# Lecture 02 Part 02 - Measures of Dispersion

Data Analysis in CJ (CJUS 6103)

9/6/2021

## Measures of Dispersion (Variability)

Outline

I. Overview of Measures of Dispersion
II. Dispersion in Qualitative Data
III. Dispersion in Quantitative Data
IV. Computational Formulas
V. Parameters v. Statistics

### Overview of Measures of Dispersion

Measures of central tendency pinpoint a single value or category around which others tend to cluster. In other words, it is a "best guess" for the value that is most reflective of the data. In addition to knowing something about central tendency, it is also important to know about how widely other scores are scattered about this central score. Measures of dispersion inform us about how different or dispersed the scores are in a distribution. In other words, they reflect the degree of uncertainty in our data. These measures answer the question, "How good of a 'guess' is it?"

An example illustrates the importance of dispersion. Let's say that you are standing on the bank of a river and need to cross to the other side, but are not a good swimmer. You know that the average depth of the river is three feet (about waist high). This is your measure of central tendency. Do you decide to cross? The answer is that it depends. Knowing that the mean depth is three feet says nothing about the depth of the river at any particular point as you are crossing. This is where measures of dispersion are informative. Suppose that I have measurements of the depth of the river at five foot intervals. Consider the following possibilities.

- Scenario #1: 3 3 3 3 3 3 3 3 3 3
- Scenario #2: 1 2 2 3 3 3 4 4 4 4
- Scenario #3: 1 1 1 1 2 2 2 2 9 9
- Scenario #4: 1 1 1 1 1 1 1 1 1 21

In each of these cases, the mean depth is three feet ($\overline{x} = 3$). However, this shows that knowing the amount of variability in the data will help me decide whether it is wise to attempt the crossing.

## Dispersion in Qualitative Data

The index of qualitative variation, or IQV, is a useful measure of variability with qualitative data. The IQV provides an estimate of how evenly or unevenly the cases are distributed across a given number of categories. I typically do not include this measure in the lecture slides because we do not often use it in practice (and our limited time is better spent on measures you will see more often). It is calculated as:

$$IQV = \frac{\text{Observed Heterogeneity}}{\text{Maximum Heterogeneity}} = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} f_i f_j}{\frac{k(k-1)}{2}\left(\frac{n}{k}\right)^2}$$

An alternative way to derive an estimate of the IQV is:

$$IQV = \frac{k(n^2 - \sum f^2)}{n^2(k-1)}$$

The IQV estimate is interpreted as the percent of maximum heterogeneity, or the proportion of the maximum amount of possible variation. It is important to remember that the IQV is useful insofar as we are comparing the amount of variation in a single variable with more than one subsample (e.g., by sex, by type of homicide, by neighborhood). By itself, the IQV for a single sample is very difficult to interpret.

The variation ratio, or VR, is another measure of variability useful with qualitative data. Whereas the IQV provides an estimate of the extent to which cases are distributed evenly across all categories, the VR provides an estimate of the extent to which cases are not concentrated in the modal category. It is calculated as:

$$VR = 1 - \left(\frac{f_{mode}}{n}\right) = 1 - p_{mode}$$

The VR estimate is interpreted as the proportion of cases that fall outside of the modal category. Like the IQV, the VR for a single sample is not easily interpreted and is thus more useful as a relative measure. Lastly, it is not identified if there is more than one mode in the data.

## Dispersion in Quantitative Data

### Range

The range is the simplest measure of variability to compute with quantitative data. It is calculated as:

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

The simplicity of the range comes at a price. The range is sensitive to outliers or extreme scores, which means that in the presence of outliers it is subject to distortion. Moreover, it provides no information about the middle of the data, only the endpoints.

### Interquartile Range (IQR)

The interquartile range, or IQR, is a somewhat less distorted variation on the range. It is much less sensitive to extreme scores. The IQR measures the range of the middle 50 percent of the distribution, or between the first and third quartiles. It is calculated as follows:

- Step 1: Arrange the data in ascending order.

- Step 2: Find the position of the median using the formula $MP = (n + 1)/2$. Note that the median is often referred to as the second quartile (Q2, or the 50th percentile). Drop off the decimal point (if any) to create the truncated median position (TMP).
- Step 3: Find the quartile position using the formula $QP = (TMP + 1)/2$. Determine the value (or midpoint of two values) associated with the first and third quartiles. This can be done by counting up from the lowest score to get the first quartile, and down from the highest score to get the third quartile.
- Step 4: Compute the interquartile range using the formula $IQR = Q3\breve{}Q1$.

**Mean Deviation**

The mean deviation, or MD, provides the average deviation of the scores about the sample mean. In other words, the mean tells us the average value, and the mean deviation tells us how far away the average value is from the mean. The steps to compute the mean deviation are as follows.

- Step 1: Calculate the sample mean.
- Step 2: Subtract the mean from each score to create a deviation.
- Step 3: Take the absolute value of the deviation.
- Step 4: Add up the deviations.
- Step 5: Divide the deviations by the sample size to compute the mean deviation.

And here is the equation:

$$MD = \frac{\sum \mid x - \overline{x} \mid}{n}$$

**Variance and Standard Deviation**

The variance relies on the "least squares" property of the mean as a measure of variability. This means that by taking deviations from the mean, we can be assured that the variance will be at a minimum. In other words, the variance is the smallest value possible when we take deviations from the mean as opposed to deviations from any other number. The problem with the variance is that, since we take the squared deviation, we change the unit of measurement and as a result make it difficult to interpret what it really means. A solution is to take the square root of the variance to create the standard deviation. This in effect converts the squared deviations back into their original unit of measurement.

Note the similarity between the calculation of the mean deviation and the variance. Why is the variance a preferable measure of variability? A primary reason is that the mean deviation applies equal weight to all deviations from the mean. By squaring the deviations, however, the variance penalizes observations that are further away from the mean (that, and squaring deviations has other valuable distributional/efficiency-related properties).

## Examples

### River Crossing Example Revisited

| **Scenario# 1** | | **Scenario# 2** | | **Scenario# 3** | | **Scenario# 4** | |
|---|---|---|---|---|---|---|---|
| $x$ | $(x-\overline{x})^2$ | $x$ | $(x-\overline{x})^2$ | $x$ | $(x-\overline{x})^2$ | $x$ | $(x-\overline{x})^2$ |
| 3 | 0 | 1 | 4 | 1 | 4 | 1 | 4 |
| 3 | 0 | 2 | 1 | 1 | 4 | 1 | 4 |
| 3 | 0 | 2 | 1 | 1 | 4 | 1 | 4 |
| 3 | 0 | 3 | 0 | 1 | 4 | 1 | 4 |
| 3 | 0 | 3 | 0 | 2 | 1 | 1 | 4 |
| 3 | 0 | 3 | 0 | 2 | 1 | 1 | 4 |
| 3 | 0 | 4 | 1 | 2 | 1 | 1 | 4 |
| 3 | 0 | 4 | 1 | 2 | 1 | 1 | 4 |
| 3 | 0 | 4 | 1 | 9 | 36 | 1 | 4 |
| 3 | 0 | 4 | 1 | 9 | 36 | 21 | 324 |
| | 0 | | 10 | | 92 | | 360 |
| $VR = 0$ | | $VR = 0.6$ | | $VR = undefined$ | | $VR = 0.1$ | |
| $IQR = 3 - 3 = 0$ | | $IQR = 4 - 2 = 2$ | | $IQR = 2 - 1 = 1$ | | $IQR = 1 - 1 = 0$ | |
| $s = \sqrt{\frac{0}{10}} = 0$ | | $s = \sqrt{\frac{10}{10}} = 1.0$ | | $s = \sqrt{\frac{92}{10}} = 3.0$ | | $s = \sqrt{\frac{360}{10}} = 6.0$ | |

### IQR with a Grouped Frequency Distribution

Consider the data in the following frequency distribution. We can obtain the IQR more easily by using the cumulative proportion rather than relying on frequencies to find the quartile positions. Specifically, the first quartile is the first value that equals or exceeds .250, and the third quartile is the first value that equals or exceeds .750.

| Score (X) | $f$ | $p$ | $cp$ |
|---|---|---|---|
| 1 | 3 | .060 | .060 |
| 2 | 4 | .080 | .140 |
| 3 | 5 | .100 | .240 |
| 4 | 10 | .200 | .440 |
| 5 | 7 | .140 | .580 |
| 6 | 6 | .120 | .700 |
| 7 | 6 | .120 | .820 |
| 8 | 5 | .100 | .920 |
| 9 | 3 | .060 | .980 |
| 10 | 1 | .020 | 1.000 |

**Mean Deviation and Standard Deviation/Variance with a Grouped Frequency Distribution**

How about if I want to compute the mean deviation and variance?

| Score (x) | $f$ | $p$ | $x - \overline{x}$ | $f \mid x - \overline{x} \mid$ | $f(x - \overline{x})^2$ |
|---|---|---|---|---|---|
| 1 | 3 | .060 | -4.12 | 12.36 | 50.92 |
| 2 | 4 | .080 | -3.12 | 12.48 | 38.94 |
| 3 | 5 | .100 | -2.12 | 10.60 | 22.47 |
| 4 | 10 | .200 | -1.12 | 11.20 | 12.54 |
| 5 | 7 | .140 | -0.12 | 0.84 | 0.10 |
| 6 | 6 | .120 | 0.88 | 5.28 | 4.65 |
| 7 | 6 | .120 | 1.88 | 11.28 | 21.21 |
| 8 | 5 | .100 | 2.88 | 14.40 | 41.47 |
| 9 | 3 | .060 | 3.88 | 11.64 | 45.16 |
| 10 | 1 | .020 | 4.88 | 4.88 | 23.81 |
| Total | 50 | 1.000 | | 94.96 | 261.28 |

The mean is 256 / 50 = 5.12. Once I calculate deviation scores, I can make appropriate modifications to our formulas to compute the mean deviation and variance:

$$MD = \frac{\sum f* \mid x - \overline{x} \mid}{\sum f} = \frac{94.96}{50} = 1.899$$

$$s^2 = \frac{\sum f * (x - \overline{x})^2}{\sum f} = \frac{261.28}{50} = 5.226$$

When I do so, I find that the mean deviation is 94.96 / 50 = 1.899, and the variance is 261.28 / 50 = 5.226, with a standard deviation of 2.286.

## Computational Formula for Variance & Standard Deviation

With a small number of values or categories, these formulas, called "definitional" formulas, are adequate. However, as we have a larger number of values to compute, we can apply a different formula to arrive at the same result. These are "computational" formulas that require less information to compute.

Requires less information (only $x$ & $x^2$):

$$s^2 = \frac{\sum(x^2) - (\sum x)^2}{n} = \frac{\sum(x^2)}{n} - \overline{x}^2; \text{ where } \overline{x} = \frac{\sum x}{n}$$

$$s^2 = \frac{\sum(w * x^2) - (\sum w * x)^2}{\sum w} = \frac{\sum(w * x^2)}{\sum w} - \overline{x}^2; \text{ where } \overline{x} = \frac{\sum w * x}{\sum w}$$

**Examples Using the Computational Formula**

- Sentence length in months for armed robbery ($n$=40)
  - 36 38 39 47 50 51 51 53 55 55
  - 56 57 60 62 63 64 64 66 67 68
  - 69 70 70 70 71 75 78 79 80 80
  - 81 83 85 86 87 89 95 98 99 99

Mode = 70

Median = 68.5

Mean = 68.7

VR = 1-(3/40) = 1-0.75 = .925

QP = (20+1)/2 = 10.5 $\rightarrow$ IQR = 80.5-55.5 = 25.0

For $s$, first **square** all raw values:

- Sentence length in months for armed robbery ($n$=40)
  - 1296 1444 1521 2209 2500 2601 2601 2809 3025 3025
  - 3136 3249 3600 3844 3969 4096 4096 4356 4489 4624
  - 4761 4900 4900 4900 5041 5625 6084 6241 6400 6400
  - 6561 6889 7225 7396 7569 7921 9025 9604 9801 9801
  - Sum = 199,534

Then, plug the relevant numbers into the formulas:

$$s^2 = \frac{\sum(x^2)}{n} - \overline{x}^2 = \frac{199534}{40} - 68.7^2 = 268.66$$

$$s = \sqrt{268.66} = 16.39$$

Here is an example using the homicide rates from Washington, D.C. and Baltimore from earlier in lecture.

| Washington, DC | Wash$^2$ | Baltimore, MD | Balt$^2$ |
|---|---|---|---|
| 23.5 | 552.25 | 27.6 | 761.76 |
| 31.0 | 961 | 30.6 | 936.36 |
| 36.2 | 1310.44 | 29.5 | 870.25 |
| 59.5 | 3540.25 | 30.6 | 936.36 |
| 71.9 | 5169.61 | 34.3 | 1176.49 |
| 77.8 | 6052.84 | 41.4 | 1713.96 |
| 80.6 | 6496.36 | 40.6 | 1648.36 |
| 75.2 | 5655.04 | 44.3 | 1962.49 |
| 78.5 | 6162.25 | 48.2 | 2323.24 |
| 70.0 | 4900 | 43.4 | 1883.56 |
| 65.2 | 4251.04 | 45.2 | 2043.04 |
| | 45051.08 | | 16255.87 |

$$\begin{aligned} s_{Wash} &= \sqrt{\frac{45051.08}{11} - 60.85^2} \\ &= \sqrt{392.830} = 19.82 \\ s_{Balt} &= \sqrt{\frac{16255.87}{11} - 37.79^2} \\ &= \sqrt{49.722} = 7.05 \end{aligned} \tag{1}$$

## Population Parameters and Sample Statistics

It is important to establish that there is a fundamental difference between what we call, for example, a population mean (a parameter) and what we call a sample mean (a statistics). Generally, parameters are constant and unobservable, while statistics are variable and observable. This distinction is important because we can only compute statistics and assume that they match parameters.

### Population Parameters

In statistics, we generally care about making some inference about an average value and the variation around it, so I will focus here on the differences between population and sample values for means and standard deviations, but please note that the difference between parameters and statistics applies generally to any measure of central tendency or dispersion that we calculate using data from a sample drawn from the population.

The *population* mean and standard deviation are represented by different symbols. The mean is represented by $\mu$ while the standard deviation is represented by $\sigma$. We refer to these as **population parameters**. The formulas are as follows:

$$\mu = \frac{\sum x}{N};\ \sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Meanwhile the sample mean is represented by $\overline{x}$ while the sample standard deviation is represented by $s$. We refer to these as **sample statistics**. The formulas are as follows:

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

We use sample statistics to make some inference about the likely values of population parameters, asking ourselves if $\overline{x}$ is a valid estimate for the true value of $\mu$ or if $s$ is a valid estimate for the true value of $\sigma$.

Fortunately, $\overline{x}$ is an unbiased estimator for $\mu$ such that, accounting for sampling error:

$$\overline{x} = \hat{\mu}$$

*Mu hat* is used because the hat signifies an estimate of the true quantity of interest.

However, for reasons we will discuss later in the semester $s$ is a biased estimate $\sigma$:

$$s \neq \hat{\sigma}$$

We call the term on the right-hand side of the equality **sigma hat**. An unbiased estimate of $\sigma$ substitutes $n-1$ in the denominator rather than $n$.

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