Esssentials of Applied Data Analysis
IPSA-USP Summer School 2017

Building a fake dataset

Leonardo Sangali Barone
leonardo.barone@usp.br

jan/17

Activity 3 - Variance "by hand"

Download the at

https : *//raw.githubusercontent.com/leobarone/IPSA_USP_EADA*2017/*master/Data/fake_data.c*

. Open it using MS Excel, Libre Officce or any other spreadsheet manager.

Calculating the variance "by hand"

The variance is the "Sum of the squared deviations from the mean weighted by the probability of occurence" (or frequency in the sample case).

Let's break the sentence into parts. First, using parenthesis:

"(Sum of the ((squared (deviations from the (mean))) weighted by the probability of occurence))"

From inside to outside, the first "operation" we need to do is (1) calculate the mean (expected value) of the variable. Start by summing the values for all the observations and divide it by the total number of observations. For example, in the case of age:

$$E[X] = \frac{x_1 + x_2 + ... + x_{29} + x_{30}}{30} = \frac{37 + 26 + ... + 32 + 23}{30} = 34.067$$

Once we have the mean, we can (2) calculate the "deviation from the mean". In math notation, deviations from the mean are $x_i - E[X]$. This can be done by subtracting the mean from the values of each observation. To simplify the math, let's pretend $E[X]$ is exactly 34 (when in fact is 34.067).

x_i	$x_i - E[X]$
37	$37 - 34 = 3$
26	$26 - 34 = -8$
...	...
32	$32 - 34 = -2$
26	$23 - 34 = -11$

Since deviations can be positive or negative, the trick is to "square" them, because the square of a number is always positive. The notation for the squared deviations is: $(x_i - E[X])^2$. On Ms. Excel, just multiply the column for itself:

x_i	$x_i - E[X]$	$(x_i - E[X])^2$
37	$37 - 34 = 3$	$3^2 = 9$
26	$26 - 34 = -8$	$(-8)^2 = 64$
...	...	...
32	$32 - 34 = -2$	$(-2)^2 = 4$
26	$23 - 34 = -23$	$(-11)^2 = 121$

Now that we have the squared deviations, we need to weight them by their probability $P(X = x_i)$ or frequency $f(x_i)$of occurence and sum them. Since this is a sample, and all of the observations have the same frequency, the weight is $\frac{1}{n} = \frac{1}{30}$. So we have.

$$Var[x] = \sum_{i=1}^{30} (x_i - E[X])^2 * f(x_1) = 9 * \frac{1}{30} + 64 * \frac{1}{30} + ... + 4 * \frac{1}{30} + 121 * \frac{1}{30} = \frac{9 + 64 + ... + 4 + 121}{30} = 27.098$$

Now, try it by yourself with some other variable using MS Excel and Fake Data.