Measures of Variation and Relative Standing

Math 130: Introductory Statistics

1

1.1 Goals

Goals of this class meeting

- Learn about standard measures of variation.
- Learn how to use concepts of center and variability to measure relative standing.

2 Different measures of variation

2.1 Thinking about variation

Thinking about variation

- What does it matter?
- One easy (and pretty useless) measure: range
  - Range = (highest value) - (lowest value)
- Better idea: Measure how far is *each* observation from the mean.

2.2 Deviations from the mean

Deviations from the mean

- Some made up data: 1 6 3 5 9 11.
- What is the mean?

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i$$

$$= \frac{1}{6}(1+6+3+5+9+11)$$

$$= \frac{1}{6}(35) = 5.83$$

Deviations from the mean

- 1 6 3 5 9 11.
- Compute deviations from the mean:

x_i	$x_i - \bar{x}$
1	-4.8333
6	0.1667
3	-2.8333
5	-0.8333
9	3.1667
11	5.1667

- What might be a good overall measure of the amount of variation?

Average deviations from the mean

x_i	$x_i - \bar{x}$
1	-4.8333
6	0.1667
3	-2.8333
5	-0.8333
9	3.1667
11	5.1667

- Average the deviations from the mean?

$$v = \frac{1}{6}\sum_{i=1}^{6}(x_i - \bar{x})$$

$$= \tfrac{1}{6}(-4.8333 + 0.1667 - 2.8333 - 0.8333 + 3.1667 + 5.1667)$$

$$= \frac{1}{6}(0) = 0$$

- Not too useful.

Absolute deviations from the mean

x_i	$x_i - \bar{x}$
1	-4.8333
6	0.1667
3	-2.8333
5	-0.8333
9	3.1667
11	5.1667

- Average the *absolute value* of the deviations from the mean?

$$\tilde{s} = \frac{1}{6}\sum_{i=1}^{6}|x_i - \bar{x}|$$

$$= \begin{array}{c} \frac{1}{6}(4.8333 + 0.1667 + 2.8333 \\ +0.8333 + 3.1667 + 5.1667) \end{array}$$

$$= \frac{1}{6}(17) = 2.833$$

Squared deviations from the mean

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-4.8333	23.3611
6	0.1667	0.0278
3	-2.8333	8.0278
5	-0.8333	0.6944
9	3.1667	10.0278
11	5.1667	26.6944

- Average the *squares* of the deviations from the mean?

$$\sigma^2 = \frac{1}{6}\sum_{i=1}^{6}(x_i - \bar{x})^2$$

$$= \begin{array}{c} \frac{1}{6}(23.3611 + 0.0278 + 8.0278 \\ +0.6944 + 10.0278 + 26.6944) \end{array}$$

$$= \frac{1}{6}(68.833) = 11.472$$

2.3 Variance and standard deviations

Squared deviations from the mean

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-4.8333	23.3611
6	0.1667	0.0278
3	-2.8333	8.0278
5	-0.8333	0.6944
9	3.1667	10.0278
11	5.1667	26.6944

- This measure, σ^2 is known as the **variance**.

- Since using squares can lead to a large result, take square root at the end:

$$\sigma = \sqrt{\sigma^2}$$
$$= \sqrt{11.472} = 3.387$$

- This measure, σ, is known as the **standard deviation**.

Variance and standard deviation

- Formula for the variance and standard deviation of a population:

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

- Formula for the variance and standard deviation of a sample:

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

- For a sample, need to divide by $n - 1$ to make the variance *unbiased*.

Alternative formulas

- Variance of a population:

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

- Variance of a sample:

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

Coefficient of variation

- Measures that have large means tend to have large standard deviations.
- **Coefficient of variation** allows one to compare variability across samples.

$$\text{population CV} = \frac{\sigma}{\mu}$$

$$\text{sample CV} = \frac{s}{\bar{x}}$$

- This measure has no unit, so a CV can compare apples and oranges.

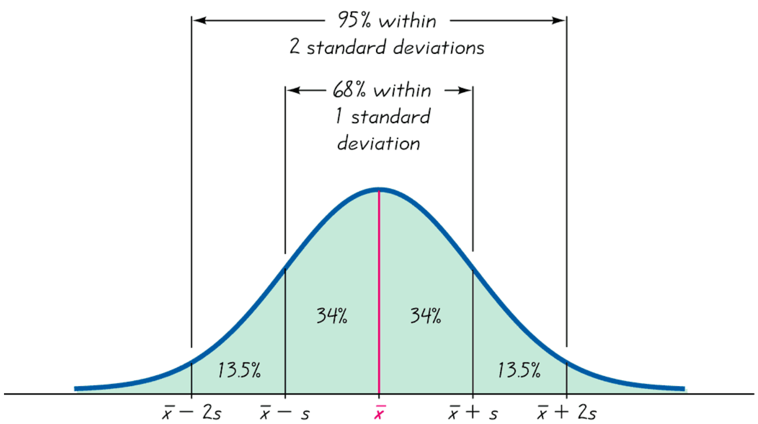
3 Measures of relative standing

3.1 Empirical rule

Measures of relative standing

- Suppose you are interested in how far away an observation is from the mean.
- Standard deviations are a natural yard stick.
- **Empirical rule**: With bell shaped curves,
  - 68% of the data lies within one standard deviation of the mean.
  - 95% of the data lies within two standard deviations of the mean.
  - 99.7% of the data lies within three standard deviations of the mean.

Empirical rule: bell shaped curve



3.2 Chebyshev's theorem

Chebyshev's theorem

Use **Chebyshev's theorem** if you are not certain the data has a bell shaped distribution

- At least $1 - 1/k^2$ of the data lies within k standard deviations of the mean.
- At least 3/4 of the data lies within 2 standard deviations of the mean.
- At least 8/9 of the data lies within 3 standard deviations of the mean.

3.3 Z-scores

Z-scores

- A **standard score** or **z-score** is the number of standard deviations that a value is below the mean.
- For a sample:
$$z = \frac{x - \bar{x}}{s}$$
- For a population:
$$z = \frac{x - \mu}{\sigma}$$

Z-scores and the normal distribution

- **Normal distribution:** very specific bell shaped curve.
- What is the probability of having a z-score above 1?
- What is the probability of having a z-score above 2?
- What about other values? Use the z-table.

Who ever figured that out?