

Exercise – Confidence Intervals

(Fall 2015)

Sources (adapted with permission)-

T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

For educational uses only - adapted from sources with permission. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission from the author/presenter.

Objectives

- Decide what tasks to complete before you analyze your data.
- Use the Summary Statistics task to produce descriptive statistics.

15

Confidence Intervals for the Mean

Point Estimates

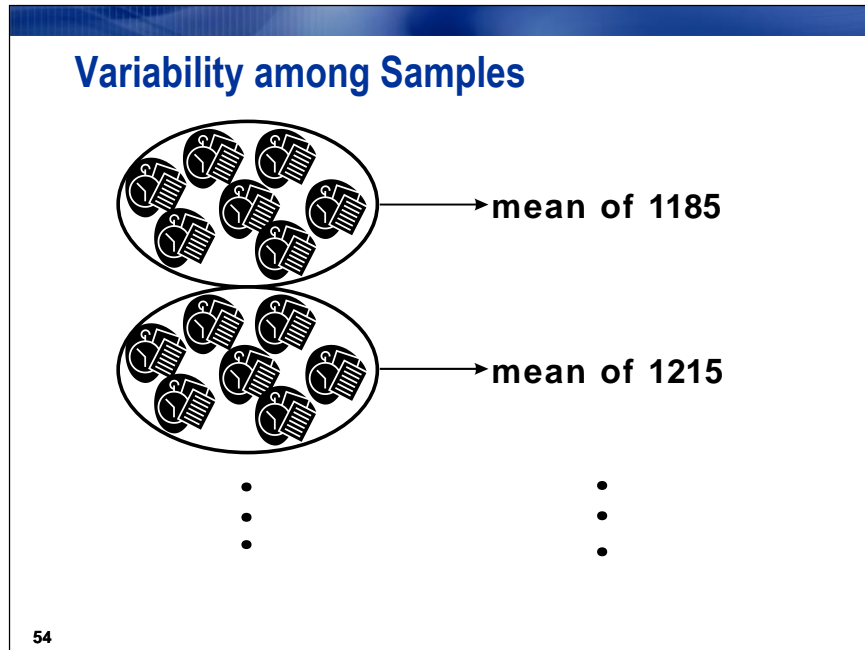
\bar{x} estimates μ

S estimates σ

53

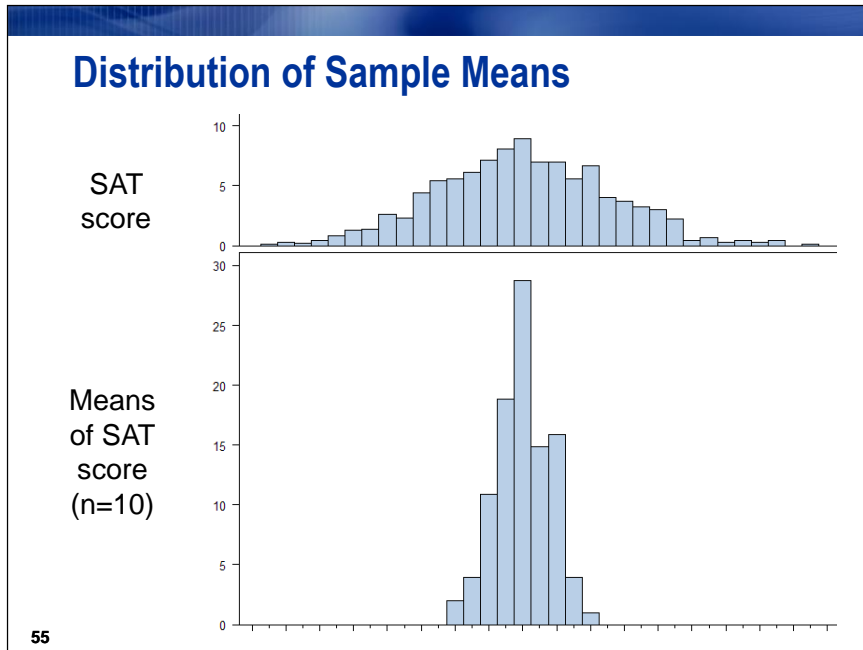
A *point estimate* is a sample statistic used to estimate a population parameter.

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



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

Different samples yield different estimates of the mean for the same population. How close on average these sample means are to one another is the variability of the estimate of the population mean.



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

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

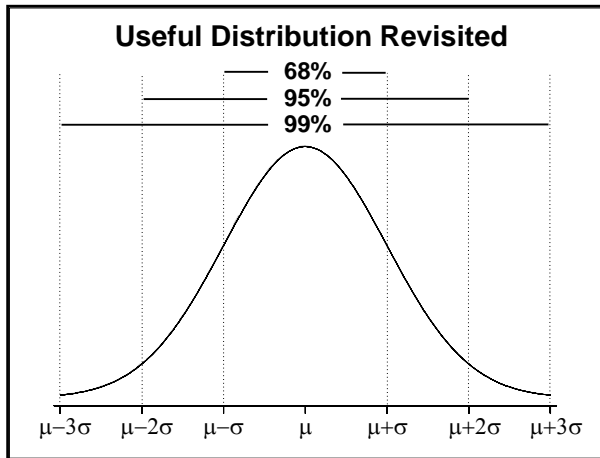
- The top histogram shows the distribution of **all 5000 observations**.
- The bottom histogram, however, represents the distribution of the **1000 sample means**.

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



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

Normal Distribution for the Mean



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

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

Standard Error of the Mean

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

It differs from the sample standard deviation because

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

$$\text{Standard error of the mean} = \frac{S}{\sqrt{n}} = S_{\bar{x}}$$

57

The standard error of the mean is computed as

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

where

s is the sample standard deviation

n is the sample size

Assume a sample size of $n= 80$ and a sample standard deviation $s = 147.058447$.

The standard error of the mean for the variable **SATScore** is $147.058447 / \sqrt{80}$, or approximately 16.44.

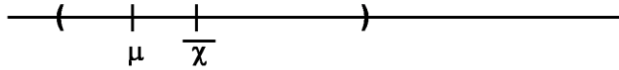
This is a measure of how much variability of sample means there is around the population mean. The smaller the standard error, the more precise your sample estimate is.



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

Confidence Intervals

95% Confidence



- A 95% confidence interval states that you are 95% certain that the true population mean lies between two calculated values.
 - In other words, if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

58

A *confidence interval*

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

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

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

Do you want to be as confident as possible?

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

Details

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

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

This is the basis of confidence intervals for the mean. If you rearrange the terms above and replace the known $\sigma_{\bar{x}}$ with the estimated standard error, $s_{\bar{x}}$, the probability is 0.95 for

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

When the values of μ and σ are unknown, one of the family of Student's t distributions is used in place of the normal (z) distribution. The value of 1.96 will be replaced by a t -value determined by the degrees of freedom. The larger the sample size, the closer that t -value will be to 1.96.

Confidence Interval for the Mean

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

where

\bar{x} is the sample mean.

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

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

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

59

Student's t distribution arises when you are 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 in the equation above refers to the number of standard deviation (or standard error) units away from the mean required to get a desired confidence in a confidence interval. That value will vary not only with the confidence that you choose, but also with the sample size. For 95% confidence, that t value will usually be approximately 2, because, as you have seen, 2 standard errors below to 2 standard errors above a mean will give you about 95% of the area under a normal distribution curve.

Normality and the Central Limit Theorem

To satisfy the assumption of normality, you can either

- verify that the population distribution is approximately normal, or
- apply the **central limit theorem**.
 - The central limit theorem states that the distribution of sample means is approximately normal, regardless of the distribution's shape, if the sample size is large enough.
 - “Large enough” is usually about 30 observations: more if the data are heavily skewed, fewer if the data are symmetrically distributed.

61

To apply the central limit theorem, your sample size should be at least 30. The central limit theorem holds even if you have no reason to believe the population distribution is not normal.

Because the sample size for the test scores example is 80, you can apply the central limit theorem and satisfy the assumption of normality for the confidence intervals.

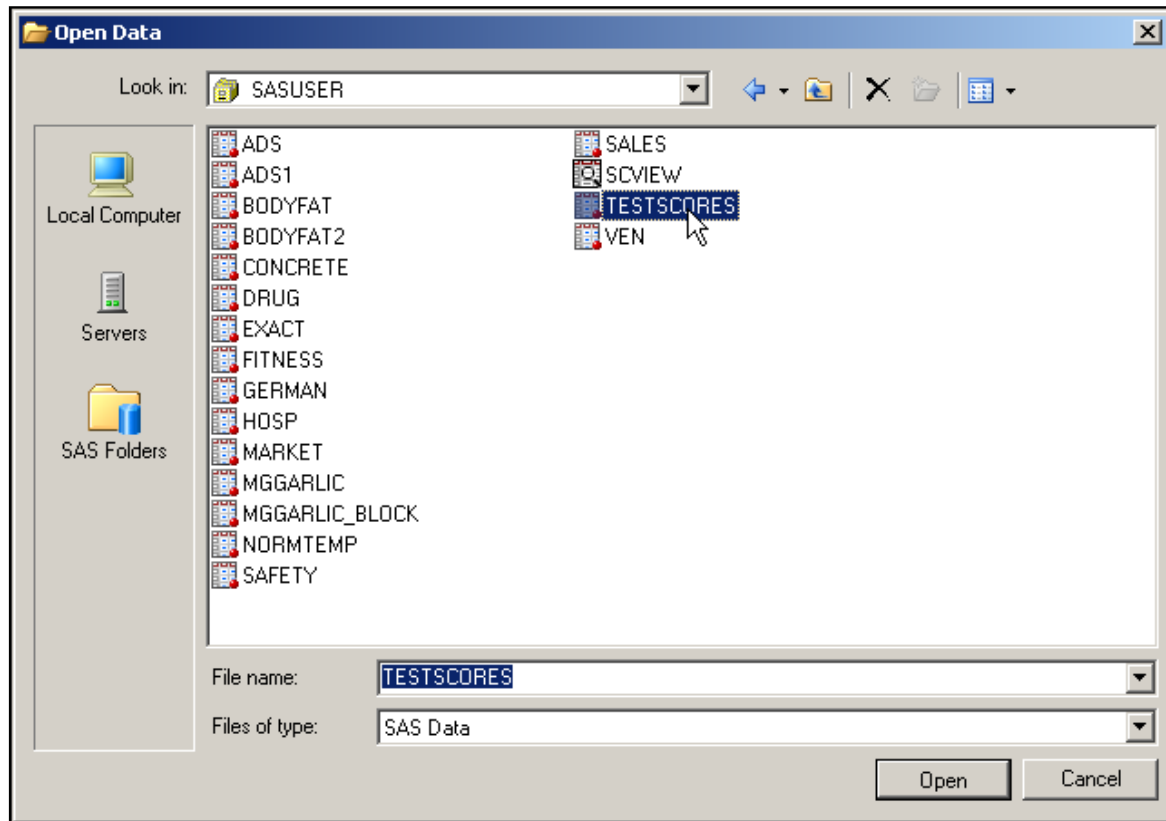
Exercise - Confidence Intervals



Use the Summary Statistics task to generate a 95% confidence interval for the mean of **SATScore** in the **testscores** data set.

1. Obtain and open **TESTSCORES** SAS Dataset.

File > Open > Data--> Servers > SASApp-->Files > D: > ISYS 5503--> ISYS 5503 Shared Datasets



The data table opens automatically. You can close it after looking at it.


Partial Listing

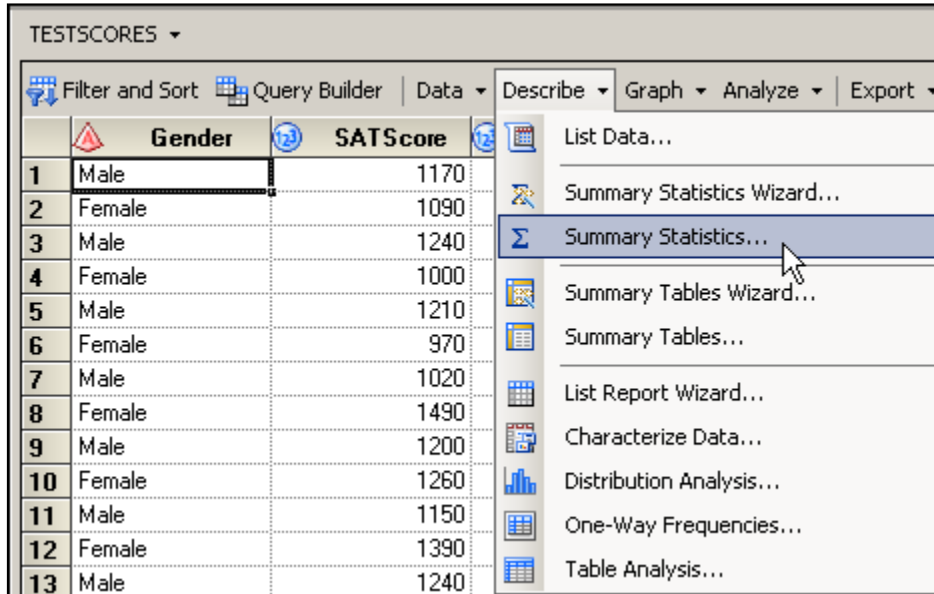
	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397

There are three variables in the **TESTSCORES** data set. One variable, **Gender**, is a character variable that contains the gender of the student. The other two variables, **SATSCORE** and **IDNumber**, are numeric variables that contain the SAT combined verbal and quantitative score and an identifying code for each student.

Create a summary statistics report for the **TESTSCORES** data set.

2. Above the data table, select **Describe** ⇒ **Summary Statistics...** from the drop-down menus.

 If you close the data table first, then you will have to click **Tasks** ⇒ **Describe** ⇒ **Summary Statistics...** from the top menu bar.



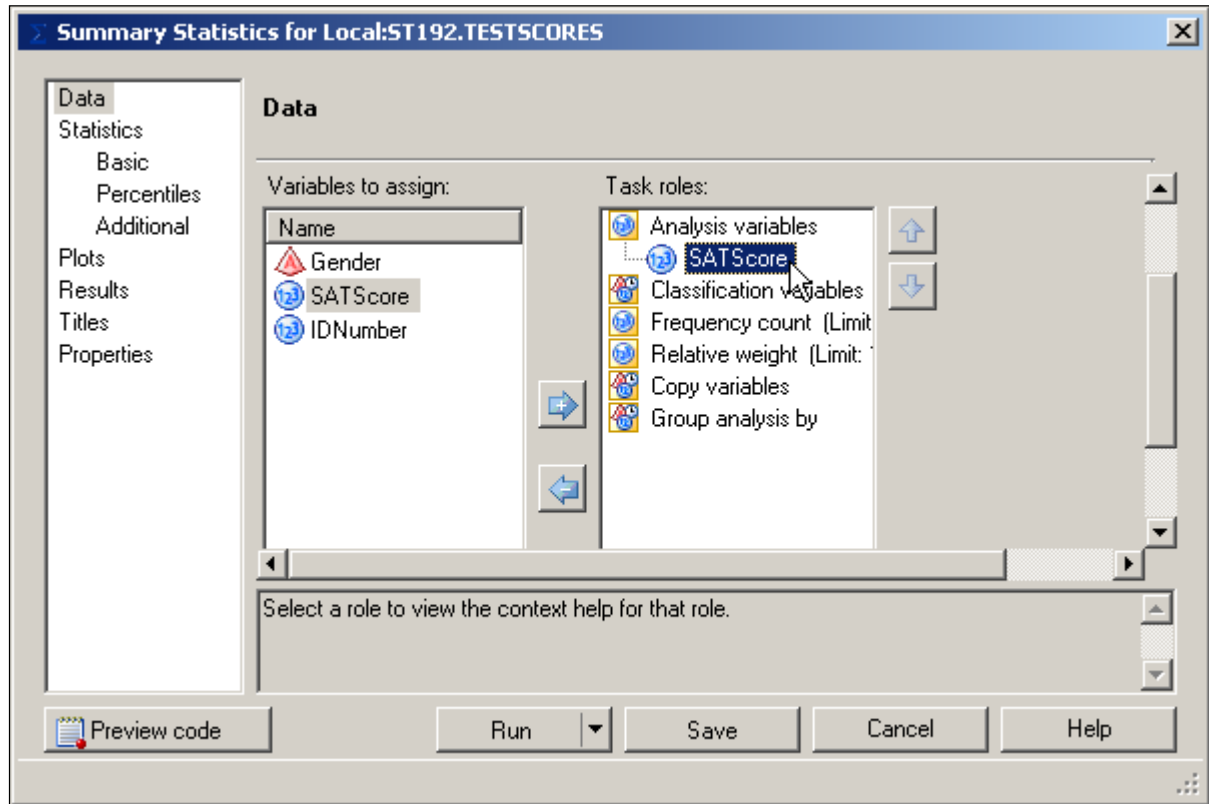
The screenshot shows a software interface with a data table and a menu. The data table has two columns: **Gender** and **SATScore**. The menu is open, showing options under the **Describe** dropdown. The **Summary Statistics...** option is highlighted by a mouse cursor.

	Gender	SATScore
1	Male	1170
2	Female	1090
3	Male	1240
4	Female	1000
5	Male	1210
6	Female	970
7	Male	1020
8	Female	1490
9	Male	1200
10	Female	1260
11	Male	1150
12	Female	1390
13	Male	1240

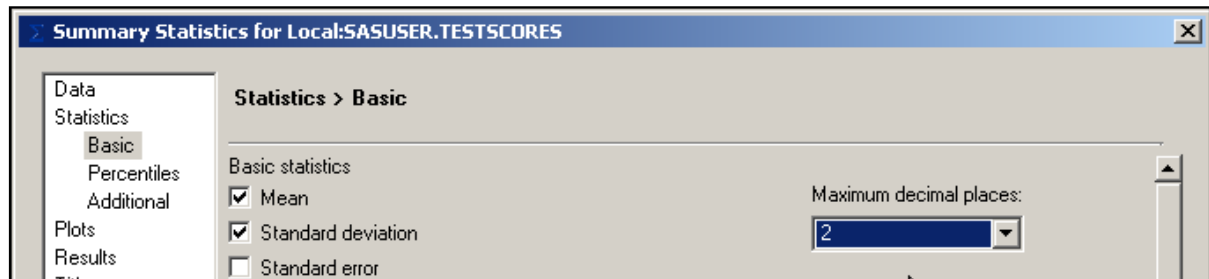
Menu options under **Describe**:

- List Data...
- Summary Statistics Wizard...
- Summary Statistics...**
- Summary Tables Wizard...
- Summary Tables...
- List Report Wizard...
- Characterize Data...
- Distribution Analysis...
- One-Way Frequencies...
- Table Analysis...

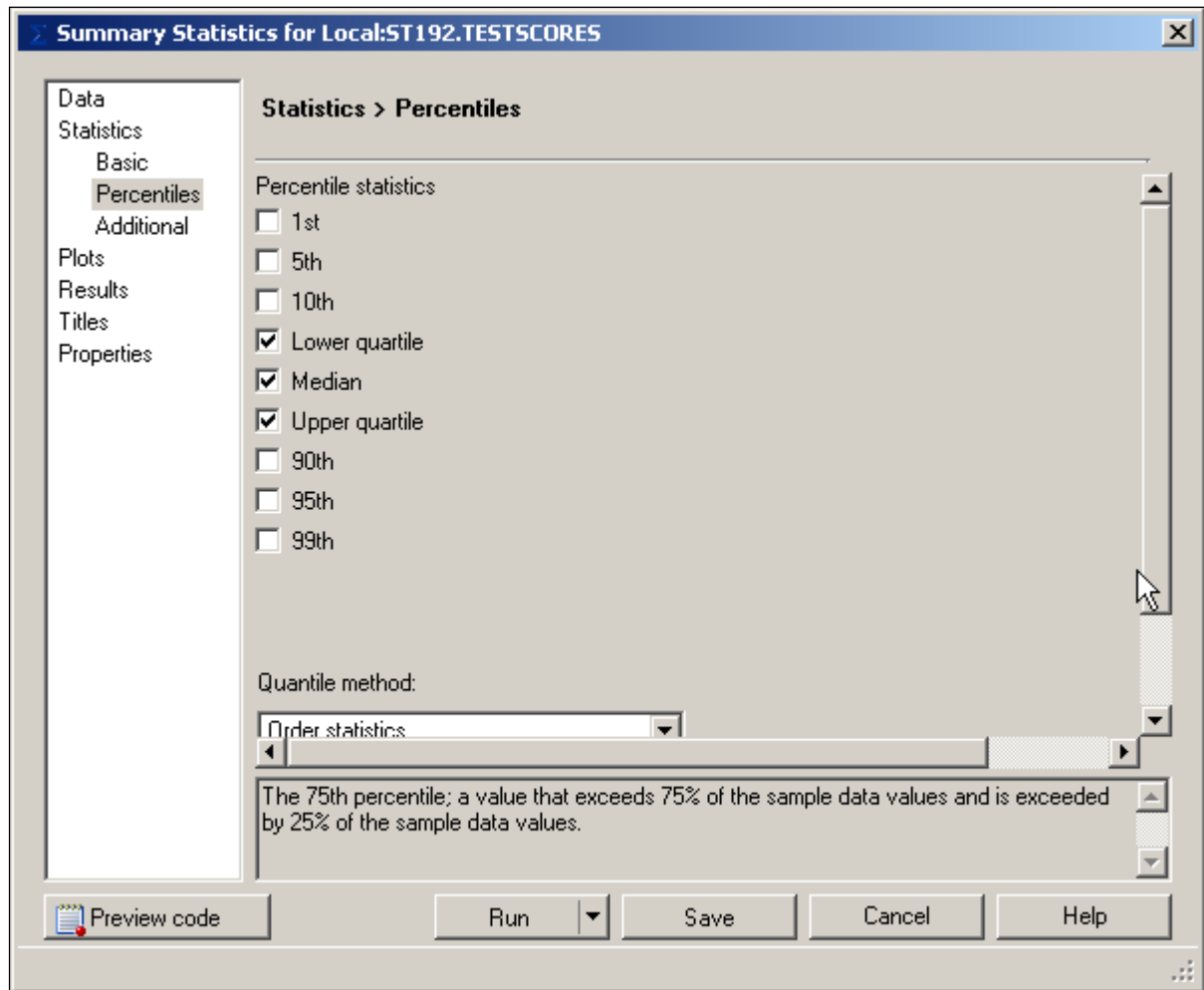
3. With **Data** selected on the left, drag the variable **SATScore** from the Variables to assign pane to the analysis variables role in the Task roles pane, as shown below:



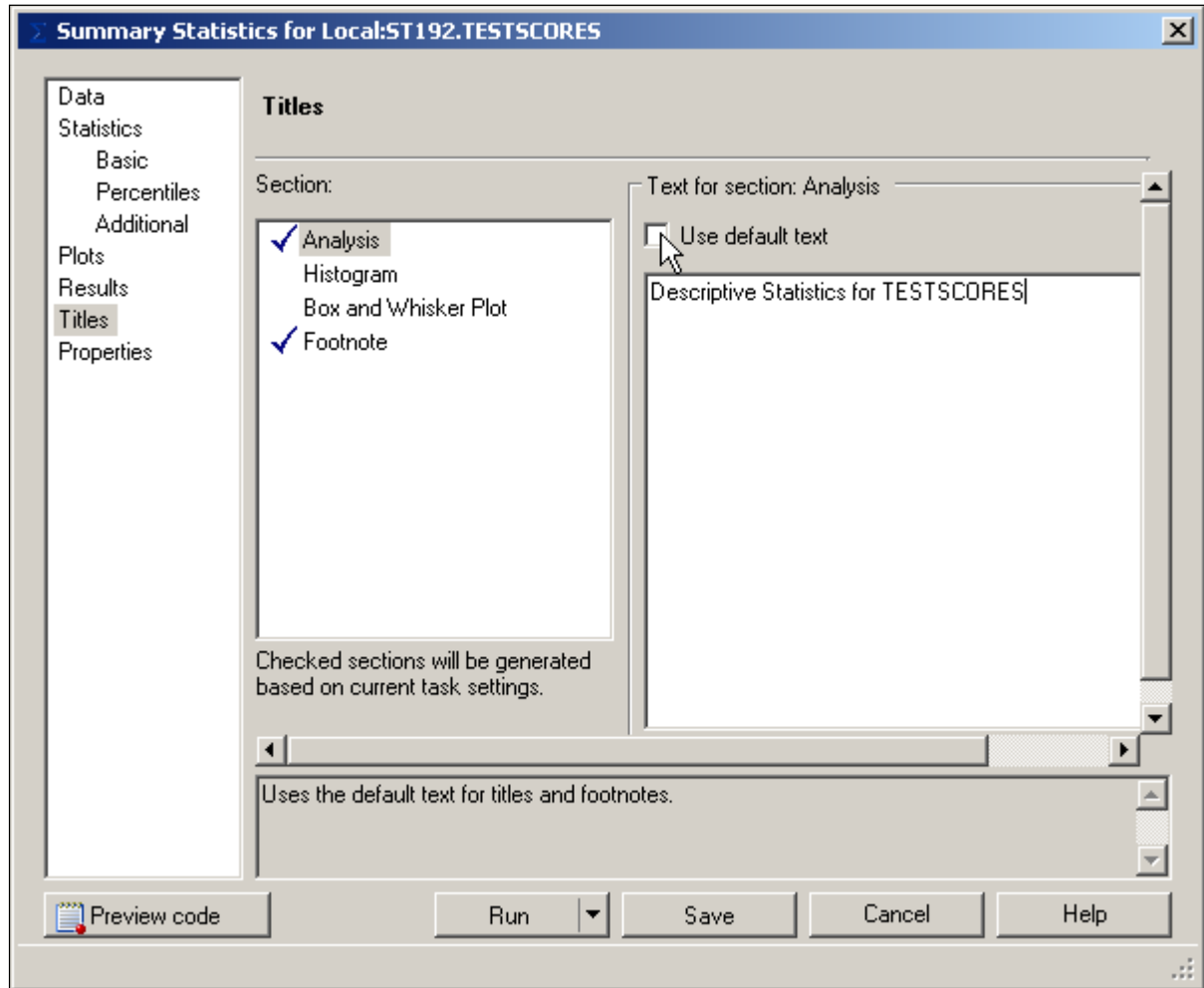
4. Select **Basic** under Statistics on the left. Leave the default basic statistics. Change Maximum decimal places to **2**.



5. Select **Percentiles** on the left. Under Percentile statistics, check the boxes for **Lower quartile**, **Median**, and **Upper quartile**.

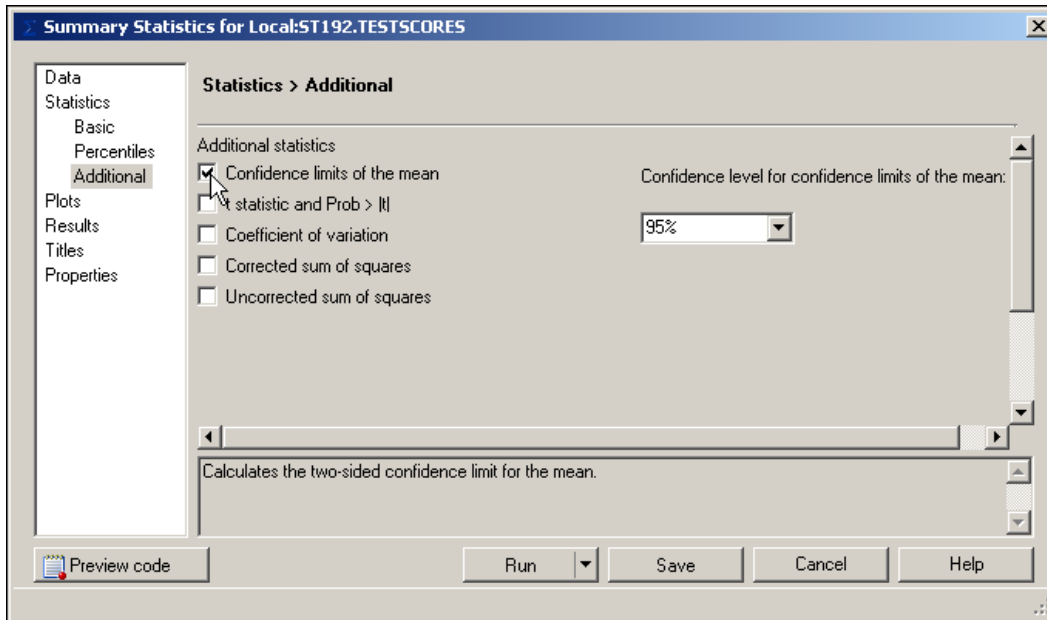


6. Select **Titles** on the left. Deselect **Use default text**. Select the default text in the box and type **Descriptive Statistics for TESTSCORES**. Leave the default footnote text.



Confidence Interval

7. Click **Additional** at left and then check **Confidence limits of the mean**. Leave the confidence level at 95%.





8. Click **Run** and then click **Yes** when asked if you want to replace the results from the previous run.

The output is shown below.

Analysis Variable : SAT Score									
Mean	Std Dev	Minimum	Maximum	N	Lower Quartile	Median	Upper Quartile	Lower 95% CL for Mean	Upper 95% CL for Mean
1190.63	147.06	890.00	1600.00	80	1085.00	1170.00	1280.00	1157.90	1223.35

In the test score example, you are 95% confident that the population mean is contained in the interval 1157.8987 and 1223.3513. Because the interval between the upper and lower limits is small from a practical point of view, you can conclude that the sample mean is a fairly precise estimate of the population mean.

 How do you increase the **precision** of your estimate using the same confidence level? If you increase your sample size, you reduce the standard error of the sample mean and therefore reduce the width of your confidence interval. Thus, your estimate will be more precise.

 *Accuracy* is the difference between a sample estimate and the true population value. *Precision* is the difference between a sample estimate and the mean of the estimates of all possible samples that can be taken from the population. For an unbiased estimator, precision and accuracy are the same.