Sources of information:

1) SAS Introduction to Statistics

2) Teach yourself statistics, https://stattrek.com/estimation/estimation-in-statistics.aspx

# SAS - **INTRODUCTION TO STATISTICS**

LESSON OVERVIEW

1 BASIC STATISTICAL CONCEPTS

1.1 Introduction

1.2 Descriptive and Inferential Statistics

1.3 Populations and Samples

1.4 Parameters and Statistics

1.5 Uses of Variables: Independent and Dependent

1.6 Types of Variables: Quantitative and Categorical

1.7 Scales of Measurement (Nominal, Ordinal, Interval and Ratio)

1.8 Statistical Methods

2 DESCRIPTIVE STATISTICS

2.1 Introduction

2.2 Describing your data

2.3 Measures of Location/Central Tendency

2.4 Percentiles

2.5 Measures of Variability/Spread/Dispersion

2.6 The MEANS Procedure in SAS

2.7 Using the PROC MEANS Procedure to Generate Descriptive Statistics

3 PICTURING YOUR DATA (Descriptive Statistics)

3.1 Introduction

3.2 Picturing your Data: Histogram

3.3 Normal Distribution

3.4 Assessing Normality (with skewness and kurtosis)

3.5 Measures of Shape: Skewness

3.6 Measures of Shape: Kurtosis

3.7 Normal Probability Plots

3.8 Box Plots

3.9 Assessing whether our data follow the normal distribution

3.10 The UNIVARIATE Procedure in SAS

3.11 Statistical Graphics Procedures in SAS

3.12 The SGPLOT Procedure in SAS

3.13 ODS Graphics Output

3.14 Using SAS to Picture your Data

4 CONFIDENCE INTERVALS FOR THE MEAN (Inferential Statistics)

4.1 Introduction

4.2 Point Estimators, Variability, and Standard Error

4.3 Distribution of Sample Means

4.4 Interval Estimators

4.5 Confidence Intervals

4.6 Normality and the Central Limit Theorem

4.7 Using PROC MEANS to Generate a Confidence Interval for the Mean

4.8 Calculating a 95% Confidence Interval

5 5 HYPOTHESIS TESTING (Inferential Statistics)

5.1 Introduction

5.2 Decision-Making Process

5.3 Steps in Hypothesis Testing

5.4 Types of Errors, Power after we completed the Hypothesis Testing

5.5 The p-Value, Effect Size and Sample Size

5.6 Summary of Hypothesis Testing

5.7 The t statistic, t Distribution and two-Sided t-test

5.8 Using the PROC UNIVARIATE Procedure in SAS to Generate a t statistic

5.9 Using the PROC UNIVARIATE Procedure in SAS to Perform a Hypothesis Test

# LESSON OVERVIEW

Statistics is the science of collecting, analyzing, and interpreting data.

Statistical methods are widely used in many areas, such as business and finance, medicine, science, and agriculture. In fact, we also encounter statistical information in many aspects of our personal lives; for example, you've probably read or heard many reports in the media that summarize economic data or the results of a health study by providing statistics.  
  
This course introduces some fundamental ideas in statistics. First, you'll learn some basic statistics terminology, as well as how samples are chosen for study and analysis. This course also shows you how variables are classified based on the type of information they contain and how they are used in an analysis.  
  
Next, you'll learn some of the most common ways to summarize different kinds of data using both numbers and pictures. You probably encounter these types of statistical summaries on a regular basis.

**Numerical summary values** that you might be familiar with include

* measures of location, such as the mean and percentiles, and
* measures of variability, such as standard deviation.

Finally, you'll learn **two of the most common tools of inferential statistics**

* confidence intervals and
* hypothesis testing.

These tools combine statistics with mathematical probability. The results enable analysts to draw formal conclusions about what the data means and to make better decisions. Don't worry if these terms are unfamiliar. You'll know what they mean by the time you finish the course!

You might be wondering what value this course has for your work. That's a reasonable question. Let's consider a couple of situations.

Perhaps you took a college-level statistics course in the past and now plan to use SAS to perform data analysis. You've found that you need a refresher on the basics before taking advanced analytics classes.

This course will provide that refresher and prepare you for the next step in your training, which might be either Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression or SAS Enterprise Guide: ANOVA, Regression, and Logistic Regression.

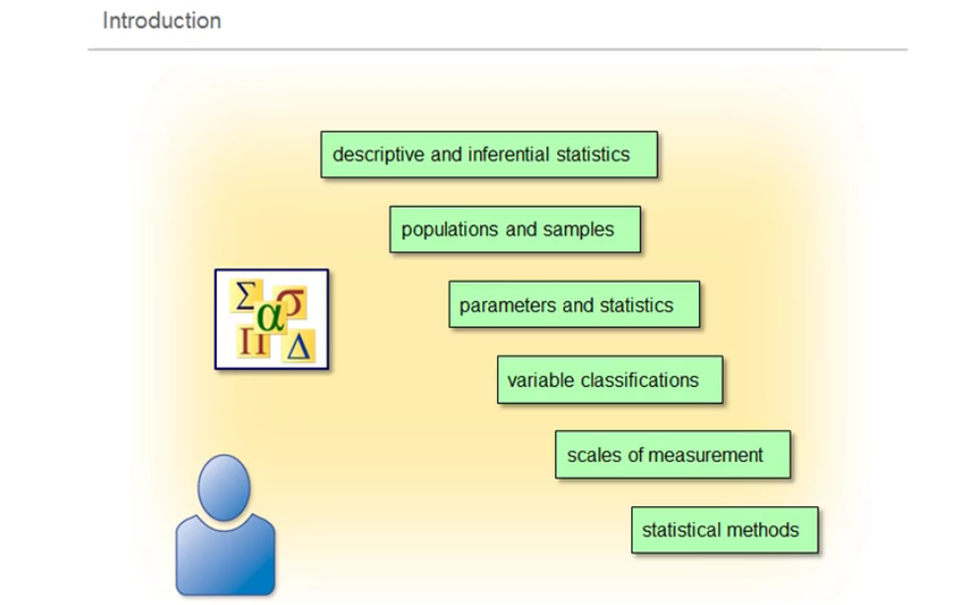
If you're a SAS programmer who works in a field that makes use of statistical methods, the concepts and terminology you learn in this course will enable you to communicate more effectively with statisticians and analysts. These concepts will also provide some insight into how data sets should be constructed in SAS for statistical analysis, and may enable you to perform simple data quality checks on your own. All of these skills will make you a better and more efficient SAS programmer.

Whether this course serves as the first step of your SAS analytics training or as a stand-alone introduction to statistics, we hope that you enjoy the course and find it useful for your work!

# 1 BASIC STATISTICAL CONCEPTS

## 1.1 Introduction

Let's start by reviewing some statistical concepts that you need in order to perform basic statistical analyses. Here's what you learn in this topic.



## 1.2 Descriptive and Inferential Statistics

Descriptive statistics organize, describe, and summarize data using numbers and graphical techniques. This branch of statistics uses a set of standard measures such as percent, averages, and variability, as well as simple graphs, charts, and tables.

Descriptive statistics help you to better understand your data by describing and summarizing its basic features.

DESCRIPTIVE STATISTICS, NUMERICAL SUMMARIES

You learn how to generate and understand numerical summaries. These include

* frequency
* measures of location, including minimum, maximum
* percentiles
* quartiles, and
* central tendency (mean, median, and mode); and
* measures of dispersion or variability, including range, interquartile range, variance, and standard deviation.

DESCRIPTIVE STATISTICS, GRAPHICAL SUMMARIES

You learn how to generate and understand graphical summaries. These include:

* histogram,
* normal probability plot, and
* box plot.

The goals when you're describing data are to:

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

Is the data as error free as possible? What unique features can you identify? Are there data values that cluster or show some unusual shape? Does your data include any possible outliers?

When you have a basic understanding of your data, then you can use inferential statistics.

INFERENTIAL STATISTICS is the branch of statistics concerned with drawing conclusions about a population from analysis of a random sample drawn from that population.

It is also concerned with the precision and reliability of those inferences.

Inferential statistics generalize from data you observe in the sample, to the population that you have not observed.

Descriptive statistics describe your sample data, but inferential statistics help you draw conclusions about the entire population of data.

## 1.3 Populations and Samples

A population is the complete set of observations or the entire group of objects that you are researching.

A sample is a subset of the population.

You gather a sample so that you don't have to obtain data for the entire population.

The sample should be representative of the population, meaning that the sample's characteristics are similar to the population's characteristics. One way to obtain a representative sample is to collect a simple random sample. With this sampling method, every possible sample of a given size in the population has an equal chance of being selected. Random sampling can help to ensure that the sample is representative of the population. You should avoid collecting your sample from a section of the population that is easily available to you. This is called convenience sampling, and it can lead to a biased sample that is not representative of the population from which it is drawn. A sample that's not representative can cause you to draw incorrect conclusions.

Let's look at an example.

Suppose a university wants to estimate the percent of its freshmen who plan to return for their sophomore year.

The population for this study is the entire set of 2,500 freshmen in attendance.

Researchers gathered a representative sample of 100 freshmen by selecting 100 student ID numbers at random from the entire set of 2,500 freshmen.

If the researchers had simply selected the first 100 freshmen who responded to an e-mail questionnaire, this would have resulted in a biased sample. This could lead to an incorrect estimate of the number who plan to return for their sophomore year

If you have a representative sample, you can make correct inferences to the entire population. In this course, we always assume that the sample is representative. Click the Information button for information on how to generate random samples.

## How to generate random samples

**Random Samples**

**Selecting Random Samples**

The SURVEYSELECT procedure 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 online 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 SURVEYSELECT procedure statement:

|  |  |
| --- | --- |
| STRATA | enables the user to specify one or more stratification variables. If no STRATA statement is specified, no stratification takes place. |

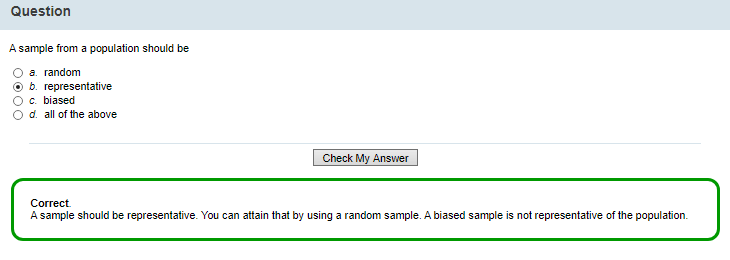
Other statements and options for the SURVEYSELECT procedure can be found in the SAS online documentation.  
  
Program A shows how to select a certain sample size using the SAMPSIZE= option.

|  |
| --- |
| /\*Program A\*/  proc surveyselect data=Statdata.cars /\* sample from data table \*/  seed=31475 /\* recommended that you use this option \*/  method=srs /\* simple random sample \*/  sampsize=12 /\* sample size \*/  out=work.CarSample12; /\* sample stored in this data set \*/  run;  proc print data=work.CarSample12;  run; |

**NOTE:** If you do not provide a seed, you cannot reproduce the sample. It is recommended that you always include a seed when using PROC SURVEYSELECT.  
  
  
Program B shows how to select a certain percentage of the original sample using the SAMPRATE= option.

|  |
| --- |
| /\*Program B\*/  proc surveyselect data=Statdata.cars /\* sample from data table \*/  seed=13094425 /\* recommended that you use this option \*/  method=srs /\* simple random sample \*/  samprate=0.05 /\* 0 < sampling rate < 1 \*/  out=work.cars12pc; /\* sample stored in this data set \*/  run;  proc print data=work.cars12pc;  run; |

## Question

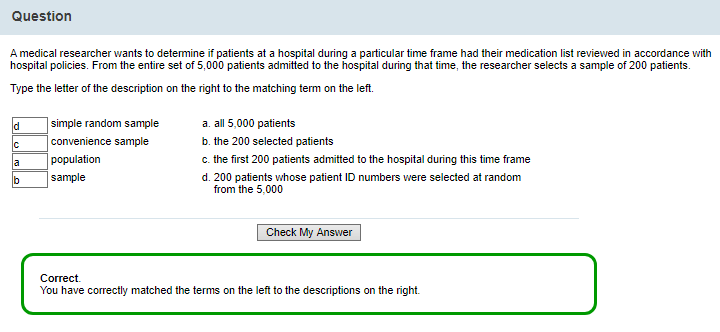


A sample should be representative.

You can attain that by using a random sample.

(A biased sample is not representative of the population.)

## Question



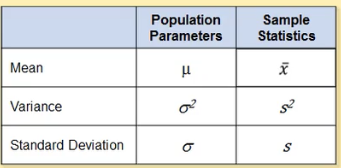
## 1.4 Parameters and Statistics

**Parameters** are numerical values that summarize characteristics of a population.

For example, the mean, or average, of some measurement on every individual in a population is a parameter. Parameter values are typically unknown, because you can't usually measure your entire population. You use Greek letters to represent population parameters.

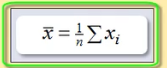
**Statistics** summarize characteristics of a sample. You use letters from the English alphabet to represent sample statistics. You can measure characteristics of your sample and provide numerical values that summarize those characteristics.

You use statistics to estimate parameters. Here is a table listing some common parameters and statistics.



Let's look at these statistics in more detail and see how you can use them to estimate parameters.

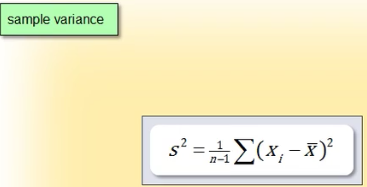
X-bar, for example, is the sample mean or sample average. You can use X-bar to estimate the population mean or population average, μ. Suppose, for example, that you have a sample x1, x2, and so on, through xn from some population. You can calculate the mean for that sample using the formula shown here.



The sample variance s2, measures the variability of your sample around the mean.

Sample variance gives you a specific measurement indicating how much your data values vary in comparison with the average value.

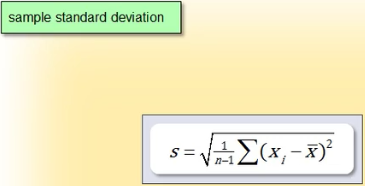
You can calculate sample variance using the formula shown here.



After you calculate your sample variance s2, you can use this statistic to estimate the population variance, σ2.

The sample standard deviation is another common measure of variability.

It is simply the square root of the variance. You calculate the sample standard deviation using the formula shown here.



Because it is the square root of the variance, the resulting measure of variability will be in the same units as the data and, therefore, the same units as the mean. You can use this statistic to estimate the standard deviation for the population.

What do we mean when we say that the sample standard deviation is a measure of variability reported in the same units as the data?

Let's look at an example.

Suppose that you are interested in knowing the average dollar amount people spend in a particular store. The unit of measurement is dollars.

The data you gather and the mean you calculate from the data will be in dollars.

The sample variance will be a measure of the spread in your data in dollars squared.

Because the standard deviation is the square root of the variance, it puts the measure of spread back on the original dollar scale.

## 1.5 Uses of Variables: Independent and Dependent

Variables are characteristics or properties of data that can take on different values or amounts for different individuals in the population.

They are data attributes such as ID number, age, gender, and body temperature.

Variables can be classified according to their function or the way they're used in a study

A variable can be independent or dependent.

* An independent variable can take different values. An independent variable affects or determines a dependent variable.
* A dependent variable can take different values in response to an independent variable.

In some contexts, a researcher selects or controls the value of an independent variable, in order to determine its relationship to the dependent variable.

For example

a researcher investigating the effect of fertilizer on plant growth could change the amount of fertilizer – the independent variable – in order to observe the effect on the plants – the dependent variable.

In other contexts, however, the independent variable's values are simply taken as given.

For example, suppose a researcher is trying to determine the effect of a variable such as incarceration rate on crime rate. The researcher can't manipulate the variable incarceration rate. It is simply observed. In this case, you might hear this variable referred to as a predictor variable.

You might also hear this type of variable referred to as an explanatory or control variable. A dependent variable is also known as a response variable or an outcome variable.

Independent Variable or Predictor Variable or Explanatory Variable or Control Variable

Dependent Variable or Response Variable or Outcome Variable

## 1.6 Types of Variables: Quantitative and Categorical

Classification of Variables according to the type of their data.

Variables can be quantitative or categorical, according to the type of their data.

**QUANTITATIVE VARIABLES** (the type of their data is quantitative)

In order to plan a statistical analysis or interpret your results, you need to know which types of variables you have. Data that consists of counts or measurements is called quantitative data. You also hear this type of data referred to as numerical data. If you can perform arithmetic operations, like addition and subtraction, or take a sample average of your data, then you know that it is quantitative. Suppose you take a survey of the buying habits of families. An example of quantitative data in your survey is the age in years of the respondents. Age is a quantitative variable because it would make sense to compute the average age of individuals in a sample.

Quantitative data can be further distinguished by two types:

* **Discrete Variables**

Discrete Variables contains data that can have only a countable number of values within a measurement range. That is, the values can be 0, 1, 2, 3, and so on. An example of discrete data is the number of children in a family. A family can have two or three children, but not 2.65 children.

* **Continuous Variables**

Continuous Variables contains data that are measured on a scale that has an infinite number of values and has no breaks or jumps. An example of a continuous variable is gas mileage. The gas mileage for a particular car might be 19 miles per gallon or 19.1 miles per gallon or 19.191034 miles per gallon, and so on. Remember that practical limitations can affect the precision of the measurement.

**CATEGORICAL VARIABLES** (the type of their data is categorical)

Categorical data consists of variables that denote groupings or labels. This type of data is also called attribute data. Categorical data can be distinguished from quantitative, because it does not make sense to perform arithmetic operations on categorical variables.

For example, your survey includes a variable for the political party affiliation of survey respondents (Democrat, Republican, Independent, other). It doesn't make sense to try to add or average the responses Republican and Democrat.

There are two main types of categorical variables:

* **Nominal data**

A nominal categorical variable exhibits no ordering within its observed levels, groups, or categories. Gender is an example of a nominal variable. There is no ordering to the groups male and female. The type of beverage you can order from a menu, such as soda, coffee, or juice, has no logical ordering to it, so it is also a nominal variable. Nominal categorical variables can be coded to appear numeric, but their numbers are meaningless. For example, the variable Gender can be coded 1 for male and 2 for female. These numbers are not inherently meaningful: they could be reversed, or replaced, by any random set of numbers.

A variable that lies on a nominal scale is sometimes called a qualitative or classification variable.

* **Ordinal data**

With ordinal categorical variables, the observed levels of the variable can be ordered in some meaningful way that implies that the differences between the groups or categories are due to magnitude. Disease condition divided into categories of low, moderate, or severe is an example of an ordinal variable. The size of beverage you can order from a menu being small, medium, or large does have a logical order to it, so it is also an ordinal variable.

## 1.7 Scales of Measurement (Nominal, Ordinal, Interval, and Ratio)

Classification of Variables according to the scale of measurement of their data.

As you've learned, variables are classified differently depending on the characteristics of that variable.

We often refer to a variable's classification as its scale of measurement. You need to know the scale of measurement for each variable in order to determine the statistical procedures appropriate for use with that variable.

SCALE OF MEASUREMENT FOR quantitative, continuous variables

There are two scales of measurement for continuous variables:

* Interval scale of measurement
* Ratio scale of measurement

**Interval scale of measurement**

Data from an interval scale can be rank-ordered like ordinal data, but it also has a sensible spacing of observations such that differences between measurements are meaningful.

For example, in measuring patient temperature, you can indicate specific differences in temperature, between the standard measurement of normal body temperature, 98.6 degrees F, and an observed body temperature of 98.2. Interval scales lack, however, the ability to calculate ratios between numbers on the scale. In the case of the Fahrenheit scale, for example, there is no true zero point. Zero does not imply the lack of temperature.

Another example of an interval scale is pH value. Sea water, which has a pH of 8, is not twice as alkaline as tomato juice, which has a pH of 4.

**Ratio scale of measurement**

Data on a ratio scale is not only rank-ordered with meaningful spacing, but it also includes a true zero point and can therefore accurately indicate the ratio of difference between two spaces on the measurement scale. For example, the Kelvin temperature scale has a true zero point. A temperature of 50 Kelvin is half as hot as 100 Kelvin. Another example of a ratio scale is money. If an individual has zero dollars, this does imply an absence of money. And one individual can have twice as much money as another.

SCALE OF MEASUREMENT FOR categorical variables

* Nominal scale of measurement (for nominal variables)

The nominal scale enables you to categorize or label variables such as gender or beverage type where there is no ordering to the levels of those variables

* Ordinal scale of measurement (for ordinal variables)

The ordinal scale indicates categories that can be ordered in a meaningful way, as in size of beverage or severity of disease.

## 1.8 Statistical Methods

The appropriate statistical method for your data **also** depends on the number of variables involved.

**Single Variable, Univariate analysis**

Univariate analysis provides techniques for analyzing and describing a single variable at a time.

Univariate analysis reveals patterns in the data, by looking at the range of values, measures of dispersion, the central tendency of the values, and frequency distribution.

It also summarizes large amounts of data and organizes data into graphs and tables so that it is more easily understood.

**Two Variables, Bivariate analysis**

Bivariate analysis describes and explains the relationship between two variables and how they change, or covary, together.

It includes techniques such as

* correlation analysis and
* chi-square tests of independence.

**Two or more Variables, Multivariate or multivariable analysis**

Multivariate or multivariable analysis examines two or more variables at the same time, in order to understand the relationships among them.

Techniques such as

* multiple linear regression and
* n-way ANOVA

are typically called multivariable analyses because there is only one response variable.

Techniques such as

* factor analysis and
* clustering

are typically called multivariate analysis because they consider more than one response variable.

Multivariate linear regression and multivariate ANOVA (MANOVA) are extensions of these techniques when there is more than one response variable. You learn many of these statistical methods in this course.

## Scenario: Exploring Your Data – Introduction

Now that you understand some basic concepts about variables and descriptive statistics, let's apply that knowledge by looking at an example.

As a project, students in Ms. Chao's statistics course gathered data on the SAT scores of high school students who attend magnet high schools in Carver County.

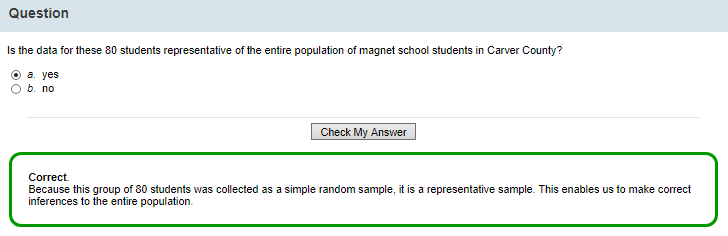
The purpose of the study is to determine if these students achieved an average combined SAT score of 1200 in math and reading, which is the goal set by the school board.

The students in Ms. Chao's class selected 80 students at random and recorded their test scores in a data file named TestScores.

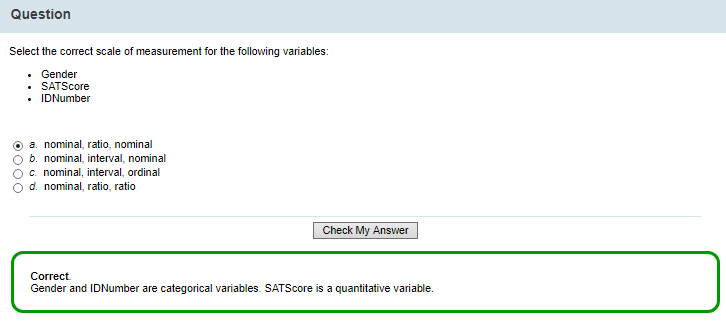
The TestScores data includes three variables.

1. The first variable is the gender of each student.
2. The second variable is each student's cumulative SAT score.
3. The last variable is IDNumber. This is a unique identifier assigned to each student.

## Question



## Question

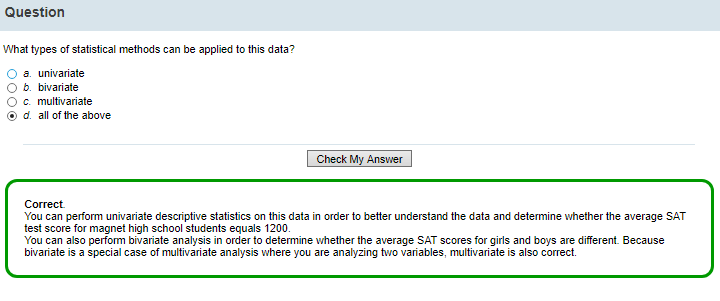


Gender (male or female), nominal variable. It makes no sense to use an order

SAT Score (number), continuous variable that uses a ratio scale of measurement (it van be zero!)

ID Number (specific numbers for identification purposes), nominal variable. It makes no sense to use an order.

## Question



Univariate statistical analysis (one variable, the SAT Score of the students)

Bivariate statistical analysis (two variables, the SAT Score and the gender)

Multivariate

## Scenario: Exploring Your Data

Now that you've examined the three variables in the TestScores data, you need to explore the particular values in this data.

You should check for unusual data points or outliers. Are there any test scores that are unusually high or low?

Are there any points that might be data entry errors? Having a unique ID number for each student is helpful if you notice any outliers. You can then look back at your data and try to understand why that point is unusual.

Before you move on to statistical analysis of your data, you must ensure that it is as error-free as possible.

For example, you should identify unique aspects of the data such as data values that cluster or show some unusual shape. If there is a possible outlier in a variable, and it goes undetected, it could cause gross errors in your interpretation of the statistics.

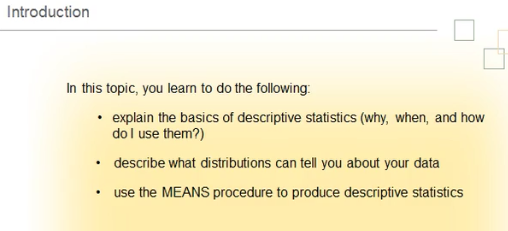
You can use descriptive statistics to look at the distribution of your data, checking the range, frequency, and shape of your data.

Then you can draw some preliminary conclusions about your data.

# 2 DESCRIPTIVE STATISTICS, NUMERICAL SUMMARIES

## 2.1 Introduction

Now you're ready to learn how to use descriptive statistics to better understand your data. Here's what you learn in this topic.



## 2.2 Describing Your Data

To summarize your data, you can construct a table of frequencies of the different values in the data.

You refer to the frequency table as the distribution of your data.

The distribution of your data tells you what values your data takes and how often it takes those values.

When you look at a data distribution, you can learn several things:

* You can look at the range of data values and demonstrate the data's spread or dispersion.
* You can verify the number of times each value appears by looking at data frequency.
* You can establish the shape of your data by determining if the data is symmetric or skewed, meaning that the distribution has a longer tail either to the left or the right.
* You can check for outliers.

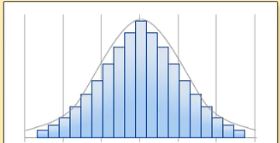


Table of Frequencies

X axis = data values

Y axis = frequency for each data value (a number indicating how many times each value appears)

Foteini Notes

Table of frequencies ( a value appears 13 times)

Table of relative frequencies (e.g. a value appears 13 times, the number of observations in the sample is 50, so the relative frequency is 13/50).

## 2.3 Measures of Location/Central Tendency

You can calculate descriptive statistics that measure locations in your data.

Statistics that locate the center of the data are called measures of central tendency.

These include mean, median, and mode.

The mean is the average of all data values. The mean is obtained by adding all the values in the data and dividing by n, the number of values. The formula is shown here. Let's take a look at an example. Suppose you have the set of 16 data values shown in this example. To calculate the mean, you add the 16 values together and divide by 16. A property of the sample mean is that the sum of the differences of each data value from the mean is always 0. This is represented by the formula shown here. The mean is useful for measuring the center of your data when the data is balanced on both sides. However, the mean is highly influenced by outliers because those unusual values are included in its calculation. If your data also included the value 35, for example, the mean would change to 75.47.

The median is another measure of central location. The median is less sensitive to the presence of outliers. The median is the middle value in the data when the data is ordered. If you have an odd number of values in your data, the median is just the middle value. If you have an even number of values in your data, the median is the average of the two middle values. In this example, there are 17 observations, so the median is the middle value. Be aware that this is the most typical way to calculate the median, but not the only one.

The mode is the data value that occurs the most. The mode in this example is 78. Be aware that the mode is not as informative in small data files. Thus, most statisticians simply compare the mean and the median when exploring their data.

**MEASURES OF LOCATION / CENTRAL TENDENCY**

1. MEAN,

The mean is the average of all data values. The mean is obtained by adding all the values in the data and dividing by n, the number of values.

In a symmetrical distribution, mean and median are very close to each other, so you can use either of them as a representation of the average.

In a positively skewed distribution, mean is greater than median, since mean is influenced by a few relatively very large scores.

In a negatively skewed distribution, mean is less than median, since mean is influenced by a few relatively very low scores.

2. MEDIAN, M

The median is the middle value in the data after the data is ordered.

For odd number of observations, M=n+1/2.

For even number of observations, M=n1+n2/2 (n1 and n2 are the two central values.

\*The median is less sensitive to the presence of outliers.

When a distribution is skewed, it is advisable to report the median

3. PERCENTILES

A percentile is the value of a variable below which a certain percentage of observations fall.

The most commonly reported percentiles are quartiles, which break the data up into quarters (Q1, Q2, Q3)

Q1, 25th percentile = value 1 (25% of the data fall at, or below this value)

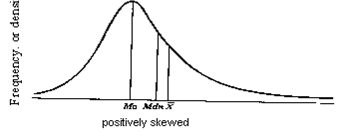
Q2, 50th percentile = value 2 (50% of the data fall at, or below this value). Also called median😊

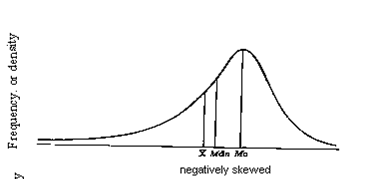
Q3, 75th percentile = value 3 (75% of the data fall at, or below this value). Also called upper quartile

4. MODE, m

The mode is the data value that occurs the most.

Be aware that the mode is not as informative in small data files





**Median is the preferred measure** of central tendency when:

There are a few extreme scores in the distribution of the data. (NOTE: Remember that a single outlier can have a great effect on the mean). b.

There are some missing or undetermined values in your data. c.

There is an open ended distribution (For example, if you have a data field which measures number of children and your options are 0, 1, 2, 3, 4, 5 or “6 or more,” then the “6 or more field” is open ended and makes calculating the mean impossible, since we do not know exact values for this field) d.

You have data measured on an ordinal scale.

**Mode is the preferred measure** when data are measured in a nominal ( and even sometimes ordinal) scale.

Foteini Notes:

We have a sample with 16 observations (n=16).

Let’s find the measures of location/central tendency.

Table 1 – Ordered values

|  |  |
| --- | --- |
| n | Values |
| 1 | 93 |
| 2 | 89 |
| 3 | 88 |
| 4 | 84 |
| 5 | 83 |
| 6 | 82 |
| 7 | 79 |
| 8 | 78 |
| 9 | 78 |
| 10 | 77 |
| 11 | 74 |
| 12 | 73 |
| 13 | 72 |
| 14 | 68 |
| 15 | 67 |
| 16 | 63 |

Mean 78

Median 78

Percentile Q1, value under which, 25% of the data lie 72.75

Percentile Q2, value under which, 50% of the data lie 78 median!

Percentile Q3, value under which, 75% of the data lie 83.25

Mode 78

Foteini Notes:

|  |
| --- |
| DESCRIPTIVE STATISTICS, NUMERICAL SUMMARIES  MEASURES OF LOCATION / CENTRAL TENDENCY define the location of the variable values   1. MEAN 2. MEDIAN 3. PERCENTILES 4. MODE   MEASURES OF DISPERSION / VARIABILITY define the way the variables values change   1. RANGE 2. INTERQUARTILE RANGE 3. VARIANCE s2 4. STANDARD DEVIATION s 5. COEFFICIENT OF VARIATION |

## 2.4 Percentiles

Percentiles are descriptive statistics that give us reference points in our data.

A percentile is the value of a variable below which a certain percentage of observations fall. The most commonly reported percentiles are quartiles, which break the data up into quarters. In this example, the 25th percentile is 72.5, which means that 25% of the data values fall at or below 72.5. You can also refer to the 25th percentile as the 1st quartile, Q1, or the lower quartile. The 50th percentile in this example is 78, which means that 50% of the data values fall at or below 78. You can also refer to the 50th percentile as the median, Q2, or the middle quartile. The 75th percentile is 83.5, which means 75% of the data values fall at or below 83.5. You can also refer to the 75th percentile as the 3rd quartile, Q3, or the upper quartile.

## 2.5 Measures of Variability/Spread/Dispersion

There are several descriptive statistics that measure the spread or dispersion of your data

Statisticians also refer to this characteristic as variability.

The range of the data is a single value that measures the difference between the maximum and minimum values. In this example, the range is 30 – that is, 93 minus 63.

The interquartile range, also known by the abbreviation IQR, is the difference between the 25th and 75th percentiles. In this example, the interquartile range is 83.5 minus 72.5, which is 11.0. The interquartile range is a robust estimate of the variability because changes in the upper and lower 25% of the data do not affect it. If there are outliers in the data, then the IQR is a more reliable measure of the spread than the overall range.

Variance is a measure of variability of the data around the mean. It is defined as the average squared difference of the observations from the mean. The formula for sample variance is shown here. Click the information button if you'd like to see how to calculate the variance for this example.

Standard deviation indicates how much variation there is from the mean, thereby measuring how spread out your data is.

Standard deviation is the square root of the variance, as shown in this formula. So for this example, the standard deviation equals 8.36. Because standard deviation is the square root of the variance, it is expressed in the same units of measurement as your data and therefore the mean.

A low standard deviation indicates that the data points tend to be very close to the mean. A high standard deviation indicates that the data is spread out over a larger range of values.

Another measure of dispersion in your data is the coefficient of variation, also referred to by the abbreviation C.V. The coefficient of variation is a measure of the standard deviation expressed as a percentage of the mean. It is useful when comparing data that has different units of measure – for example, pounds and height. It is a way of standardizing the units of measure for comparison. The formula is shown here.

MEASURES OF DISPERSION / VARIABILITY/ SPREAD

1. RANGE R

is a single value that measures the difference between the maximum and minimum values.

In this example, the range is 30 (93-63=30)

2. INTERQUARTILE RANGE IQR

is the difference between the 25th and 75th percentiles (Q3-Q1)

3. VARIANCE s2

s2=Σ(xi-)2/n-1

Variance is a measure of variability of the data around the mean.

4. STANDARD DEVIATION s, same units of measurement as the variable

s=

indicates how much variation there is from the mean, thereby measuring how spread out your data is.

A low standard deviation indicates that the data points tend to be very close to the mean.

(A high standard deviation indicates that the data is spread out over a larger range of values.)

5. COEFFICIENT OF VARIATION

The coefficient of variation is a measure of the standard deviation expressed as a percentage of the mean. It is useful when comparing data that has different units of measure

CV=s/ x100. Net number, no units of measurement

This is the standard deviation expressed as a percentage of the mean.

Foteini Notes:

We have a sample with 16 observations (n=16).

Let’s find the measures of variability / dispersion /shape

Table 1 – Ordered values

|  |  |
| --- | --- |
| n | Values |
| 1 | 93 |
| 2 | 89 |
| 3 | 88 |
| 4 | 84 |
| 5 | 83 |
| 6 | 82 |
| 7 | 79 |
| 8 | 78 |
| 9 | 78 |
| 10 | 77 |
| 11 | 74 |
| 12 | 73 |
| 13 | 72 |
| 14 | 68 |
| 15 | 67 |
| 16 | 63 |

Range R, max=93, min=63, range=max -min=30

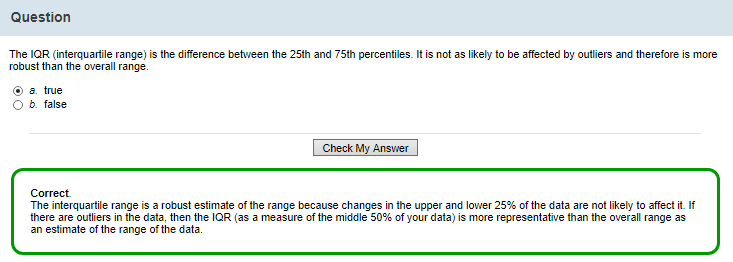
INTERQUARTILE RANGE IQR, Q3-Q1=10.5

VARIANCE s2 , 69,8666666666667

STANDARD DEVIATION s, 8,3586

COEFFICIENT OF VARIATION CV=s/ x100, 8,3586/78 x 100=10,716

## Question



The interquartile range is a robust estimate of the range because changes in the upper and lower 25% of the data are not likely to affect it.

If there are **outliers** in the data, then the IQR (as a measure of the middle 50% of your data) is more representative than the overall range as an estimate of the range of the data.

## Scenario: Using Descriptive Statistics to Answer Data Questions

As a project, students in Ms. Chao's statistics course must assess whether or not magnet school students in their district have accomplished the goal set by the school board of an average combined SAT score of 1200 in math and reading. The SAT math and reading sections have a maximum combined score of 1600. The students in Ms. Chao's statistics course selected 80 students at random from magnet school students in the district, recorded the test scores of those 80 students, and assigned each sample member an identification number. Click the information button to view the sample data.

Before the class can answer the question Ms. Chao has set for them, the class should explore the data by calculating descriptive statistics, including the sample mean and standard deviation.

Scenario: using descriptive statistics to answer data questions

**Alphabetic List of Variables and Attributes**

# Variable Type Len

1 Gender Char 6

3 IDNumber Num 8

2 SATScore Num 8

**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

11 Male 1150 38152597

12 Female 1390 99108497

## 2.6 The MEANS Procedure in SAS

Let's look at how to use the MEANS procedure to generate the descriptive statistics needed for Ms. Chao's class project.

You can use the MEANS procedure in SAS to summarize and generate descriptive statistics for your data.

PROC MEANS calculates a standard set of statistics, including the minimum, maximum, and mean data values, as well as standard deviation and n, which is the number of non missing values in the sample. You can also specify options that provide additional statistics.

Here's the syntax.

In the PROC MEANS statement, you specify your input data set.

Then you indicate any optional statistics. Other statistics that you can specify include the MEDIAN, MODE, variance (VAR), lower quartile (Q1), upper quartile (Q3), RANGE, and interquartile range (QRANGE).

Note that if you do specify any statistics in the PROC MEANS statement, they will override the default. Thus, you need to specify the entire list of statistics you want produced in the options area of the PROC MEANS statement, if you choose to list any.

In the CLASS statement, you specify the variable that the procedure uses to group the data.

In the VAR statement, you list the analysis variables for which SAS calculates descriptive statistics.

If you list no VAR statement, SAS analyzes all numeric variables in the data set.

Here's the code that you could use to generate the descriptive statistics needed for Ms. Chao's class project.

1) First, in the PROC MEANS statement, you specify the input data set, in this case, TestScores.

You can also specify options that will control which statistics SAS generates and what the output will look like.

MAXDEC=2 specifies a maximum of two decimal places for numeric values.

FW=10 specifies that the field width for all columns is 10.

The PRINTALLTYPES option displays statistics for all requested combinations of class variables – that is, for each level or occurrence of the variable and for all occurrences combined.

2) In the CLASS statement, you specify Gender as the variable that PROC MEANS will use to group the data. This is the variable referred to in the PRINTALLTYPES option. Because the Gender variable has two possible values, Male and Female, in this case, the PRINTALLTYPES option will result in three sets of descriptive statistics: one set for Male, one set for Female and a third set for all values.

3) In the VAR statement, you identify the analysis variable, in this case, SATScore.

This code also includes an optional TITLE statement so that SAS identifies your output appropriately. You end the MEANS procedure with a RUN statement.

**Program 1 (proc means, default statistics)**

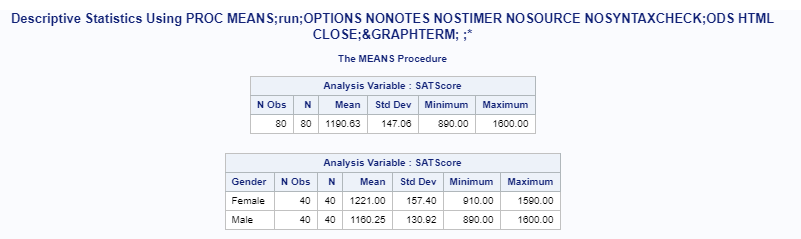
proc means data=STATDATA.TESTSCORES maxdec=2 fw=10 printalltypes;

class Gender;

var SATScore;

title'Descriptive Statistics Using PROC MEANS;

run;



**Program 2 (proc print, to see our raw data and only the first 10 observations)**

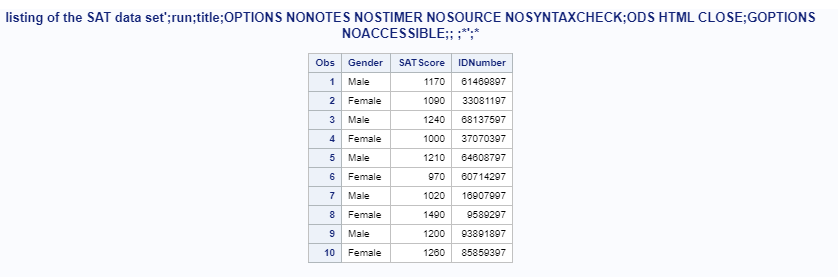
proc means data=STATDATA.TESTSCORES maxdec=2 fw=10 printalltypes;

class Gender;

var SATScore;

title'Descriptive Statistics Using PROC MEANS;

run;



**Program 3 (proc means, default statistics and also additional statistics)**

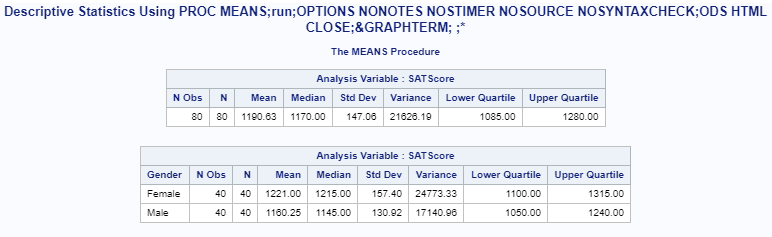
proc means data=STATDATA.TESTSCORES maxdec=2 fw=10 printalltypes n mean median std var q1 q3;

class Gender;

var SATScore;

title'Descriptive Statistics Using PROC MEANS;

run;



**Program 4 (proc means, default statistics)**

**My program**

proc means data=STATDATA.NORMTEMP maxdec=2 printalltypes n mean std QRange;

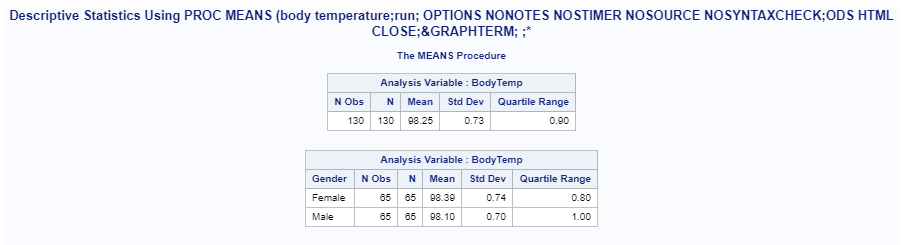
class Gender;

var BodyTemp;

title'Descriptive Statistics Using PROC MEANS (body temperature;

run;

**My results** (mind the title; in the code!



**Program suggested in the solution of this problem**

proc means data=statdata.normtemp

maxdec=2 fw=10 printalltypes

n mean std q1 q3;

var BodyTemp;

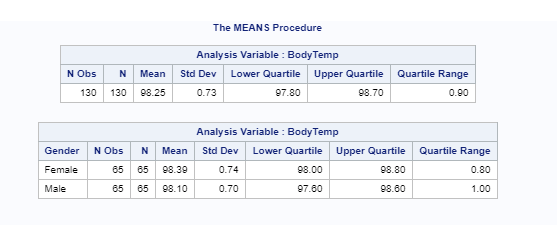
class Gender;

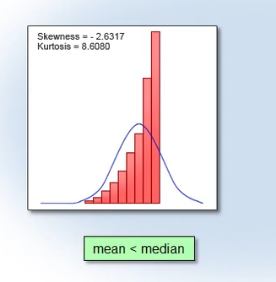
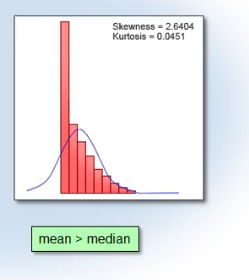
title 'Selected Descriptive Statistics for Body Temp';

run;

title;

**Results suggested in the solution of this problem**



## 2.7 Using PROC MEANS to Generate Descriptive Statistics

Before we run the MEANS procedure that we've just learned, let's use PROC PRINT to take a look at our raw data. In the PROC PRINT statement, we specify the input data set, indicating with the OBS= option that we wish to view only 10 observations, or rows, of data. The TITLE statement specifies the output heading.

proc print data=statdata.testscores (obs=10);

title 'Listing of the SAT Data Set';

run;

title;

When we submit the program, we can see the first 10 observations in our data set. The first column lists the observation numbers. The next three columns contain three variables: the gender of the student, the student's SAT score, and the unique ID number.  
  
Now that we've confirmed the basic structure of our data, let's use PROC MEANS to generate descriptive statistics for the variable SATScore.

proc means data=statdata.testscores maxdec=2 fw=10 printalltypes;

class Gender;

var SATScore;

title 'Descriptive Statistics Using PROC MEANS';

run;

title;

When we run the program, we can see how using the CLASS statement produced separate statistics for all 80 observations, then for just females and then just males. It is interesting to see that the average mean SATScore for all 80 students is 1190.6. When we look at just females, the mean is 1221 and the mean for males is 1160. Because we didn't specify which statistics we wanted, SAS generated the default statistics for PROC MEANS. The output provides the sample size (that is, the number of non-missing values), the mean, standard deviation, the minimum, and the maximum values of SATScore.  
  
Now let's modify our PROC MEANS statement and specify some different statistics. Remember that SAS will override the default statistics, so we must specify all statistics that we want included in the output. Let's request the sample size (N), the mean and median values, standard deviation and variance, in addition to the lower quartile (Q1) and upper quartile (Q3). Let's go ahead and submit our program.

proc means data=statdata.testscores maxdec=2 fw=10 printalltypes

n mean median std var q1 q3;

class Gender;

var SATScore;

title 'Selected Descriptive Statistics for SAT Scores';

run;

title;

Notice that the output is still broken out with all of the observations and then by gender because we specified Gender in the CLASS statement and PRINTALLTYPES in the PROC MEANS statement. Now we can see the values of the mean compared to the median. So, although the average score for all students was 1190.63, the median or middle value in the data was 1170. We have the values for variance as well as standard deviation. We'll talk more about how these values relate to the data later. This output also includes the values for the 25th and 75th percentiles. These figures tell us that 25% of the data values for all students fell below 1085 and that the top 25% were above 1280.

## Practice

Practice: Calculate Basic Statistics Using PROC MEANS

Task

You want to understand if true mean body temperature is 98.6 and whether women's body temperatures are the same as men's body temperatures. The Statdata.NormTemp data set contains the data that you need.

Reminder: Make sure you've defined the Statdata library.

Use PROC MEANS to determine the overall mean and standard deviation of the variable BodyTemp in the data set NormTemp. Ensure that SAS displays statistics for all requested combinations of the class variable.

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

What is the interquartile range of body temperature?

SOLUTION

Practice: Calculate Basic Statistics Using PROC MEANS

Task

You want to understand if true mean body temperature is 98.6 and whether women's body temperatures are the same as men's body temperatures. The **[Statdata.NormTemp](javascript:%20openOther('eclibjr/NormTemp.htm');)** data set contains the data that you need.

**Reminder**: Make sure you've [defined the **Statdata** library](javascript:%20openOther('eclibjr/statdata_lib.htm')).

1. Use PROC MEANS to determine the overall mean and standard deviation of the variable **BodyTemp** in the data set **NormTemp**. Ensure that SAS displays statistics for all requested combinations of the class variable.  
     
   Submit the following code.
2. proc means data=statdata.normtemp
3. maxdec=2 fw=10 printalltypes
4. n mean std q1 q3;
5. var BodyTemp;
6. class Gender;
7. title 'Selected Descriptive Statistics for Body Temp';
8. run;
9. title;

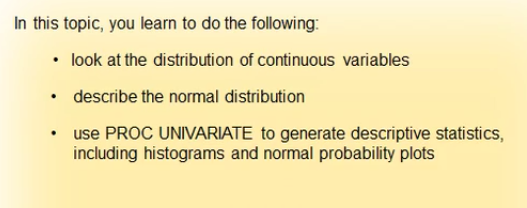
Review the Summary Statistics output. The overall mean is 98.25. The standard deviation is 0.73.

1. Do the mean values seem to differ between men and women?  
     
   The values do differ somewhat.
2. What is the interquartile range of body temperature?  
     
   The interquartile range is 0.90 (98.70 - 97.80). Another option for determining the interquartile range is to include the keyword QRANGE in the list of statistics specified in the PROC MEANS statement. That way SAS will calculate this statistic for you.

# 3 DESCRIPTIVE STATISTICS, GRAPHICAL SUMMARIES

## 3.1 Introduction

Another way to understand your data is to display it using graphs and plots. Here's what you learn in this topic.



## 3.2 Picturing Your Data: Histogram

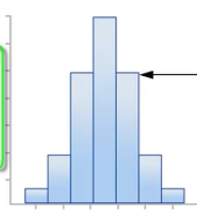
There are several ways to visualize the distribution of your data.

One of the most effective techniques is to plot a histogram.

A histogram is a visual representation of the frequency distribution of your data. The frequencies are represented by bars. Each bar in the histogram represents a group of values, sometimes referred to as a bin. The heights of the bars represent the frequency or count of values in the bins. In this example, the height of the bar represents a percentage value.

X axis: group of values (or bin)

Y axis: frequency at which we observe each group of values



## 3.3 Normal Distribution

The normal distribution is a common theoretical distribution in statistics.

The normal distribution is shaped like a bell, with values concentrated near the mean.

(Underlying the normal distribution is a mathematical function called the probability density function.)

The height of the function at any point on the horizontal axis is the probability density at that point.

The shape of a normal distribution depends on the value of two parameters, the mean (μ) and the standard deviation (σ).

* The mean (μ) determines the center of the distribution.

The normal distribution is symmetric around the mean.

* The standard deviation (σ) determines how variable the distribution is.

A larger standard deviation implies a wider normal distribution.

Foteini Notes

* Mean= determines the center of the distribution
* St Dev = determines the width of the distribution

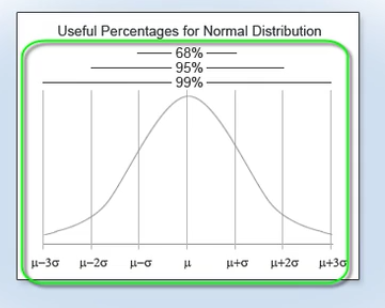
(low st dev means narrow width, the values vary only a little and close to the mean. High st dev means big width, the values vary far from the mean

You can calculate probabilities using the normal probability density function by calculating areas under the bell-shaped curve.

* Approximately 68% of the area under the bell curve falls within 1 standard deviation of the mean.

This means that approximately 68% of the data values fall within this area.

* Approximately 95% of the area under the curve falls within 2 standard deviations of the mean. This means that approximately 95% of data values fall within this area.
* Finally, approximately 99% of the area under the curve falls within 3 standard deviations of the mean, meaning that approximately 99% of the data values fall within this area.



Statisticians often consider values that are more than 2 standard deviations from the mean as unusual.

Now you can see why. Only about 5% of all values are that far away from the mean. Depending on the context, sometimes statisticians treat only values more than 3 standard deviations away from the mean as unusual.

Because the normal distribution has many useful mathematical properties, statistical procedures for data based on a random sample often assume the normal distribution. So it's important to know how to check this assumption for your data.

## 3.4 Assessing Normality (with skewness and kurtosis)

To check the assumption that your random sample has a normal distribution, it can be useful to plot a histogram.

The sample of data in this histogram is from a normal distribution.

The grey curve in this diagram represents the normal distribution.

You can see that the histogram follows the shape of the curve fairly closely.

In addition to plotting a histogram, you can also look at statistical summaries of your data in order to check the assumption that your data is normally distributed.

Two such summaries are skewness and kurtosis.

Skewness and kurtosis measure certain aspects of the shape of a distribution.

The closer these statistics are to 0 for your sample, the closer your data is shaped like the normal distribution.

Foteini Notes

**How can we examine if the data of our random sample follow the normal distribution?**

With 2 ways

1) make a plot of the data and see if the shape of the data corresponds to the bell shape of the normal distribution

2) measure the skewness and the kurtosis parameters, if they are close to 0 then our data follow the normal distribution

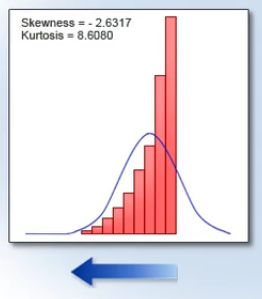
## 3.5 Measures of Shape: Skewness

**Skewness**

* Is a measure of the asymmetry of the distribution.
* measures the tendency of your data to be more spread out on one side of the mean than on the other.
* You can think of the direction of skewness as the direction the data is trailing off to.
* The closer the skewness statistic is to 0, the more normal or symmetric the data.

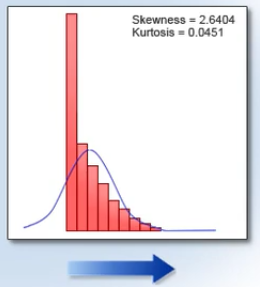
**Left – skewed distribution**

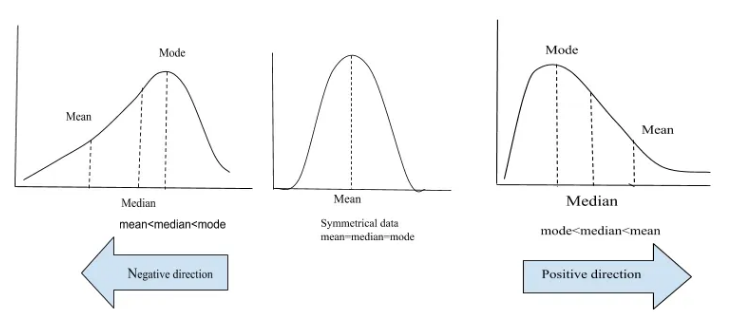
* the skewness statistic is negative
* the mean is less than the median, μ<Median



**Right – skewed distribution**

* the skewness statistic is positive
* the mean is greater than the median, μ>Median





## 3.6 Measures of Shape: Kurtosis

**Kurtosis**

* is a measure of tail thickness
* measures the tendency of your data to be concentrated toward the center or toward the tails of the distribution.
* The closer the kurtosis statistic is to 0, the closer the tails of your data resemble the tail thickness of the normal distribution.

!Be aware that normal distribution actually has a kurtosis value of positive 3.

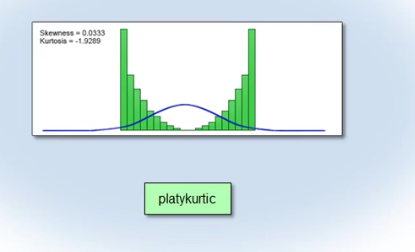
*SAS has standardized it to zero, so that when you are assessing normality, the values against which to compare the sample skewness and kurtosis are the same.*

**Platykurtic distribution/light-tailed distribution THIN TAILS**

* negative kurtosis statistic

the data has lighter tails than in a normal distribution. Data is less heavily concentrated about the mean.

* flatter than the normal distribution, less peaked



In this example, the platykurtic distribution is symmetric.

A symmetric platykurtic distribution is characterized by being flatter than the normal distribution, that is, less peaked, with heavier flanks and thinner tails.

A distribution with two peaks is also referred to as a **bimodal distribution**. (Rectangular, bimodal, and multimodal distributions tend to have low values of kurtosis.)

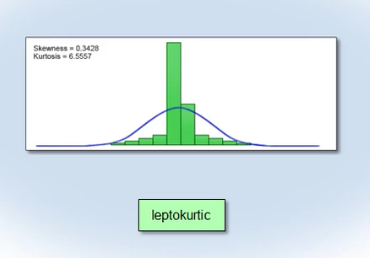
Flank = side, πλευρά?

Tail= edge, άκρο/

**Leptokurtic / heavy-tailed / outlier-prone distribution HEAVY TAILS**

* positive kurtosis statistic
* tends to have a higher peak than the normal distribution

the data has heavier tails and is more concentrated about the mean than a normal distribution. If you were to zoom in on the histogram shown here, you'd see that the data extends well beyond 2 standard deviations of the mean. You refer to this type of distribution as leptokurtic. A leptokurtic distribution is often referred to as heavy-tailed and might sometimes also be referred to as an outlier-prone distribution.



If the distribution is symmetric, a leptokurtic distribution tends to have a higher peak than the normal, with an excess of values near the mean and in the tails, but with thinner flanks. 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. Kurtosis can be difficult to assess visually.

## Question



1 normal distribution

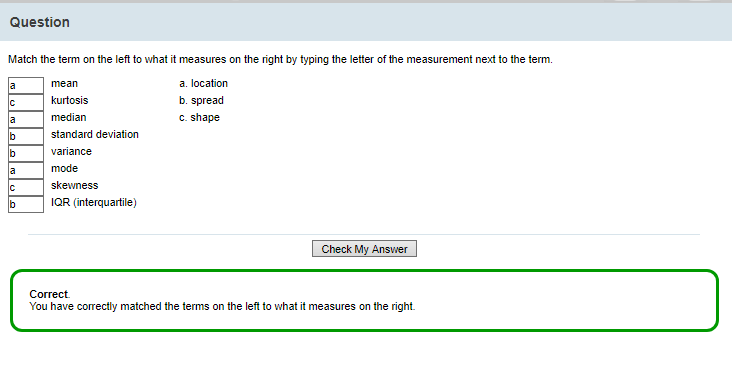
2 right-skewed distribution

3 left-skewed distribution

4 leptokurtic (or heavy-tailed) distribution

5 platykurtic (or light-tailed) distribution

## Question



## SUMMARY for measures of location, spread, shape

### Measures of location/central tendency

1. mean, x ̅
2. median, M
3. percentiles
4. mode, m

### Measures of variability/spread/dispersion

1. range R
2. interquartile range IQR
3. variance s2
4. standard deviation s
5. coefficient of variation

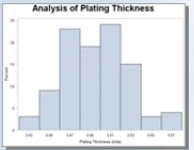
### Measures of shape

1. Skewness
2. kurtosis

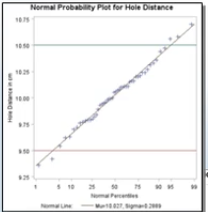
## 3.7 Normal Probability Plots

In addition to plotting your data using a histogram, there are other ways to visualize and assess the distribution of your data.

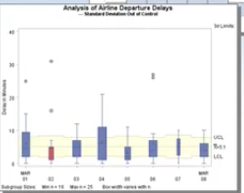
HISTOGRAM



NORMAL PROBABILITY PLOT



BOX PLOT



MORE ON NORMAL PROBABILITY PLOTS

You can use normal probability plots and box plots to compare your data with the normal distribution.

A normal probability plot is a visual method for determining whether or not your data comes from a distribution that is approximately normal.

Y AXIS The vertical axis represents the actual data values, and

X AXIS the horizontal axis displays the expected percentiles from a standard normal distribution.

The normal reference line along the diagonal is where your data would fall if it were perfectly normal.

The points hovering around the reference line are the actual data 😊

Let's look at some sample plots.

In the first plot, the distribution of data points follows the normal reference line quite closely. Thus, this data distribution looks fairly normal.

In the second plot, the distribution of data is skewed to the right of the reference line.

In the third plot, the distribution of data is skewed to the left of the reference line.

The fourth plot shows a light-tailed distribution. Remember that the vertical axis is your ordered response values, so the mean of your data is a horizontal line running approximately through the middle of the plot. If you examine your data points paying particular attention to the tails, the data in the tails looks like it is being squashed from the top and the bottom, toward the mean. In other words, the data points are closer to the mean than they should be if they were normal, meaning that the tails of the distribution are light.

The fifth plot shows a heavy-tailed distribution. If you examine your data points paying particular attention to the tails, the data in the tails looks like it is being stretched in the vertical direction, away from the mean. In other words, the data points are further from the mean than they should be if they were normal, meaning the tails of the distribution are heavy.

The main point here is that the non-normal data distributions in the second through fifth examples are visually very obvious.

## 3.8 Box Plots

Another way to examine the distribution of your data is to create a box plot.

This type of graph makes it easy to see how spread out your data is and if there are any outliers.

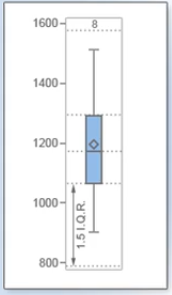
* The box represents the middle 50% of your data, which is the interquartile range.
* The bottom of the box indicates the 25th percentile.
* The middle line represents the 50th percentile, or the median, and
* the top line indicates the 75th percentile.
* The diamond denotes the mean. (Thus, you are able to visualize very easily how close the mean is to the median.)
* The whiskers extend from the box as far as the data extends, to a maximum length of 1.5 times the interquartile range (IQR) above the 75th percentile.

This is referred to as 1.5 interquartile units.

Any data points farther than 1.5 interquartile units away from the box are considered possible **outliers** and are represented in this plot as circles.

You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as by assessing the symmetry of the box and whiskers around the median line.

This box plot shows a data distribution that is approximately symmetric.



## 3.9 Assessing whether our data follow the normal distribution

In assessing whether your data comes from a normal distribution, you should consider the following points.

* Compare the mean and the median.

If they are nearly equal, that is an indicator of symmetry, which is one requirement for normality.

* Check if the skewness and kurtosis statistics are close to 0.

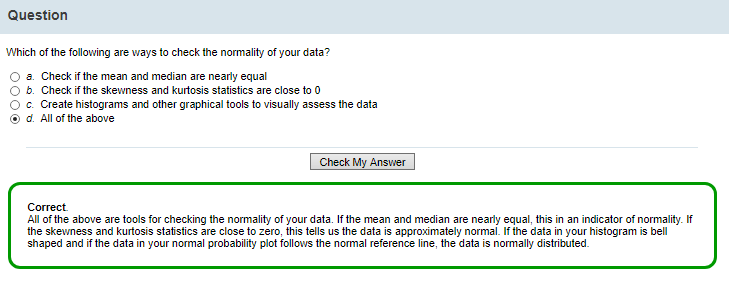
If the mean and the median coincide and the skewness and kurtosis statistics are close to 0, this tells us the data is approximately normal.

As a loose rule of thumb, if either the skewness or kurtosis statistic is greater than 1 or less than -1, many researchers conclude that the data is not normal.

Other researchers conclude that the data is not normal if either statistic is greater than 2 or less than -2.

* It's also useful to visually assess your data by creating histograms, box plots, and normal probability plots. If the data in your histogram is bell shaped and if the data in your normal probability plot follows the normal reference line, the data is normally distributed.

## Question



## Scenario: Assessing Normality (An Example)

You know several ways to visualize the distribution of your data using histograms, normal probability plots, and box plots.

Let's apply these techniques to an example.

You need to assess SAT score data for a sample of 80 students from magnet schools in Carver County. Before you can answer specific questions about this data, you need to ensure that you understand the distribution and other features of the SAT score data.

You can use descriptive statistics, including graphical techniques, to look at the distribution of your data, checking the range, frequency, and shape of your data.

Does it exhibit a normal distribution?

What are the measures of the mean and median, skewness and kurtosis, and what do those statistics tell us about the data?

Also, you should check for unusual data points or outliers. Are there test scores that are unusually high or low, or are there any data points that might be data entry errors?

After you complete this assessment, you're ready to answer specific questions about this data by applying additional statistical techniques.

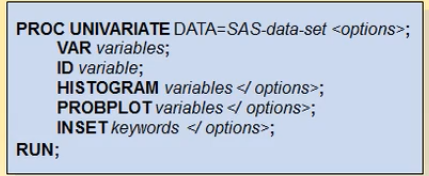
Click the information button if you'd like to review the sample data.

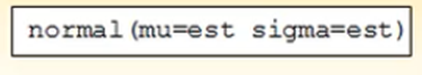
## 3.10 The UNIVARIATE Procedure in SAS

You can use PROC UNIVARIATE to generate descriptive statistics, including

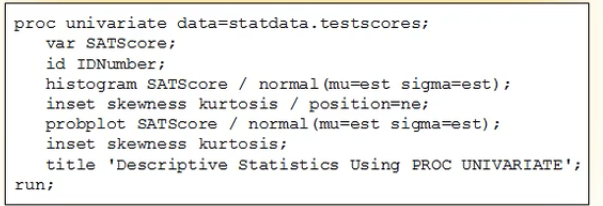
* skewness and kurtosis,
* quantiles or percentiles,
* frequency tables and
* extreme values.
* It also generates histograms and normal probability plots that assist you in assessing the distribution of your data.

Here's the syntax.



* In the PROC UNIVARIATE statement, you specify your SAS data set.
* In the VAR statement, you specify the analysis variables. If you do not include a VAR statement, SAS analyzes all numeric variables in the data set.
* Following the VAR statement, you can provide an ID statement. Here you list the variable or variables that SAS should label in the table of extreme observations and identify as outliers in the graphs. By default, SAS displays the five lowest and five highest observations in the table of extreme observations.
* Then in the HISTOGRAM statement, you list the variable or variables you want histograms created for. You can add additional options to the HISTOGRAM statement. The NORMAL option creates a normal curve overlay to the histogram using the estimates of the population mean and standard deviation. 
* In the PROBPLOT statement, you list the variables you want probability plots created for. You can add additional options to the PROBPLOT statement. The NORMAL option creates a diagonal reference line based on the estimates of the population mean and standard deviation. 
* You use the INSET statement to create a box of summary statistics directly on the graphs. The INSET statement must follow the line of code that creates the plot you want augmented. Specify INSET, followed by keywords for the summary statistics you want included. You can add additional options, such as specifying the placement of the box in the graph window.

Here's the code for using PROC UNIVARIATE to produce descriptive statistics, a histogram, and a normal probability plot for the TestScores data.



* In the PROC UNIVARIATE statement, you specify the TestScores data set.
* And in the VAR statement, identify SATScore as the analysis variable.
* The ID statement specifies that SAS will use the IDNumber variable as a label in the table of extreme observations and as an identifier for any extreme observations in the graphs.
* In the HISTOGRAM statement, you specify that you want SAS to create a histogram of the continuous variable SATScore with a normal curve superimposed using estimates of the population mean and the population standard deviation. Note that SAS determines the width and number of bins automatically.
* The INSET statement with the option POSITION=NE specifies that SAS prints a box with the skewness and kurtosis statistics in the northeast corner of the histogram.
* The PROBPLOT statement creates a normal probability plot of SATScore with the overlay of a normal reference line created using estimates of the population mean and the population standard deviation.
* A second INSET statement with no option added specifies a box with the skewness and kurtosis statistics in the default position of the plot, which is the northwest corner.

## 3.11 Statistical Graphics Procedures in SAS

In addition to the statistical graphics available to you with PROC UNIVARIATE, you might want to use PROC SGSCATTER,

PROC SGPLOT,

PROC SGPANEL, and

PROC SGRENDER to produce a wide variety of additional plot types.

You can use PROC SGSCATTER to produce several types of scatter plots. You can create a single-cell (or a simple Y by X) scatter plot, a multi-cell scatter plot with multiple independent scatter plots in a grid, and a scatter plot matrix, which produces a matrix of scatter plots comparing multiple variables.

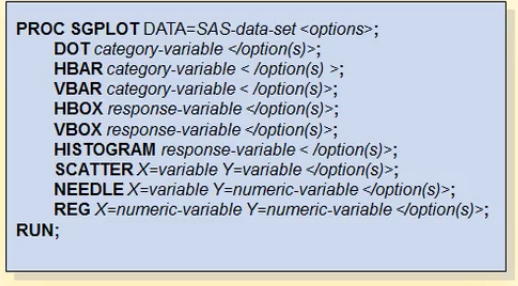
Using **PROC SGPLOT**, you can produce a wide variety of plot types, including scatter plots, line graphs, histograms with overlaid distribution curves, regression lines with confidence and prediction bands, dot plots, box plots, bar charts, and so on. You can also overlay plots together to produce many different types of graphs. Plus, you can use statements and options to control graph appearance and add features such as legends and reference lines.

Using PROC SGPANEL, you can produce panels of plots for different levels of a factor or several different time periods, depending on the classification variable. For example, you can create several bar charts side by side, each representing a different quarter of the year. You can also produce side-by-side histograms, one representing males and the other representing females for whatever type of data you are looking at. This enables you to quickly provide a visual comparison for your data.

PROC SGRENDER enables you to create plots from graph templates you have modified or written yourself. For more information about each of these procedures, consult the SAS documentation.

## 3.12 The SGPLOT Procedure in SAS

Let's look at the SGPLOT procedure in detail.



In the PROC SGPLOT statement, you first specify your input data set and any options.

You can then use the

* DOT statement to generate dot plots
* HBAR statement to generate horizontal bar charts
* VBAR statement to generate vertical bar charts

that summarize the values of a category variable.

You can generate several other types of distribution plots using PROC SGPLOT: histograms, box plots, and density curves. You use the

HBOX statement to create a horizontal box plot that shows the distribution of your data.

VBOX statement generates a vertical box plot.

HISTOGRAM statement creates a histogram that displays the frequency distribution of a continuous variable.

You can use PROC SGPLOT to generate other basic plots as well:

scatter plots, SCATTER statement to create a scatter plot of two continuous variables

series plots,

band plots,

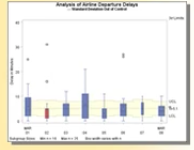
needle plots, and NEEDLE statement creates a plot with needles connecting each point to the baseline

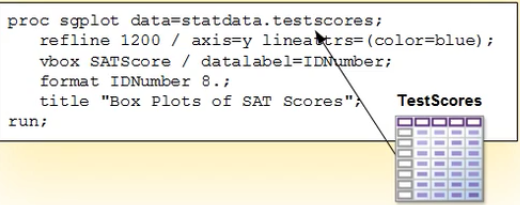
vector plots.

REG statement generates a fitted regression line or curve.

REFLINE statement to create a horizontal or vertical reference line on the plot. If you are creating a horizontally oriented plot, the reference line will be vertical. And if the plot is vertical in orientation, the reference line will be horizontal. The REFLINE statement can be placed anywhere in the procedure.

**Let's use PROG SGPLOT to create a vertical boxplot of the variable SATScore for the TestScores data.**





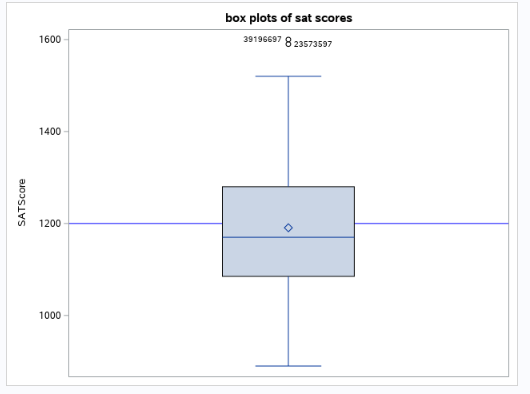
In the PROC SGPLOT statement, you specify the TestScores data set.

The REFLINE statement with additional options creates a blue horizontal reference line (through the y axis) at the data value 1200.

In the VBOX statement you specify the creation of a vertical box plot using the continuous variable SATScore.

The DATALABEL= option specifies IDNumber as the label for any outliers. If there are no outliers, then this option has no effect.

The FORMAT statement formats the variable IDNumber to 8 units.



## Code challenge



## 3.13 ODS Graphics Output

ODS Graphics is an extension of the SAS Output Delivery System.

With ODS Graphics, statistical procedures produce graphs as automatically as they produce tables, and graphs are integrated with tables in the ODS output. You can find a list of the graphs available for each SAS procedure in the SAS documentation.  
  
You can use ODS statements to control your statistical graph output as easily as your tabular output. To specify options for graphs, you submit the ODS GRAPHICS statement. You can submit this statement with or without the keyword ON. For example, you can use the WIDTH= option to specify the width of the graph.  
  
ODS statistical graphics are ODS output objects, just like tables that are produced by SAS procedures. Procedures assign a name to each output object, which enables you to specify individual statistical graphs and tables in ODS statements.  
  
To select or exclude specific test results, graphs, or tables from your output, you can use the ODS SELECT and ODS EXCLUDE statements.  
  
Another way to control your output is to use PLOTS=, an option that is usually available in the procedure statement. This option enables you to specify which graphs SAS should create, either in addition to or instead of the default plots.

  
  
In addition, there are several ways you can control the layout of your graphics output.  
  
You can use ODS templates to modify the layout and details of each graph.  
  
You can use ODS styles to control the general appearance and consistency of your graphs and tables. By default, SAS displays your ODS Graphics output using the HTMLBLUE style.  
  
In this course, you learn to control your statistical graphics in some of these ways. To learn more about ODS Graphics features that are not covered in this course, see the SAS/STAT documentation.

## 3.14 Using SAS to Picture Your Data

The program in the editor uses PROC UNIVARIATE to produce descriptive statistics, a histogram and a normal probability plot for the variable **SATScore**.

Let's specify the WIDTH= option so that our graphics are sized correctly.

Let's submit the program.

ods graphics on/width=600;

proc univariate data=statdata.testscores;

var SATScore;

id idnumber;

histogram SATScore / normal(mu=est sigma=est);

inset skewness kurtosis / position=ne;

probplot SATScore / normal(mu=est sigma=est);

inset skewness kurtosis;

title 'Descriptive Statistics Using PROC UNIVARIATE';

run;

title;

The log shows that SAS processed the code without errors. When we look at our results, the tabular output tells us quite a lot about our data.

* The sample size (N) is 80. The mean is 1190.6, which is roughly equivalent to the median of 1170 in the quantiles report.
* The standard deviation is approximately 147, which is a measure of the average variability around the mean. The variance is the standard deviation squared.
* The skewness statistic is 0.642. It is positive, but close to 0, so the distribution is slightly right-skewed.
* The kurtosis statistic is 0.424. It is positive, but also close to 0, so the distribution is slightly heavy-tailed.
* The coefficient of variation is 12.35. This is the standard deviation expressed as a percentage of the mean. This is useful if you need to compare data with different units of measurements – for example, inches to centimeters. It's a way of standardizing units of measurement.
* The standard error of the mean of 16.4 measures the variability of the mean.
* In the quantiles report, SAS provides various reference points in the data set. We can see that the minimum for SATScore is 890 and the maximum is 1600. We are also given other percentile values, such as the 25th, the median, and the 75th.
* In the extreme observations output, SAS specifies the five lowest values of SATScores and the five highest values. Remember that we specified IDNumber as an identifier in our code.

Next, let's scroll down and review our graphs.

* This histogram provides some additional information about the TestScores data. The bin identified with the midpoint of 1100 has approximately 33% of the values. The inset box displays the skewness and kurtosis values. The data looks approximately normal.
* Let's review the normal probability plot. The diagonal reference line here represents where the data values would fall if they came from a normal distribution. The circles represent the observed data values. Because the circles closely follow the diagonal reference line in the graph, you can conclude that there does not appear to be a significant departure from normality.
* Now we use PROC SGPLOT to create a boxplot of the variable SATScore. Let's specify a reference line at 1200. This is the SAT test score goal for magnet schools in the Carver County school district. We also specify the DATALABEL option in the VBOX statement and ask SAS to label any outliers with their IDNumber values. Let's submit this code.

proc sgplot data=statdata.testscores;

refline 1200 / axis=y lineattrs=(color=blue);

vbox SATScore / datalabel=IDNumber;

format IDNumber 8.;

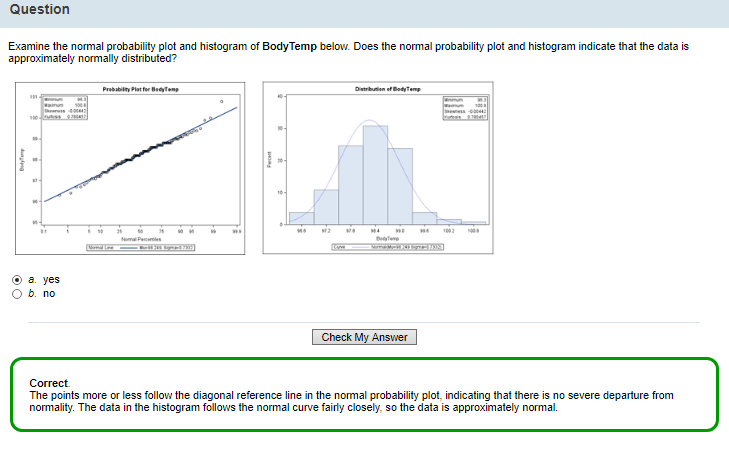
title "Box Plots of SAT Scores";

run;

title;

The log verifies that the code ran successfully. Now let's look at our output. The top whisker represents the largest point up to 1.5 interquartile units from the box. The top line of the box represents the 75th percentile. The horizontal line inside the box represents the median or 50th percentile. The bottom line of the box represents the 25th percentile. The bottom whisker represents the smallest point up to 1.5 interquartile units from the box. The diamond represents the mean. The blue horizontal line is the reference line we added where SATScore is 1200. Note that there are two outliers, values beyond 1.5 interquartile units. SAS displays their IDNumber values as we had requested.

## Question



## Practice

Practice: Producing Descriptive Statistics and Box Plots

Task

You need to determine if the variables BodyTemp and HeartRate in the Statdata.NormTemp data set are normally distributed and if average body temperature is truly 98.6 degrees.

Reminder: Make sure you've defined the Statdata library.

1) Determine the minimum, maximum, mean, and standard deviation for the variables BodyTemp and HeartRate in the NormTemp data set. Also calculate the skewness and kurtosis statistics. Do the variables appear to be normally distributed?

1. proc univariate data=statdata.normtemp noprint;
2. var BodyTemp HeartRate;
3. histogram BodyTemp HeartRate / normal(mu=est sigma=est noprint);
4. inset min max skewness kurtosis / position=ne;
5. probplot BodyTemp HeartRate / normal(mu=est sigma=est);
6. inset min max skewness kurtosis;
7. title 'Descriptive Statistics Using PROC UNIVARIATE';
8. run;
9. title;

I have 2 variables

I create a histogram for each variable (with mean=mu and stdev=sigma and also with min, max, skewness, kurtosis)

I create a normal probability plot for each variable

ANSWER= The distributions for both variables look approximately normal. I see a histogram and a normal probability distribution for the first variable and another histogram and normal probability distribution for the second variable

2) Create box plots for the BodyTemp and HeartRate variables.

Use ID to identify outliers.

For BodyTemp, display a reference line at 98.6 degrees.

Does the average body temperature seem to be 98.6 degrees?

1. proc sgplot data=statdata.normtemp;
2. refline 98.6 / axis=y lineattrs=(color=blue);
3. vbox BodyTemp / datalabel=ID;
4. format ID 3.;
5. title "Box Plots of Body Temps";
6. run;
7. proc sgplot data=statdata.normtemp;
8. vbox HeartRate / datalabel=ID;
9. format ID 3.;
10. title "Box Plots of Heart Rate";
11. run;
12. title;

I run two programs, one for each variable.

The diamond denotes the mean. The diamond is placed lower than the reference line we set, so the average boytemp (mean) is less than 98.6

3) In the NormTemp data set, which of the following phrases best describes the distributions of BodyTemp and HeartRate?

1. pretty close to normal
2. left-skewed
3. right-skewed
4. to have high positive kurtosis
5. to have high negative kurtosis.

The correct answer is a. Because the histograms are bell shaped and the data follows the diagonal reference lines in the normal probability plots, the variables BodyTemp and HeartRate are both normally distributed. The skewness and kurtosis statistics are fairly close to zero for both variables as well, which tells us that BodyTemp and HeartRate are approximately normal.

# 4 INFERENTIAL STATISTICS, Confidence intervals for the mean

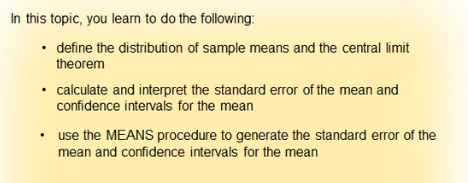
## 4.1 Introduction

You know how to assess normality and use descriptive statistics to better understand your data.

Now you want to use INFERENTIAL STATISTICS to calculate

* the **standard error of the mean** and
* **confidence intervals for the mean**.

Here's what you learn in this topic.



## 4.2 Point Estimators, Variability, and Standard Error

When you gather a sample of data, you calculate sample statistics to estimate parameters from the population.

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

You're already familiar with several point estimators.

For example, you know that you can use X-bar, the sample mean, to estimate μ, the population mean. Similarly, you can use s, the sample standard deviation, to estimate σ, the population standard deviation.

An estimator takes on different values from sample to sample, so it's important to know the variance of an estimator.

**Let's look at an example**. A research company is interested in whether drivers adhere to the posted speed limit of 45 mph on a particular street. They collect data for 25 drivers.

Here's a question.

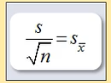
If you look at a sample from this data that has a mean of 45.56 mph, why are you not absolutely certain that this is the mean speed for the population?

The answer is because this sample mean is only an estimate of the population mean. If you collected another sample of drivers, you'd likely obtain a different estimate of the mean. Different samples yield different estimates of the mean for the same population.

The variance of the sample mean refers to how much the value of the sample mean varies from sample to sample. Any sample statistic has some variability. The key is to understand and estimate this variability.

**A STATISTIC THAT MEASURES THE VARIABILITY OF YOUR ESTIMATOR IS THE STANDARD ERROR**.

Standard error differs from the sample standard deviation because sample standard deviation deals with the variability of your data whereas standard error deals with the variability of your sample statistic.

, s= sample standard deviation , n=sample size

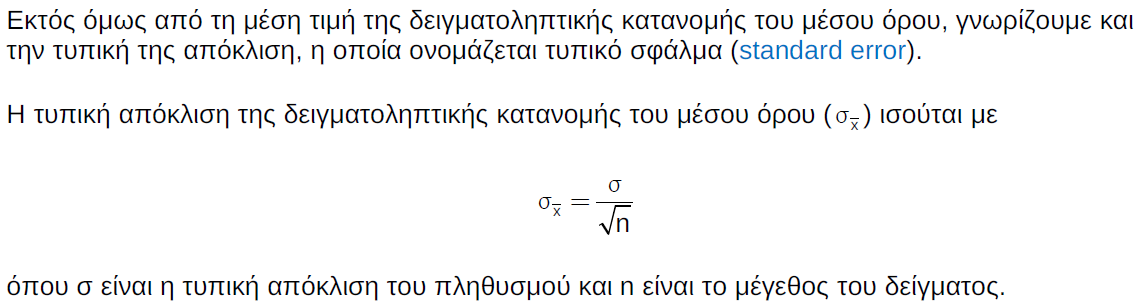
An example is the standard error of the mean, which measures the variability of your sample mean.

It's an estimate of how much you can expect the sample mean to vary from sample to sample.

You calculate the standard error of the mean using the formula shown here, where s is the sample standard deviation and n is the sample size.

The larger the sample size, the smaller the standard error of the mean will be.

The smaller the standard error of the sample mean, the more precise the sample mean is as an estimator of the population mean.



**Use of the standard error:**

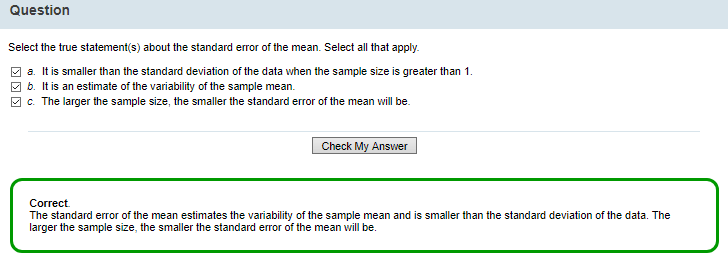
If it is low, then the standard deviation of the means is low, then the statistic we have calculated is precise and close to the true mean μ

Foteini Notes:

The standard error is an estimate of the [standard deviation](https://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) of a [statistic](https://stattrek.com/Help/Glossary.aspx?Target=Statistic).

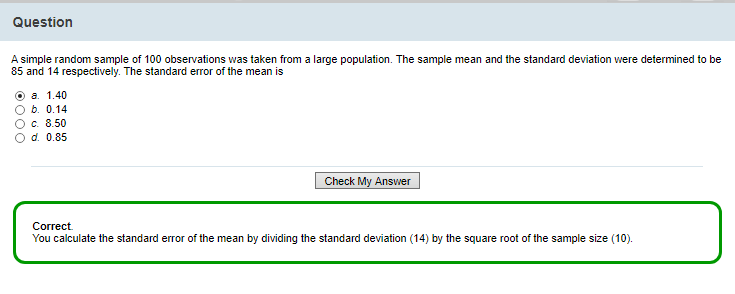
(The standard error is important because it is used to compute other measures, like [confidence intervals](https://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) and [margins of error](https://stattrek.com/Help/Glossary.aspx?Target=margin-of-error).)

## Question



The standard error of the mean estimates the variability of the sample mean and is smaller than the standard deviation of the data. The larger the sample size, the smaller the standard error of the mean will be.

## Question



## 4.3 Distribution of Sample Means (logical step from Descriptive to Inferential statistics)

What happens if you look at the distribution of ALL possible sample means from the population?

This is called the distribution of sample means.

Let's look at the car example.

Suppose you take 500 random samples, all with a sample size of 10, from an identified population size of 5,000 and calculate a mean for each sample.

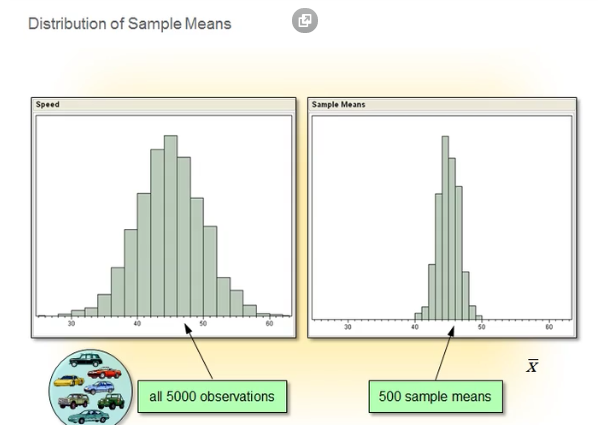
The histogram on the left shows the distribution of all 5000 observations.

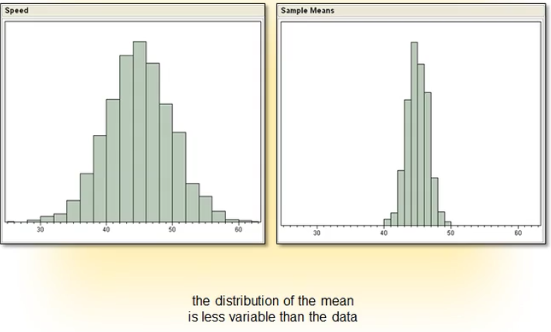
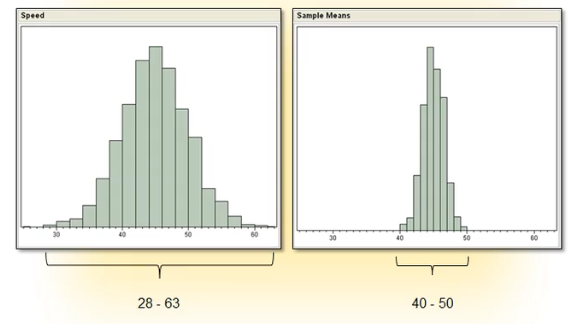
The histogram on the right represents the distribution of the 500 sample means.

**The variability of the distribution of sample means is smaller than the variability of the distribution of the 5000 observations**. That should make sense.

With a hypothesized mean of 45, it seems relatively likely to find one driver with a speed of 62. But it's not likely that the mean of a sample of 10 drivers would be 62.

The distribution of the means is always less variable than the data. (it has a smaller standard deviation s)





Άμα σκεφτούμε προσεκτικά, αντιλαμβανόμαστε ότι οι μέσοι όροι των δειγμάτων έχουν και αυτοί τη

δική τους κατανομή, όπως και τα πρωτογενή δεδομένα ενός δείγματος.

Η κατανομή αυτή ονομάζεται δειγματοληπτική κατανομή (sampling distribution) του μέσου όρου και είναι εξαιρετικά σημαντική έννοια, που μας επιτρέπει να κάνουμε το λογικό άλμα από την περιγραφική στην επαγωγική στατιστική.

## 4.4 Interval Estimators

Because we know that point estimators vary from sample to sample, it would be nice to have an estimator of the mean that directly accounts for this natural variability.

An interval estimator is another way to estimate the true population mean. The interval estimator gives us a range of values that is likely to contain the population mean. It incorporates the uncertainty that arises from random variability. The interval is centered at the sample mean. The margin of error is how far the interval extends on each side of the sample mean. You calculate the interval from the standard error of the mean that we just talked about as well as a value that is determined by the degree of certainty we require.

**Foteini Notes:**

We can use an interval estimator instead of a point estimator because point estimators change from sample to sample.

An estimate of a population parameter may be expressed in two ways:

(estimate = statistic. When referring to sample, use the term statistic (­ ), when referring to population use the term parameter (μ))

* **Point estimate**. A point estimate of a population parameter is a single value of a statistic.

For example, the sample mean is a point estimate of the population mean μ.

Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.

* **Interval estimate**. An interval estimate is defined by two numbers, between which a population parameter is said to lie.

For example, *a* < x < *b* is an interval estimate of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*.

We can use an interval estimator instead of a point estimator because

* The point estimator changes from sample to sample and also provides only a single value of the sample statistic
* The interval estimator takes into account the variability of the statistic

## 4.5 Confidence Intervals

Confidence intervals for the mean are interval estimators of the population mean.

They take into account the variability of the sample statistic - in this case, the sample mean

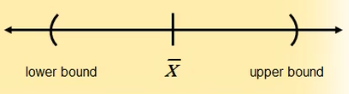
Whereas the point estimate gives a single best estimate of the population mean, the confidence interval shows a range of plausible values for the unknown population mean.

It's described by reporting an upper and a lower bound.

**Foteini Notes:**

the interval estimator provides a range of plausible values for the population mean μ and takes into account the variability of the sample statistic **and therefore is better than**

the point estimator which gives a single value for the sample statistic .

Confidence interval = interval estimator of the mean μ 

Sample mean = point estimator of the mean μ

Here is a good way to think about confidence intervals.

If you were to draw infinitely many samples and estimate your confidence interval exactly the same way each time, your confidence level (for example, 95%) would represent the percentage of those intervals that would contain the true population mean.

Of course, nobody draws infinitely many samples. It's rare that you would draw more than one.

Confidence represents something to do with the method for calculating these intervals, rather than something about any particular sample interval.

An important thing to remember is that the confidence interval calculated for a particular sample might or might not contain the value of the true population mean. Thus, for any given sample, you'll never know whether it does or it doesn't.

**You get to choose your desired degree of confidence, or confidence level.**

It's expressed as the percentage of those theoretically infinite number of intervals that should contain the true population mean.

A typical confidence level is 95%.

(It's most often linked to a value important for hypothesis testing, called the **significance level**. That is described shortly).

Here's the formula for calculating the confidence interval for the mean:



X-bar is the sample mean

t is the t quantile value that is determined by the confidence level and the sample size

s sub x-bar is the standard error of the mean.

The formula calculates both the upper and lower bounds, equidistant from the sample mean estimate in the middle.

To **construct a confidence interval for the mean**,

you must first select a significance level so that the appropriate value of t is used in the formula.

The numerical width of the confidence interval is partially determined by the confidence level.

Higher confidence levels are associated with higher t values, which, in turn, result in wider intervals. The formula calculates both the upper and lower bounds, equidistant from the sample mean estimate in the middle.

To use the speed example, if you were to draw 100 different samples from the same population of cars and calculate the 100 intervals, you'd expect that approximately 95 of them will contain the true population mean, but 5 would not.

(**Foteini Notes**: I choose a confidence level of 95%, then 100 samples, 100 sample means , 100 confidence intervals ± margin of error. So, 95% of these confidence intervals (95) contain the true population mean and the remaining 5 confidence intervals do not)

Here's a question. Why not raise the confidence level to, let's say, 99.9%, so that you could feel more secure that any confidence interval you calculate will contain the true value of the population mean?

As you increase the confidence level, the width of the interval increases, making it less informative. In the extreme, a 100% confidence interval for the mean for any sample ranges from negative infinity to positive infinity, which would tell you nothing useful about where the true population mean lies.

When the confidence level is high then the

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

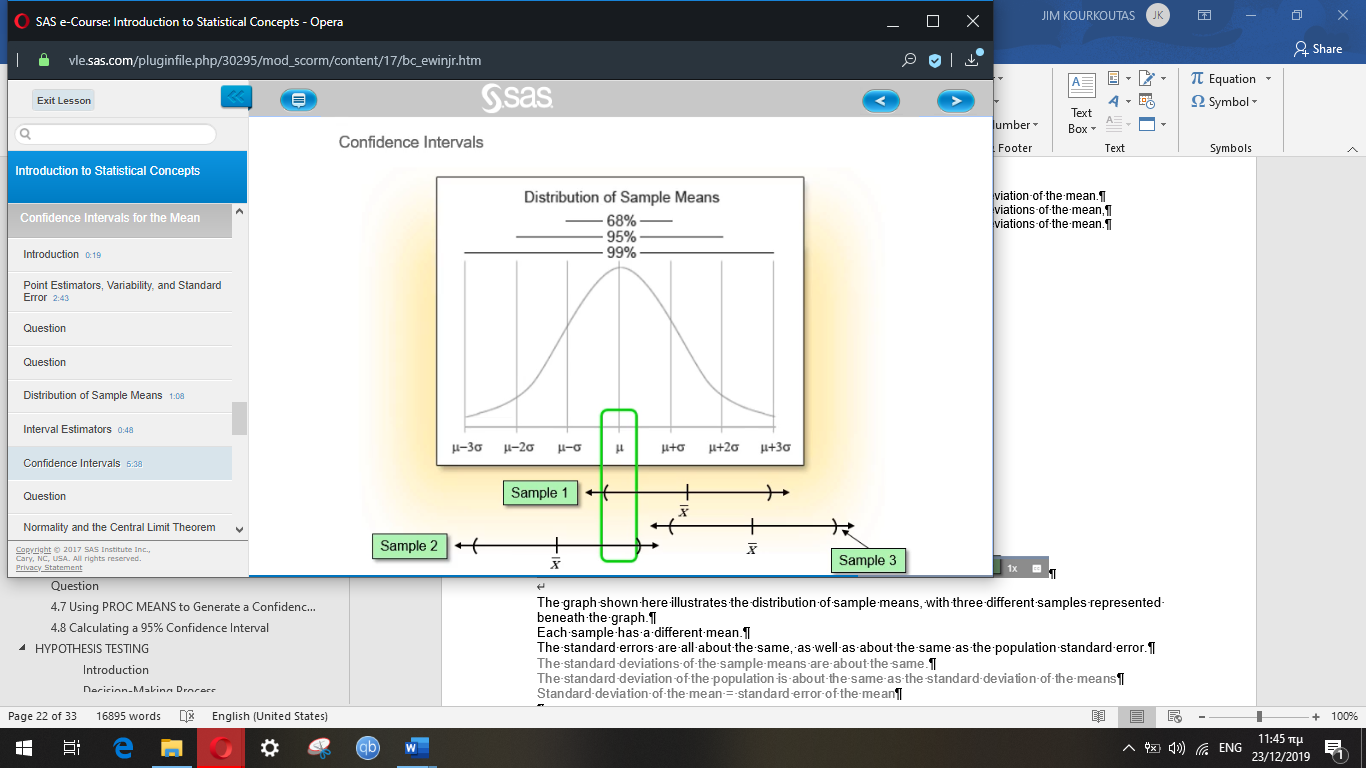
ATTENTION. the confidence intervals in this course assume that the sample means are normally distributed.

**Remember that in a NORMAL DISTRIBUTION:**

approximately 68% of the data values fall within 1 standard deviation of the mean (μ-σ up to μ+σ)

approximately 95% of the data values fall within 2 standard deviations of the mean (μ-2σ up to μ+2σ)

approximately 99% of the data values fall within 3 standard deviations of the mean (μ-3σ up to μ+3σ



The graph shown here illustrates the distribution of sample means, with three different samples represented beneath the graph.

Each sample has a different mean.

The standard errors are all about the same, as well as about the same as the population standard error.

The standard deviations of the sample means are about the same.

The standard deviation of the population is about the same as the standard deviation of the means

Standard deviation of the mean = standard error of the mean

The double-headed arrows around each of the means measure about 2 standard errors to each side of each sample mean. (**t is approximately 2**.)

Sample 1: the mean falls within 2 standard errors from the population mean, by chance

Sample 2: the mean falls within 2 standard errors from the population mean, by chance

In fact, you expect 95% of all sample means to fall within 2 standard errors of the population mean.

Sample 3: the mean falls outside of 2 standard errors

Sample 3 was among the 5% whose mean, by chance alone, fell outside of 2 standard errors.

Even though this sample was collected in the same manner as the others, because the sample mean was more than 2 standard errors from the population mean, the confidence interval did not extend far enough to include the true mean.

Degree of confidence = confidence level

**Foteini Notes**

The confidence interval is used to express the precision and uncertainty associated with a particular sampling method.

A **CONFIDENCE INTERVAL** consists of three parts.

1. **A confidence level**

The confidence level has to do with the sampling method.

The confidence level describes the **uncertainty** of a sampling method.

The confidence level is the probability part of a confidence interval

The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

The confidence level is not affected by the margin of error

Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample.

Some confidence intervals would include the true population parameter; others would not.

A 95% confidence level means that 95% of the intervals contain the true population parameter

a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

We get to choose the confidence level

1. **A statistic.**

The statistic and the margin of error define an interval estimate that describes the **precision** of the method.



1. **A margin of error.**

The statistic and the margin of error define an interval estimate that describes the **precision** of the method.

The margin of error is the range of values above and below the sample statistic.

Margin of error = 

Higher confidence levels result in wider intervals

The two margins of error (negative and positive are equidistant from the statistic)

Example:

A newspaper conducts a survey.

It announces that a candidate will receive 30% of the vote, with a margin of error of 5% and a confidence level of 95%.

That means:

We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Attention! Many surveys report the statistic (30%) and the margin of error (± 5%) but do not report the confidence level. But this is wrong because we cannot interpret survey results we need the margin of error and the confidence level.

And we are more likely to accept a high confidence level than a low one (e.g. we prefer 95% rather than 50%)

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

The interval estimate of a confidence interval is defined by the sample statistic + margin of error.

Interval estimate = sample statistic, **+** margin of error

Confidence interval = sample statistic, **+** margin of error, and of course the confidence level

Example

For example, suppose we compute an interval estimate of a population parameter.

We might describe this interval estimate as a 95% confidence interval.

Definition of the 95% confidence interval

This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the sample statistic **+** margin of error, 95% of the time.

Confidence intervals are preferred to point estimates\*, because confidence intervals indicate

(a) the precision of the estimate and

(b) the uncertainty of the estimate.

*\* the confidence interval is a more “advanced” interval estimate*

*The interval estimate is better than the point estimate (see above), and the confidence interval is better than the point estimate*

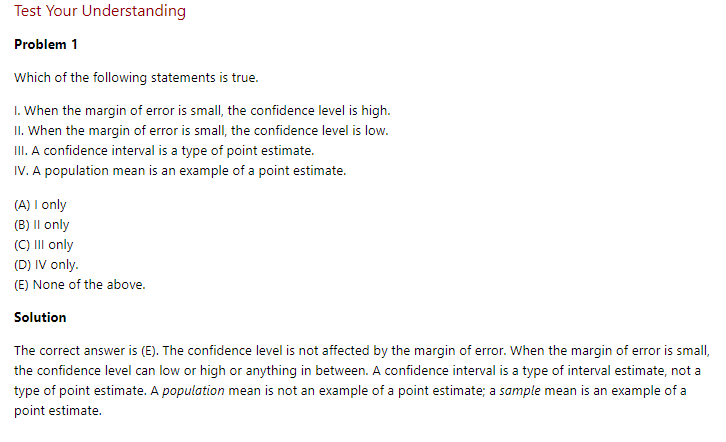
|  |  |
| --- | --- |
| INTERVAL ESTIMATOR | CONFIDENCE INTERVAL  (it is an interval estimator too, but also incorporates the confidence level) |
| Statistic | Confidence level (we choose this first) |
| Margin of error | Statistic |
|  | Margin of error |



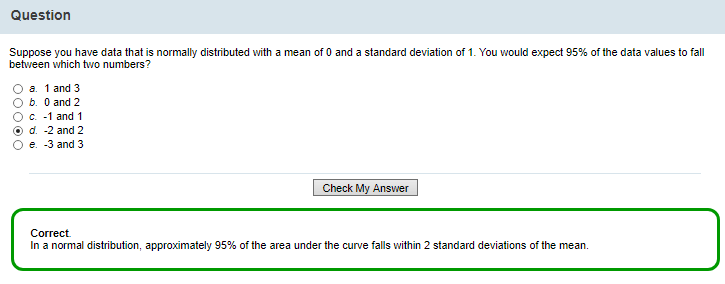
Foteini Notes:

Διαστήματα Εμπιστοσύνης

Είναι ο κλάδος της Στατιστικής Συμπερασματολογίας που ασχολείται με τον προσδιορισμό διαστημάτων, περιοχών γενικότερα, που περιέχουν με μεγάλη πιθανότητα άγνωστες παραμέτρους κάποιας κατανομής



## Question



In a normal distribution,

68% of data values lie within μ-σ, μ+σ close to the true mean 1 standard deviation

95% of data values lie within μ-2σ, μ+2σ 2 standard deviation

99% data values lie within μ-3σ, μ+3σ 3 standard deviations

## 4.6 Normality and the Central Limit Theorem

What if our data doesn't have a normal distribution?

We have 2 options:

1) We can either verify that the population distribution is approximately normal or

2) apply the central limit theorem.

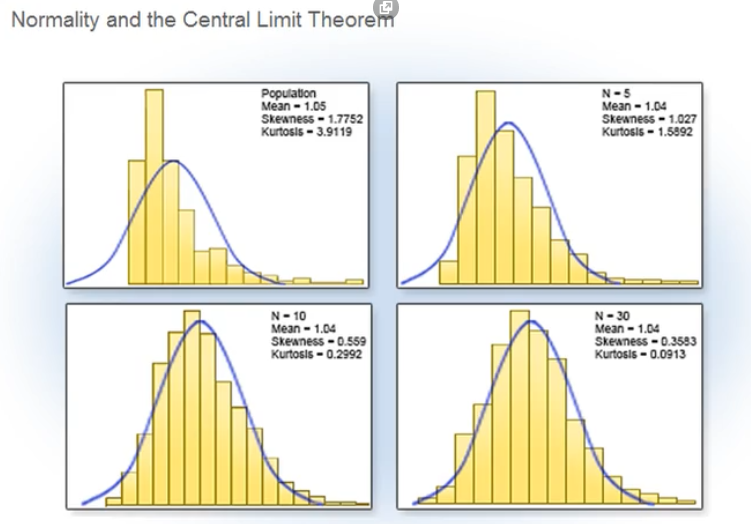
The central limit theorem states that **the distribution of sample means is approximately normal, REGARDLESS of the population distribution's shape**( if the sample size is large enough).

"Large enough" is usually about 30 observations.

It's more if the data is heavily skewed, fewer if the data is symmetric.

How does this work?

Let's see how a distribution of sample means approaches normality as the sample size increases.



Here's a histogram of data values drawn from an exponential distribution, which is a right-skewed population.

The remaining charts display histograms of the sample means for samples of differing sizes drawn from the same exponential distribution.

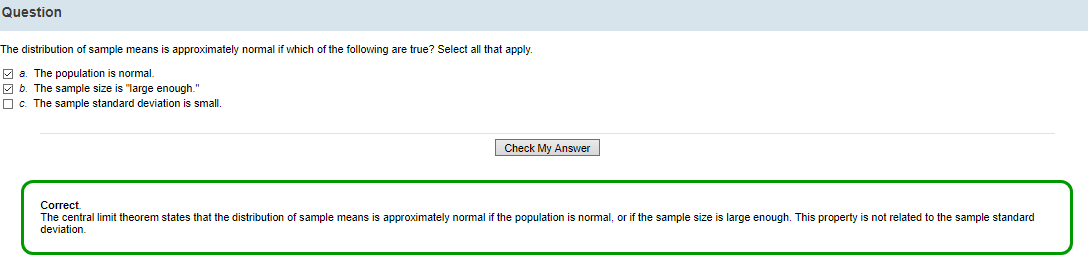
**The second chart** displays the distribution of sample means from 1000 samples, each of size 5, from the same right-skewed population. n=5.000

**The third chart** displays the distribution of sample means from 1000 samples, each of size 10, from the same right-skewed population. Notice that the sample is moving closer to normal. n=10.000

**The fourth chart** displays the distribution of sample means from 1000 samples, each of size 30, from the same right-skewed population. This distribution is approximately bell-shaped and symmetric. N=30.000

You can see that when the sample size is large enough, it doesn't matter what your original population looked like. The distribution of sample means is close enough to normal.

## Question



The central limit theorem states that:

The distribution of the sample mean is normal if

a) the population is normal

or

b) the sample size is large enough

Normality is not related to the sample standard deviation.

**Foteini Notes:**

The central limit theorem, links the normal distribution to any other distribution even if it is not a normal one.

That is the reason for which the normal distribution principles are applied to so many experiments and phenomenons.

To prove that the values of a variable are following the normal distribution, we have to do measurements that prove it.

Random errors (not systematic errors) that appear in various measurements, follow the normal distribution to a great extent.

Standard normal distribution has a mean of 0 and standard deviation of 1. A random variable that follow the standardized normal distribution is symbolized with a z.

## Scenario: Calculating Confidence Intervals for the Mean

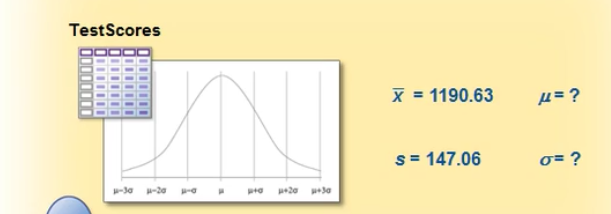
You need to assess whether or not the average combined SAT score in math and reading of magnet high school students in Carver County is 1200, which is the goal set by the school board.

The TestScores data includes information on 80 of these students selected at random.

You already know that this sample data is approximately normal.

You've calculated the sample mean and sample standard deviation.

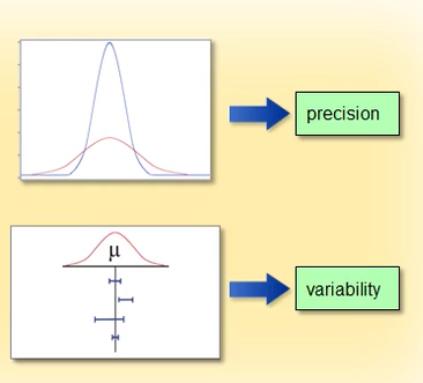
However, you don't know the actual **population mean μ** and **population standard deviation σ**. 😊😊



You need to calculate the **standard error of the mean, s** for the TestScores data so that you can assess how precise your sample estimate is.

(if we have a low standard error, then we have calculated the mean with good precision!)

You also need to calculate the **confidence interval for the mean** so that you can estimate the population mean while taking into account the variability of the sample statistic.



* by calculating STANDARD ERROR, we see if we have good PRECISION when calculating the sample mean
* by calculating the CONFIDENCE INTERVAL, we see the VARIABILITY of the sample statistic

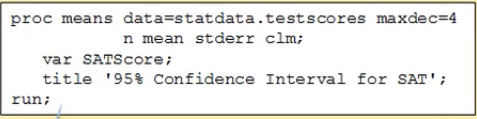
## 4.7 Using PROC MEANS to Generate a Confidence Interval for the Mean

You can use the PROC MEANS procedure to generate a 95% confidence interval for the mean.

You're already familiar with the syntax for PROC MEANS.

You can use the CLM option in the PROC MEANS statement to calculate the confidence limits for the mean.

Here's the code that you can use to calculate a 95% confidence interval for the mean of the variable SATScore in the TestScores data.



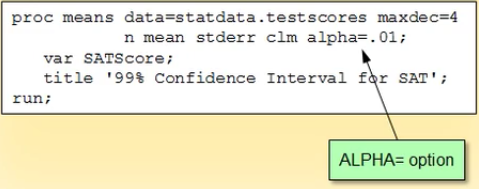
You want to display

* the number of observations,
* the mean of the data,
* the standard error of the mean, and
* the confidence interval for the mean.

The CLM option calculates the confidence limits for the mean.

SAS defaults to a 0.05 α, or 95% confidence level.

If you want to construct confidence intervals with a different confidence level, you can add the ALPHA= option to the PROC MEANS statement. For example, here the ALPHA= option specifies a 99% confidence level.



Foteini Notes

Confidence limits = either of the extreme values of a confidence interval.

## 4.8 Calculating a 95% Confidence Interval

Let's use PROC MEANS to calculate a 95% confidence interval for the mean of the variable SATScore in the TestScores data set.

The MAXDEC= option rounds the values in the table to four decimal places.

We specify that SAS should display the number of observations, the mean, the standard error of the mean, and the confidence limits of the mean at the default 95% confidence level.

proc means data=statdata.testscores maxdec=4 n mean stderr clm;

var SATScore;

title '95% Confidence Interval for SAT';

run;

title;

Let's look at the output.

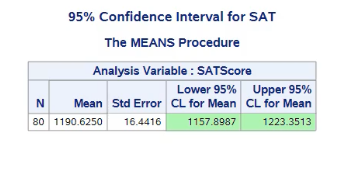
The sample size is 80.

The mean of SATScore is 1190.625.

The standard error of the mean is 16.4416.

The standard error measures the variability of the sample mean or **how much variability we expect in the distribution of all means of samples of size n**. (n=80 in this example.)

The 95% confidence interval of the mean is 1157.9 to 1223.4.



How does this relate to your original question?

You want to know whether the average SAT score for the Carver County magnet high schools is different from the standard of 1200 set by the school board.

Even though the sample mean of 1190.6 is not exactly 1200, the confidence interval of 1157.9 to 1223.4 contains the value 1200.

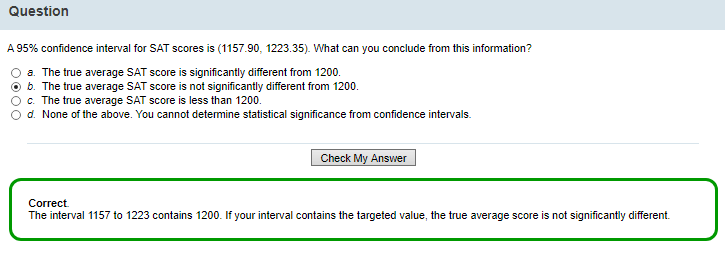
This is evidence that the value of 1200 is at least a reasonable candidate for the true population mean.

Hence, you can conclude that 1190.6 is a reasonably likely sample mean for a random sample drawn from a population centered around 1200.

Of course, this logic applies only if you can assume that the population of means is normally distributed, and the standard deviation of the sample is a good estimate of the standard deviation of the population.

This concept is discussed further in the section on hypothesis testing.

## Question



A 95% confidence interval for SAT scores is (1157.90, 1223.35). What can you conclude from this information?

That the true average SAT score is not significantly different from 1200.

**That is Correct because:** The interval 1157 to 1223 contains 1200. If your interval contains the targeted value, the true average score is not significantly different.

## Practice

Practice: Using PROC MEANS to Generate a 95% Confidence Interval for the Mean

Task

You need to generate a 95% confidence interval for the mean of the variable **BodyTemp** in the **[Statdata.NormTemp](javascript:%20openOther('eclibjr/NormTemp.htm');)** data set.

**Reminder**: Make sure you've [defined the **Statdata** library](javascript:%20openOther('eclibjr/statdata_lib.htm')).

1. Use PROC MEANS to generate a 95% confidence interval for the mean of **BodyTemp** in the **NormTemp** data set. Is the assumption of normality met to produce a confidence interval for these data?  
   yes the sample is over 30 observations
2. What is the confidence interval?  
   98.1220 to 98.3765
3. How do you interpret this interval with regards to the true population mean for body temperature?

It contains the true population mean

You are 95% confident that the true mean body temperature for the population of all people in the world is somewhere between 98.12 and 98.38 degrees.

# 5 INFERENTIAL STATISTICS, Hypothesis Testing

## 5.1 Introduction

Now you're ready to learn a key technique of inferential statistics: hypothesis testing.

A hypothesis test uses sample data to evaluate a question about a population.

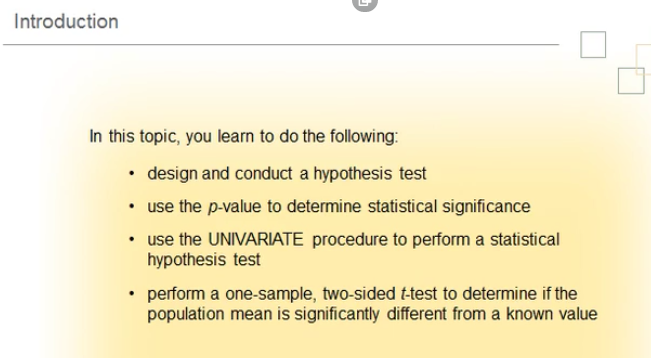
It provides us with a way to make inferences about a population based on sample data.

We don't usually know the true value of population parameters.

We estimate these values based on statistics from a sample.

In hypothesis testing, we reject OR fail to reject a statistical hypothesis that is a statement about the value of a population parameter.

Here's what you learn in this topic.



## 5.2 Decision-Making Process

To help you understand how to perform a hypothesis test, it's helpful to look at a decision-making process that's similar.

Statistical hypothesis testing is like a court trial. In a criminal court, you put defendants on trial because you suspect that they're guilty of a crime. But how does the trial proceed?

1) First you must determine your hypotheses, referred to in statistics as the null and alternative hypotheses.

* The NULL HYPOTHESIS is what you assume to be true when you start your analysis. So, in this case, the null hypothesis is that the defendant is not guilty.
* The ALTERNATIVE HYPOTHESIS is your initial research hypothesis, that is, your proposed explanation. In this case, you propose that the defendant is guilty.

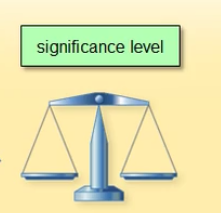


2) Then you select the significance level.

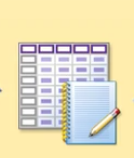
In this example, this is the amount of evidence needed to convict the defendant.

In a criminal court of law, the evidence must prove guilt "beyond a reasonable doubt."

In a civil court, there must be a preponderance of evidence. Preponderance=υπεροχή, excess, superiority



3) Next, you gather your data. In this case, the police and investigators collect evidence. Lawyers present the evidence to the jury.



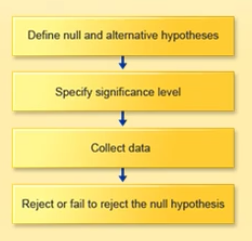
4) Finally, the jury makes a judgment using a decision rule.

* If the evidence is sufficiently strong, the jury rejects the null hypothesis that the defendant is not guilty.
* If the evidence is not strong enough, the jury fails to reject the null hypothesis that the defendant is not guilty.



Note that failing to prove guilt does not prove that the defendant is innocent.

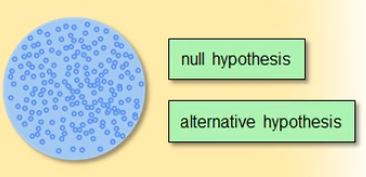
Statistical hypothesis testing follows this same basic path.



## 5.3 Steps in Hypothesis Testing

As in a court trial, there are four steps in conducting a hypothesis test.

1) The first step is to identify the population of interest and determine the null and alternative hypotheses.



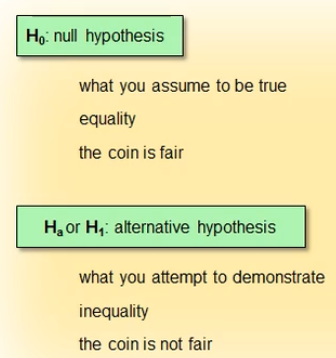
Let's look at an example.

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

The null hypothesis, which statisticians designate H0, is what you assume to be true, unless proven otherwise. The null hypothesis is usually a hypothesis of equality. Your null hypothesis in this case is that the coin is fair.

Suppose you suspect that the coin is not fair.

This is the alternative hypothesis; which statisticians designate Ha or H1. The alternative hypothesis is typically what you suspect, or are attempting to demonstrate. The alternative hypothesis is usually a hypothesis of inequality.



2) The second step in hypothesis testing is to select the significance level.

This is the amount of evidence needed to reject the null hypothesis.

A common significance level is 0.05 (1 chance in 20). This is what you'll use in this course.

*If you require a stricter cutoff, you might consider lowering your significance level when planning your own analysis.*

For this example

* if you observe five heads or five tails in a row (1 chance in 16\*, ), you conclude the coin is not fair and you reject the null hypothesis (since we require 1 chance in 20, 0.05!)
* If you get a mix of heads and tails, you decide there isn't enough evidence to determine that the coin is not fair. To understand why the probability is 1 out of 16, click the information button.

**Why is the probability 1 out of 16?**

There are 5 tosses, each has probability 1/2 of being a head.

Tosses are independent, therefore the probability of 5 heads is (1/2)5 (or 1 out of 32 or 1/32).

Similarly, the probability of 5 tails is 1 out of 32.

These probabilities can be added to give the probability of 5 heads or 5 tails as 2 out of 32 or 1 out of 16.

**Γιατί είναι η πιθανότητα 1 στις 16?**

Έχουμε 5 ρίψεις του νομίσματος.

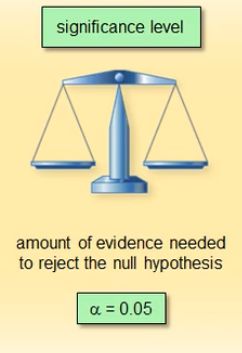
Σε κάθε ρίψη, η πιθανότητα να εμφανιστεί η κορώνα είναι ½.

Οι ρίψεις είναι ανεξάρτητες, άρα η πιθανότητα για εμφάνιση κορώνας 5 φορές είναι (½ )5

(δηλαδή ½ x ½ x ½ x ½ x ½ ). Η πιθανότητα είναι 1/32 ή 1 στις 32.

Ομοίως, η πιθανότητα εμφάνισης για γράμματα 5 φορές είναι 1/32 ή 1 στις 32.

Αυτές πιθανότητες μπορούν να προστεθούν για να δώσουν την πιθανότητα 5 κορωνών στη σειρά (ή 5 γραμμάτων) ως 2/32 ή 1/16.



3) Your third step is to collect the data.

In this case, you flip the coin five times and record each outcome as a head or tail.

You compute the sample statistic by counting the total number of heads.

4) The fourth step in hypothesis testing is to evaluate the data using your decision rule.

You decide whether or not there is enough evidence to reject the null hypothesis.

In this example, you conclude whether the coin is fair.

## 5.4 Types of Errors, Power after we completed the Hypothesis Testing

You perform a hypothesis test and make a decision.

But is that decision correct?

You start by assuming that the null hypothesis is true. In the coin example, you start by assuming that the coin is fair. However, there's always a risk that you're wrong.

* If you reject the null hypothesis when it's actually true, you've made a Type I error.

The probability of committing a Type I error is α.

α is the significance level of a test.

In the coin example, it's the probability that you conclude that the coin is not fair when it is fair.

* If you fail to reject the null hypothesis and it's actually false, you've made a Type II error.

The probability of committing a Type II error is β.

In the coin example, it's the probability that you fail to find that the coin is not fair when it is not fair.

Type I and II errors are inversely related. As one type increases, the other decreases.

Power is the probability that you correctly reject the null hypothesis. It is the ability of the statistical test to detect a true difference, or the ability to successfully reject a false null hypothesis.

The power of a statistical test is equal to (1 – β), where β is the Type II error rate.

Errors

|  |  |  |
| --- | --- | --- |
| **Type I Error**  The null hypothesis is true | We reject it | The probability of committing Type I error is **α** |
| **Type II Error**  The alternative hypothesis is true | We reject it | The probability of committing Type II error is **β** |
| **Power of a statistical test**  It is the ability of the statistical test to detect a true difference, or the ability to successfully reject a false null hypothesis. | ability to successfully reject a false null hypothesis. | **(1-β)** |
|  |  |  |

Type I error and Type II error are inversely related (when Type I increases then Type II decreases).

Για να ελέγξουμε μια υπόθεση την οποία συνήθως ονομάζουμε

μηδενική υπόθεση (null hypothesis) και την οποία συμβολίζουμε με

το Η0, χρειαζόμαστε μια εναλλακτική υπόθεση (alternative hypothesis)

σε αντιπαράθεση προς την οποία ελέγχεται η Η0. Η εναλλακτική

υπόθεση συμβολίζεται συνήθως με Η1 ή με ΗA.

σφάλμα τύπου Ι: απόρριψη της H0 ενώ ισχύει η H0, πιθανότητα α

σφάλμα τύπου ΙI: αποδοχή της H0 ενώ ισχύει η H1, πιθανότητα β

Power (1-β)

Το συμπλήρωμα της ποσότητας β, η πιθανότητα δηλαδή να απορριφθεί η μηδενική υπόθεση όταν πράγματι η μηδενική υπόθεση δεν ισχύει, είναι ένα μέτρο της ικανότητας του ελέγχου που έχουμε επιλέξει να λειτουργεί σωστά. Η συνάρτηση αυτή ονομάζεται ισχύς (power) του στατιστικού ελέγχου.

To a (επίπεδο σηµαντικότητας) είναι προεπιλεγµένο, συνήθως α = 0.05 ή α = 0.01

Η τιμή της πιθανότητας α ονομάζεται επίπεδο

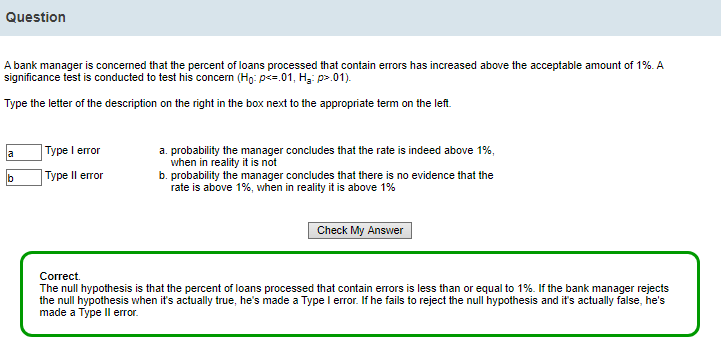
σημαντικότητας (level of significance).

Η τιμή αυτή δηλώνει τη μέγιστη πιθανότητα που ο ερευνητής

επιτρέπει στον εαυτό του να κάνει λάθος τύπου I. Συνήθως, η τιμή

του α επιλέγεται από τον ίδιο τον ερευνητή.

## Question



Null hypothesis Ho: p<0.01 we consider this as TRUE.

we would like to test this and prove it is wrong

Alternative hypothesis Ha: p>0.01 we consider this as FALSE.

we suspect this is in reality true

The null hypothesis is that the percent of loans processed that contain errors is less than or equal to 1%. If the bank manager rejects the null hypothesis when it's actually true, he's made a Type I error. If he fails to reject the null hypothesis and it's actually false, he's made a Type II error.

## 5.5 The *p*-Value, Effect Size, and Sample Size

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.

You'd be more skeptical that the coin is fair if you observe 37 or 63 heads and highly skeptical that the coin is fair if you observe 15 or 85 heads.

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

Statisticians refer to the difference between the observed statistic and the hypothesized value as the EFFECT SIZE.

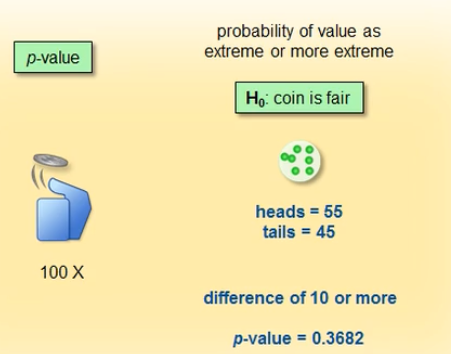
A **p-value** measures the probability of observing a value as extreme as the one observed or more extreme, assuming that the null hypothesis is true.

For example, if your null hypothesis is that the coin is fair

but you observe 55 heads and 45 tails,

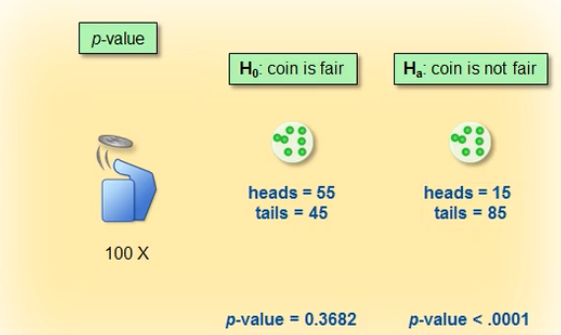
the p-value is the probability of observing a difference in the number of heads and tails of 10 or more from a fair coin tossed 100 times.

(The p-value makes the null hypothesis we assumed as true, look “ridiculous” 😊😊)



When the p-value is large, you would often see a difference this large in experiments with a fair coin.

When the p-value is small, however, you would rarely see differences this large with a fair coin.



In the latter situation (the p-value is small), you have evidence that the coin is not fair.

* A large p-value indicates a high probability of observing your results, or more extreme results, (often see a difference this large in experiments with a fair coin) given that the null hypothesis is true.

So, it is reasonable to assume that the null hypothesis is true.

* A small p-value indicates a low probability of observing your results, or more extreme results, (rarely see differences this large with a fair coin) given that the null hypothesis is true.

So, it is no longer reasonable to assume that the null hypothesis is true.

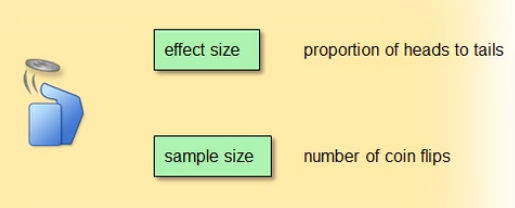
α, significance level

p-value, observed significance level

So, the p-value is used to determine statistical significance.

It helps you assess whether you should reject the null hypothesis.

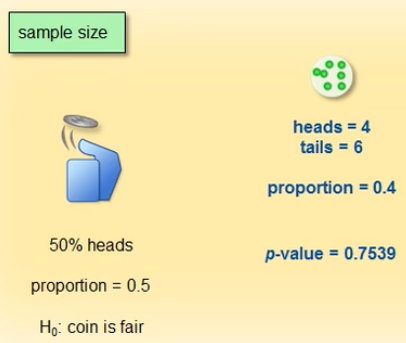
A p-value is not only affected by the EFFECT SIZE (in this example, the observed proportion of heads). It is also affected by the SAMPLE SIZE (here, the number of coin flips).



For a fair coin, you'd expect 50% of the flips to turn up heads.

Let's say that you flip the coin 10 times and observe 4 heads.

In this example, the proportion of heads is 0.4. This value is different from the 0.5 you'd expect according to the null hypothesis.



The evidence becomes stronger as the number of trials on which the proportion is based becomes greater.

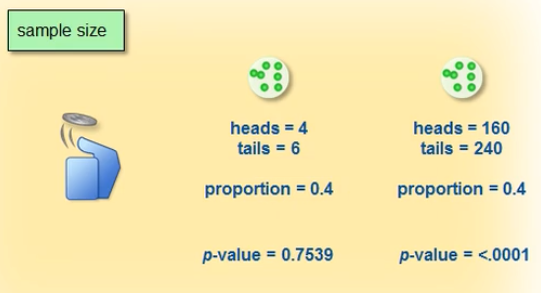
As you saw in the section on confidence intervals, the variability around a mean estimate, gets smaller as the sample size gets larger.

For larger sample sizes, you can measure means more precisely.

Therefore, 40% heads out of 400 flips makes you more sure that this was not just a chance difference from 50% than 40% out of 10 flips.

The smaller p-values reflect this confidence.

The p-value here (p<0.001) assesses how likely this difference is from 50%.



Proportion= difference between number of heads and number of tails or else heads/tails

Ορισμός: Ορίζουμε ως παρατηρούμενο επίπεδο σημαντικότητας ή τιμή πιθανότητας ή p-τιμή (observed level of significance ή probability value ή p-value)

την πιθανότητα ή στατιστική συνάρτηση ελέγχου να πάρει μία τιμή τόσο ακραία ή περισσότερο ακραία από αυτήν που πήρε για το συγκεκριμένο δείγμα, κάτω από την μηδενική υπόθεση.

Σημείωση: Η χρησιμοποίηση της p-τιμής αντί του επιπέδου σημαντικότητας α στην αντιμετώπιση ενός προβλήματος στατιστικού ελέγχου υποθέσεως δεν μεταβάλλει την κλασσική στατιστική μεθοδολογία που αναλύθηκε νωρίτερα.

Απλά εκείνο που συμβαίνει είναι ότι ο ερευνητής αναφέρει στον ενδιαφερόμενο την p-τιμή (το παρατηρούμενο επίπεδο σημαντικότητας) και αφήνει την επιλογή του κατά πόσο θα πρέπει να απορριφθεί ή όχι η μηδενική υπόθεση στον ενδιαφερόμενο.

Μεταφέρεται δηλαδή η ευθύνη επιλογής του α από τον ερευνητή στον ενδιαφερόμενο.

https://towardsdatascience.com/p-values-explained-by-data-scientist-f40a746cfc8

Hypothesis testing is used to test the validity of a claim (null hypothesis) that is made about a population using sample data. The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue.

In other words, we’ll make a claim (null hypothesis) and use a sample data to check if the claim is valid. If the claim isn’t valid, then we’ll choose our alternative hypothesis instead. Simple as that.

Our ultimate goal is to determine the statistical significance of our results.

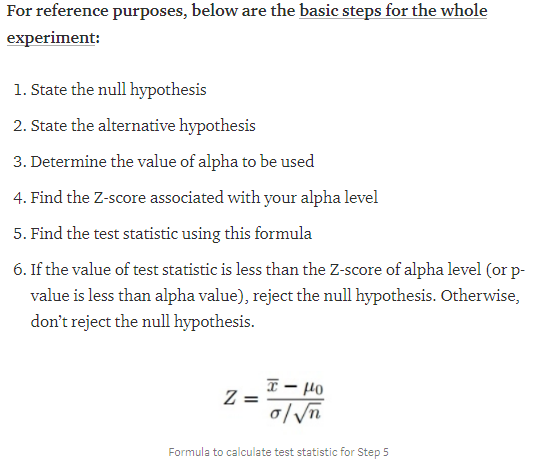
And statistical significance is built on these 3 simple ideas:

* Hypothesis testing
* Normal distribution
* P-values

To know if a claim is valid or not, we’ll use a **p-value** to weigh the strength of the evidence to see if it’s statistically significant. If the evidence supports the alternative hypothesis, then we’ll reject the null hypothesis and accept the alternative hypothesis.

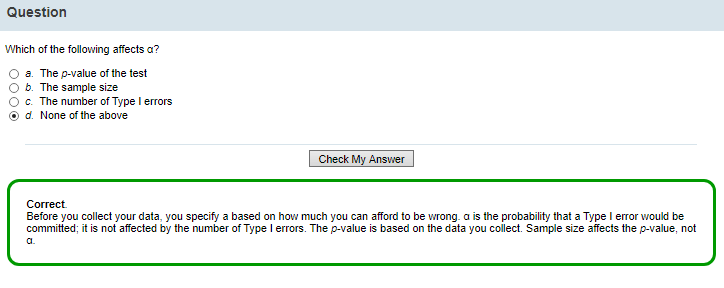
**If the p-value is lower than a predetermined significance level α, then we reject the null hypothesis**.

p-values are used as a tool to challenge our initial belief (null hypothesis) when the result is statistically significant.



Practical Statistics for Data Scientists: 50 Essential Concepts 1st Edition

## Question



We decide the level of significance α, we decide how much we afford to be wrong!

The level of significance is not affected by p-value, sample size, number of Type I errors

P value is affected by the sample size and the effect size

You measure the strength of the evidence by using a p-value.

You calculate the p-value from the collected data.

• The decision rule is that

p => α then you don’t reject the null hypothesis

p < α You reject the null hypothesis

**A large p value is required for the null hypothesis to be true? Yes when p=>a we fail to reject the null hypothesis**

## 5.6 Summary of Statistical Hypothesis Testing

So, let's summarize what we know about statistical hypothesis testing.

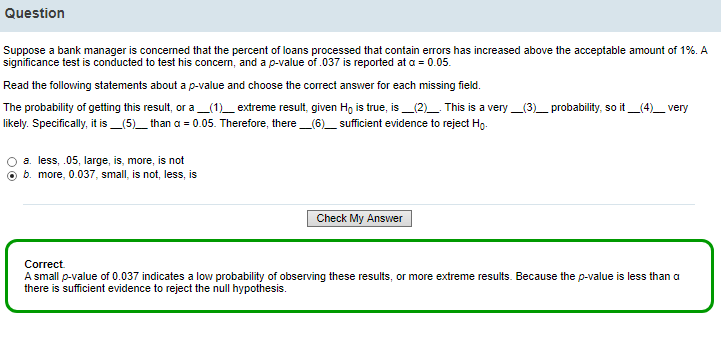
In statistics,

* the null hypothesis is denoted as H0. This is your initial assumption, usually one of equality or no relationship.
* The alternative hypothesis is denoted as Ha or H1. This is what you suspect or are trying to demonstrate, and it is typically an expression of an inequality or a relationship.
* You specify the significance level, which is also the Type I error rate, before collecting the data. It is a function of both your knowledge of the data and theoretical considerations. Statisticians refer to the significance level as α. You choose the level of α based on the cost of making a Type I error.
* You collect the data. You measure the strength of the evidence by using a p-value. You calculate the p-value from the collected data.
* The decision rule is that
* you don’t reject the null hypothesis if the p-value is greater than or equal to α.
* You reject the null hypothesis if the p-value is less than α.

You never conclude that two things are the same or have no relationship.

You can only fail to show a difference or a relationship.

## Question



Ho: p<0.01

Ha: p>0.01

a= 0.05

p-value= 0.037

so, p-value<a. So, we have a low p-value.

“The probability of getting this result or a more extreme result, given Ho is true, is 0.037.

This is a very small probability, so it is not very likely.

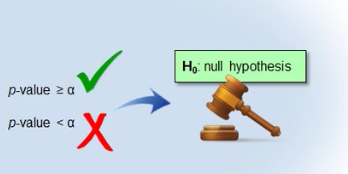
Specifically, it is less than a=0.05.

Therefore, there is sufficient evidence to reject Ho.”

Because the p-value is less than α there is sufficient evidence to reject the null hypothesis.

## 5.7 The *t* Statistic, *t* Distribution, and Two-Sided *t*-Test

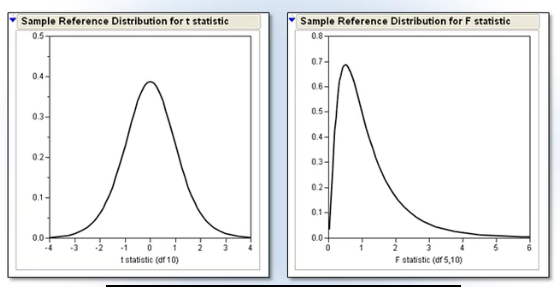
You know that you can compare α with the associated p-value in order to make a decision about the null hypothesis.



Another way to measure statistical significance and test the null hypothesis is to compute a **test statistic.**

Two common reference distributions for statistical hypothesis testing are

* the t distribution and
* the F distribution.



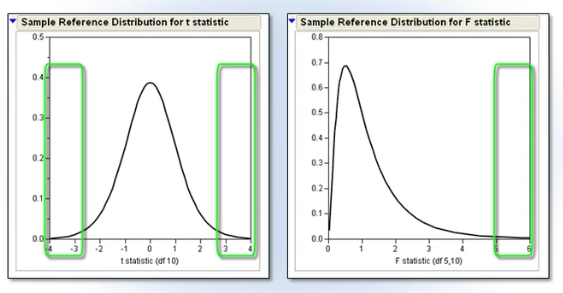
A reference distribution enables you to quantify the probability of observing a particular outcome (*the calculated sample statistic*) or a more extreme outcome, if the null hypothesis is true.

That probability is the p-value you're already familiar with.

The t distribution and F distribution are characterized by the degrees of freedom associated with your data.

Depending on the type of analysis you're performing, you'd use your data to calculate a sample t statistic or a sample F statistic and compare that statistic to the reference distribution.

Values that fall in the tails of the distribution are values that are possible under the null hypothesis, but unlikely.

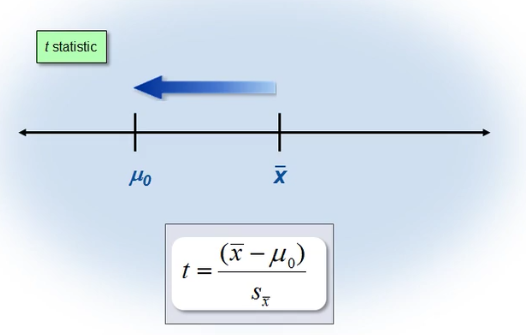


The **t distribution** arises when you're making inferences about a population mean and (as in nearly all practical statistical work) the population standard deviation (and therefore, standard error) is unknown and has to be estimated from the data. It is approximately normal as the sample size grows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

THE **T STATISTIC** MEASURES HOW FAR X-BAR, THE SAMPLE MEAN, IS FROM THE HYPOTHESIZED MEAN, Μ0.

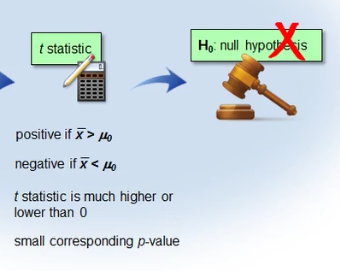
It is the number of standard errors that the sample mean is from the hypothesized mean.

The formula is shown here.



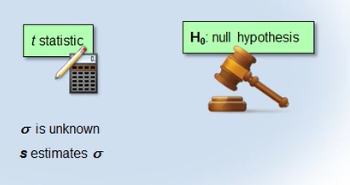
The t statistic is positive when the sample mean is larger than the hypothesized mean, or negative when the sample mean is less than the hypothesized mean.

If the t statistic is much higher or lower than 0 and has a small corresponding p-value, this indicates that the sample mean is quite different from the hypothesized mean, and you would reject the null hypothesis.



When do we use the t statistic?

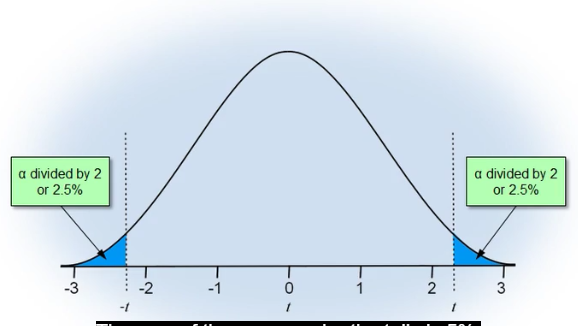
You use the t statistic when you don't know the true population standard deviation, σ, and you have only an estimated standard deviation, s.



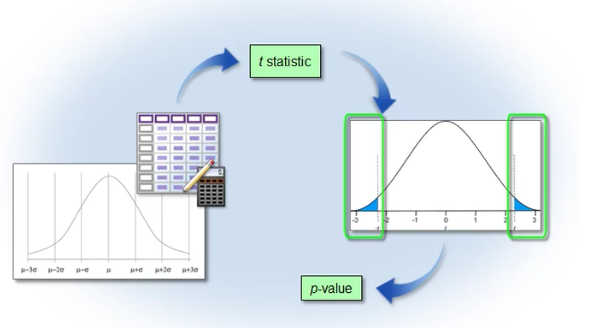
This graph shows the **t critical region** in relation to the t distribution.

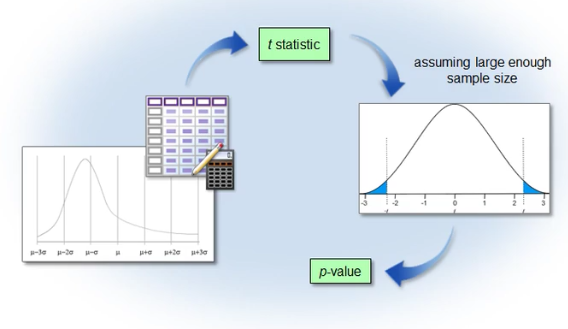
If the t statistic and corresponding p-value falls in the critical region (the shaded region on the graph), then you reject the null hypothesis. Otherwise, you fail to reject the null hypothesis.

The area in each of the tails corresponds to α divided by 2 [α/2], or 2.5%. The sum of the areas under the tails is 5%, which is α. Because the rejection region for the t statistic is contained in both tails of the data distribution, statisticians refer to this type of hypothesis test as a two-sided t-test.



* If the data comes from a normal population, then the t statistic calculated from the data follows a t distribution, and we can use the t distribution to compute p-values. The t distribution is a symmetric distribution like the normal distribution, except that the t distribution has thicker tails than a normal distribution.



* If the data does not come from a normal distribution, then the t statistic approximately follows a t distribution as long as the sample size is large, and we can still use the t distribution to calculate a p-value. This is another application of the central limit theorem
* 

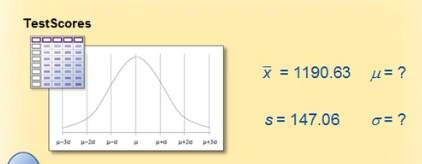
## Scenario: Comparing Sample and Hypothesized Means

You need to assess whether or not the average combined SAT score in math and reading of magnet high school students in Carver County is 1200, which is the goal set by the school board.

The TestScores data includes information on 80 of these students selected at random.

You already know that this sample data is approximately normal. You've calculated the sample mean and sample standard deviation.

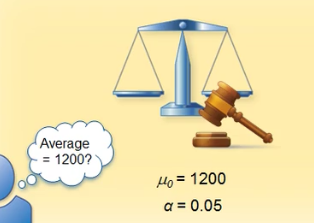
However, you don't know the actual population mean and population standard deviation.



You can perform a statistical hypothesis test to see if the sample mean is different from the standard of 1200 set by the school board.

1200 is your hypothesized mean, μ0.

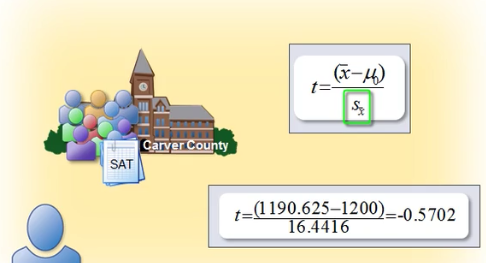
You'll use a significance level of 0.05.



You can calculate the t statistic using the formula you learned earlier.

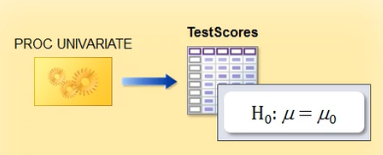
X-bar is the sample mean SAT score of students selected from the school district.

s-sub-x-bar is the standard error of the mean that you calculated earlier in this lesson.

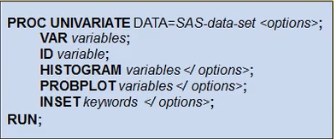


## 5.8 Using the PROC UNIVARIATE Procedure in SAS to Generate a t statistic

You can use PROC UNIVARIATE to test the hypothesis that the mean of the SAT score is equal to 1200.



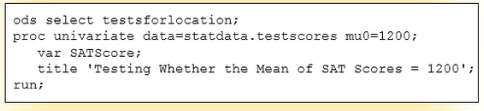
You're already familiar with the syntax for the UNIVARIATE procedure.



By default, PROC UNIVARIATE sets the value of μ0 to zero.

You can use the MU0= option in the PROC UNIVARIATE statement to specify the value of the hypothesized mean.

Here's the code that you can use to test the hypothesis that the mean of the SAT score is equal to 1200.



You use the ODS SELECT statement to specify that SAS provide output only for the TestsForLocation table.

This statement produces three tests for location, one of them being the t-test.

All three tests produce a test statistic for the null hypothesis, indicating that the mean or median is equal to a given value μ0 against the two-sided alternative that the mean or median is not equal to μ0.

By the way, if you don't know which table name to specify for the output you need, you can browse details for the UNIVARIATE procedure in the SAS documentation.

Then in the PROC UNIVARIATE statement, you specify the input data set, TestScores.

You specify the MU0= option to be 1200. This program uses the default α of 0.05. You can change it in your own analysis if you want, by adding the ALPHA= option to the PROC UNIVARIATE statement.

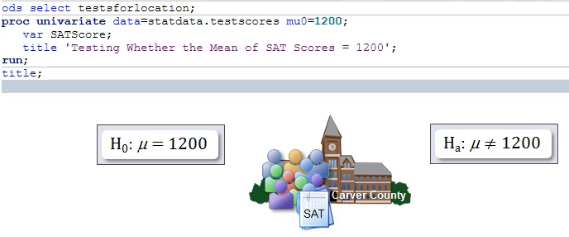
Then in the VAR statement, you identify SATScore as the analysis variable.

## 5.9 Using the PROC UNIVARIATE Procedure in SAS to Perform a Hypothesis Test

This program uses PROC UNIVARIATE to test the hypothesis that the mean of SATScore is equal to 1200.

Your null hypothesis is that the population's mean SAT score for the Carver County magnet high schools is 1200.

Your alternative hypothesis is that the population's mean SAT score is not 1200.



Let's submit this program and take a look at the output SAS produces.

ods select testsforlocation;

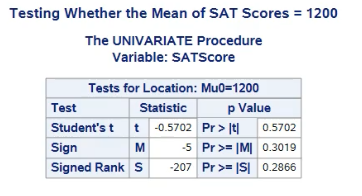
proc univariate data=statdata.testscores mu0=1200;

var SATScore;

title 'Testing Whether the Mean of SAT Scores = 1200';

run;

title;



The Tests for Location table provides the *t* statistic, labeled Student's t, and the corresponding *p*-value.

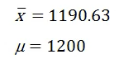
The *p*-value is greater than the significance level, or α, of 0.05 that we had set.

*Note by the way that it is a coincidence that the t statistic and p-value have the same numeric value (although one is positive and the other negative).*

Because the *p*-value is greater than alpha, we fail to reject the null hypothesis.



Therefore, we believe that **there is no statistical difference** between the sample mean of 1190 and the hypothesized mean of 1200.



Here's another way to look at it.

If the null hypothesis is true, how likely are we to see a *t* statistic with an absolute value of .5702 or greater?

Well, about 57% of the time. This value confirms that we do not have enough evidence to reject the null hypothesis.

To summarize: the original question was whether the mean SAT score for Carver County magnet high school students equals 1200.

From this hypothesis test, we conclude that there is not enough evidence to say that the sample mean score of 1190.625 is statistically different from 1200.

Foteini Notes

Στατιστική Σηµαντικότητα

* Όταν το αποτέλεσµα δεν έχει προέλθει από τυχαίους παράγοντες, τότε λέµε ότι είναι στατιστικά σηµαντικό
* Όταν το αποτέλεσµα έχει προέλθει από τυχαίους παράγοντες, τότε λέµε ότι είναι στατιστικά µη σηµαντικό

Τα επίπεδα επίπεδα σηµαντικότητας

Η ποσοστιαία πιθανότητα να έχει εµφανιστεί το αποτέλεσµα από τυχαίους παράγοντες

Έχουµε διάφορα επίπεδα σηµαντικότητας: 0,05 (5%) 0,01 (1%) 0,001 (1‰)

## Practice

Practice: Using PROC Univariate to Perform a One-Sample t-Test

Task

You need to perform a one-sample t-test for the variable BodyTemp in the Statdata.NormTemp data set to confirm whether average body temperature is truly 98.6 degrees.

Task description goes right here. Reminder: Make sure you've defined the Statdata library.

Use PROC UNIVARIATE to perform a one-sample t-test to determine whether the mean of the variable BodyTemp in the data set NormTemp is truly the value 98.6. What is the value of the t statistic and the corresponding p-value?

What is the null hypothesis?

What is the alternative hypothesis?

Using a 0.05 alpha, do you reject or fail to reject the null hypothesis?

# 6 LESSON REVIEW

This summary contains topic summaries, syntax, and sample programs.

**Topic Summaries**

Basic Statistical Concepts

Descriptive statistics organizes, describes, and summarizes data using numbers and graphical techniques. Inferential statistics is concerned with drawing conclusions about a population from the analysis of a random sample drawn from that population. Inferential statistics is also concerned with the precision and reliability of those inferences.

A population is the complete set of observations or the entire group of objects that you are researching. A sample is a subset of the population. The sample should be representative of the population. You can obtain a representative sample by collecting a simple random sample.

Parameters are numerical values that summarize characteristics of a population. Parameter values are typically unknown and are represented by Greek letters. Statistics summarize characteristics of a sample. You use letters from the English alphabet to represent sample statistics. You can measure characteristics of your sample and provide numerical values that summarize those characteristics. You use statistics to estimate parameters.

Variables are characteristics or properties of data that take on different values or amounts. A variable can be independent or dependent. In some contexts, you select the value of an independent variable, in order to determine its relationship to the dependent variable. In other contexts, the independent variable’s values are simply taken as given.

Variables are also classified according to their characteristics. They can be quantitative or categorical. Data that consists of counts or measurements is quantitative. Quantitative data can be further distinguished by two types: discrete and continuous. Discrete data takes on only a finite, or countable, number of values. Continuous data has an infinite number of values and no breaks or jumps.

Categorical or attribute data consists of variables that denote groupings or labels. There are two main types: nominal and ordinal. A nominal categorical variable exhibits no ordering within its groups or categories. With ordinal categorical variables, the observed levels of the variable can be ordered in a meaningful way that implies differences due to magnitude.

A variable’s classification is its scale of measurement. There are two scales of measurement for categorical variables: nominal and ordinal. There are two scales of measurement for continuous variables: interval and ratio. Data from an interval scale can be rank-ordered and has a sensible spacing of observations such that differences between measurements are meaningful. However, interval scales lack the ability to calculate ratios between numbers on the scale because there is no true zero point. Data on a ratio scale includes a true zero point and can therefore accurately indicate the ratio of difference between two spaces on the measurement scale.

The appropriate statistical method for your data also depends on the number of variables involved. Univariate analysis provides techniques for analyzing and describing a single variable at a time. Bivariate analysis describes and explains the relationship between two variables and how they change, or covary, together. Multivariate analysis examines two or more variables at the same time, in order to understand the relationships among them.

Descriptive Statistics

A data's distribution tells you what values your data takes and how often it takes those values.

You can calculate descriptive statistics that measure locations in your data. Statistics that locate the center of the data are measures of central tendency. These include mean, median, and mode.

Percentiles are descriptive statistics that give you reference points in your data. A percentile is the value of a variable below which a certain percentage of observations fall. The most commonly reported percentiles are quartiles, which break the data into quarters.

There are several descriptive statistics that measure the variability of your data: range, interquartile range (IQR), variance, standard deviation, and coefficient of variation (C.V.).

To summarize and generate descriptive statistics, you use the MEANS procedure. PROC MEANS calculates a standard set of statistics, including the minimum, maximum, and mean data values, as well as standard deviation and n. The PRINTALLTYPES option displays statistics for all requested combinations of class variables

Picturing Your Data

A histogram is a visual representation of the frequency distribution of your data. The frequencies are represented by bars.

The normal distribution is a common theoretical distribution in statistics. It is bell-shaped, with values concentrated near the mean, and it is symmetric around the mean. The standard deviation (σ) determines how variable the distribution is. Underlying the normal distribution is a mathematical function named the probability density function.

To check the assumption that your random sample has a normal distribution, you can plot a histogram. You can also look at statistical summaries of your data. The closer skewness and kurtosis are to 0, the closer your data is shaped like the normal distribution.

Skewness measures the tendency of your data to be more spread out on one side of the mean than on the other. It measures the asymmetry of the distribution. The direction of skewness is the direction to which the data is trailing off. The closer the skewness is to 0, the more normal or symmetric the data.

Kurtosis measures the tendency of data to be concentrated toward the center or toward the tails of the distribution. The closer kurtosis is to 0, the closer the tails of the data resemble the tail thickness of the normal distribution. Kurtosis can be difficult to assess visually.

A negative kurtosis statistic means that the data has lighter tails than in a normal distribution and is less heavily concentrated about the mean. This is a platykurtic distribution.

A positive kurtosis statistic means that the data has heavier tails and is more concentrated about the mean than a normal distribution. This is a leptokurtic distribution, which is often referred to as heavy-tailed and also as an outlier-prone distribution.

A normal probability plot is another way to visualize and assess the distribution of your data. The vertical axis represents the actual data values. The horizontal axis displays the expected percentiles from a standard normal distribution. The normal reference line along the diagonal indicates where the data would fall if it were perfectly normal.

A box plot makes it easy to see how spread out the data is and if there are any outliers.

You can use PROC UNIVARIATE to generate descriptive statistics, histograms, and normal probability plots.

In the ID statement, you list the variable or variables that SAS should label in the table of extreme observations and identify as outliers in the graphs.

You can add additional options to the HISTOGRAM and PROBPLOT statements. The NORMAL option uses estimates of the population mean and standard deviation to add a normal curve overlay to the histogram and a diagonal reference line to the normal probability plot.

You can use the INSET statement to create a box of summary statistics directly on the graphs.

In addition to the statistical graphics available to you with PROC UNIVARIATE, you might want to use PROC SGSCATTER, PROC SGPLOT, PROC SGPANEL, and PROC SGRENDER to produce a wide variety of additional plot types.

You can use PROC SGPLOT to generate dot plots, horizontal and vertical bar charts, histograms, box plots, density curves, scatter plots, series plots, band plots, needle plots, and vector plots. The REG statement generates a fitted regression line or curve. You use a REFLINE statement to create a horizontal or vertical reference line on the plot.

ODS Graphics is an extension of the SAS Output Delivery System. With ODS Graphics, statistical procedures produce graphs as automatically as they produce tables, and graphs are integrated with tables in the ODS output. You can find a list of the graphs available for each SAS procedure in the SAS documentation.

Confidence Intervals for the Mean

A point estimator is a sample statistic used to estimate a population parameter. A statistic that measures the variability of your estimator is the standard error.

The standard error of the mean measures the variability of your sample mean. It’s an estimate of how much you can expect the sample mean to vary from sample to sample.

The distribution of sample means is the distribution of all possible sample means from the population. The distribution of the mean is always less variable than the data.

An interval estimator is another way to estimate a population parameter. It incorporates the uncertainty that arises from random variability.

Confidence intervals are a type of interval estimator used to estimate the population mean, while taking into account the variability of the sample mean.

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.

You can use the MEANS procedure to generate a 95% confidence interval for the mean.

You can use the CLM option in the PROC MEANS statement to calculate the confidence limits for the mean.

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

Hypothesis Testing

A hypothesis test uses sample data to evaluate a question about a population. It provides a way to make inferences about a population, based on sample data.

There are four steps in conducting a hypothesis test. The first step is to identify the population of interest and determine the null and alternative hypotheses. The null hypothesis, H0, is what you assume to be true, unless proven otherwise. It is usually a hypothesis of equality. The alternative hypothesis, Ha or H1, is typically what you suspect, or are attempting to demonstrate. It is usually a hypothesis of inequality.

The second step in hypothesis testing is to select the significance level. This is the amount of evidence needed to reject the null hypothesis. A common significance level is 0.05 (1 chance in 20).

The third step is to collect the data. The fourth step is to use a decision rule to evaluate the data. You decide whether or not there is enough evidence to reject the null hypothesis.

If you reject the null hypothesis when it's actually true, you've made a Type I error. The probability of committing a Type I error is α. α is the significance level of a test. If you fail to reject the null hypothesis and it's actually false, you've made a Type II error. The probability of committing a Type II error is β. Type I and II errors are inversely related.The power of a statistical test is equal to 1 minus beta (1 – β),

The difference between the observed statistic and the hypothesized value is the effect size. A p-value measures the probability of observing a value as extreme as the one observed or more extreme. A p-value is not only affected by the effect size, but also by the sample size.

The t statistic measures how far X-bar, the sample mean, is from the hypothesized mean, μ0. If the t statistic is much higher or lower than 0 and has a small corresponding p-value, this indicates that the sample mean is quite different from the hypothesized mean, and you would reject the null hypothesis.

You can use PROC UNIVARIATE to perform a statistical hypothesis test. You use the MU0= option to specify the value of the hypothesized mean, μ0. You can use the ALPHA= option to change the significance level.

Syntax

To go to the movie where you learned a statement or option, select a link.

PROC MEANS DATA=SAS-data-set <options>;

CLASS variables;

VAR variables;

RUN;

PROC UNIVARIATE DATA=SAS-data-set <options>;

VAR variables;

ID variables;

HISTOGRAM variables </options>;

PROBPPLOT variables </options>;

INSET keywords </options>;

RUN;

PROC SGPLOT DATA=SAS-data-set<options>;

DOT category-variable </option(s)>;

HBAR category-variable </option(s)>;

VBAR category-variable </option(s)>;

HBOX response-variable </option(s)>;

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

HISTOGRAM response-variable </option(s)>;

SCATTER X=variable Y=variable </option(s)>;

NEEDLE X=variable Y=numeric-variable </option(s)>;

REG X=numeric-variable Y=numeric-variable </option(s)>;

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

RUN;

ODS GRAPHICS ON <options>;

statistical procedure code

ODS GRAPHICS OFF;

Sample Programs

Using PROC MEANS to Generate Descriptive Statistics

proc means data=statdata.testscores maxdec=2 fw=10 printalltypes

n mean median std var q1 q3;

class Gender;

var SATScore;

title 'Selected Descriptive Statistics for SAT Scores';

run;

title;

Using SAS to Picture Your Data

proc univariate data=statdata.testscores;

var SATScore;

id idnumber;

histogram SATScore / normal(mu=est sigma=est);

inset skewness kurtosis / position=ne;

probplot SATScore / normal(mu=est sigma=est);

inset skewness kurtosis;

title 'Descriptive Statistics Using PROC UNIVARIATE';

run;

title;

proc sgplot data=statdata.testscores;

refline 1200 / axis=y lineattrs=(color=blue);

vbox SATScore / datalabel=IDNumber;

format IDNumber 8.;

title "Box Plots of SAT Scores";

run;

title;

Calculating a 95% Confidence Interval

proc means data=statdata.testscores maxdec=4

n mean stderr clm;

var SATScore;

title '95% Confidence Interval for SAT';

run;

title;

Using PROC UNIVARIATE to Perform a Hypothesis Test

ods select testsforlocation;

proc univariate data=statdata.testscores mu0=1200;

var SATScore;

title 'Testing Whether the Mean of SAT Scores = 1200';

run;

title;

For an asymmetric (or skewed) distribution, which of the following statistics is a good measure for the middle of the data?

|  |  |  |
| --- | --- | --- |
|  | *a.* | mean |
|  | *b.* | median |
|  | *c.* | either mean or median |

2

Which of the following code examples correctly calculates descriptive statistics of popcorn yield (**Yield**) for each level of the class variable (**Type**) in the data set **Statdata.Popcorn**, as well as statistics for all levels combined?

The output should include the following statistics: sample size, mean, median, standard deviation, variance, range, and interquartile range.

|  |  |  |
| --- | --- | --- |
|  | *a.* | proc means data=statdata.popcorn  maxdec=2 fw=10  n mean median std var  range qrange;  class Type;  var Yield;  run; |
|  | *b.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std var  range qrange;  class Yield;  var Class;  run; |
|  | *c.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std var  range qrange;  class Type;  var Yield;  run; |
|  | *d.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std range  IQR;  class Type;  var Yield;  run; |

3

Read the following statement about the central limit theorem and choose the answer that contains the correct values for all of the missing fields.

The central limit theorem states that the distribution of sample \_\_(1)\_\_ is approximately \_\_(2)\_\_, regardless of the distribution of the population data, and this approximation improves as the sample size gets \_\_(3)\_\_.

|  |  |  |
| --- | --- | --- |
|  | *a.* | means, skewed, larger |
|  | *b.* | variance, equal, smaller |
|  | *c.* | means, normal, larger |
|  | *d.* | proportions, equal, smaller |

4

Psychologists at a college want to know if students are sleeping more or less than the recommended average of 8 hours a day.

Which of the following code choices correctly tests the null hypothesis?

|  |  |  |
| --- | --- | --- |
|  | *a.* | proc univariate data=statdata.sleep mu0<>8;  var hours;  run; |
|  | *b.* | proc univariate data=statdata.sleep;  var hours / mu0=8;  run; |
|  | *c.* | proc univariate data=statdata.sleep;  var hours / mu0<>8;  run; |
|  | *d.* | proc univariate data=statdata.sleep mu0=8;  var hours;  run; |

How do you define the term *power*?

|  |  |  |
| --- | --- | --- |
|  | *a.* | the measure of the ability of the statistical hypothesis test to reject the null hypothesis when it is actually false |
|  | *b.* | the probability of committing a Type I error |
|  | *c.* | the probability of failing to reject the null hypothesis when it is actually false |

Power of a statistical test

It is the ability of the statistical test to detect a true difference, or the **ability to successfully reject a false null hypothesis.**

ability to successfully reject a false null hypothesis.

(1-β)

Select the choice that lists only continuous variables.

|  |  |  |
| --- | --- | --- |
|  | *a.* | body temperature, number of children, gender, beverage size |
|  | *b.* | age, body temperature, gas mileage, income |
|  | *c.* | number of children, gender, gas mileage, income |
|  | *d.* | gender, gas mileage, beverage size, income |

Which of the following code choices creates a histogram for the variable **Speed** from the data set **SpeedTest**with a normal curve overlay and a box with the skewness and kurtosis statistics printed in the northeast corner?

|  |  |  |
| --- | --- | --- |
|  | *a.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(mu=est sigma=est);  inset skewness kurtosis;  run; |
|  | *b.* | proc univariate data=statdata.speedtest;  histogram Speed / normal;  inset skewness kurtosis / position=ne;  run; |
|  | *c.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(mu=est sigma=est);  inset skewness kurtosis / position=ne;  run; |
|  | *d.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(skewness kurtosis);  run; |

Select the statement below that **incorrectly** interprets a 95% confidence interval (15.02, 15.04) for the population mean, if the sample mean is 15.03 ounces of cereal.

|  |  |  |
| --- | --- | --- |
|  | *a.* | You are 95% confident that the true average weight for a box of cereal is between 15.02 and 15.04 ounces. |
|  | *b.* | The probability is .95 that the true average weight is between 15.02 and 15.04 ounces. |
|  | *c.* | In the long run, approximately 95% of the intervals calculated with this procedure will capture the true average weight.  . |

The location and spread of a normal distribution depends on the value of which two parameters?

|  |  |  |
| --- | --- | --- |
|  | *a.* | the mean (x̄) and the standard deviation (s) |
|  | *b.* | the standard deviation (σ) and the variance (σ²) |
|  | *c.* | the mean (µ) and the standard deviation (σ) |
|  | *d.* | none of the above |

The standard error of the mean is

|  |  |  |
| --- | --- | --- |
|  | *a.* | used to calculate confidence intervals of the mean. |
|  | *b.* | always normally distributed. |
|  | *c.* | sometimes less than 0. |
|  | *d.* | none of the above |

**7 Quiz, Introduction to Statistical Concepts**

|  |  |
| --- | --- |
| Your Score: 80% | Congratulations! Your score of 80% indicates that you've mastered the topics in this lesson. If you'd like, you can review the feedback for each question. |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1. For an asymmetric (or skewed) distribution, which of the following statistics is a good measure for the middle of the data?  |  |  | | --- | --- | | *a.* | mean | | *b.* | median | | *c.* | either mean or median |   **Your answer: b Correct answer: b**  The median is not affected by outliers and is less affected by the skewness. The mean, on the other hand, averages in any outliers that might be in your data. |
|  | 1. Which of the following code examples correctly calculates descriptive statistics of popcorn yield (**Yield**) for each level of the class variable (**Type**) in the data set **Statdata.Popcorn**, as well as statistics for all levels combined?   The output should include the following statistics: sample size, mean, median, standard deviation, variance, range, and interquartile range.   |  |  | | --- | --- | | *a.* | proc means data=statdata.popcorn  maxdec=2 fw=10  n mean median std var  range qrange;  class Type;  var Yield;  run; | | *b.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std var  range qrange;  class Yield;  var Class;  run; | | *c.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std var  range qrange;  class Type;  var Yield;  run; | | *d.* | proc means data=statdata.popcorn  maxdec=2 fw=10  printalltypes  n mean median std range  IQR;  class Type;  var Yield;  run; |   **Your answer: c Correct answer: c**  The PROC MEANS statement must include the option PRINTALLTYPES in order for SAS to display statistics for all requested combinations of class variables – that is, for each level or occurrence of the variable and for all occurrences combined. The statistics specified on the second line must include the keywords N MEAN MEDIAN STD VAR RANGE QRANGE. The code must specify**Type** as the class variable and **Yield** as the analysis variable. |
|  | 1. Read the following statement about the central limit theorem and choose the answer that contains the correct values for all of the missing fields.   The central limit theorem states that the distribution of sample \_\_(1)\_\_ is approximately \_\_(2)\_\_, regardless of the distribution of the population data, and this approximation improves as the sample size gets \_\_(3)\_\_.   |  |  | | --- | --- | | *a.* | means, skewed, larger | | *b.* | variance, equal, smaller | | *c.* | means, normal, larger | | *d.* | proportions, equal, smaller |   **Your answer: c Correct answer: c**  The central limit theorem states that the distribution of sample means is approximately normal, regardless of the distribution of the population data, and this approximation improves as the sample size gets larger. |
|  | 1. Psychologists at a college want to know if students are sleeping more or less than the recommended average of 8 hours a day.   Which of the following code choices correctly tests the null hypothesis?   |  |  | | --- | --- | | *a.* | proc univariate data=statdata.sleep mu0<>8;  var hours;  run; | | *b.* | proc univariate data=statdata.sleep;  var hours / mu0=8;  run; | | *c.* | proc univariate data=statdata.sleep;  var hours / mu0<>8;  run; | | *d.* | proc univariate data=statdata.sleep mu0=8;  var hours;  run; |   **Your answer: d Correct answer: d**  You specify the MU0= option as part of the PROC UNIVARIATE statement to indicate the test value of the null hypothesis. The alternative hypothesis is that μ is not equal to 8 hours, but this does not need to be specified in the PROC UNIVARIATE code. |
|  | 1. How do you define the term *power*?  |  |  | | --- | --- | | *a.* | the measure of the ability of the statistical hypothesis test to reject the null hypothesis when it is actually false | | *b.* | the probability of committing a Type I error | | *c.* | the probability of failing to reject the null hypothesis when it is actually false |   **Your answer: a Correct answer: a**  Power is the ability of the statistical test to detect a true difference, or the ability to successfully reject a false null hypothesis. The probability of committing a Type I error is α. The probability of failing to reject the null hypothesis when it is actually false is a Type II error. |
|  | 1. Select the choice that lists only continuous variables.  |  |  | | --- | --- | | *a.* | body temperature, number of children, gender, beverage size | | *b.* | age, body temperature, gas mileage, income | | *c.* | number of children, gender, gas mileage, income | | *d.* | gender, gas mileage, beverage size, income |   **Your answer: b Correct answer: b**  The continuous variables are age, body temperature, gas mileage, and income. |
|  | 1. Which of the following code choices creates a histogram for the variable **Speed** from the data set **SpeedTest**with a normal curve overlay and a box with the skewness and kurtosis statistics printed in the northeast corner?  |  |  | | --- | --- | | *a.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(mu=est sigma=est);  inset skewness kurtosis;  run; | | *b.* | proc univariate data=statdata.speedtest;  histogram Speed / normal;  inset skewness kurtosis / position=ne;  run; | | *c.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(mu=est sigma=est);  inset skewness kurtosis / position=ne;  run; | | *d.* | proc univariate data=statdata.speedtest;  histogram Speed / normal(skewness kurtosis);  run; |   **Your answer: c Correct answer: c**  In the HISTOGRAM statement, you specify the **Speed** variable and the NORMAL option using estimates of the population mean and the population standard deviation. In the INSET statement, you specify the keywords SKEWNESS and KURTOSIS, as well as the POSITION=NE option. |
|  | 1. Select the statement below that **incorrectly** interprets a 95% confidence interval (15.02, 15.04) for the population mean, if the sample mean is 15.03 ounces of cereal.  |  |  | | --- | --- | | *a.* | You are 95% confident that the true average weight for a box of cereal is between 15.02 and 15.04 ounces. | | *b.* | The probability is .95 that the true average weight is between 15.02 and 15.04 ounces. | | *c.* | In the long run, approximately 95% of the intervals calculated with this procedure will capture the true average weight.  . |   **Your answer: a Correct answer: b**  A 95% confidence interval means that you are 95% confident that the interval contains the true population mean. If you sample repeatedly and calculate a confidence interval for each sample mean, 95% of the time your confidence interval will contain the true population mean. A confidence interval is not a probability. When a confidence interval is calculated, the true mean is in the interval or it is not. There is no probability associated with it. |
|  | 1. The location and spread of a normal distribution depend on the value of which two parameters?  |  |  | | --- | --- | | *a.* | the mean (x̄) and the standard deviation (s) | | *b.* | the standard deviation (σ) and the variance (σ²) | | *c.* | the mean (µ) and the standard deviation (σ) | | *d.* | none of the above |   **Your answer: a Correct answer: c**  The location and spread of a normal distribution depend on the value of two parameters, the mean (µ) and the standard deviation (σ). |
|  | 1. The standard error of the mean is  |  |  | | --- | --- | | *a.* | used to calculate confidence intervals of the mean. | | *b.* | always normally distributed. | | *c.* | sometimes less than 0. | | *d.* | none of the above |   **Your answer: a Correct answer: a**  The standard error of the mean is part of the equation used to calculate a confidence interval of the mean. It is not normally distributed, and it is never less than 0. |