurtosis statistic is negative, the distribution is said to be *platykurtic*. If the distribution is both symmetric and platykurtic, then there tends to be a smaller-than-normal proportion of observations in the tails and/or a somewhat flat peak. Rectangular, bimodal, and multimodal distributions tend to have low (negative) values of kurtosis.

If the value of the kurtosis statistic is positive, the distribution is said to be *leptokurtic*. If the distribution is both symmetric and leptokurtic, then there tends to be a larger-than-normal proportion of observations in the extreme tails and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.

Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

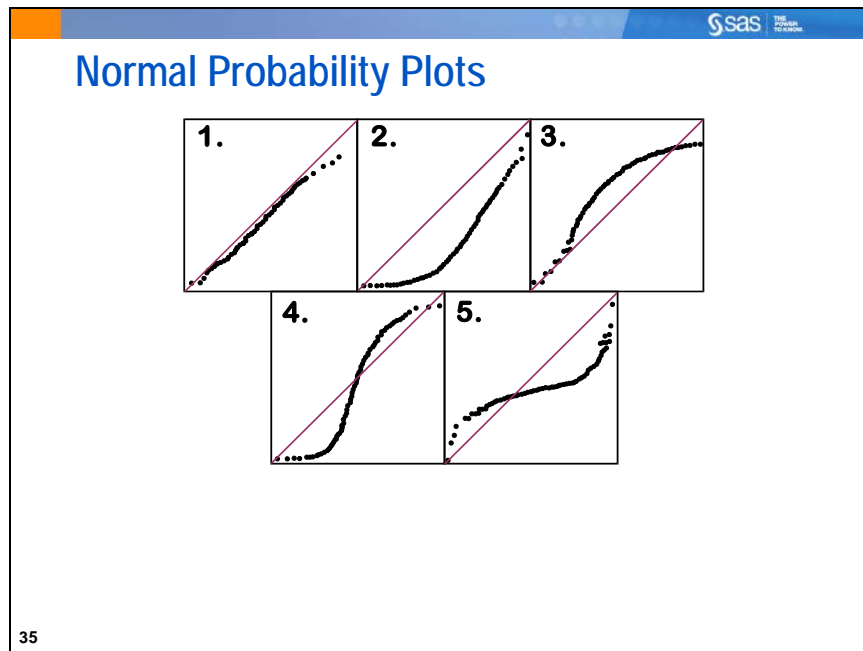


The normal distribution actually has a kurtosis value of 3, but SAS subtracts a constant of 3 from all reported values of kurtosis, making the constant-modified value for the normal distribution 0 in SAS output. That is the value against which to compare a sample kurtosis value in SAS when assessing normality. This value is often referred to as *relative kurtosis*.

Graphical Displays of Distributions

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

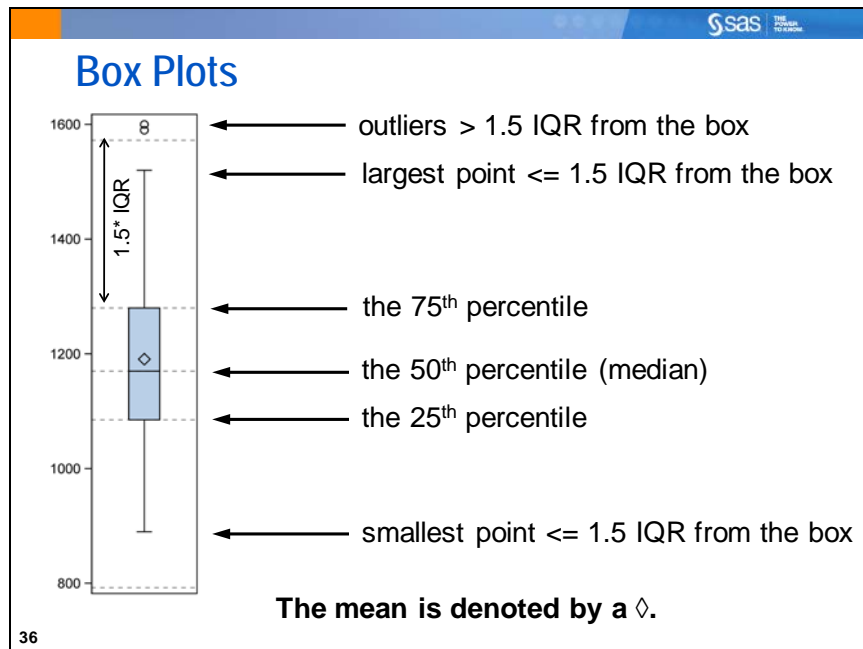
- histograms
- normal probability plots
- box plots



A *normal probability plot* is a visual method for determining whether your data come from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.

The above diagrams illustrate some possible normal probability plots for data from the following:

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



Box plots (Tukey 1977) (sometimes referred to as *box-and-whisker plots*) provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values). You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile range (IQR) units. If any values lay more than 1.5 IQR units from either end of the box, they are represented in SAS by individual plot symbols.

The plot above shows that the data are approximately symmetric.

ODS Graphics Output

- Some graphs are created by default.
- Procedure options (such as PLOTS=) are used to specify which graphs to create.
- You can specify where you want your graphs displayed by using ODS destination statements (for example, LISTING, HTML, RTF).
- ODS SELECT and ODS EXCLUDE statements can be used to select and exclude information from your output.

37

In this course, you use ODS Statistical Graphics for graphical analysis of data. ODS Statistical Graphics were first made available in SAS 9.2. In SAS 9.3, statistical graphics from SAS statistical procedures are produced unless the ODS GRAPHICS OFF statement is submitted. This statement only needs to be submitted once within an interactive SAS session (or batch job) and remains in effect until the ODS GRAPHICS ON (ODS GRAPHICS) statement is submitted.

The SAS documentation lists the available graphics in the description of the SAS procedure.

ODS templates can be used to modify the layout and details of each graph.



ODS Statistical Graphics can also be toggled on and off by checking and unchecking the appropriate box by selecting **Tools** ⇒ **Options** ⇒ **Preferences** ⇒ **Results**.

Some Recommended ODS Styles	
Style	Description
HTMLBLUE	This lighter color scheme for HTML content is the default for the HTML destination.
STATISTICAL	Color style recommended for output in Web pages or color print media.
ANALYSIS	Color style with a somewhat different appearance from STATISTICAL.
JOURNAL and JOURNAL2	Gray-scale and pure black-and-white styles, respectively; recommended for graphs in black-and-white publications.
RTF	Used to produce graphs to insert into a Microsoft Word document or a Microsoft PowerPoint slide.

38

ODS styles are used to control the general appearance and consistency of all graphs and tables. (You can use a variety of styles and destinations throughout this course.)

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

39

The UNIVARIATE Procedure

General form of the UNIVARIATE procedure:

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

40

The UNIVARIATE procedure not only computes descriptive statistics, but also provides greater detail about the distributions of the variables.

Selected UNIVARIATE procedure statements:

VAR	specifies numeric variables to analyze. If no VAR statement appears, then 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.
INSET	places a box or table of summary statistics, called an <i>inset</i> , directly in a graph created with a CDFPLOT, HISTOGRAM, PPLOT, PROBPLOT, or QQPLOT statement. The INSET statement must follow the PLOT statement that creates the plot that you want to augment.

Selected option for HISTOGRAM and PROBPLOT statements:

NORMAL<(options)>	creates a normal probability plot. Options (MU= SIGMA=) determine the mean and standard deviation of the normal distribution used to create reference lines (normal curve overlay in HISTOGRAM and diagonal reference line in PROBPLOT).
-------------------	--

The SGPLOT Procedure

General form of the SGPLOT procedure:

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

41

The SGPLOT procedure creates one or more plots and overlays them on a single set of axes. You can use the SGPLOT procedure to create statistical graphics such as histograms and regression plots, in addition to simple graphics such as box plots, scatter plots, and line plots.

Selected SGPLOT procedure statements:

VBOX creates a vertical box plot that shows the distribution of your data.



Examining Distributions

```
/*st101d02.sas*/  /*Part A*/
proc univariate data=sasuser.testscores;
  var SATScore;
  histogram SATScore / normal(mu=est sigma=est) kernel;
  inset skewness kurtosis / position=ne;
  probplot SATScore / normal(mu=est sigma=est);
  inset skewness kurtosis;
  title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```

Selected HISTOGRAM statement options:

KERNEL	superimposes kernel density estimates on the histogram.
NORMAL	displays fitted normal density curves on the histogram. MU=__ specifies mean μ for normal curve. SIGMA=__ specifies standard deviation σ for normal curve. The EST option requests that the value be estimated from the data.

Optional ODS statement:

ODS LISTING (*action*) opens, manages, or closes the LISTING destination.

GPATH= *file-specification* <(url='Uniform-Resource-Locator' | NONE)>
specifies the location for all graphics output that is generated while the destination is open.



By default, output goes to the HTML destination. Other options are RTF, LISTING, and PDF destinations, which can also be opened, managed, and closed by ODS RTF, ODS LISTING, and ODS PDF, respectively. If graphical output is requested for either HTML or LISTING destinations, it is sent to the user's default location. You can select a different location with the GPATH= option.

Selected Output

Moments			
N	80	Sum Weights	80
Mean	1190.625	Sum Observations	95250
Std Deviation	147.058447	Variance	21626.1867
Skewness	0.64202018	Kurtosis	0.42409987
Uncorrected SS	115115500	Corrected SS	1708468.75
Coeff Variation	12.3513656	Std Error Mean	16.4416342

Basic Statistical Measures			
Location		Variability	
Mean	1190.625	Std Deviation	147.05845
Median	1170.000	Variance	21626
Mode	1050.000	Range	710.00000
		Interquartile Range	195.00000


Tests for Location: Mu0=0			
Test	Statistic		p Value
Student's t	t	72.41525	Pr > t <.0001
Sign	M	40	Pr >= M <.0001
Signed Rank	S	1620	Pr >= S <.0001


Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1600
99%	1600
95%	1505
90%	1375
75% Q3	1280
50% Median	1170
25% Q1	1085
10%	1020
5%	995
1%	890
0% Min	890

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
890	69	1490	8
910	74	1520	42
970	6	1520	54
990	51	1590	70
1000	4	1600	25

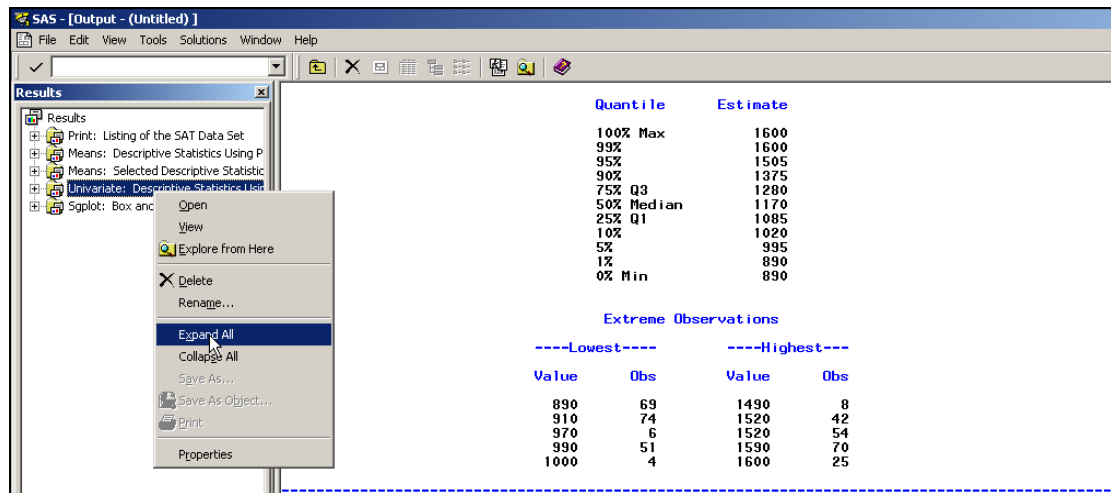
The tabular output indicates the following:

- The mean of the data is 1190.625. This is approximately equal to the median (1170), which indicates the distribution is fairly symmetric.
- The standard deviation is 147.058447, which means that the average variability around the mean is approximately 147 points.
- The distribution is slightly skewed to the right.
- The distribution has slightly heavier tails than the normal distribution.
- The student with the lowest score is observation 69, with a score of 890. The student with the highest score is number 25, with a score of 1600 (highest possible score for the SAT).

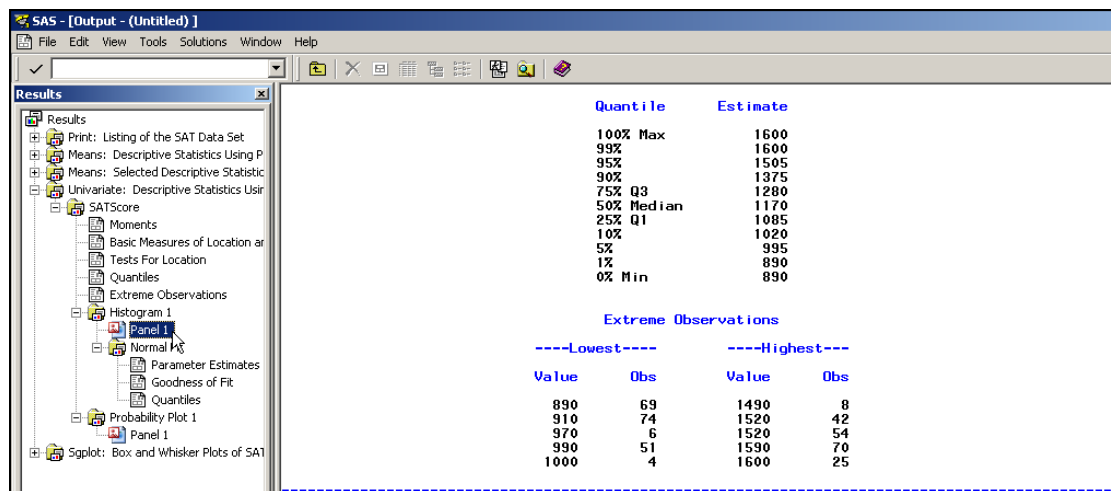
 In the Quantiles table, Definition 5 indicates that PROC UNIVARIATE uses the default definition for calculating percentile values. You can use the PCTLDEF= option in the PROC UNIVARIATE statement to specify one of five methods. These methods are listed in an appendix.

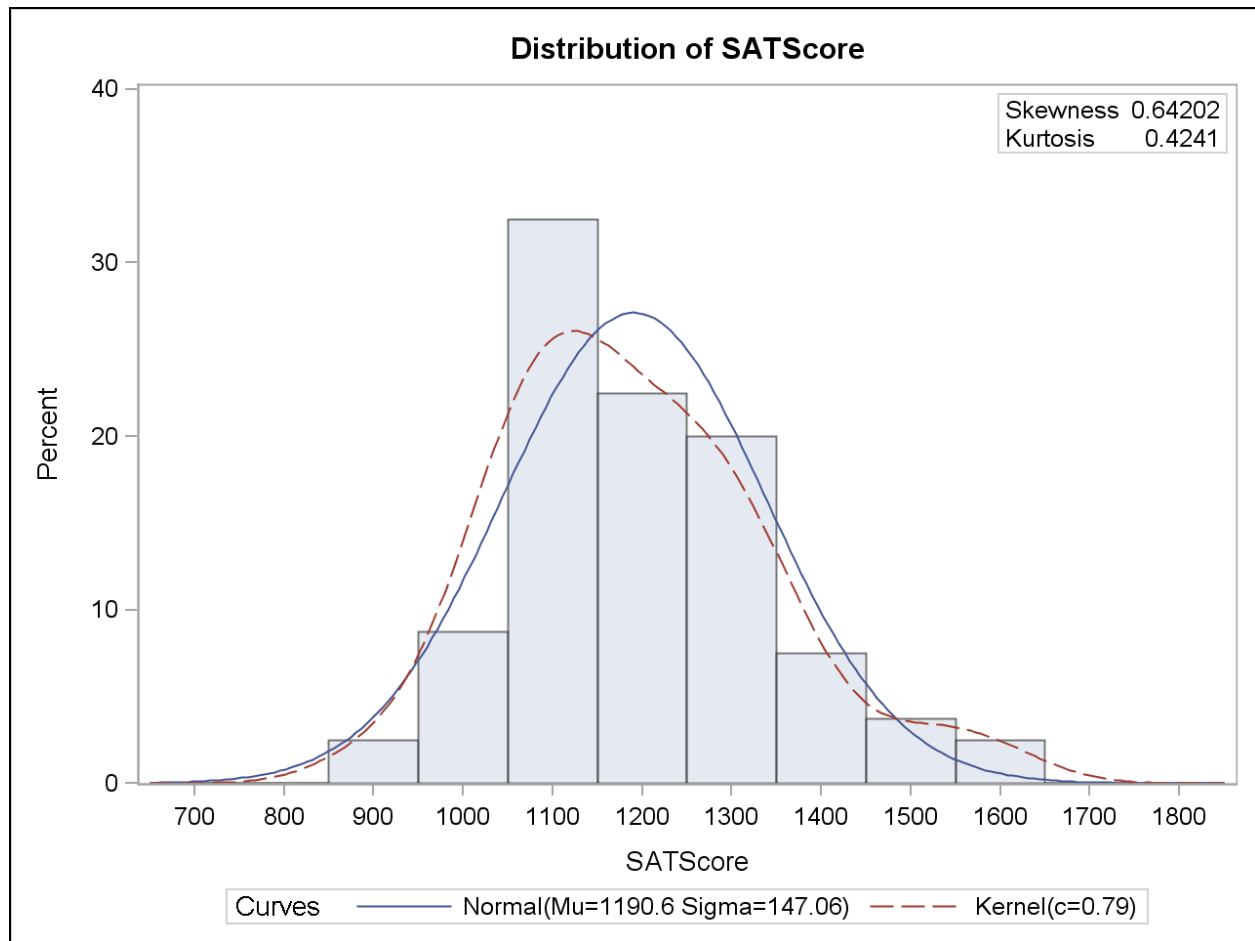
 In order to view the graphical output when you use the SAS windowing environment with the Listing destination active, follow these steps:

1. Expand the output from the Results window by right-clicking on the name of the procedure in the Results window and selecting **Expand All** in the drop-down menu.

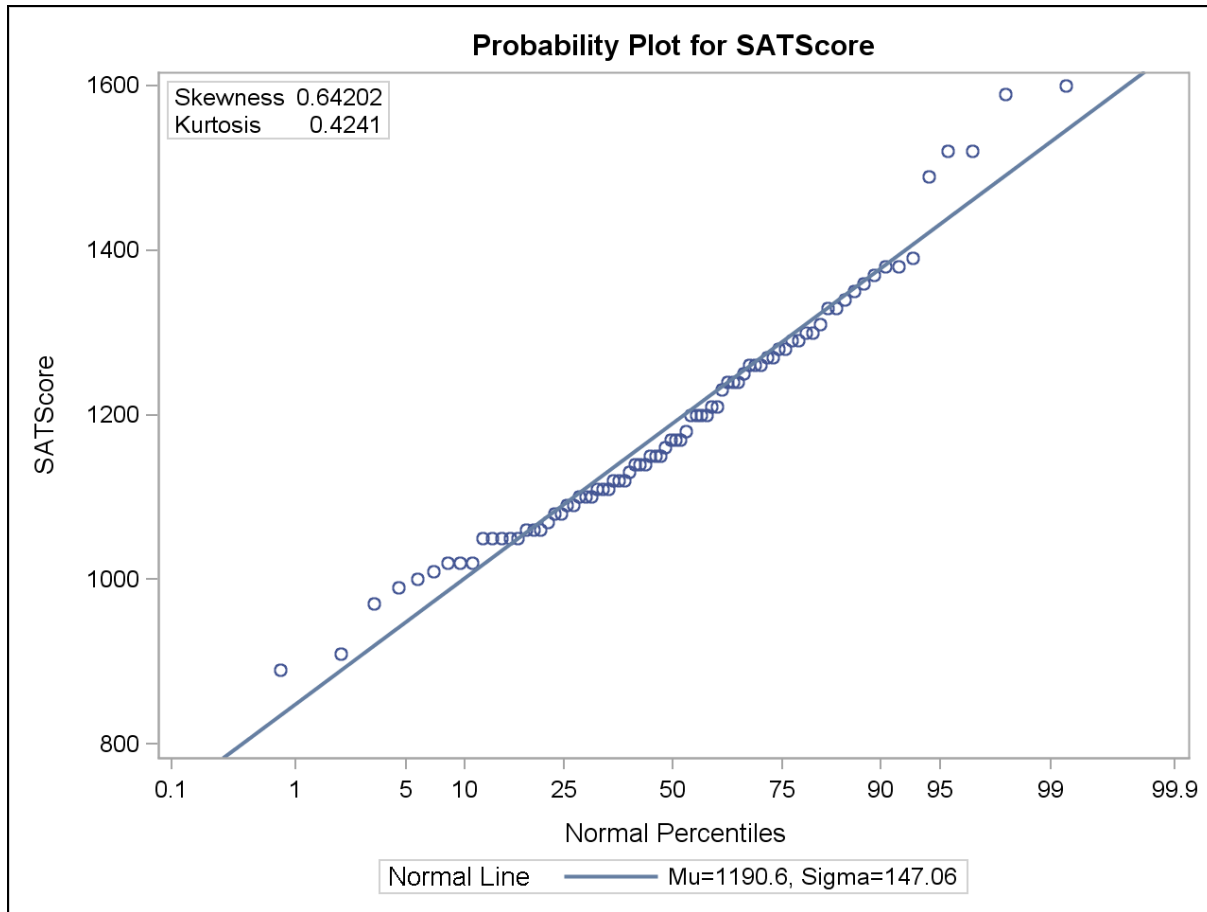


2. Double-click on the image icon to open the image in the user's default graphics software window.





The bin identified with the midpoint of 1100 has approximately 33% of the values. The skewness and kurtosis values are reported in the inset. The kernel density curve is a smoothed version of the histogram and can be used to compare the approximate sample distribution to a normal distribution. In this case, the distribution of the observed data seems to approach normality.



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

Asking for the normal reference curve for the histogram also produces a set of tables relating to assessing whether the distribution is normal or not. There is a table with three tests presented: Kolmogorov-Smirnov, Anderson-Darling, and Cramér-von Mises. In each case, the null hypothesis is that the distribution is normal. Therefore, high p -values are desirable.

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	1190.625
Std Dev	Sigma	147.0584

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08382224	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.09964577	Pr > W-Sq	0.114
Anderson-Darling	A-Sq	0.70124822	Pr > A-Sq	0.068

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
1.0	890.000	848.516
5.0	995.000	948.735
10.0	1020.000	1002.162
25.0	1085.000	1091.436
50.0	1170.000	1190.625
75.0	1280.000	1289.814
90.0	1375.000	1379.088
95.0	1505.000	1432.515
99.0	1600.000	1532.734

All three tests have high p -values (greater than 0.05). This is known as the *alpha level* of the test and is explained in a later section. The high p -values imply that the distribution of **SATScore** is approximately normal.

```
/*st101d02.sas*/ /*Part B*/
proc sgplot data=sasuser.testscores;
  vbox SATScore / datalabel=IDNumber;
  format IDNumber 8.;
  reflate 1200 / axis=y label;
  title "Box-and-Whisker Plots of SAT Scores";
run;
```

Selected PROC SGPLOT statements and options:

VBOX *response-variable* *</ option(s)>*;

creates a vertical box plot that shows the distribution of your data.

[VBOX Statement] **DATALABEL**= *option*

adds data labels for the outlier markers. If you specify a variable, then the values for that variable are used as data labels. If you do not specify a variable, then the values of the response variable are used.

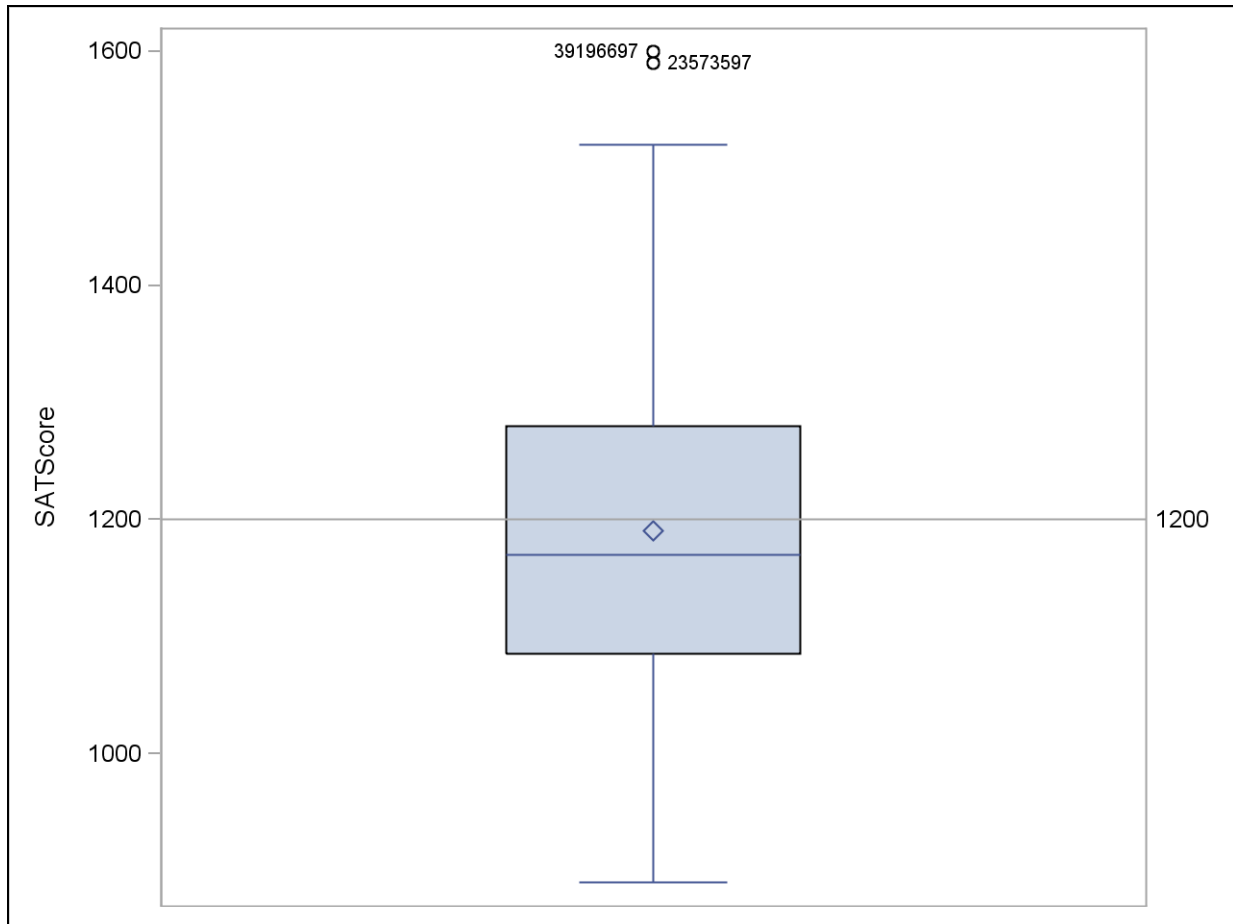
REFLINE *variable* | *value-1* *<... value-n>* *</ option(s)>*;

creates a horizontal or vertical reference line.

[REFLINE Statement] **LABEL**= *option*

creates labels for each reference line. If you do not specify a label for a line, the reference value for that line is used as the label.

A reference line is requested at 1200 on the Y axis. Because this is a vertical box plot, the Y-axis is the **SATScore** axis. A **DATALABEL** option is used to identify potential outliers. If there are no outliers, that option has no effect.



There are two outliers (values beyond 1.5 interquartile units from the box). The **IDnumber** values are displayed.



Exercises

2. Producing Descriptive Statistics

Use the **sasuser.NormTemp** data set to answer the following:

- a. What are the minimum, the maximum, the mean, and the standard deviation for **BodyTemp**? Does the variable appear to be normally distributed?

	BodyTemp
Minimum	
Maximum	
Mean	
Standard Deviation	
Skewness	
Kurtosis	
Distribution: Normal	Yes/No

- b. Create box plots for **BodyTemp**. Use **ID** to identify outliers. Display a reference line at 98.6 degrees. Does the average body temperature seem to be 98.6 degrees?

1.02 Multiple Choice Poll

In the **NormTemp** data set, the distribution of **BodyTemp** seemed to be which of the following?

- a. Close to normal
- b. Left skewed
- c. Right skewed
- d. Having high positive kurtosis
- e. Having high negative kurtosis

1.3 Confidence Intervals for the Mean

Objectives

- Explain and interpret the confidence intervals for the mean.
- Explain the central limit theorem.
- Use PROC MEANS to calculate confidence intervals.

48

Point Estimates

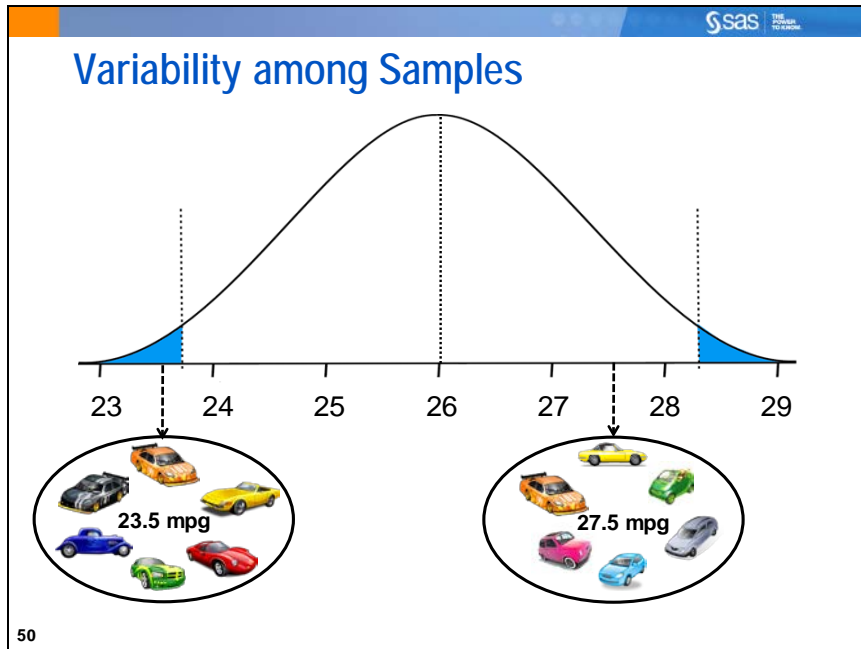
\bar{x} estimates μ

S estimates σ

49

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

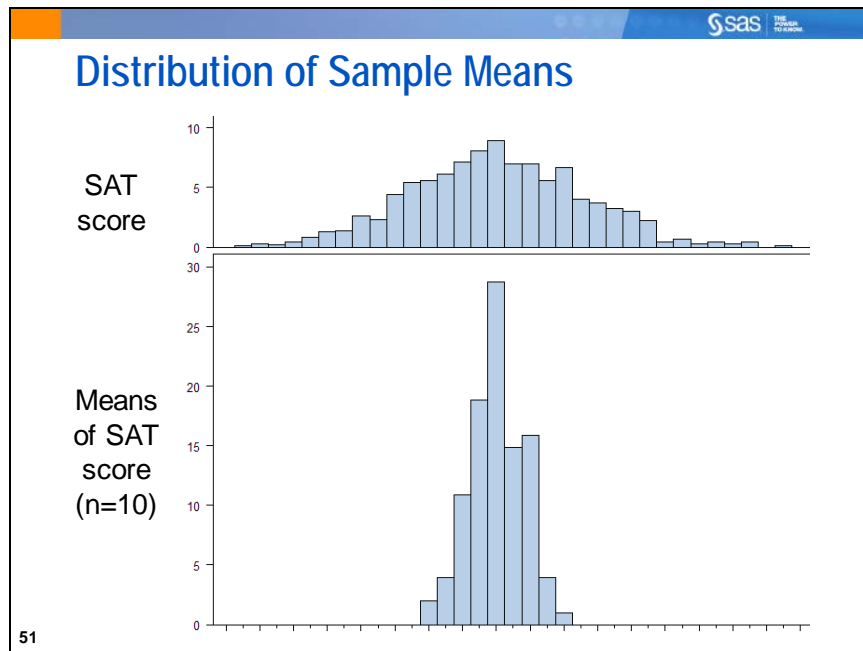
- An estimate of the average **SATScore** is 1190.6, and an estimate of the standard deviation is 147.06.
- Because you only have an estimate of the unknown population mean, you need to know the variability of your estimate.



Why can you not be absolutely certain that the average SAT Math+Verbal score for students in Carver County magnet schools is 1190.6? The answer is because the sample mean is only an estimate of the population mean. If you collected another sample of students, you would likely obtain another estimate of the mean.

Different samples yield different estimates of the mean for the same population. Your mean can be thought of as a selection from a distribution of all possible means. Another sample would likely yield a different value from that distribution.

For example, you could take a random sample of size 6 of cars in your town and measure highway gas mileage. The sample that you choose today might have a mean of 23.5 miles per gallon. Tomorrow's sample from the same population might result in a mean of 27.5.



What is a distribution of sample means? It is a distribution of many mean values, each of a common sample size.

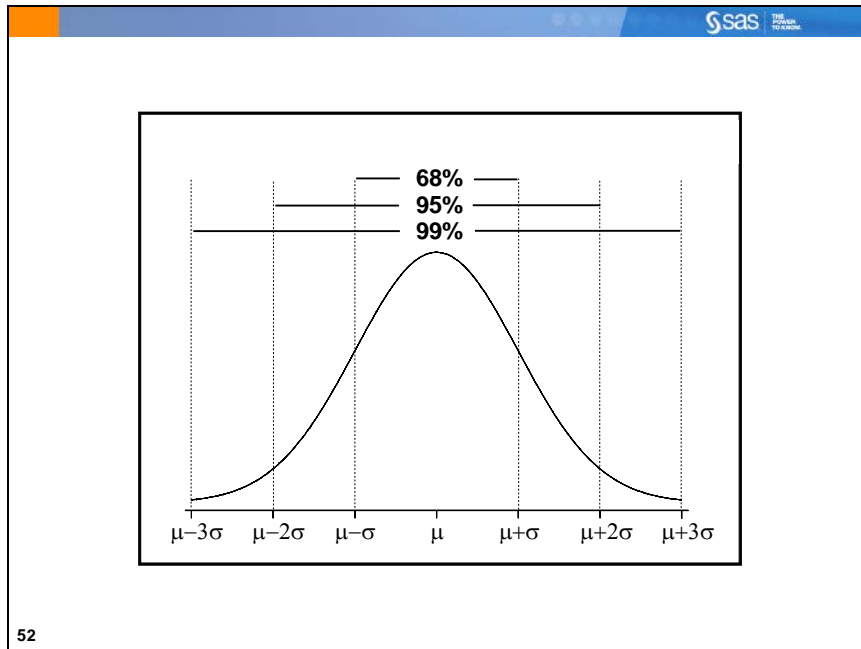
Suppose 1000 random samples, all with the same sample size of 10, are taken from an identified population.

- The top histogram shows the distribution of all 5000 *observations*.
- The bottom histogram, however, represents the distribution of the 1000 *sample means*.

The variability of the distribution of sample means is smaller than the variability of the distribution of the 5000 observations. That should make sense. It seems relatively likely to find one student with an SAT score of 1550 (out of a maximum of 1600), but not likely that a mean of a sample of 10 students would be 1550.



The samples in the 1000 are assumed to be taken with replacement, meaning that after 10 student values are taken, all 10 of those students can be chosen again in subsequent samples.



For purposes of finding confidence limits for parameters (such as a mean), you might make assumptions about a theoretical population distribution. You might, for example, assume normality of sample means.

Standard Error of the Mean

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

It differs from the sample standard deviation because

- the sample standard deviation is a measure of the variability of data
- the standard error of the mean is a measure of the variability of sample means.

– Standard error of the mean = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$

53

The standard error of the mean is computed as follows:

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

where

s is the sample standard deviation.

n is the sample size.

The standard error of the mean for the variable **SATScore** is $147.058447/\sqrt{80}$, or approximately 16.44. This is a measure of how much variability of sample means there is around the population mean. The smaller the standard error, the more precise your sample estimate is.

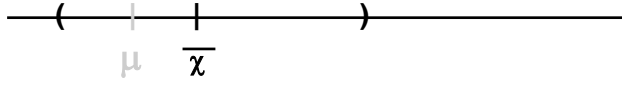


You can improve the precision of an estimate by increasing the sample size.

sas THE SAS INSTITUTE

Confidence Intervals

95% Confidence



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

54

A confidence interval

- is a range of values that you believe to contain the population parameter of interest
- is defined by 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 test score example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean SAT test score 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.

Confidence Interval for the Mean

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

where

\bar{x} is the sample mean.

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

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

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

55

Student's t distribution arises when you make inferences about a population mean and (as in nearly all practical statistical work) the population standard deviation (and therefore, standard error) is unknown and must be estimated from the data. It is approximately normal as the sample size grows larger.

The t in the equation above refers to the number of standard deviation (or standard error) units away from the mean required to get a desired confidence in a confidence interval. That value varies not only with the confidence that you choose, but also with the sample size. For 95% confidence, that t value is usually approximately 2, because, as you have seen, two standard errors below to two standard errors above a mean gives you approximately 95% of the area under a normal distribution curve.

Details

In any normal distribution of sample means with parameters μ and σ , over samples of size n , the probability is 0.95 for the following:

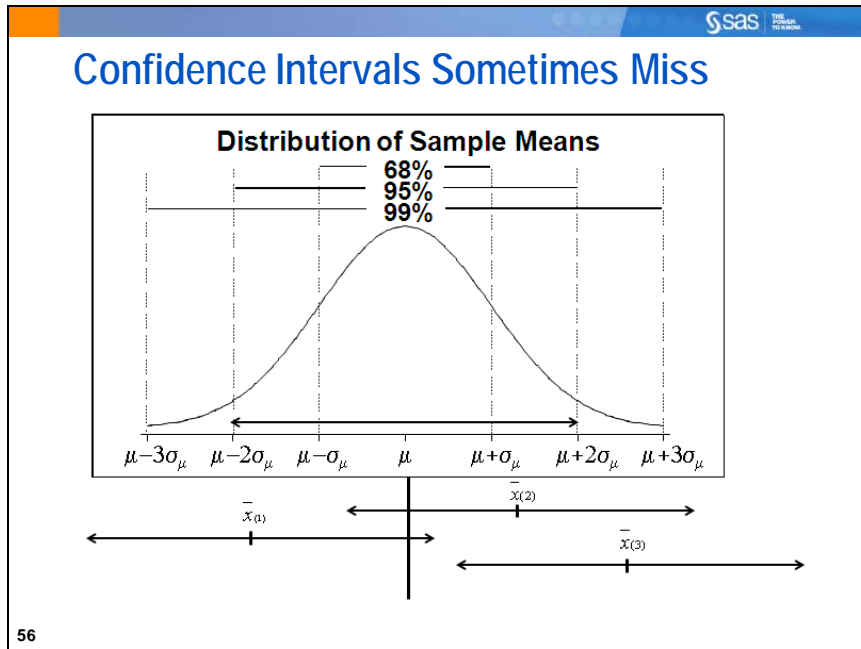
$$-1.96\sigma_{\mu} \leq \bar{x} - \mu \leq 1.96\sigma_{\mu}$$

This is the basis-of-confidence intervals for the mean. If you rearrange the terms above the probability is 0.95 for as shown below:

$$\bar{x} - 1.96\sigma_{\mu} \leq \mu \leq \bar{x} + 1.96\sigma_{\mu}$$

When the value of σ is unknown, one of the family of Student's t distributions is used in place of 1.96 (a value that comes from the normal (z) distribution). The value of 1.96 is replaced by a t -value determined by the desired confidence and the degrees of freedom. When the sample size is larger, the t -value is closer to 1.96. Then also you must replace the known σ_{μ} with the estimated standard error, $s_{\bar{x}}$:

$$\bar{x} - t^* s_{\bar{x}} \leq \mu \leq \bar{x} + t^* s_{\bar{x}}$$



The graph above is the distribution of sample means. You typically take only one sample from that distribution, but in this picture you see that three researchers each took a sample from the same population. Each sample had a different mean. The standard errors are all approximately the same and approximately the same as the population standard error.

The double-headed arrows around each of the means (for researcher 1, 2, and 3) measure approximately two standard errors to each side of the sample mean. (The t value is approximately 2 for these researchers.) The sample means for researcher 1 and 2 fell within two standard errors from the (unknown) population mean, by good luck. Actually, 95% of all researchers should have equivalent “luck.” Researcher number 3 was in the unlucky 5%. He did his work as well and rigorously and then blissfully reported his sample mean and confidence interval. Because his sample mean was more than two standard errors from the (unknown) population mean, his confidence interval did not extend far enough to include that true mean.

If the confidence interval is faithfully calculated using the formula shown earlier and assumptions are met, 95% of the time they include the true mean. Unfortunately, there is no way to know whether yours is in the 95% group or the 5% group.



The observed value of t (the number of standard errors your observed mean is away from a hypothesized mean) is related to a specific probability, known in statistics as a p -value.

Normality and the Central Limit Theorem

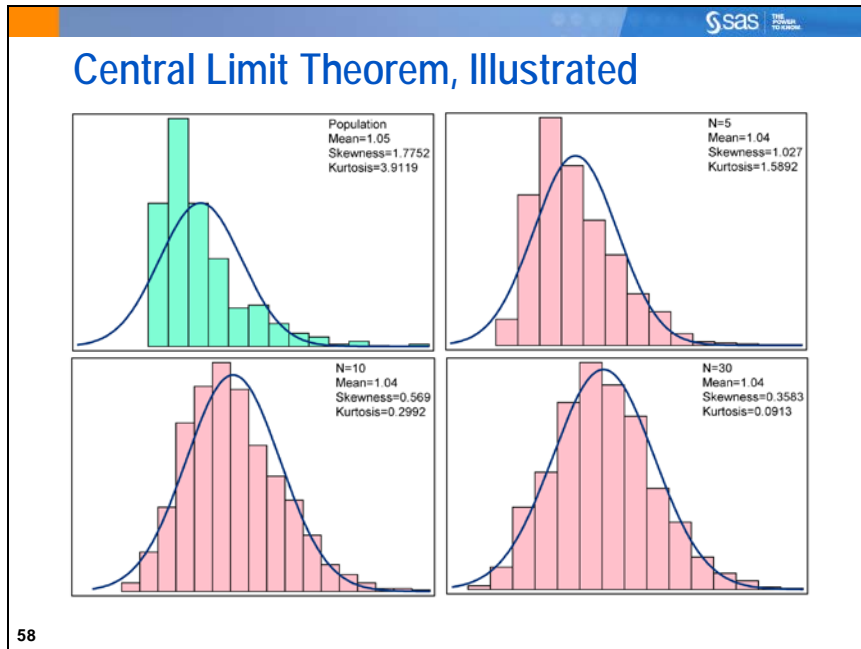
To satisfy the assumption of normality, you can do one of the following:

- Verify that the population distribution is approximately normal.
- Apply the **central limit theorem**.
 - The central limit theorem states that the distribution of sample means is approximately normal, regardless of the population distribution's shape, if the sample size is large enough.
 - “Large enough” is usually about 30 observations. It is more if the data are heavily skewed, and fewer if the data are symmetric.

57

To apply the central limit theorem, the standard rule of thumb with a relatively symmetric population is that your sample size should be at least 30. For skewed populations, the sample size should be greater. The central limit theorem applies even if you have no reason to believe that the population distribution is normal.

Because the sample size for the test scores example is 80 and the random sample implies that the population is relatively symmetric, you can apply the central limit theorem and satisfy the assumption of normality for the confidence intervals of the sample mean.



The 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 different sizes drawn from the same exponential distribution.

1. Data from an 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 are highly skewed. The number 30 is not a magic number, but a common rule of thumb.



Confidence Intervals

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

```
/*st101d03.sas*/
proc means data=sasuser.testscores maxdec=2
          n mean std stderr clm;
  var SATScore;
  title '95% Confidence Interval for SAT';
run;
```

Selected PROC MEANS statement options:

CLM requests confidence limits for the mean.

STDERR requests the standard error of the mean.

The output is shown below.

Analysis Variable : SATScore					
N	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
80	1190.63	147.06	16.44	1157.90	1223.35

In the test score example, you are 95% confident that the population mean is contained in the interval 1157.90 and 1223.35. 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 precise estimate of the population mean.



How do you increase the precision 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 is more precise.



You can use the **ALPHA=** option in the PROC MEANS statement to construct confidence intervals with a different confidence level. Choose $(1.00 - \text{Confidence}/100)$ as your ALPHA level. By default, **ALPHA=0.05** $(1.00 - 95/100)$.




Exercises

3. Producing Confidence Intervals

Generate the 95% confidence interval for the mean of **BodyTemp** in the **sasuser.NormTemp** data set.

- a. Is the assumption of normality met to produce a confidence interval for this data?
- b. What are the bounds of the confidence interval?



1.03 Multiple Answer Poll

The distribution of sample means is approximately normal if which of the following are true?

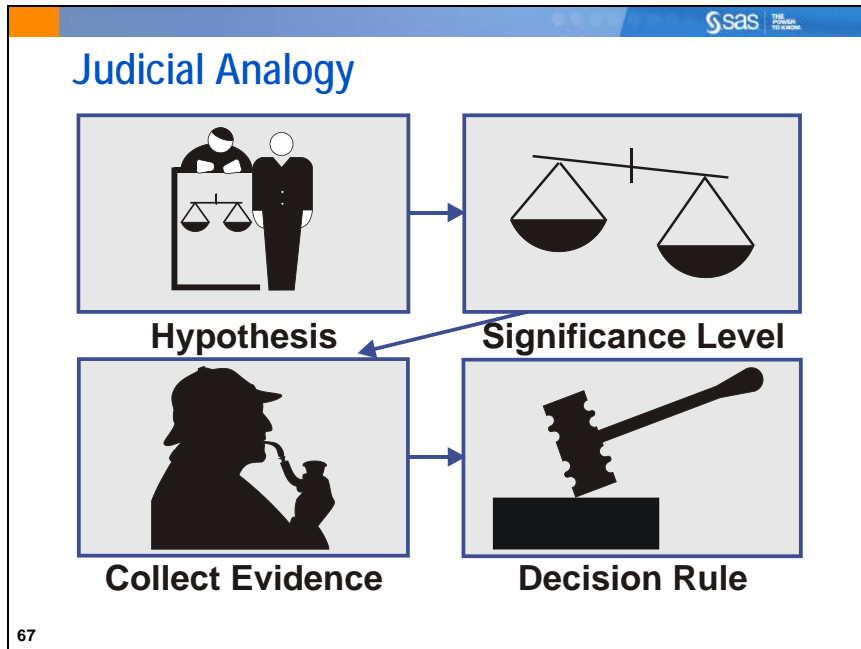
- a. The population is normal.
- b. The sample size is “large enough.”
- c. The sample standard deviation is small.

63

1.4 Hypothesis Testing

Objectives

- Define some common terminology related to hypothesis testing.
- Perform hypothesis testing using the UNIVARIATE and TTEST procedures.



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

First, the two sides need to be determined. In order to relate this to statistical hypothesis testing, name these two sides the null and alternative hypotheses. In criminal court, there exists a presumption of innocence and the defense attorney presents that side. This can be called the *null hypothesis* for a criminal court case. The *alternative hypothesis* is typically your initial research hypothesis (the defendant is guilty). The prosecuting attorney (or the statistical researcher) argues that the presumption of innocence is wrong. The alternative is the logical opposite of the null hypothesis. You generally start with the assumption that the null hypothesis is true, even if your research aims are to disprove the null.

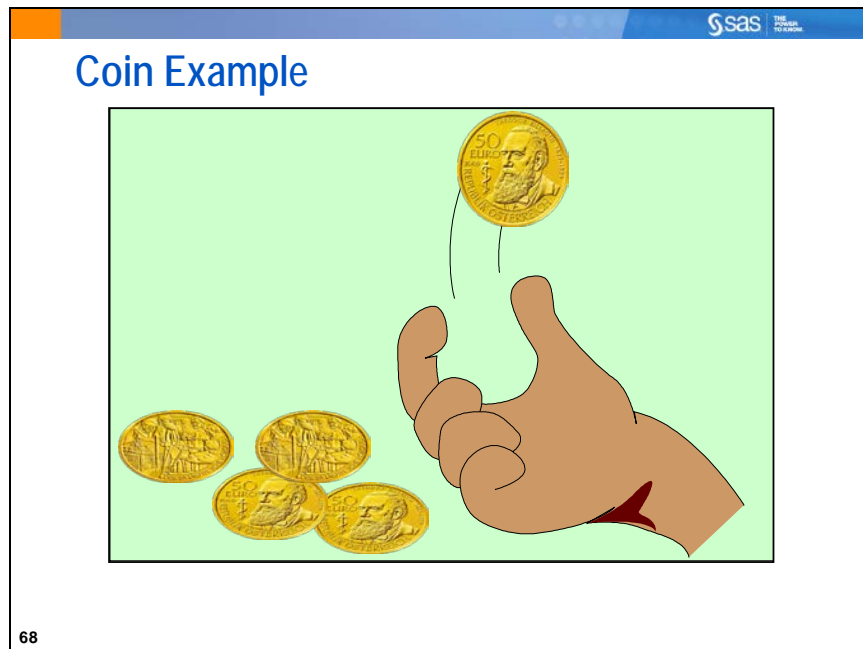
Select a *significance level* as the amount of evidence needed to convict. In a criminal court, the evidence must prove guilt “beyond a reasonable doubt.” In a civil court, the plaintiff must prove his or her case by the “preponderance of the evidence.” In either case, the burden of proof is specified before the trial.

Collect evidence. More accurately, present the collected evidence to the judge and jury.

Use a *decision rule* to make a judgment. If the evidence contradicting the null hypothesis is

- sufficiently strong to meet the burden of proof (significance level), then reject the null hypothesis.
- not strong enough to meet the burden of proof, then fail to reject the null hypothesis. Be aware that failing to prove guilt does not mean that the defendant is **proven** innocent. It could mean that the prosecuting attorney did not build a strong enough case to meet the burden of proof.

Statistical hypothesis testing follows this same basic path.



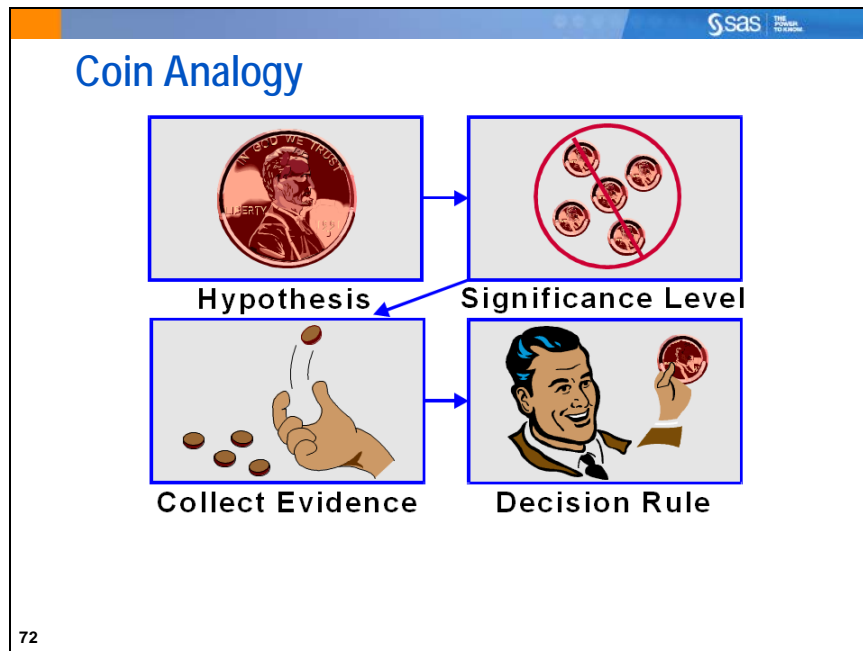
Suppose you want to know whether a coin is fair. You cannot flip it forever, so you decide to take a sample. Flip it five times and count the number of heads and tails.

1.04 Poll

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


- ☐ Yes
- ☐ No

70



Test whether a coin is fair.

1. You suspect that the coin is **not** fair, but recall the legal example and begin by assuming that the coin is fair. In other words, you assume that the null hypothesis is true.
2. You select a significance level. If you observe five heads in a row or five tails in a row, you conclude that the coin is not fair. Otherwise, you decide that there is not enough evidence to show that the coin is not fair.
3. In order to collect evidence, 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 is
 - enough evidence to reject the assumption that the coin is fair
 - 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?

DECISION \ ACTUAL	H ₀ Is True	H ₀ Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

73

Recall that you start by assuming that the coin is fair.

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.

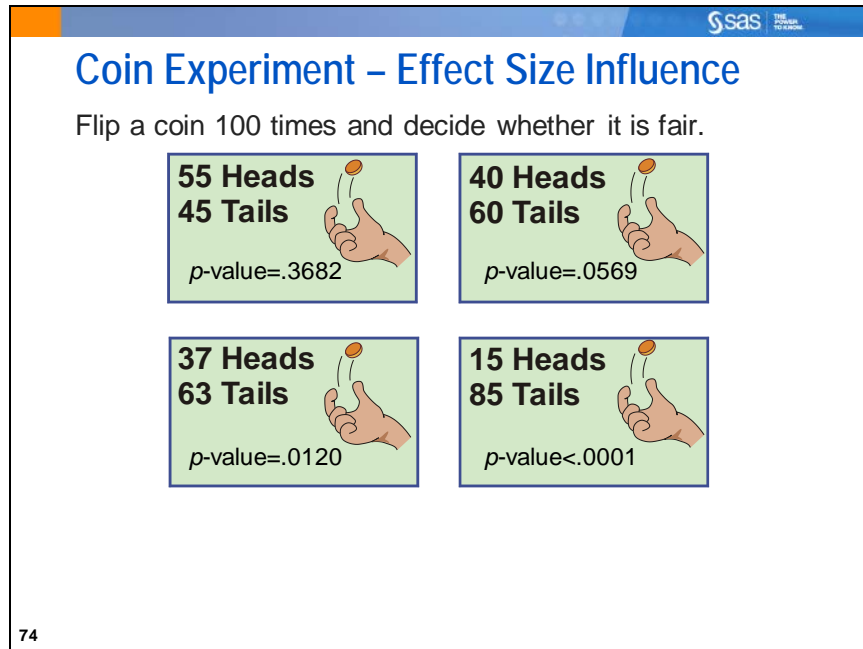
- In the legal example, it is the probability that you conclude that the person is guilty when he or she is innocent.
- In the coin example, it is the probability that you conclude that the coin is not fair when it is fair.

The probability of a Type II error, often denoted β , is the probability that 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 that the coin is not fair when it is not fair.



The *power* of a statistical test is equal to $1-\beta$, where β is the Type II error rate. This is the probability that you correctly reject the null hypothesis, given some assumed values of the true population mean and standard deviation in the population and the sample size.



The *effect size* refers to the magnitude of the difference in sampled population from the null hypothesis. In this example, the null hypothesis of a fair coin suggests 50% heads and 50% tails. If the true coin flipped were actually weighted to give 55% heads, the effect size would be 5%.


If you flip a coin 100 times and count the number of heads, you do not doubt that 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 situation, as the difference between the number of heads and tails increases, you have more evidence 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, simply by chance, given that 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.

A large *p-value* means that you would often see a test statistic value this large in experiments with a fair coin. A small *p-value* means that you would rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair, because if the null hypothesis were true, a random sample selected from it would not likely have the observed statistic values.



Coin Experiment – Sample Size Influence

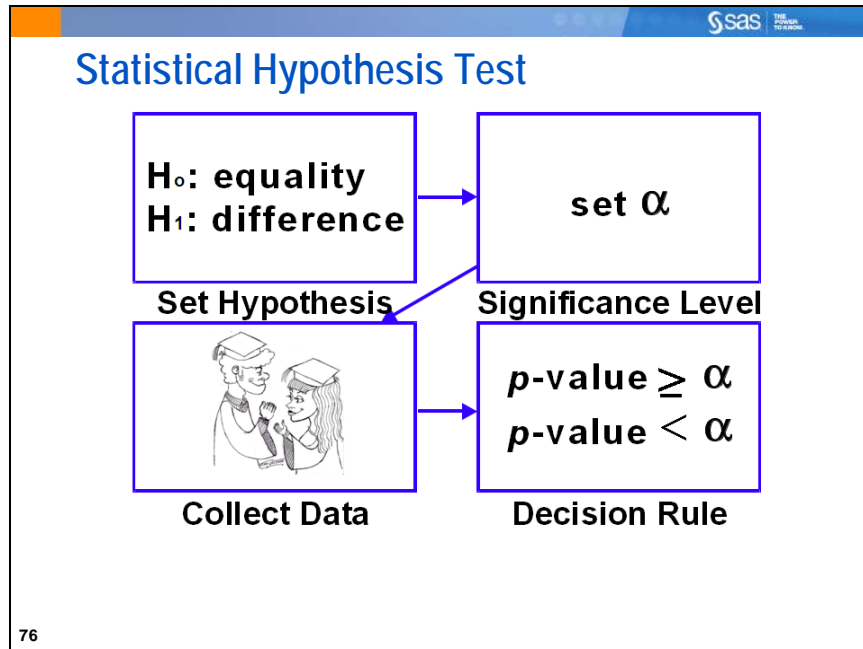
Flip a coin and get 40% heads and decide whether it is fair.

4 Heads 6 Tails $p\text{-value}=.7539$ 	16 Heads 24 Tails $p\text{-value}=.2682$ 
40 Heads 60 Tails $p\text{-value}=.0569$ 	160 Heads 240 Tails $p\text{-value}<.0001$ 

75

A p -value is not only affected by the effect size. It is also affected by the sample size (number of coin flips, k).

For a fair coin, you would expect 50% of k flips to be heads. In this example, in each case, the observed proportion of heads from k flips was 0.4. This value is different from the 0.5 you would expect under H_0 . The evidence is stronger, when the number of trials (k) on which the proportion is based increases. As you saw in the section about confidence intervals, the variability around a mean estimate is smaller, when the sample size is larger. For larger sample sizes, you can measure means more precisely. Therefore, 40% of the heads out of 400 flips would make you more certain that this was not a chance difference from 50% than would 40% out of 10 flips. The smaller p -values reflect this confidence. The p -value here assesses the probability that this difference from 50% occurred purely by chance.




In statistics, the following rules apply:

1. The null hypothesis, denoted H_0 , is your initial assumption and is usually one of equality or no relationship. For the test score example, H_0 is that the mean combined Math and Verbal SAT score is 1200. The alternative hypothesis, H_1 , is the logical opposite of the null, namely that the combined Math and Verbal SAT score is *not* 1200.
2. The significance level is usually denoted by α , the Type I error rate.
3. The strength of the evidence is measured by a p -value.
4. The decision rule is
 - fail to reject the null hypothesis if the p -value is greater than or equal to α
 - reject the null hypothesis if the p -value is less than α .



From a single hypothesis test, you would not conclude that two things are the same or have no relationship; you can only fail to show a difference or a relationship.



Comparing α and the p -Value

In general, you do one of the following:

- reject the null hypothesis if $p\text{-value} < \alpha$
- fail to reject the null hypothesis if $p\text{-value} \geq \alpha$.

77

It is important to clarify the following:

- The value of 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 that α is set before data collection based on the circumstances of the experiment. The level of α is chosen based on the cost of making a Type I error. It is also a function of your knowledge of the data and theoretical considerations.

For the test score example, α was set to 0.05, based on the consequences of making a Type I error (the error of concluding that the mean SAT combined score is not 1200 when it really is 1200). If making a Type I error is especially egregious, you might consider choosing a lower significance level when planning your analysis.

1.05 Multiple Choice Poll

Which of the following affects alpha?

- a. The p -value of the test
- b. The sample size
- c. The number of Type I errors
- d. All of the above
- e. Answers a and b only
- f. None of the above

79

Performing a Hypothesis Test

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

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

For the test score example:

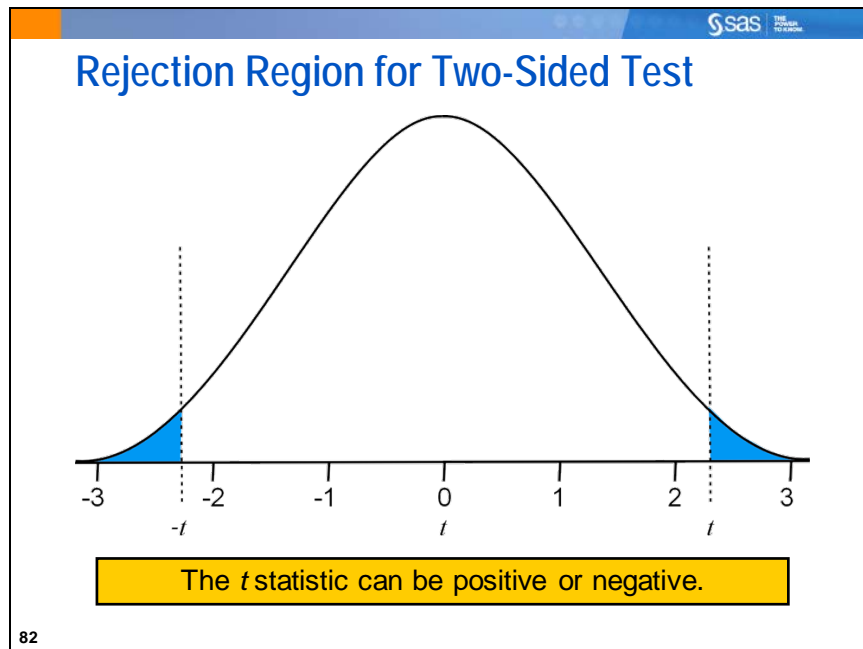
$$t = \frac{(1190.625 - 1200)}{16.4416} = -0.5702$$

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

81

For the test score example, μ_0 is the hypothesized value of 1200, \bar{x} is the sample mean SAT score of students selected from the school district, 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 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.

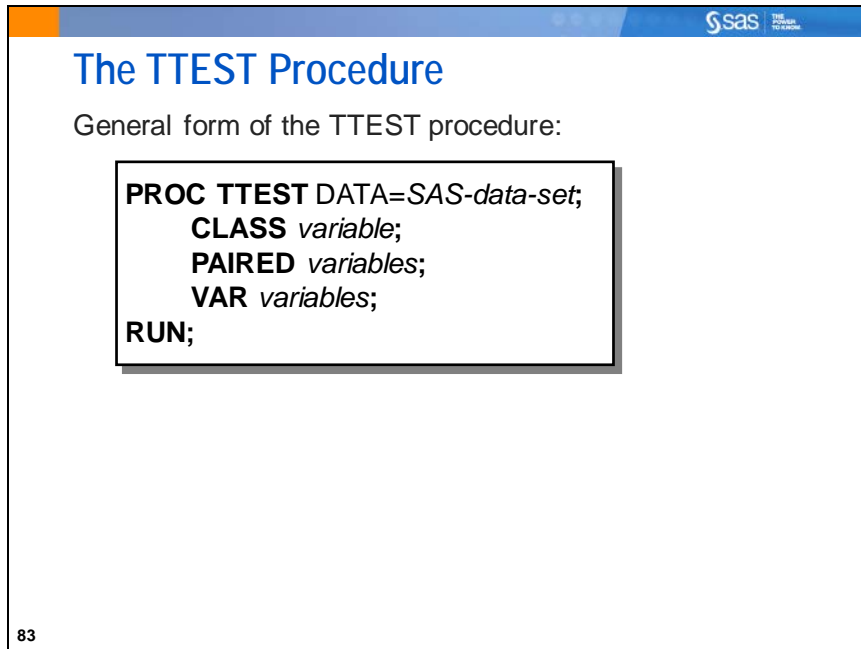


For a two-sided test of a hypothesis, the rejection region is contained in both tails of the t distribution. If the t statistic falls in the rejection region (in the shaded region in the graph above), then you reject the null hypothesis. Otherwise, you fail to reject the null hypothesis.

The area in each of the tails corresponds to $\alpha/2$ or 2.5%. The sum of the areas under the tails is 5%, which is alpha.



The alpha and t -distribution mentioned here are the same as those in the section about confidence intervals. In fact, there is a direct relationship. The rejection region based on α begins at the point where the $(1.00-\alpha)\%$ confidence interval no longer includes the true value of μ_0 .

The image is a screenshot of a presentation slide from SAS. The slide has a blue header bar with the SAS logo and the text 'THE POWER OF DATA'. The main title of the slide is 'The TTEST Procedure' in a large blue font. Below the title, it says 'General form of the TTEST procedure:'. In the center, there is a white box with a black border containing the SAS code for the TTEST procedure:

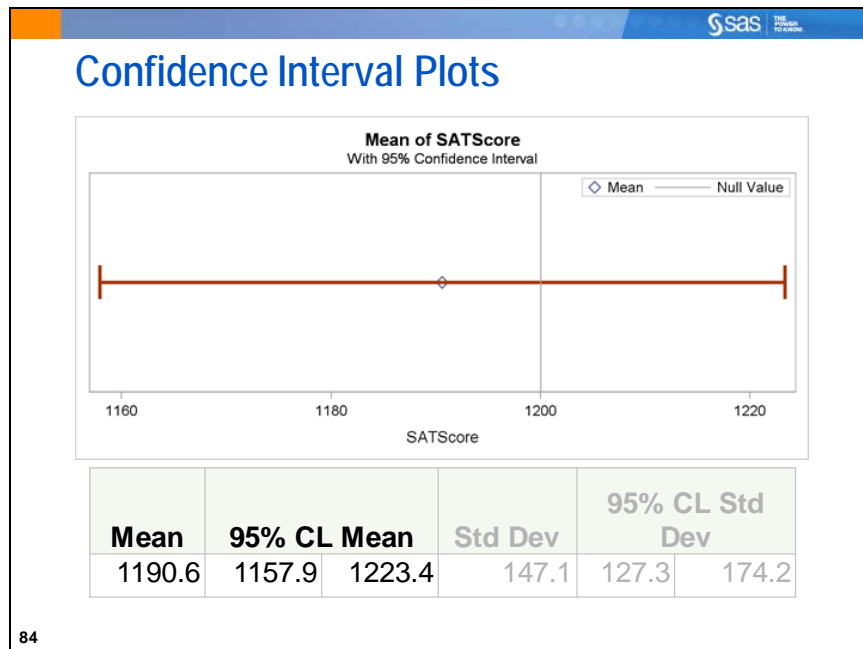
```
PROC TTEST DATA=SAS-data-set;  
  CLASS variable;  
  PAIRED variables;  
  VAR variables;  
RUN;
```

 The slide number '83' is in the bottom left corner.

The TTEST procedure performs t tests and computes confidence limits for one sample, paired observations, two independent samples, and the AB/BA crossover design. With ODS Statistical Graphics, PROC TTEST can also be used to produce histograms, Quantile-Quantile plots, box plots, and confidence limit plots.

Selected TTEST procedure statements:

CLASS	specifies the two-level variable for the analysis. Only one variable is allowed in the CLASS statement. If no CLASS statement is included, a one-sample t test is performed.
PAIRED <i>PairLists</i> ;	specifies the <i>PairLists</i> to identify the variables to be compared in paired comparisons. You can use one or more PairLists.
VAR	specifies <i>numeric</i> 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.



A *confidence interval plot* is a visual display of the sample statistic value (of the mean, in this case) and the confidence interval calculated from the data. If there is a null hypothesized value for the parameter, it can be drawn on the plot as a reference line. In this way, the statistical significance of a test can be visually assessed. If the $(1.00-\alpha)\%$ confidence interval does not include the null hypothesis value, then that implies that the null hypothesis can be rejected at the α significance level. If the confidence interval includes the null hypothesis value, then that implies that the null hypothesis cannot be rejected at that significance level.



Hypothesis Testing

Example: Use the MU0= option in the UNIVARIATE procedure to test the hypothesis that the mean of SAT Math+Verbal score is equal to 1200.

```
/*st101d04.sas*/ /*Part A*/
ods graphics off;
proc univariate data=sasuser.testscores mu0=1200;
    var SATScore;
    title 'Testing Whether the Mean of SAT Scores = 1200';
run;
ods graphics on;
```

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=1200				
Test	Statistic	p Value		
Student's t	t	-0.5702	Pr > t	0.5702
Sign	M	-5	Pr >= M	0.3019
Signed Rank	S	-207	Pr >= S	0.2866

The t statistic and p -value are labeled Student's t and $\text{Pr} > |t|$, respectively.

- The t statistic value is -0.5702 and the p -value is .5702.
- Therefore, you cannot reject the null hypothesis at the 0.05 level. Thus, even though the mean of the student scores in this sample (1190.625) is slightly lower than the magnet school goal of 1200, there is not enough evidence to reject the hypothesis that the population mean of all magnet school students in the district is equal to 1200.



The Sign and Signed Rank tests are known as *nonparametric tests for location*. If the normality assumption cannot be met, then these tests can be used to test slightly modified tests about the central location of the population.

Use the H0= option in the TTEST procedure to test the hypothesis that the mean of the SAT Math+Verbal score is equal to 1200.

```
/*st101d04.sas*/ /*Part B*/
proc ttest data=sasuser.testscores h0=1200
    plots(shownull)=interval;
    var SATScore;
    title 'Testing Whether the Mean of SAT Scores = 1200 '
        'Using PROC TTEST';
run;
```

Selected PROC TTEST statement options:

H0= specifies the value of the mean or location parameter in the null hypothesis for tests of location (H0=0 by default).

PLOTS(SHOWNULL)=INTERVAL includes a plot of confidence intervals of the mean. **SHOWNULL** places a vertical reference line at the mean value of the null hypothesis.

PROC TTEST Output

N	Mean	Std Dev	Std Err	Minimum	Maximum
80	1190.6	147.1	16.4416	890.0	1600.0

Summary statistics are reported. These are the same values that were obtained using both PROC MEANS and PROC UNIVARIATE.

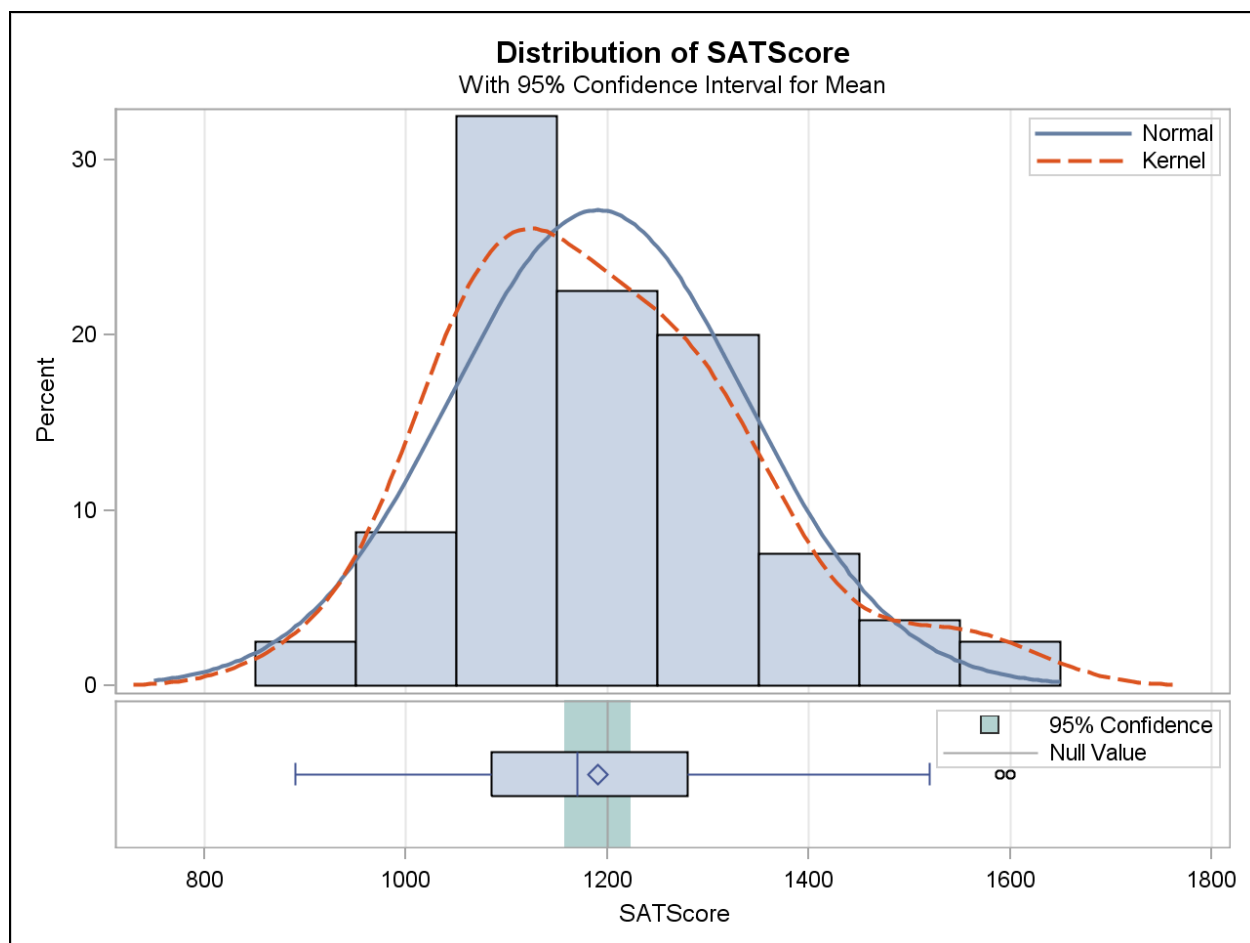
Mean	95% CL Mean	Std Dev	95% CL Std Dev
1190.6	1157.9 1223.4	147.1	127.3 174.2

The confidence interval around the sample mean and sample standard deviation are reported. Notice that the 95% confidence interval around the mean includes the null hypothesis value of 1200. This implies a lack of statistical significance at the $\alpha=0.05$ significance level.

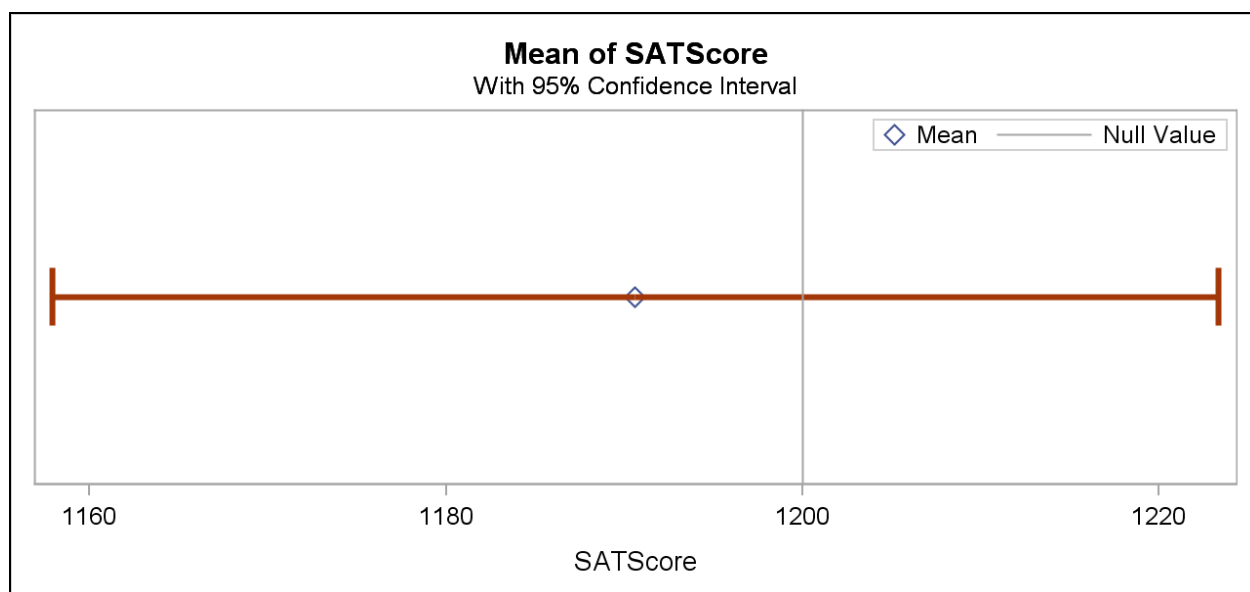
DF	t Value	Pr > t
79	-0.57	0.5702

The p -value is 0.5702, which is the same as was calculated using PROC UNIVARIATE. As the confidence interval information suggested, this value is not statistically significant at the $\alpha=0.05$ significance level.

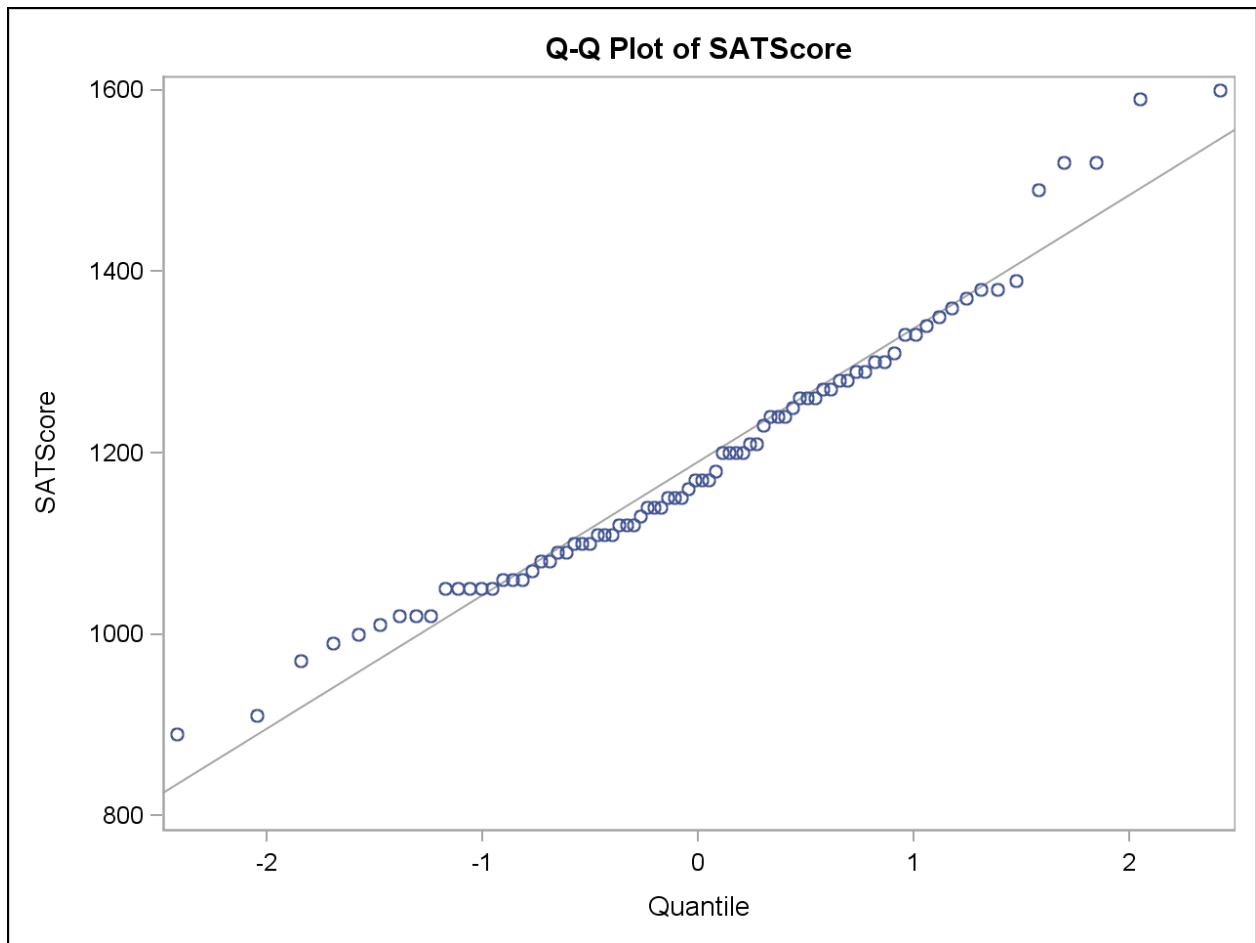
Two default plots are produced along with the confidence interval plot that was requested in the code.



The histogram along with the normal and kernel density curves are produced on one plot, along with a horizontal box plot with a shaded confidence interval for the mean included.



The confidence interval plot gives a visual display of the confidence interval table from above.



The normal quantile-quantile plot shows that the distribution of **SATScore** is approximately normal in this sample.



Exercises

4. Performing a One-Sample t Test

Perform a one-sample t test to determine whether the mean of body temperatures (the variable **BodyTemp** in **sasuser.NormTemp**) is truly 98.6.

- What is the value of the t statistic and the corresponding p -value?
- Produce a confidence interval plot of **BodyTemp** with the value 98.6 used as a reference.
- Do you reject or fail to reject the null hypothesis at the 0.05 level that the average temperature is 98.6 degrees?

1.06 Multiple Choice Poll

A 95% confidence interval for SAT scores is (1157.90, 1223.35). From this, what can you conclude, at $\alpha=0.05$?

- The true average SAT score is significantly different from 1200.
- The true average SAT score is not significantly different from 1200.
- The true average SAT score is less than 1200.
- None of the above – You cannot determine statistical significance from confidence intervals.

1.5 Solutions

Solutions to Exercises

1. Calculating Basic Statistics in PROC MEANS

```
/*st101s01.sas*/ /*Parts a and b*/
proc means data=sasuser.NormTemp
    maxdec=2
    n mean std q1 q3 qrange;
var BodyTemp;
title 'Selected Descriptive Statistics for Body Temp';
run;
```

PROC MEANS Output

Analysis Variable : BodyTemp						
	N	Mean	Std Dev	Lower Quartile	Upper Quartile	Quartile Range
	130	98.25	0.73	97.80	98.70	0.90

- a. What is the overall mean and standard deviation of body temperature in the sample?

The overall mean is 98.25.

- b. What is the interquartile range of body temperature?

The interquartile range is 0.90 (98.70 – 97.80).

- c. Do the mean values seem to differ between men and women?

```
/*st101s01.sas*/ /*Part c*/
proc means data=sasuser.NormTemp
    maxdec=2
    n mean std q1 q3 qrange;
var BodyTemp;
class Gender;
title 'Selected Descriptive Statistics for Body Temp';
run;
```

Analysis Variable : BodyTemp							
	N				Lower	Upper	Quartile
Gender	Obs	N	Mean	Std Dev	Quartile	Quartile	Range
Female	65	65	98.39	0.74	98.00	98.80	0.80
Male	65	65	98.10	0.70	97.60	98.60	1.00

The values differ somewhat.

2. Producing Descriptive Statistics

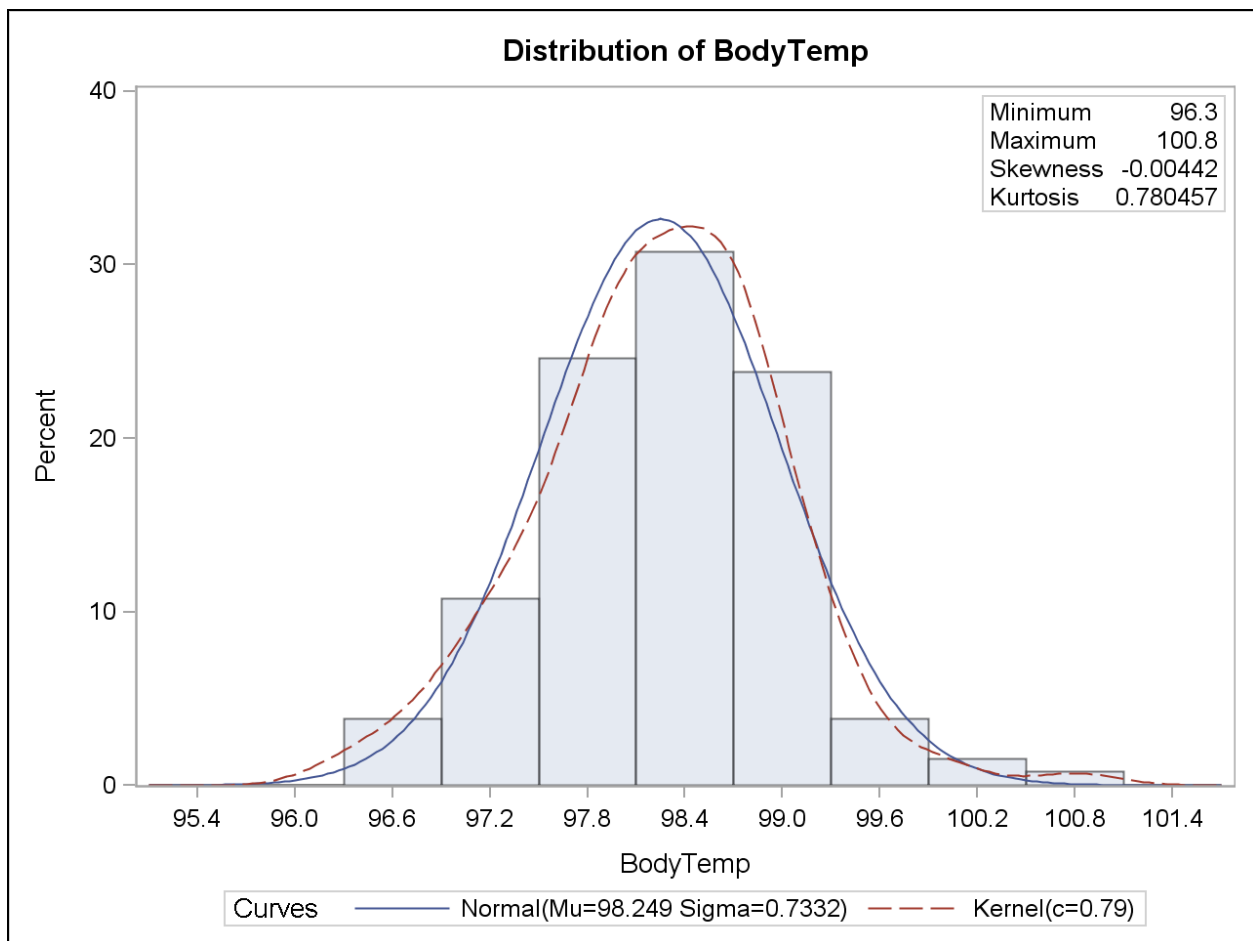
Use the **sasuser.NormTemp** data set to answer the following:

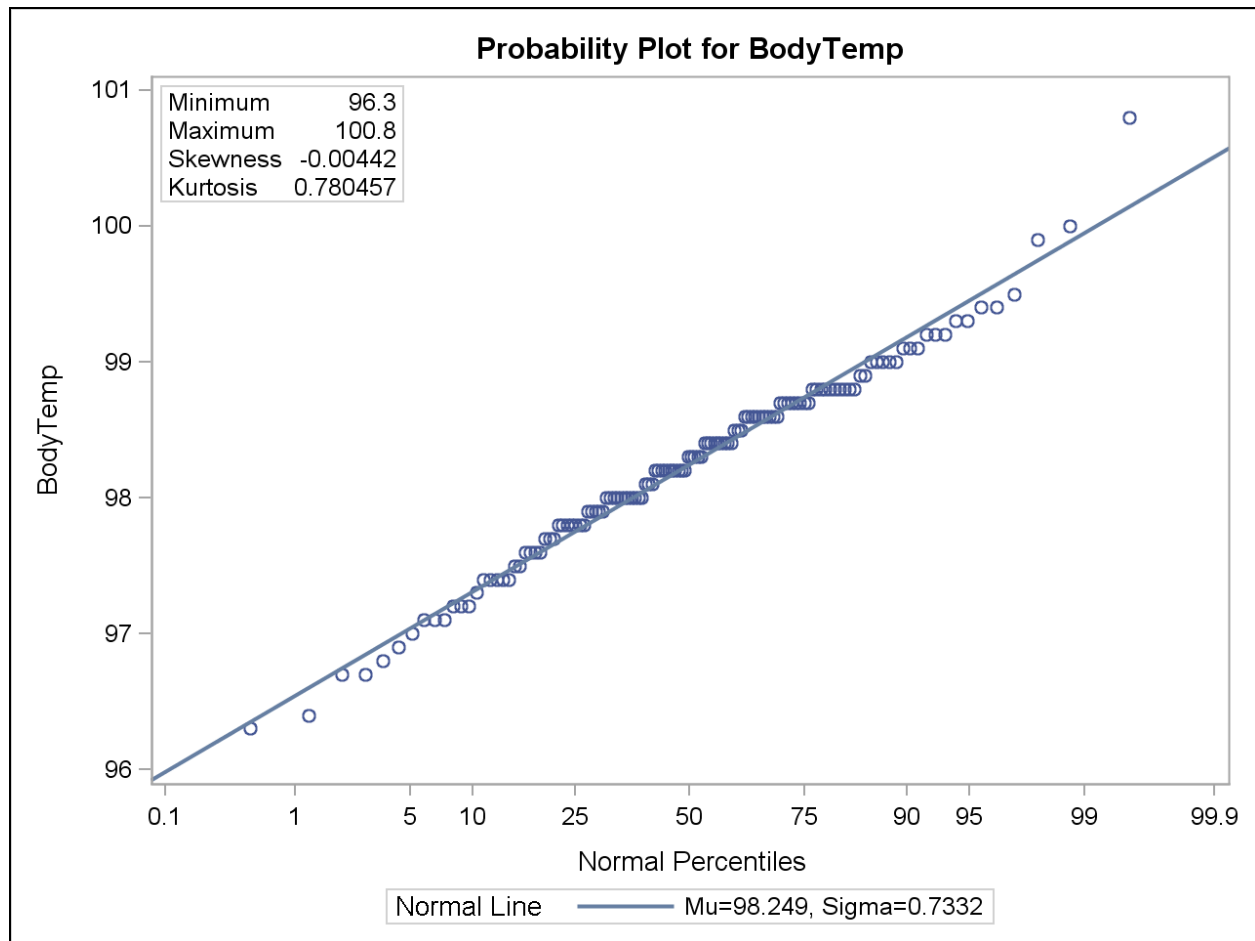
- a. What are the minimum, the maximum, the mean, and the standard deviation for **BodyTemp**? Does the variable appear to be normally distributed?

```
/*st101s02.sas*/ /*Part a*/
proc univariate data=sasuser.NormTemp noprint;
  var BodyTemp;
  histogram BodyTemp / normal(mu=est sigma=est noprint) kernel;
  inset min max skewness kurtosis / position=ne;
  probplot BodyTemp / normal(mu=est sigma=est);
  inset min max skewness kurtosis;
  title 'Descriptive Statistics Using PROC UNIVARIATE';
run;
```



The NOPRINT option in both the PROC UNIVARIATE and HISTOGRAM statements suppresses the printing of the tabular output. Because the statistics are being reported in the insets of the plots, they are not needed in the output tables.

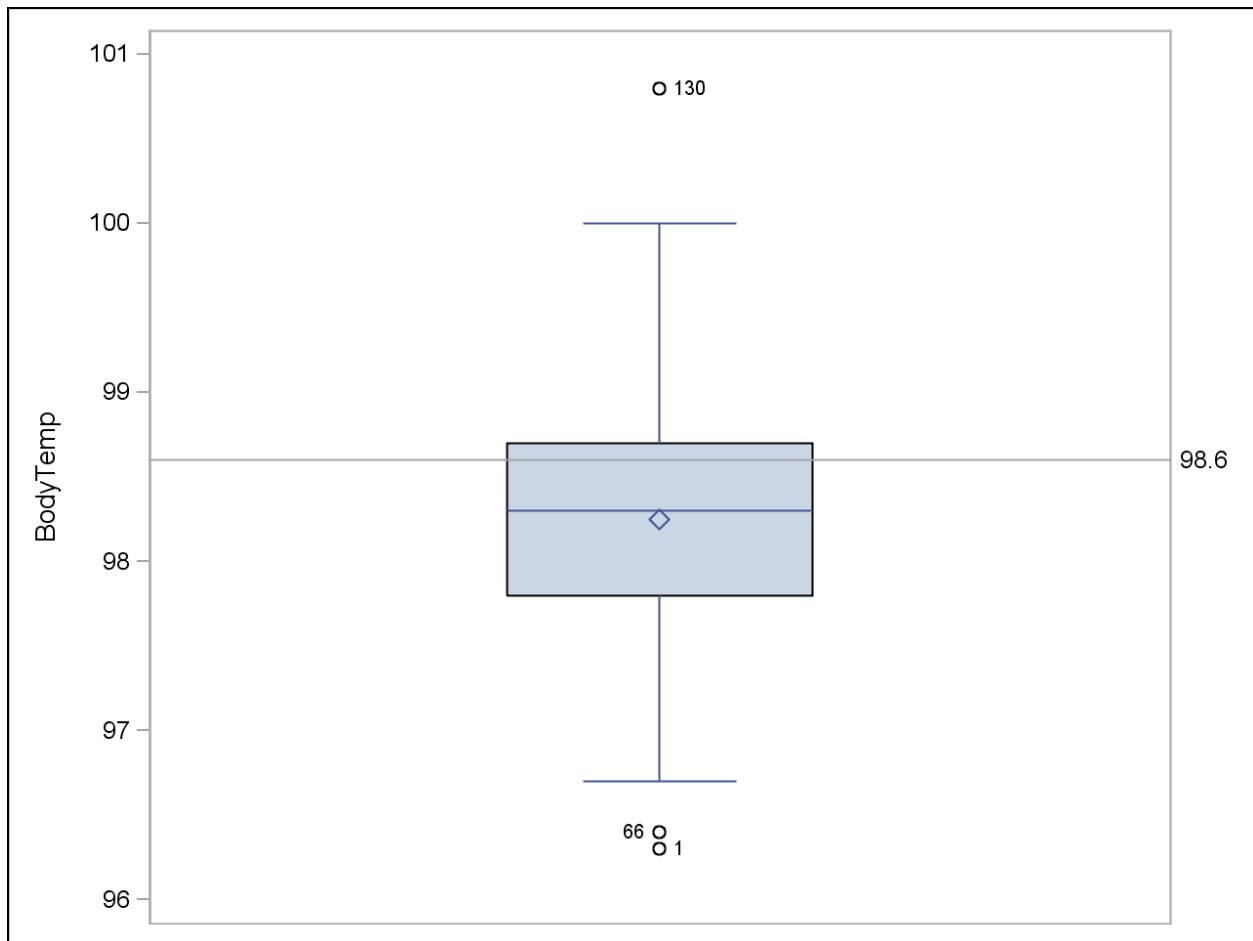




The distribution appears approximately normal.

- b. Create box plots for **BodyTemp**. Use **ID** to identify outliers. Display a reference line at 98.6 degrees. Does the average body temperature seem to be 98.6 degrees?

```
/*st101s02.sas*/ /*Part b*/
proc sgplot data=sasuser.NormTemp;
  vbox BodyTemp / datalabel=ID;
  format ID 3.;
  refline 98.6 / axis=y label;
  title "Box-and-whisker Plots of Body Temp";
run;
```



The average body temperature seems to be somewhat less than 98.6 degrees, as was seen in the tabular output.

3. Producing Confidence Intervals

Generate the 95% confidence interval for the mean of **BodyTemp** in the **sasuser.NormTemp** data set.

```
/*st101s03.sas*/
proc means data=sasuser.NormTemp maxdec=2
      n mean std stderr clm;
  var BodyTemp;
  title '95% Confidence Interval for Body Temp';
run;
```

Analysis Variable : BodyTemp					
N	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
130	98.25	0.73	0.06	98.12	98.38

- a. Is the assumption of normality met to produce a confidence interval for this data?

Yes. Because the sample size is large enough and because the data values seemed to be normally distributed, the normality assumption seems to hold.

- b. What are the bounds of the confidence interval?

The 95% confidence interval is 98.12 to 98.38 degrees Fahrenheit.

4. Performing a One-Sample t Test

Perform a one-sample t test to determine whether the mean of body temperatures (the variable **BodyTemp** in **sasuser.NormTemp**) is truly 98.6.

```
/*st101s04.sas*/  /*PROC UNIVARIATE*/
proc univariate data=sasuser.NormTemp mu0=98.6;
  var BodyTemp;
  title 'Testing Whether the Mean Body Temperature = 98.6 ';
    'Using PROC UNIVARIATE';
run;
```

Partial Output

Tests for Location: Mu0=98.6				
Test	Statistic		p Value	
Student's t	t	-5.45482	Pr > t	<.0001
Sign	M	-21	Pr >= M	0.0002
Signed Rank	S	-1963	Pr >= S	<.0001

```
/*st101s04.sas*/  /*PROC TTEST*/
proc ttest data=sasuser.NormTemp h0=98.6
  plots(shownull)=interval;
  var BodyTemp;
  title 'Testing Whether the Mean Body Temperature = 98.6 '
    'Using PROC TTEST';
run;
```

Partial Output

DF	t Value	Pr > t
129	-5.45	<.0001

- a. What is the value of the t statistic and the corresponding p -value?

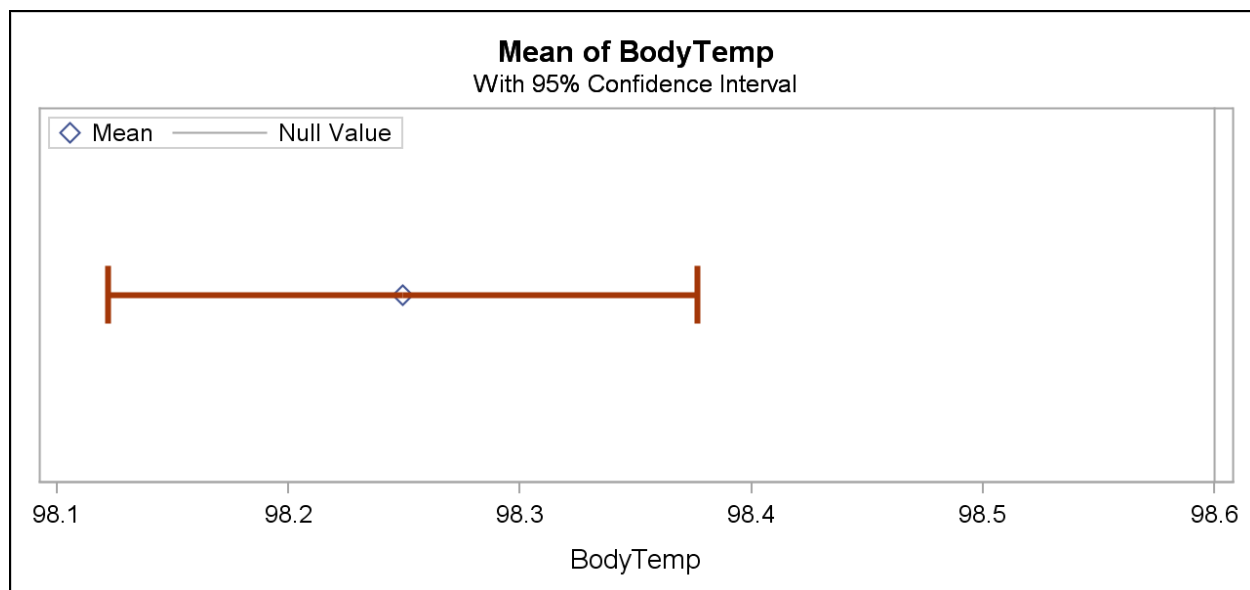
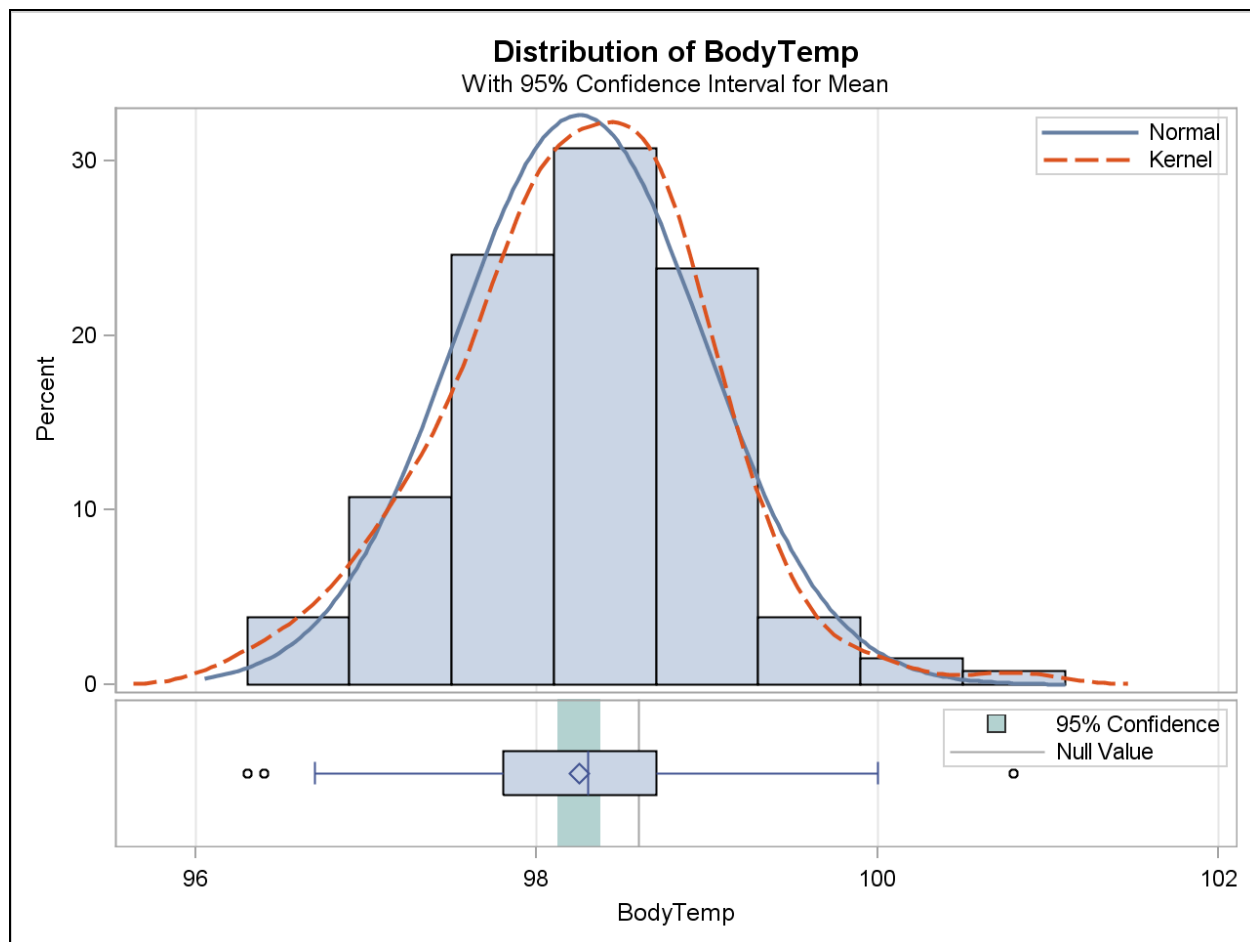
They are -5.45 and <.0001, respectively.

- b. Produce a confidence interval plot of **BodyTemp** with the value 98.6 used as a reference.

N	Mean	Std Dev	Std Err	Minimum	Maximum
130	98.2492	0.7332	0.0643	96.3000	100.8

Mean	95% CL Mean	Std Dev	95% CL Std Dev
98.2492	98.1220 98.3765	0.7332	0.6536 0.8350


DF	t Value	Pr > t
129	-5.45	<.0001



- c. Do you reject or fail to reject the null hypothesis at the 0.05 level that the average temperature is 98.6 degrees?

Because the p -value is less than the stated alpha level of 0.05, you do reject the null hypothesis. The confidence limit plot can be used to reach the same conclusion. The 95% confidence interval does not contain the value 98.6. Therefore, you can reject the null hypothesis that the true population mean body temperature is 98.6 degrees Fahrenheit.

Solutions to Student Activities (Polls/Quizzes)




1.01 Multiple Answer Poll – Correct Answers

A sample from a population should be which of the following?

- a. Random
- ☒ b. Representative
- c. Normal

15



1.02 Multiple Choice Poll – Correct Answers

In the **NormTemp** data set, the distribution of **BodyTemp** seemed to be which of the following?

- ☒ a. Close to normal
- b. Left skewed
- c. Right skewed
- d. Having high positive kurtosis
- e. Having high negative kurtosis

47

1.03 Multiple Answer Poll – Correct Answers

The distribution of sample means is approximately normal if which of the following are true?

- ☒ a. The population is normal.
- ☒ b. The sample size is “large enough.”
- c. The sample standard deviation is small.

64

1.04 Poll – Correct Answer

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

- ☒ Yes
- ☐ No

71

1.05 Multiple Choice Poll – Correct Answer

Which of the following affects alpha?

- a. The p -value of the test
- b. The sample size
- c. The number of Type I errors
- d. All of the above
- e. Answers a and b only
- ☒ f. None of the above

80

1.06 Multiple Choice Poll – Correct Answer

A 95% confidence interval for SAT scores is (1157.90, 1223.35). From this, what can you conclude, at $\alpha=0.05$?

- a. The true average SAT score is significantly different from 1200.
- ☒ b. The true average SAT score is not significantly different from 1200.
- c. The true average SAT score is less than 1200.
- d. None of the above – You cannot determine statistical significance from confidence intervals.

90