Measures of Variation

In the preceding sections several measures which are used to describe the central tendency of a distribution were considered. While the mean, median, etc. give useful information about the center of the data, we also need to know how "spread out" the numbers are about the center.

Consider the following data sets:

								Mean
Set 1:	60	40	30	50	60	40	70	50
Set 2:	50	49	49	51	48	50	53	50

The two data sets given above have a mean of 50, but obviously set 1 is more "spread out" than set 2. How do we express this numerically? The object of measuring this scatter or dispersion is to obtain a single summary figure which adequately exhibits whether the distribution is compact or spread out.

Some of the commonly used measures of dispersion (variation) are: **Range**, **interquartile range**, **variance**, **standard deviation** and **coefficient of variation.**

1. Range

The range is defined as the difference between the highest and smallest observation in the data. It is the crudest measure of dispersion. The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units.

$$\text{Range} = X_{max} - X_{min}$$

Where , X_{max} = highest (maximum) value in the given distribution.

X_{min} = lowest (minimum) value in the given distribution.

In our example given above (the two data sets)

* The range of data in set 1 is 70-30 =40

* The range of data in set 2 is 53-48 =5

2. Quantiles

Another approach that addresses some of the shortcomings of the range is in quantifying the spread in the data set is the use of quantiles or percentiles. Intuitively, the P^{th} percentile is the value Vp such that p percent of the sample points are less than or equal to Vp.

The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the p^{th} percentile, depending on whether np/100 is an integer or not.

Definition: The p^{th} percentile is defined by

(1) The $(k+1)^{th}$ largest sample point if np/100 is not an integer (where k is the largest integer less than np/100)

(2) The average of the $(np/100)^{th}$ and $(np/100 + 1)^{th}$ largest observation is np/100 is an integer.

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentages have the advantage over the range of being less sensitive to outliers and of not being much affected by the sample size (n).

Example: Compute the 10^{th} and 90^{th} percentile for the birth weight data.

Solution: Since $20\times0.1=2$ and $20\times0.9=18$ are integers, the 10^{th} and 90^{th} percentiles are defined by 10^{th} percentile = the average of the 2^{nd} and 3^{rd} largest values = (2581+2759)/2 = 2670 g

90^{th} percentile=the average of the18^{th} and 19^{th} largest values =(3609+3649)/2 = 3629 grams.

We would estimate that 80 percent of birth weights would fall between 2670 g and 3629 g, which gives us an overall feel for the spread of the distribution.

Other quantlies which are particularly useful are the **quartiles** of the distribution. The quartiles divide the distribution into four equal parts.

The second quartile is the median. The interquartile range is the difference between the first and the third quartiles.

To compute it, we first sort the data, in ascending order, then find the data values corresponding to the first quarter of the numbers (first quartile), and then the third quartile. The interquartile range (IQR) is the distance (difference) between these quartiles.

E.g. Given the following data set (age of patients): 18 , 59 , 24 , 42 , 21 , 23 , 24 , 32 find the interquartile range?

1. sort the data from lowest to highest

2. find the bottom and the top quarters of the data

3. find the difference (interquartile range) between the two quartiles.

18 21 23 24 24 32 42 59

1st quartile = The {(n+1)/4}th observation = (2.25)th observation = 21 + (23-21)x .25 = 21.5

3rd quartile = {3/4 (n+1)}th observation = (6.75)th observation =32 + (42-32)x .75 = 39.5

Hence, **IQR** = 39.5 - 21.5 = 18

The interquartile range is a preferable measure to the range. Because it is less prone to distortion by a single large or small value. That is, outliers in the data do not affect the interquartile range. Also, it can be computed when the distribution has open-end classes.

3. Standard Deviation and Variance

Definition: The sample and population standard deviations denoted by S and σ (by convention) respectively are defined as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} = \sqrt{\text{sample variance}} = \text{sample standard deviation}$$

This measure of variation is universally used to show the scatter of the individual measurements around the mean of all the measurements in a given distribution.

Note that the sum of the deviations of the individual observations of a sample about the sample mean is always 0.

The square of the standard deviation is called the **variance**. The variance is a very useful measure of variability because it uses the

information provided by every observation in the sample and also it is very easy to handle mathematically. Its main disadvantage is that the units of variance are the square of the units of the original observations.

Thus if the original observations were, for example, heights in cm then the units of variance of the heights are cm2. The easiest way around this difficulty is to use the square root of the variance (i.e., standard deviation) as a measure of variability.

Weighted Mean of Sample Means and Pooled Standard Deviation

When averaging quantities, it is often necessary to account for the fact that not all of them are equally important in the phenomenon being described. In order to give quantities being averaged there proper degree of importance, it is necessary to assign them relative importance called ***weights***, and then calculate a weighted mean. In general, the weighted mean $\overline{X}w$ of a set of numbers X1, X2, … and Xn, whose relative importance is expressed numerically by a corresponding set of numbers w1, w2, … and wn, is given by

$$\overline{X}_w = \frac{w_1X_1 + w_2X_2 + \ldots + w_nX_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w \times X}{\sum w}$$

On the other hand, where several means (X 's) and standard

deviations (S's) for a variable are available and if we need to compute the overall mean and standard deviation, the weighted mean ($\overline{X}$ w) and pooled standard deviation (Sp) of the entire group consisting of all the samples may be computed as:

$$\overline{X}_w = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \ldots + n_k\overline{X}_k}{n_1 + n_2 + \ldots + n_k}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_k-1)S_k^2}{(n_1-1) + (n_2-1) + \ldots + (n_k-1)}}$$

where ni ,X*i* and Si represent number of observations, mean and standard deviation of each single sample, respectively.

Example: The mean systolic blood pressure was found to be 129.4 and 133.6 mm Hg with standard deviations of 10.6 and 15.2 mm Hg, respectively, for two groups of 12 and 15 men. What is the mean systolic pressure of all the 27 men?

Solution: Given: Group 1: $\overline{X_1}$ = 129.4, S_1= 10.6 and n_1 = 12

Group 2: $\overline{X_2}$ = 133.6, S_2= 15.2 and n_2 = 15

The mean of the 27 men is given by the weighted mean of the two groups.

$$\overline{X}_W = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} = \frac{12(129.4) + 15(133.6)}{12+15} = 131.73 \text{ mm Hg}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1) + (n_2-1)}} = \sqrt{\frac{11\times 10.6^2 + 14\times 15.2^2}{11+14}} = 13.37 \text{ mm Hg}$$

The coefficient of variation

The standard deviation is an absolute measure of deviation of observations around their mean and is expressed with the same unit of the

data. Due to this nature of the standard deviation it is not directly used for comparison purposes with respect to variability. Therefore, it is useful to relate the arithmetic mean and SD together, since, for example, a standard deviation of 10 would mean something different conceptually if the arithmetic mean were 10 than if it were 1000. A special measure called the coefficient of variation, is often used for this purpose.

Definition: The coefficient of variation (CV) is defined by:

$$100\% \times \frac{S}{\overline{X}}$$

The coefficient of variation is most useful in comparing the variability of several different samples, each with different means. This is because a higher variability is usually expected when the mean increases, and the CV is a measure that accounts for this variability.

The coefficient of variation is also useful for comparing the reproducibility of different variables. CV is a relative measure free from unit of measurement. CV remains the same regardless of what units are used, because if the units are changed by a factor C, both the mean and SD change by the factor C; the CV, which is the ratio between them, remains unchanged.

Example: Compute the CV for the birth weight data when they are expressed in either grams or ounces.

Solution: in grams $\overline{X}$ = 3166.9 g, S = 445.3 g,

$$CV=100\% \times \frac{S}{\overline{X}} = 100\% \times \frac{445.3}{3166.9} = 14.1\%$$

If the data were expressed in ounces, $\overline{X}$ =111.71 oz, S=15.7 oz, then

$$CV = 100\% \times \frac{S}{\overline{X}} = 100\% \times \frac{15.7}{111.71} = 14.1\%$$

The Standard error

The standard deviation of such a distribution of means is referred to as **the standard error** of the mean because it represents the distribution of errors (or random fluctuations) in estimating the population mean.

Thus the standard error of the mean is the standard deviation for the distribution of errors or random fluctuations that are likely to occur in estimating the population mean from sample means in a particular situation. The standard error of the mean from a ***single sample:***

$$Se_{\bar{y}} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}}$$

The standard error of the difference between two means (from a ***two sample):.***

$$Se_{(\bar{y}i. - \bar{y}j.)} = \sqrt{\frac{2\,mse}{r}}$$