3.2: Dispersion- Variance and Standard Deviation

Variance, and its square root standard deviation, measure how "wide" or "spread out" a data distribution is. We begin by using the formula definitions; they are slightly different for populations and samples.

***1. Population* Formulae** :

Variance :

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

where N is the size of the population, μ is the mean of the population and x_i is an individual value from the population.

Standard Deviation :

$$\sigma = \sqrt{\sigma^2} \tag{3.2.1}$$

The standard deviation, σ, is a population parameter, we will learn about how to make inferences about population parameters using statistics from samples.

2. Sample Formulae :

Variance :

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

where n = sample size (number of data points), $n-1$ = degrees of freedom for the given sample, $\bar{x}$ and x_i is a data value.

Standard Deviation :

$$s = \sqrt{s^2} \tag{3.2.2}$$

Equations (3.3) and (3.4) are the definitions of variance as the second moment about the mean; you need to determine the means (μ or $\bar{x}$) before you can compute variance with those formulae. They are algebraically equivalent to a "short cut" formula that allow you to compute the variance directly from sums and sums of squares of the data without computing the mean first. For the sample standard deviation (the useful one) the short cut formula is

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)}{n-1}$$

At this point you should figure out how to compute $\bar{x}$, s and σ on your calculator for a given set of data.

Fact (not proved here) : The sample standard deviation s is the "optimal unbiased estimate" of the population standard deviation σ. s is a statistic", the best statistic it turns out, that is used to estimate the population parameter σ. It is the $n-1$ in the denominator that makes s the optimal unbiased estimator of σ. We won't prove that here but we will try and build up a little intuition about what that should be so — why dividing by $n-1$ should be better than dividing by n. ($n-1$ is known as the degrees of freedom of the estimator s). First notice that you can't guess or estimate a value for σ (i.e. compute s) with only one data point. There is no spread of values in a data set of one point! This is part of the reason why the degrees of freedom is $n-1$ and not n. A more direct reason is that you need to remove one piece of information (the mean) from your sample before you can guess σ (compute s).

Coefficient of Variation

The coefficient of variation, CVar, is a "normalized" measure of data spread. It will not be useful for any inferential statistics that we will be doing. It is a pure descriptive statistic. As such it can be useful as a dependent variable but we treat it here as a descriptive statistic that combines the mean and standard deviation. The definition is :

$$\boxed{\mbox{CVar} \ = \ \frac{s}{\bar{x}} \times 100\% \ \mbox{\ \ \ samples}}$$

$$\boxed{\mbox{CVar} \ = \ \frac{\sigma}{\mu} \times 100 \ \% \ \mbox{\ \ \ population}}$$

Example 3.9 : In this example we take the data given in the following table as representing the whole *population* of size $N = 6$. So we use the formula of Equation (3.3) which requires us to sum $(x_i - \mu)^2$.

x_i	$(x_i - \mu)^2$
10	$(10-35)^2$
60	$(60-35)^2$
50	$(50-35)^2$
30	$(30-35)^2$
40	$(40-35)^2$
20	$(20-35)^2$
$\sum x_i = 210$	$\sum (x_i - \mu)^2 = 1750$

Using the sum in the first column we compute the mean :

$$\mu = \frac{\sum x_i}{N} = \frac{210}{6} = 35. \tag{3.2.3}$$

Then with that mean we compute the quantities in the second (calculation) column above and sum them. And then we may compute the variance :

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{1750}{6} = 291.7 \tag{3.2.4}$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{291.7} = 17.1. \tag{3.2.5}$$

Finally, because we can, we compute the coefficient of variation:

$$\text{CVar} = \frac{\sigma}{\mu} \times 100\% = \frac{17.1}{35} \times 100\% = 48.9\%. \tag{3.2.6}$$

□

Example 3.10 : In this example, we have a *sample*. This is the usual circumstance under which we would compute variance and sample standard deviation. We can use either Equation (3.4) or (3.5). Using Equation (3.4) follows the sample procedure that is given in Example 3.9 and we'll leave that as an exercise. Below we'll apply the short-cut formula and see how s may be computed without knowing $\bar{x}$. The dataset is given in the table below in the column to the left of the double line. The columns to the right of the double line are, as usual, our calculation columns. The size of the sample is $n = 6$.

x_i	$(x_i - \bar{x})^2$	x_i^2
11.2		11.2^2 = 125.44
11.9		11.9^2 = 141.161
12.0	exercise	12.0^2 = 144
12.8		12.8^2 = 163.84
13.4		13.4^2 = 179.56
14.3		14.3^2 = 204.49
$\sum x_i = 75.6$		$\sum x_i^2 = 958.94$

To find s compute

$$s^2 = \frac{\sum x_i^2 - (\frac{(\sum x_i)^2}{n})}{n-1} = \frac{958.94 - (\frac{5715.36}{6})}{6-1} = \frac{958.94 - 952.56}{5} = 1.28 \tag{3.2.7}$$

So

$$s = \sqrt{s^2} = \sqrt{1.28} = 1.13. \tag{3.2.8}$$

Note that s^2 is never negative! If it were then you couldn't take the square root to find s. Also not that we have not yet determined the mean. We can do that now:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{75.6}{6} = 12.60. \tag{3.2.9}$$

And with the mean we can then compute

$$\text{CVar } = \frac{s}{\bar{x}} = \frac{1.13}{12.6} \times 100\% = 9.0\% \tag{3.2.10}$$

□

Grouped Sample Formula for Variance

As with the mean, we can compute an approximation of the data variance from frequency table, histogram, data. And again this computation is precise for probability distributions with class widths of one. The grouped sample formula for variance is

$$s^2 = \frac{\sum_{i=1}^{G}(f_i \cdot x_{m_i}^2) - [\frac{(\sum_{i=1}^{G} f_i \cdot x_{m_i})^2}{n}]}{n-1}$$

where G is the number of groups or classes, x_{m_i} is the class center of group i, f_i is the frequency of group i and

$$n = \sum_{i=1}^{G} f_i \tag{3.2.11}$$

is the sample size. Equation (3.6) the short-cut version of the formula. We can also write

$$s^2 = \frac{\sum_{i=1}^{G} f_i (x_{m_i} - \mu)^2}{n-1} \tag{3.2.12}$$

or if we are dealing with a population, and the class width is one so that the class center $X_{m_i} = X_i$,

$$\sigma^2 = \frac{\sum_{i=1}^{G} f_i (X_{m_i} - \mu)^2}{N} \tag{3.2.13}$$

which will be useful when we talk about probability distributions. In fact, let's look ahead a bit and make the frequentist definition for the probability for X_i as $P(X_i) = f_i/N$ (which is the relative frequency of class i) so that

$$\sigma^2 = \sum_{i=1}^{G} P(X_i)(X_i - \mu)^2.$$

If we make the same substitution $P(X_i) = f_i/N$ in the grouped mean formula, Equation (3.1) with population items X and N in place of the sample items x and n, then it becomes

$$\mu = \sum_{i=1}^{G} P(X_i) X_i.$$

More on probability distributions later, for now let's see how we use Equation (3.6) for frequency table data.

Example 3.11 : Given the frequency table data to the left of the double dividing line in the table below, compute the variance and standard deviation of the data using the grouped data formula.

Class	Class Boundaries	Freq, f_i	Class Centre x_{m_i}	$f_i \cdot x_{m_i}$	$x_{m_i}^2$	$f_i \cdot x_{m_i}^2$
1	5.5 – 10.5	1	8	$1 \cdot 8 = 8$	$8^2 = 64$	$1 \cdot 64 = 64$
2	10.5 – 15.5	2	13	$2 \cdot 13 = 26$	$13^2 = 169$	$2 \cdot 169 = 338$
3	15.5 – 20.5	3	18	$3 \cdot 18 = 54$	$18^2 = 324$	$3 \cdot 324 = 972$
4	20.5 – 25.5	5	23	$5 \cdot 23 = 115$	$23^2 = 529$	$5 \cdot 529 = 2645$
5	25.5 – 30.5	4	28	$4 \cdot 28 = 112$	$28^2 = 784$	$4 \cdot 784 = 3136$
6	30.5 – 35.5	3	33	$3 \cdot 33 = 99$	$33^2 = 1089$	$3 \cdot 1089 = 3267$
7	35.5 – 40.5	2	38	$2 \cdot 38 = 76$	$38^2 = 1444$	$2 \cdot 1444 = 2888$

		$\sum f = 20$		$\sum f x_m = 490$		$\sum f x_m^2 = 13310$

The formula

$$s^2 = \frac{\sum (f x_m^2) - [\frac{(\sum f x_m)^2}{n}]}{n-1}$$

tells us that we need the sums of $f x_m^2$ and $f x_m$ after we compute the class centres x_m and their squares x_m^2 — these calculations we do in the columns added to the right of the double bar in the table above. With the sums we compute

\begin{equation*} s^{2} &=& \frac{\sum f_{i} x^2_{m_{i}} - [\frac{(\sum f_{i} x_{m_{i}})^2}{n}]}{n-1} \\ &=& \frac{13310 - [\frac{490^{2}}{20}]}{20-1} = \frac{13310-12005}{19} \\ &=& 68.7. \end{equation*}

So

$$s = \sqrt{s^2} = \sqrt{68.7} = 8.3.$$

The mean, from one of the sums already finished is

$$\bar{x} = \frac{\sum f_i x_{m_i}}{n} = \frac{490}{20} = 24.5$$

and the coefficient of variation is

$$\text{CVar} = \frac{s}{\bar{x}} \times 100\% = \frac{8.3}{24.5} \times 100\% = 33.9\%$$

□

Now is a good time to figure out how to compute $\bar{x}$ and s (and σ) on your calculators.
