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

Copyright © 2013 ISYS 5503 Decision Support and Analytics, Information Systems; Timothy Paul Cronan. 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

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 $x_1, x_2, ..., x_n$, a sample from some population.

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

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

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

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

the variance measures the sample variability.

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.

	🔌 Gender	3 SATScore	IDNumber
1	Male	1170	6146989
2	Female	1090	3308119
3	Male	1240	6813759
4	Female	1000	3707039
5	Male	1210	6460879
6	Female	970	6071429
7	Male	1020	1690799
8	Female	1490	958929
9	Male	1200	9389189
10	Female	1260	8585939

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?



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

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

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 $\frac{s}{\overline{x}} \times 100$.

∑ Summary Statis	tics for Local:ST192.TESTSCORES
Data Statistics	Data
Percentiles Additional Plots Hesuits Titles Properties	Variables to assign: Name Analysis variables SATScore SATScore SATScore Frequency count (Limit: Copy variables Copy variables
	Select a role to view the context help for that role.

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 <u>New Project</u>.

🐼 Welcome to SAS Enterprise Guide	×
Select one of these options to get started:	
Open a project	
🚔 More projects	
New	
🐨 New Project	
🔀 New SAS Program	
🧱 New Data	
Assistance	
2 Tutorial: Getting Started with SAS Enterprise Guide	
Don't show this window again	//

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

🞯 SAS Enter	rprise Gu	ide				
File Edit	View	Tasks	Program	Tools	Help	1 🛅
Project Tree			•	×F	Process Fl	ow 🔻
		Run Pro Schedu Create Paste Delete Renam	e cess Flow le Process F Stored Proc	low ess		St

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

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

<i>ि</i> Open Data			×
Look in:	SASUSER	💌 🔶 🗲	🗙 🦢 🖬 -
Local Computer	ADS ADS BODYFAT BODYFAT2 CONCRETE DRUG EXACT FITNESS GERMAN HOSP MARKET MGGARLIC MGGARLIC MGGARLIC NORMTEMP SAFETY	SALES SCVIEW TESTSCORES VEN	
	File name:	TESTSCORES	•
	Files of type:	SAS Data	•
			Open Cancel

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

Partial Listing

TESTSCORES -								
💞 Filter and Sort 🟥 Query Builder Data 🔹 Describe 👻 Graph 👻 Analyze 👻 Export 👻 Send To 👻 📝								
	🔌 Gender	😥 S#	ATScore	12	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, TESTSCORES select **Describe** ⇒ **Summary Statistics...** 📆 Filter and Sort 🕮 Query Builder | Data Describe - Graph - Analyze -Export -Ŧ from the drop-down 62 à **b** SATScore TE List Data... Gender menus. Male 1170 1 20 Summary Statistics Wizard... 2 1090 Female If you close the Σ Summary Statistics... 3 Male 1240 data table first, Female 1000 4 Summary Tables Wizard... then you will have 5 Male 1210 Summary Tables... to click Tasks ⇒ 6 Female 970 7 Male 1020 <u>Describe</u> ⇒ List Report Wizard... 8 Female 1490 **Summary** Characterize Data... 9 Male 1200 Statistics... from 10 Female 1260 đh Distribution Analysis... the top menu bar. 11 Male 1150 One-Way Frequencies... 12 Female 1390 Table Analysis... 1240 13 Male
- 5. With **<u>Data</u>** 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:

🔰 Summary Statis	tics for Local:ST192.TESTSCORES	×
Data Statistics Basic	Data Variables to assign: Task roles:	
Percentiles Additional Plots Results Titles Properties	Name Image: Analysis variables Image: Analysis variables Image: Analysis variables Image: Analysis	
	Select a role to view the context help for that role.	
Preview code	Run 👻 Save Cancel	Help:

6. Select <u>Basic</u> under Statistics on the left. Leave the default basic statistics. Change Maximum decimal places to <u>2</u>.

Σ	Summary Statis	tics for Local:SASUSER.TESTSCORES		×
	Data Statistics Basic	Statistics > Basic		
	Percentiles Additional	Basic statistics Mean	Maximum decimal places:	-
	Plots Results	Standard deviation	2	

7. Select <u>Percentiles</u> on the left. Under Percentile statistics, check the boxes for <u>Lower quartile</u>, <u>Median</u>, and <u>Upper quartile</u>.

🔀 Summary Statis	tics for Local:ST192.TESTSCORES	×
Data Statistics	Statistics > Percentiles	
Basic Percentiles Additional Plots Results Titles Properties	Percentile statistics 1 1st 5th 10th ✓ Lower quartile ✓ Median ✓ Upper quartile 90th 95th 99th	
Preview code	Quantile method:	

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

S	AS Output	∑ Summary Statis	stics for Local:ST192.TESTSCORES	×
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 ($1280 - 1170$).	Data Statistics Basic Percentiles Additional Plots Results Titles Properties	Titles Section: Analysis Histogram Box and Whisker Plot Footnote Descriptive Statistics for TESTSCORES Checked sections will be generated based on current task settings. Image: Save Cancel Help	

Descriptive Statistics for TESTSCORES

The MEANS Procedure

Analysis Variable : SATScore							
Mean	Std Dev	Minimum	Maximum	Ν	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 <u>File</u> \Rightarrow <u>Save EGBS</u> or use . Picturing Distributions