STAT 2593

Lecture 004 - Measures of Variability

Dylan Spicker

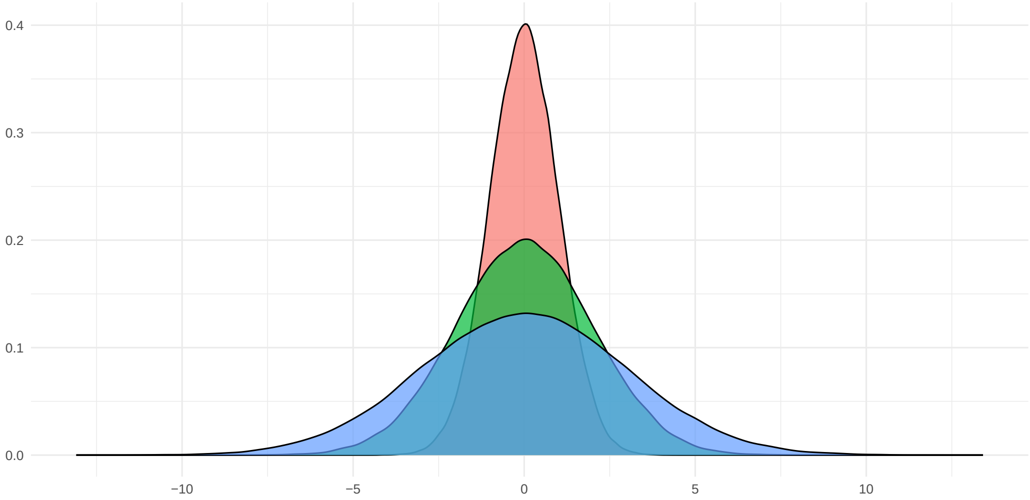
Measures of Variability

Learning Objectives

1. Understand and interpret the range, variance (and standard deviation), and percentiles

2. Understand and interpret boxplots

What does the Location Miss?



When some data are more *spread out* than others, we say that they have higher **variability**.

There is less concentration around the measures of location.

Measures of Variability

- The simplest measure of variability is the **range**, given by

$$\text{range} = x_{\max} - x_{\min}.$$

Measures of Variability

- The simplest measure of variability is the **range**, given by

$$\text{range} = x_{\text{max}} - x_{\text{min}}.$$

- The **sample variance** is given by squared deviations from the mean.

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

Measures of Variability

- The simplest measure of variability is the **range**, given by

$$\text{range} = x_{\text{max}} - x_{\text{min}}.$$

- The **sample variance** is given by squared deviations from the mean.

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

  - Note, the formula you will often see will be *slightly* different. Ignore this for now!

Measures of Variability

- The simplest measure of variability is the **range**, given by

$$\text{range} = x_{\text{max}} - x_{\text{min}}.$$

- The **sample variance** is given by squared deviations from the mean.

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

  - Note, the formula you will often see will be *slightly* different. Ignore this for now!
  - The square root of the sample variance is called the **standard deviation**, $s = \sqrt{s^2}$.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.
- We have that $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.
- We have that $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.
- Adding constants to all of the data will not change the variance.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.
- We have that $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.
- Adding constants to all of the data will not change the variance.
- Multiplying all of the data by a constant, c, multiplies the variance by c^2

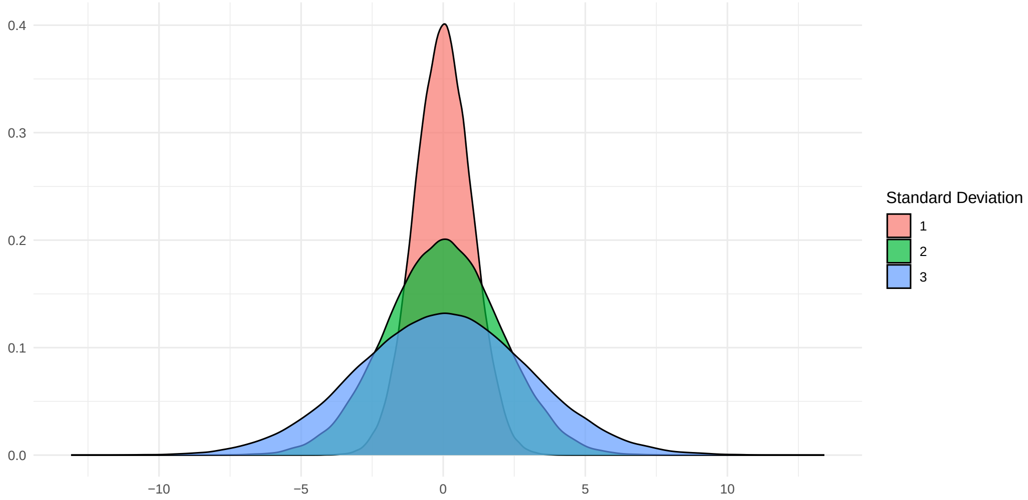
Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.
- We have that $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.
- Adding constants to all of the data will not change the variance.
- Multiplying all of the data by a constant, c, multiplies the variance by c^2
  - The standard deviation will be multiplied by $|c|$.

Properties of the Variance and Standard Deviation

- We have both $s^2 \geq 0$ and $s \geq 0$, with equality only in constant data.
- The standard deviation makes most sense to discuss in conjunction with the mean.
- We define $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, so that $s^2 = \frac{S_{xx}}{n}$.
- We have that $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.
- Adding constants to all of the data will not change the variance.
- Multiplying all of the data by a constant, c, multiplies the variance by c^2
  - The standard deviation will be multiplied by $|c|$.
  - This can be useful for unit conversions.

What does the Location Miss? (Variation!)



Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?
  - This quantity is called the p-**th percentile**.

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?
  - This quantity is called the p-**th percentile**.
  - The median is the 50-th percentile.

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?
  - This quantity is called the p-**th percentile**.
  - The median is the 50-th percentile.
- We call the 25th percentile Q1, and the 75th percentile Q3.

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?
  - This quantity is called the p-**th percentile**.
  - The median is the 50-th percentile.
- We call the 25th percentile Q1, and the 75th percentile Q3.
  - This stands for **q**uartile 1 and 3.

Generalizing from the Median

- Recall that the median divides the data so that 50% is above it and 50% is below it.
- What if we swapped from 50% to p% (below, and $(100 - p)$% above)?
  - This quantity is called the p-**th percentile**.
  - The median is the 50-th percentile.
- We call the 25th percentile Q1, and the 75th percentile Q3.
  - This stands for **q**uartile 1 and 3.
  - These can be computed as the median of the lower and upper half of the data.

Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as $\text{IQR} = Q3 - Q1$.

Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as $\text{IQR} = Q3 - Q1$.
  - IQR is a measure of spread, related to the median.

Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as $\text{IQR} = Q3 - Q1$.
  - IQR is a measure of spread, related to the median.
  - Useful for detecting outliers.

Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as $\text{IQR} = Q3 - Q1$.
  - IQR is a measure of spread, related to the median.
  - Useful for detecting outliers.
  - Data which are $1.5 \times \text{IQR}$ away from the nearest quartile are mild outliers; more than 3 times are extreme outliers.

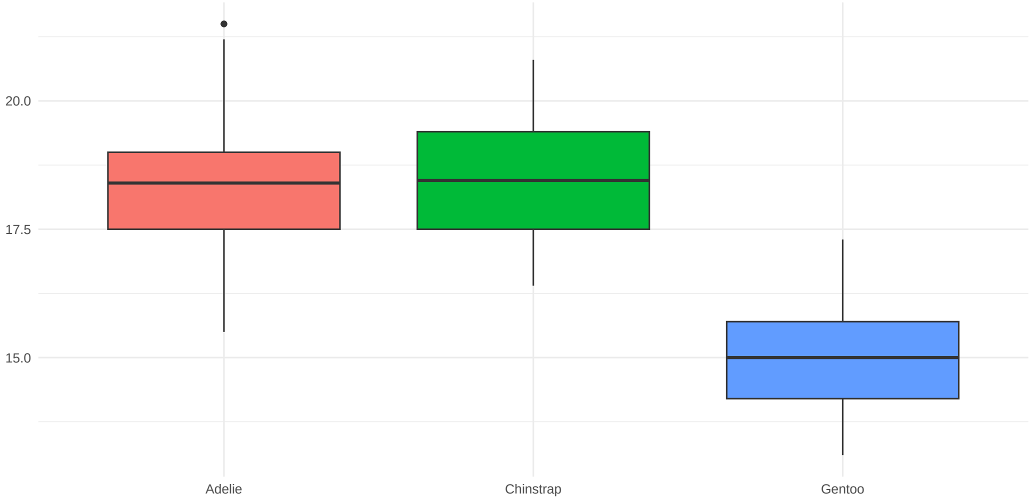
Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as $\text{IQR} = Q3 - Q1$.
  - IQR is a measure of spread, related to the median.
  - Useful for detecting outliers.
  - Data which are $1.5 \times \text{IQR}$ away from the nearest quartile are mild outliers; more than 3 times are extreme outliers.
- If we list min, $Q1$, median, $Q3$, max for data, this is the **five number summary**.

Interquartile Range and Five Number Summary

- The interquartile range, or IQR, is calculated as IQR $= Q3 - Q1$.
  - IQR is a measure of spread, related to the median.
  - Useful for detecting outliers.
  - Data which are $1.5 \times$ IQR away from the nearest quartile are mild outliers; more than 3 times are extreme outliers.
- If we list min, $Q1$, median, $Q3$, max for data, this is the **five number summary**.
  - We can display the five number summary using a **box plot**

Boxplots



Summary

- Location does not capture all the nuance of a particular distribution.
- It is important to consider the spread, or variability as well.
- The range, standard deviation, and variance are all common methods for measuring variability.
- Medians can be generalized to arbitrary values, called percentiles.
- Percentiles are used to form the IQR and the five number summary.
- The five number summary can be graphically represented through box plots.