SASEG 1 Exercise – Fundamental Summary Analytics

(Fall 2017)

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

Objectives

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

15

Defining the Problem

The purpose of the study is to determine whether or not the average combined Math and Verbal scores on the Scholastic Aptitude Test (SAT) at Carver County magnet high schools is 1200 – the goal set by the school board.

16

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

A *population* is a collection of all objects about which information is desired. In this example, the population is all Carver County magnet school seniors.

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

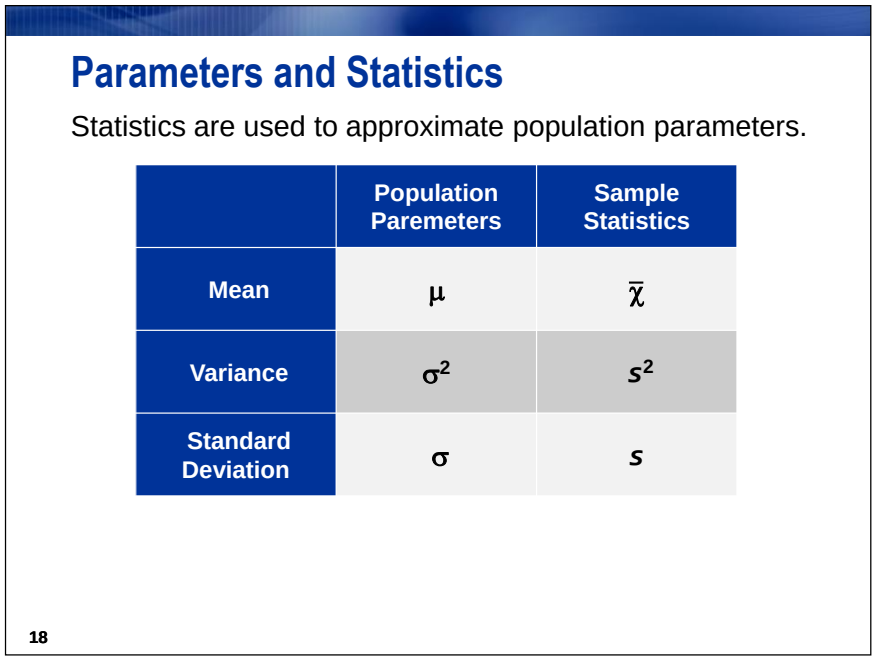
Simple random sampling, a technique in which each member of the population has an equal probability of being selected, is used by Ms. Chao's students. Random sampling can help to ensure that the sample is representative of the population.

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

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

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

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



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

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

$\bar{x} = \frac{1}{n}\sum x_i$ the mean is an average, a typical value in the distribution.

$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ the variance measures the sample variability.

$s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$ the standard deviation measures variability. It is reported in the same units as the mean.

Descriptive Statistics

The goals when you are describing data are to

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

Inferential Statistics

The goals for statistical inference are to

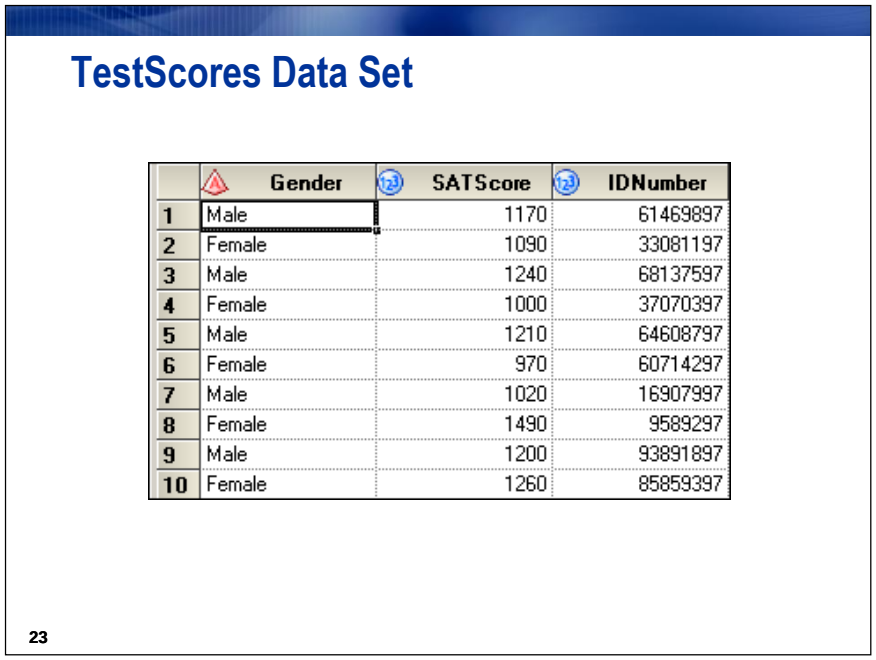
- estimate or predict unknown parameter values from a population, using a sample
- make probabilistic statements about population attributes.

19

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

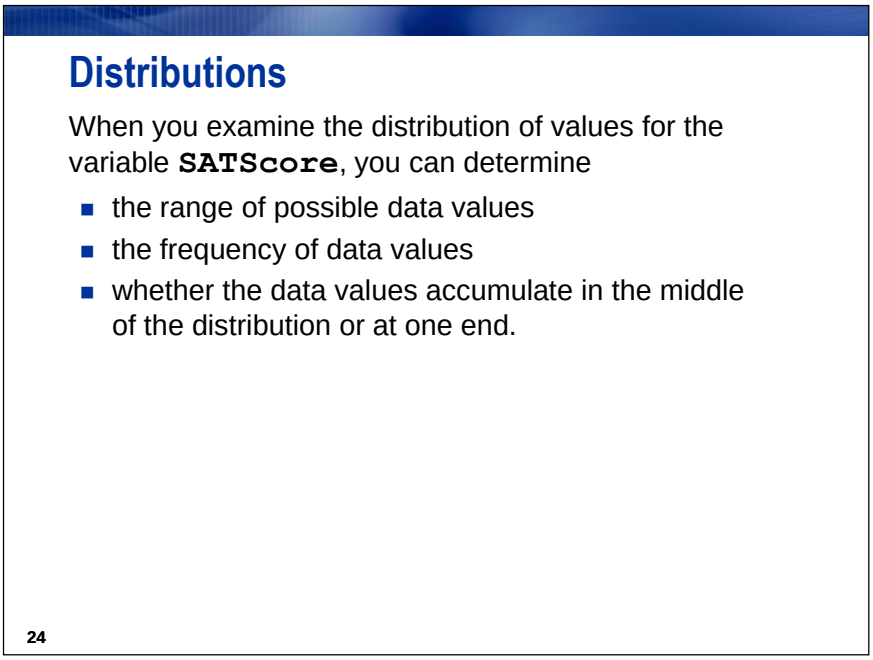
Why?

- Data must be as error free as possible.
- Unique aspects, such as data values that cluster or show some unusual shape, must be identified.
- An extreme value of a variable, if not detected, could cause gross errors in the interpretation of the statistics.



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

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



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

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

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

You can answer these questions using descriptive statistics.

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

"Typical Values" in a Distribution

- Mean: the sum of all the values in the data set divided by the number of values

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

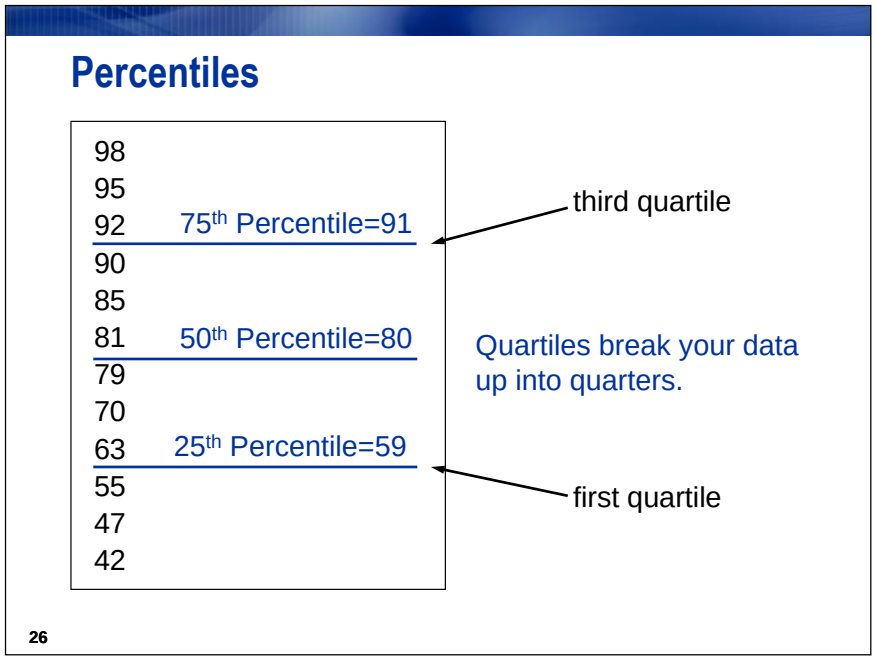
- Median: the middle value (also known as the 50th percentile)
- Mode: the most common or frequent data value

25

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

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

The mean is the physical balancing point of your data.



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

Commonly reported percentile values are

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

The Spread of a Distribution: Dispersion

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

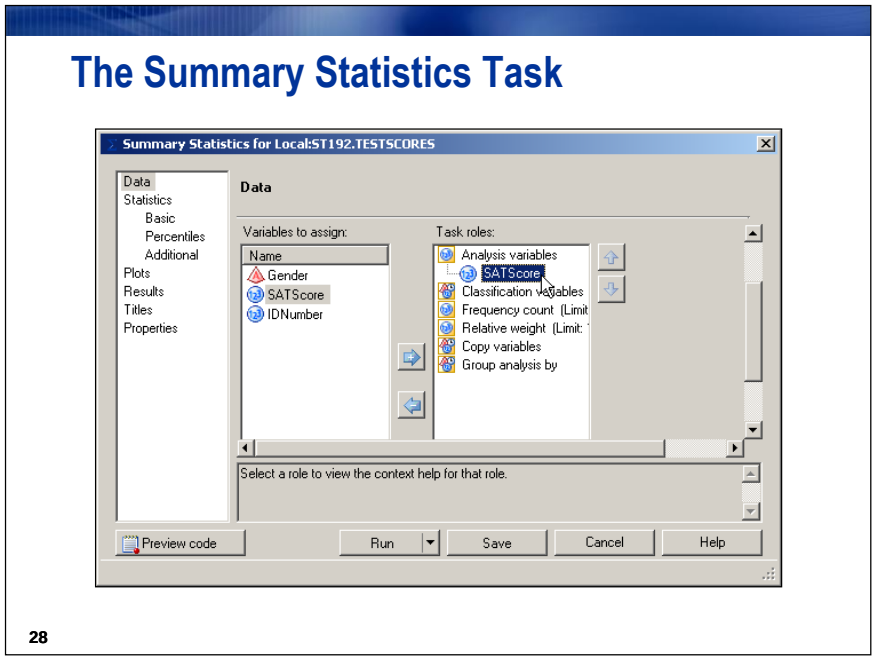
27

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

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

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

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

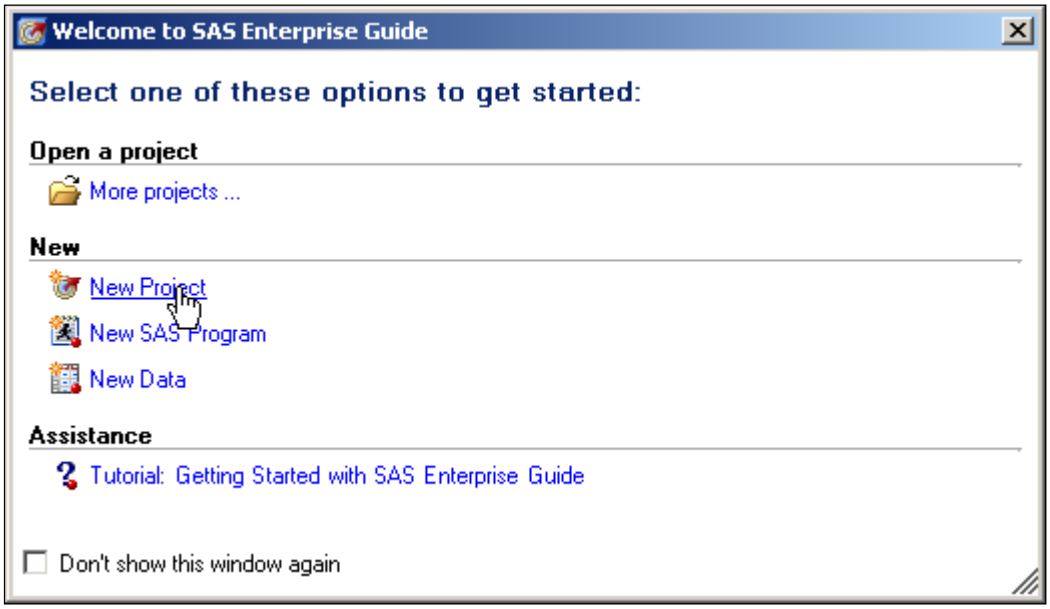


The Summary Statistics task is used for generating descriptive statistics for your data.

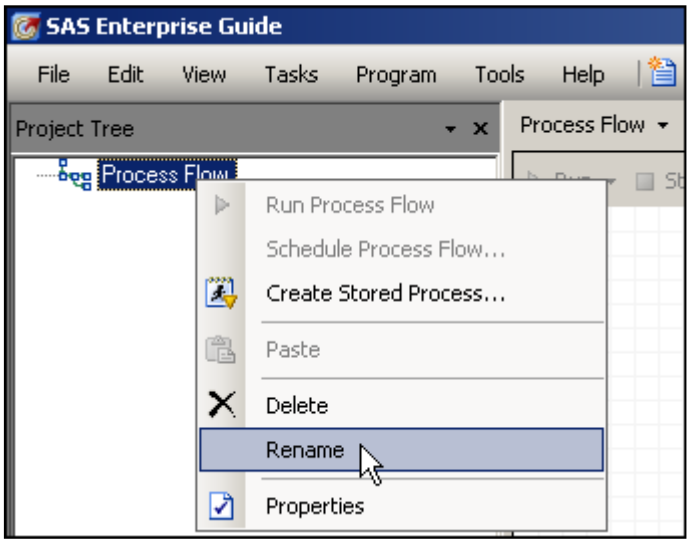
Exercise - Descriptive Statistics

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

1. When you open SAS Enterprise Guide, you see a dialog box that gives you several options. Select **New Project**.

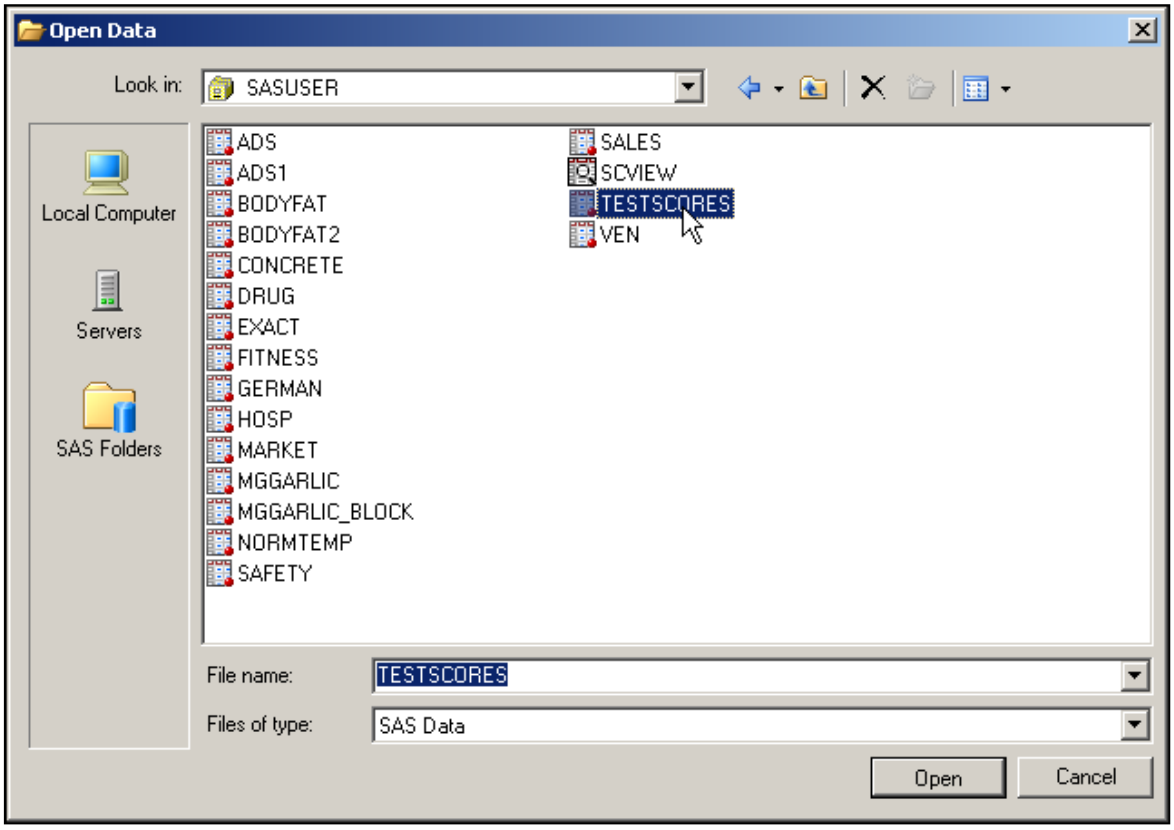


2. Rename the process flow by right-clicking the **Process Flow** icon in the Project Tree Pane and clicking **Rename** in the drop-down menu.



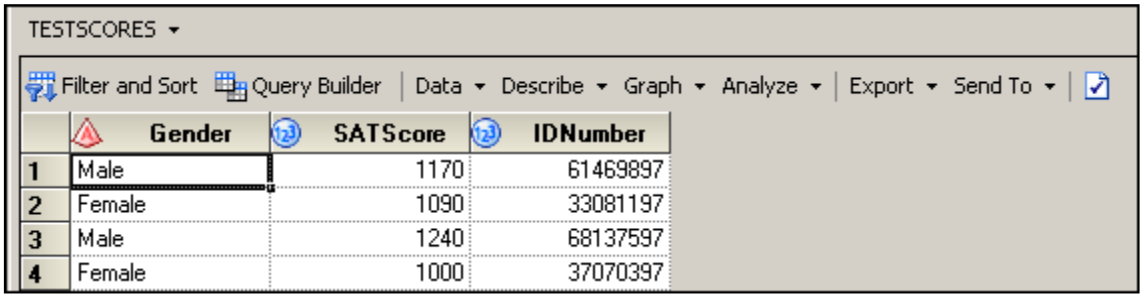
3. Obtain and open **`TESTSCORES`** SAS Dataset.

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



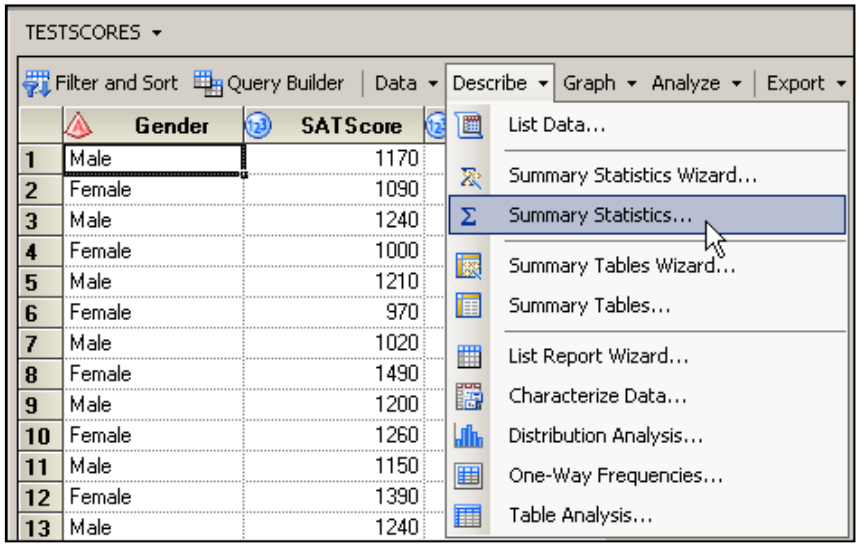
	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397

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 Summary Statistics

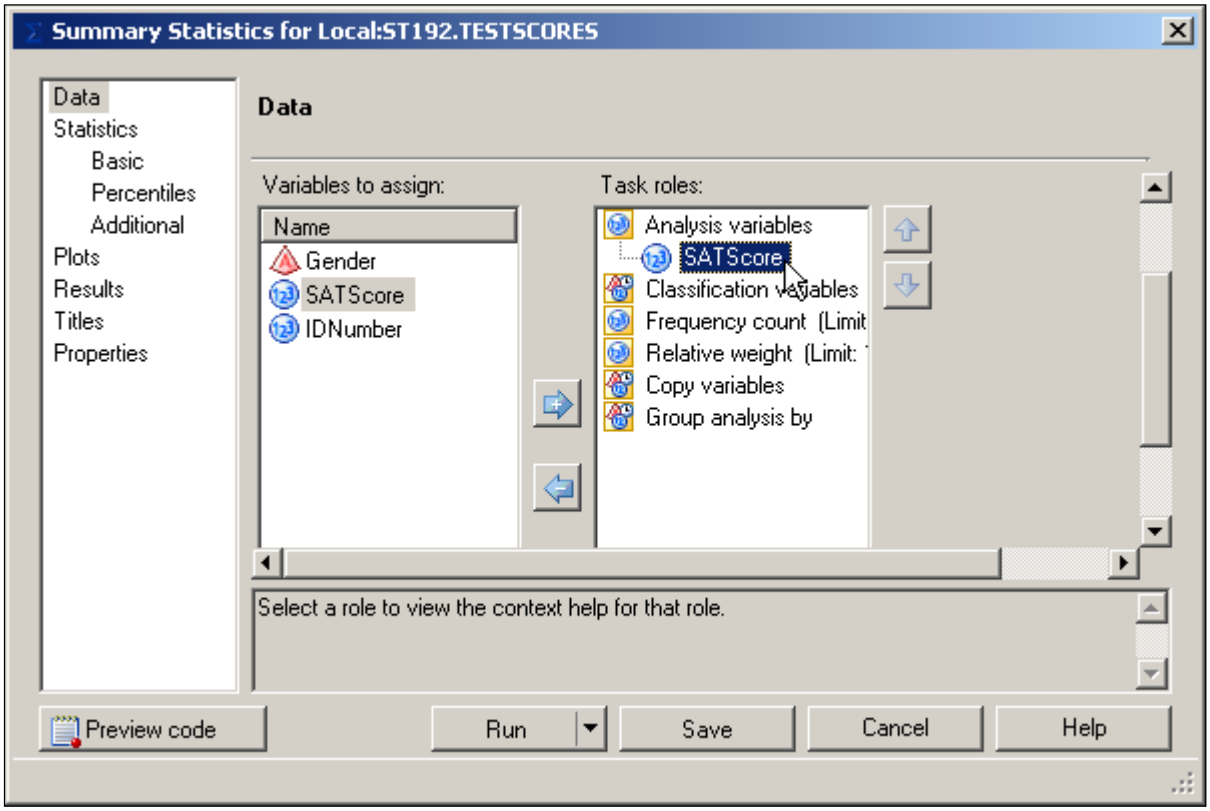
Create a summary statistics report for the **TESTSCORES** data set.

4. Above the data table, select **Describe** ⇨ **Summary Statistics…** from the drop-down menus.

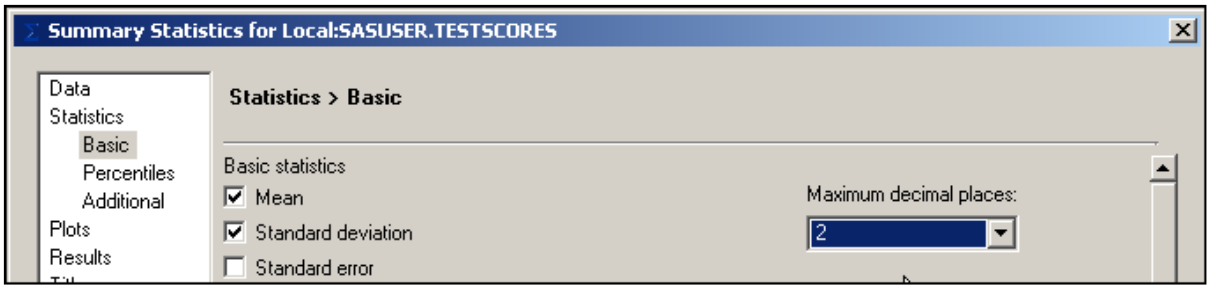


> If you close the data table first, then you will have to click **Tasks** ⇨ **Describe** ⇨ **Summary Statistics…** from the top menu bar.

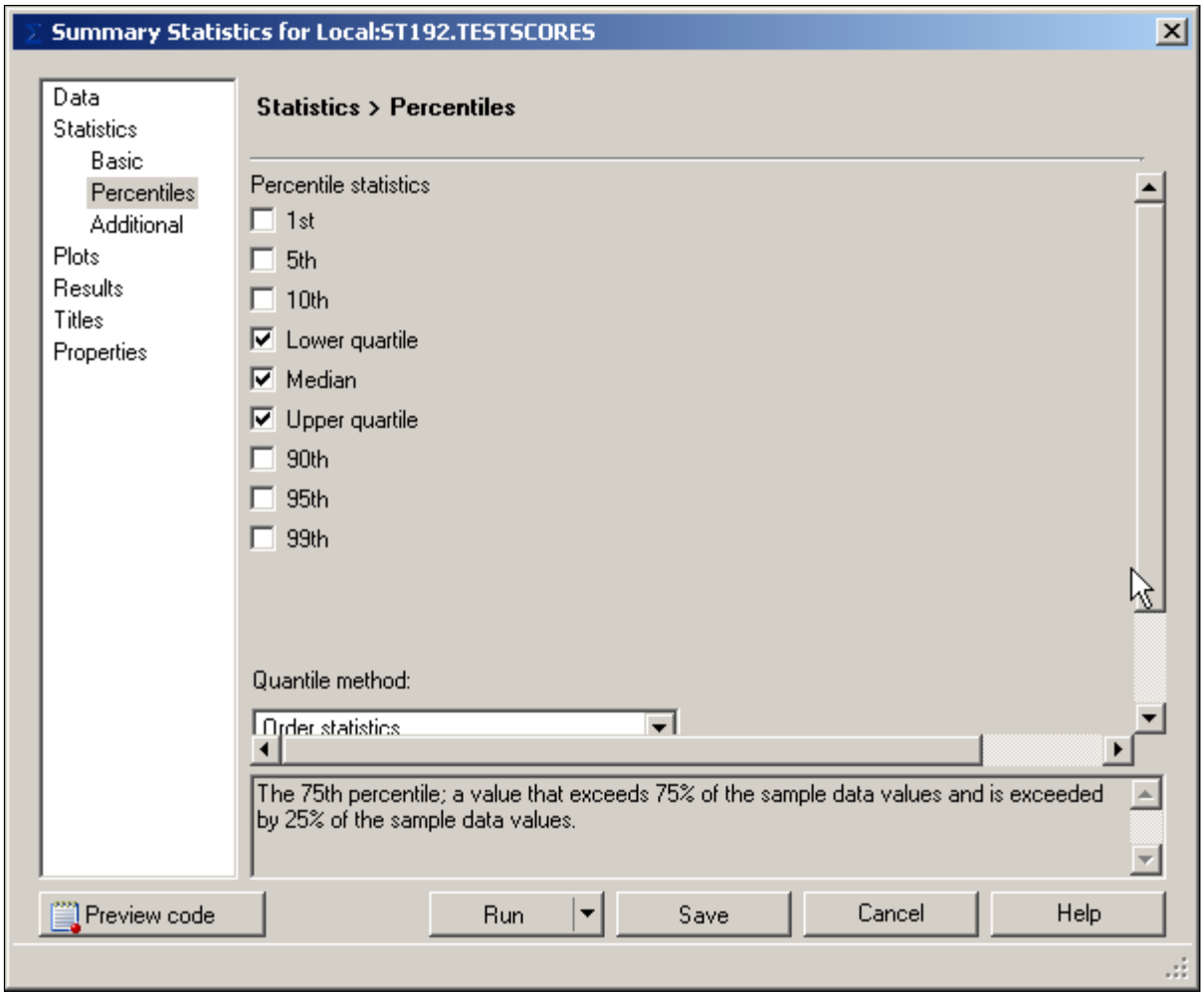
5. With **Data** selected on the left, drag the variable **SATScore** from the `Variables to assign` pane to the analysis variables role in the `Task roles` pane, as shown below:



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



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



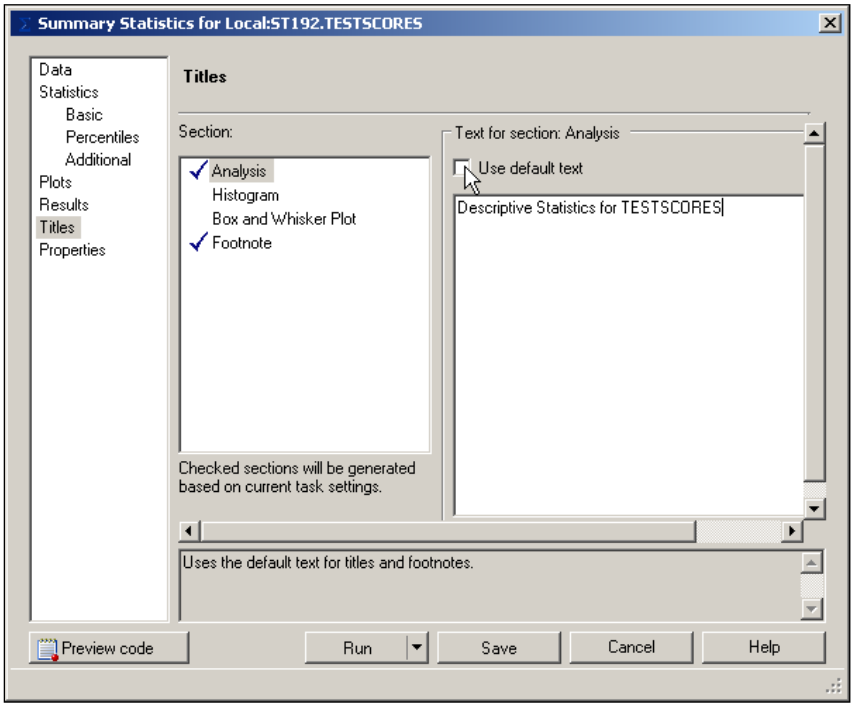
8. Select **Titles** on the left. Deselect **Use default text**. Select the default text in the box and type **`Descriptive Statistics for TESTSCORES`**. Leave the default footnote text.

SAS Output

9. Select Run to run the analysis.

   The report is shown below:

   The mean is 1190.63, which is not exactly the 1200 that the school board had set as a goal. The standard deviation is 147.06. The range is 710 (1600 – 890) and the interquartile range is 110 (1280 – 1170).



Descriptive Statistics for TESTSCORES

The MEANS Procedure

Analysis Variable : SATScore							
Mean	Std Dev	Minimum	Maximum	N	Lower Quartile	Median	Upper Quartile
1190.63	147.06	890.00	1600.00	80	1085.00	1170.00	1280.00

10. Save the project by selecting **File** ⇨ **Save EGBS** or use  .Picturing Distributions