SPSS Tutorial 4: Mean, Variance & Standard Deviation

Goals

- To calculate the mean, variance and standard deviation (SD) using the conceptual formula, but using SPSS (instead of a calculator) to do the calculations
- To calculate the mean, variance and SD using the MEANS procedure

Reading the data into SPSS

For this tutorial, we will use the data from Problem 2 on page 67 of *Bare Essentials* (2nd Edition). You can read the data into and assign variable and value labels using the following syntax. (NOTE: To save keystrokes, you can copy the syntax from this file and paste it into your syntax file.)

```
DATA LIST LIST /subj grp hairs (3f2.0).
BEGIN DATA.
1 1 12
2 1 14
3 1 28
4 1 3
5 1 22
6 0 5
7 0 10
8 0 20
9 0 2
10 0 12
END DATA.

var lab grp 'Group'.
val lab grp
 0 'Placebo'
 1 'Drug'.
```

Calculating the Mean using the Conceptual Formula

Use the MEANS procedure to obtain the SUM of variable HAIRS for each group, and the overall sum, as well as the group Ns and the overall N. Then use this information in COMPUTE statements to calculate the group means and the overall mean for HAIRS. Bear in mind that the values for variable GRP are 0 = Placebo, 1 = Treatment. If you get stuck, review Tutorial 3. [1]

[1] If everything worked out properly, you should have the following means: 9.80 for Placebo, 15.80 fro Drug, and 12.80 overall.

Calculating the Variance using the Conceptual Formula

In most introductory statistics books, you will encounter two symbols for variance of X: σ_X^2 and s_X^2. It is conventional in statistics to use Greek letters to represent *parameters*, and Arabic letters to represent *statistics*.[2] So σ_X^2 is used to symbolize the variance of a **population**, and s_X^2 to represent the variance of a **sample**. The conceptual formulae for σ_X^2 and s_X^2 are given below.

$$\sigma_X^2 = \frac{\sum (X - \mu_X)^2}{N} = \text{Population Variance}^{3}$$

$$s_X^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \text{Sample Variance}$$

Finally, recall that the sample variance provides an unbiased estimate[4] of the population variance. In symbols, $s_X^2 = \hat{\sigma}_X^2$, where the little hat over σ_X^2 means *is an estimate of*.

Let us now calculate the variance of all of the scores (the X's) in our data file. First, use the following COMPUTE statements to calclate the deviation and squared deviation of each score from the grand mean. Of course, you will have to substitute the variable name you used for the mean of all the scores where I used `mean.all`.

```
compute dev = hairs - mean.all.
compute sqdev = dev**2.
exe.
```

The `**2` you see in the second COMPUTE statement means squared. In general, `X**Y` means *X to the power of Y*.

Then format the new variables to display 4 decimal places, and add some variable labels, as follows:

```
format dev sqdev (f8.4).
var lab
 dev      'Deviation of score from grand mean'
 sqdev    'Squared deviation of score from grand mean'.
```

[2] A parameter is calculated using population data, a statistic using sample data.
[3] Note that Norman and Streiner do not follow the convention of using Greek letters for paramters when they use s^2 and s for the population variance and standard deviation in Equations 3-7 and 3-8. To avoid confusion over whether they are referring to a population or sample variance, you will have to pay attention to the denominator of the equation. The denominator = N for population variances and SDs, and n-1 for sample variances and SDs.
[4] Imagine drawing all possible samples of a given size from some population, and for each sample calculating the sample variance. The mean of those sample variances would be equal to the population variance. If the sample variance was a biased estimator, the mean of the sample variances would **not** be equal to the population variance.

Variable **SQDEV** has the squared deviation of each score from the grand mean of all the scores. Use the MEANS procedure to get the *sum of these squared deviations*. Then use COMPUTE statements to create new variables **SS.TOT** (with variable label 'Total Sum of Squares') and **N.TOT** (with variable label 'Total N'). Set **SS.TOT** equal to the sum of the squared deviations about the grand mean, and **N.TOT** equal to the total number of observations.

Finally, use two COMPUTE statements to calculate variables VAR.POP and VAR.SAMP, then look at your working data file.

```
compute var.pop = ss.tot / n.tot.
compute var.samp = ss.tot / (n.tot - 1).
exe.
format var.pop var.samp (f8.4).
```

Your data file should now show that VAR.POP = 65.16, and VAR.SAMP = 72.40.

Calculating the Standard Deviation

The standard deviation is simply the square root of the variance. Compute new variables SD.POP and SD.SAMP to hold the population and sample standard deviations.

Computing the Variance & SD using the DESCRIPTIVES Procedure

- Click on Analyze→Descriptive Statistics→Descriptives
- Move HAIRS into the Variable(s) box
- Click on Options, and check the Variance box
- Uncheck Minimum and Maximum, then click Continue
- Click PASTE, and run the syntax

The output you generate should look like this:

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
HAIRS	10	12.80	8.509	72.400
Valid N (listwise)	10			

Notice that the Variance and SD calculated by SPSS match the *sample variance* and *sample SD* you calculated using the conceptual formula.

SPSS has no built in procedure for calculating the population variance or standard deviation, because it is assumed that you will always be working with a sample, not an entire population. If you do have the entire population of interest, and want the population variance, you will have to take the variance SPSS calculates, and convert it as follows:

$$\text{Population Variance} = \text{Variance from SPSS} \times \left(\frac{N-1}{N} \right)$$

Multiplying by n-1 gives you the sum of squares (SS), and dividing SS by n gives the population variance. If you are working with the SD:

$$\text{Population SD} = \text{SD from SPSS} \times \sqrt{\frac{N-1}{N}}$$

The SPSS COMPUTE statements to carry out these calculations look like this (assuming `var.spss` and `sd.spss` are the variance and SD computed by SPSS):

```
compute var.pop = var.spss * (n.tot-1)/n.tot.
compute sd.pop = sd.spss * sqrt((n.tot-1)/n.tot).
exe.
```

Calculating the Variance and Standard Deviation for each Group

Using the things you have learned so far, you should now be able to do the following:

- calculate the mean, variance and standard deviation for each group separately using the conceputal formula (calculate both versions of variance and SD—i.e., population and sample)
- calculate the mean, variance and standard deviation for each group separately using one of the procedures built into SPSS (e.g., MEANS); remember that variances and SDs computed by SPSS procedures will be sample statistics (with division by n-1), not population parameters

When you have finished, remember to save your syntax, as you may wish to refer to it later.
