More on Variance and Standard Deviation

Grinnell College

March 26, 2025

Review

On Monday we introduced the **Normal distribution**

The normal distribution is completely determined by two things

- mean (measure of center)
- standard deviation (measure of spread)

Variance

We are going to spend a bit of time today getting a better understanding of variability, which applies to samples in general and not *just* the normal distribution

- How is it defined
- Relationship between variance and standard deviation
- What is it used for?
  - Dispersion
  - Uncertainty
  - Prediction

Later on the idea of variance is going to help us quantify statements such as, "this is the *best guess* we have"

Definitions

Variance

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

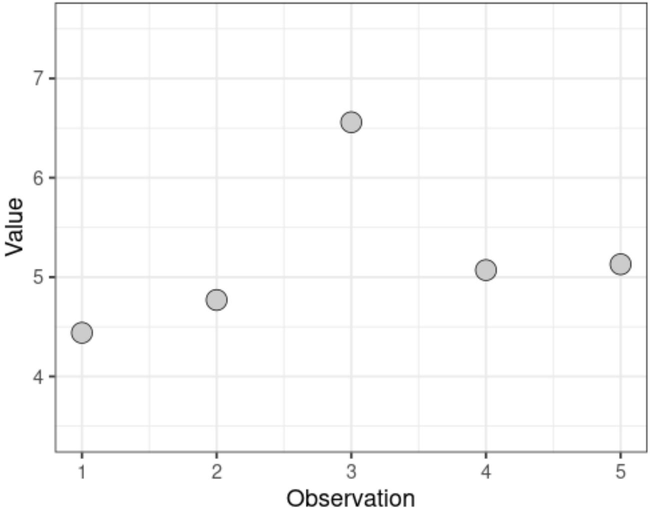
Standard Deviation

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

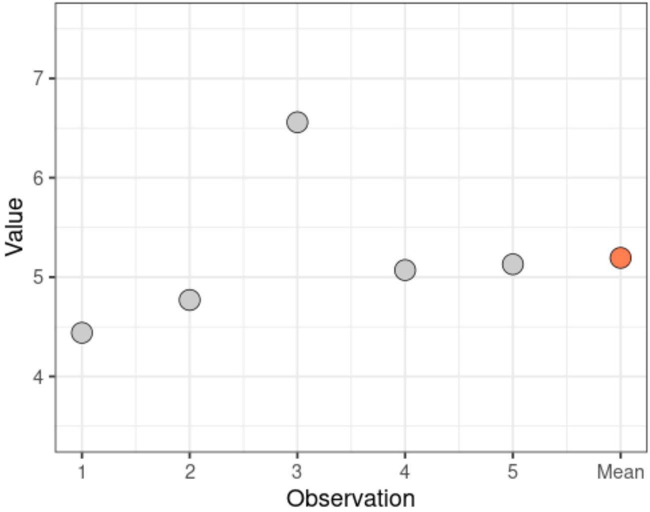
When standard deviation (variance) is calculated from a sample of data, we will refer to it is the *sample* standard deviation (variance). When standard deviation (variance) is referring to the entire population, we will refer to it is the *population* standard deviation (variance)

The *pop.* std. dev. is most often denoted with σ. Since we may want to use the *sample* std. dev. to estimate the population std. dev., sometimes we will use $\hat{\sigma}$ instead of s

Just points

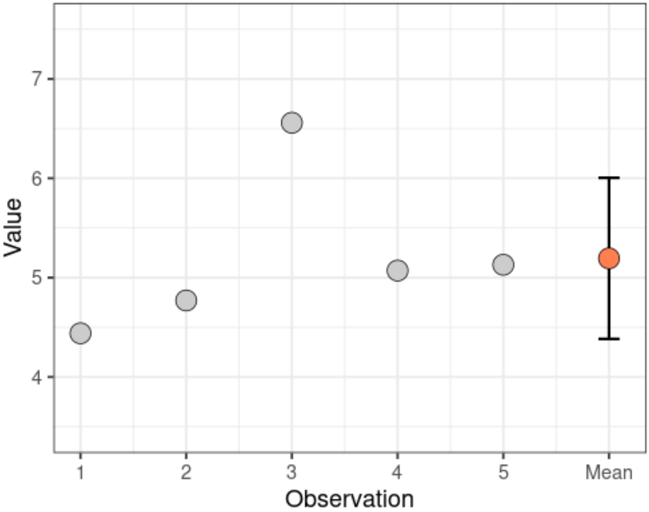


Just points



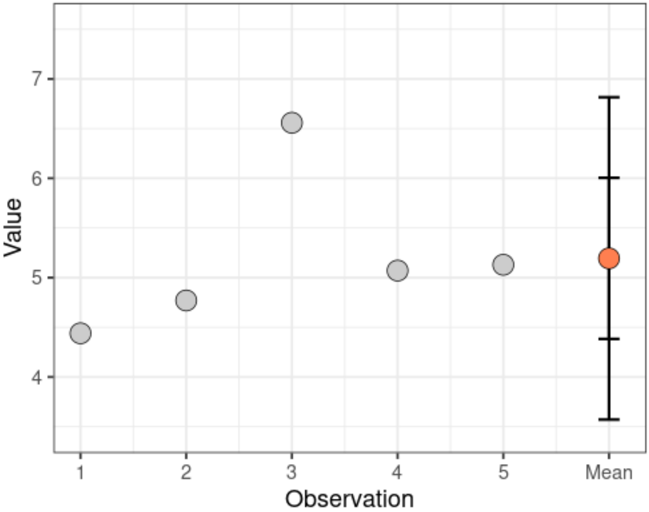
Just points

Here $n = 5$, $\overline{x} = 5.19$ and $\hat{\sigma} = s = 0.81$



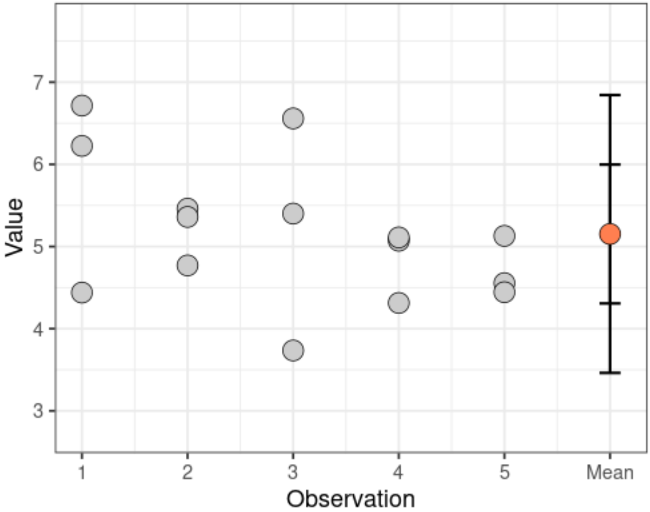
Just points

Here $n = 5$, $\overline{x} = 5.19$ and $\hat{\sigma} = s = 0.81$



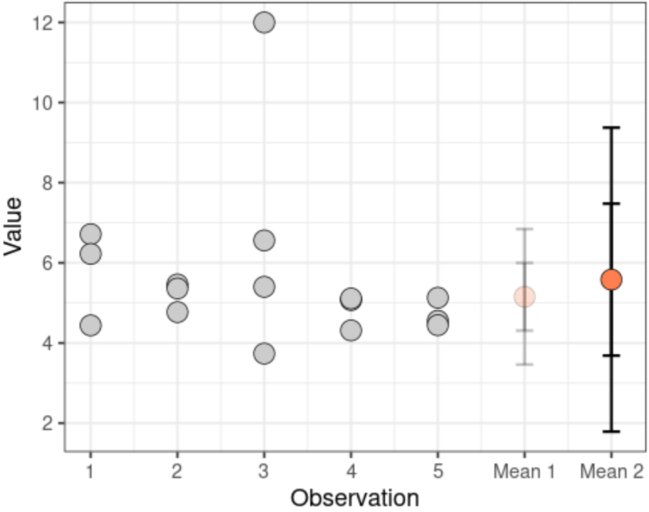
Just points

Note that the standard deviation is not necessarily affected by the number of observations. Here $n = 10$, $\overline{x} = 5.15$, $\hat{\sigma} = 0.83$ ($\approx$ same $\overline{x}$, $\hat{\sigma}$ as before)



Outlier

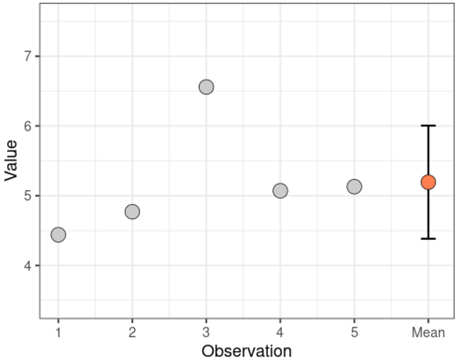
Outliers make the standard deviation larger. Same data + outlier
Now $n = 11$, $\overline{x} = 5.6$ and $\hat{\sigma} = 1.9$



Interpretation

The direct interpretation of standard deviation is "the average deviation/distance of observations to the mean"

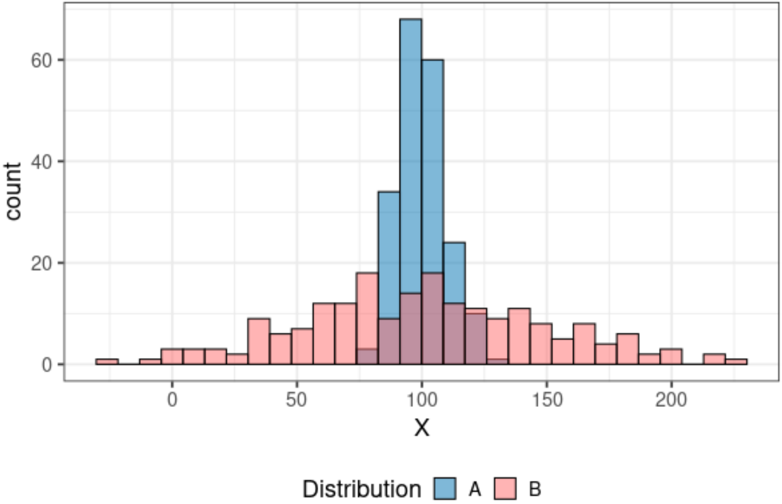
$n = 5$, $\overline{x} = 5.19$ and $\hat{\sigma} = s = 0.81$



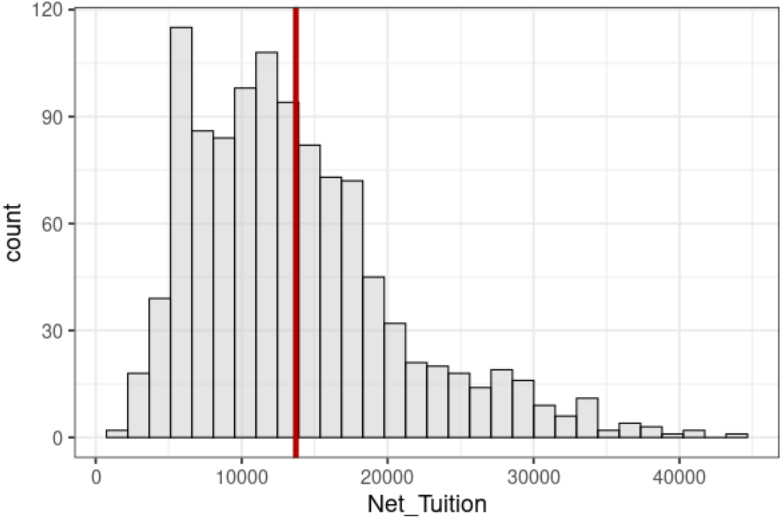
s = .81 $\rightarrow$ average deviation of observations from the mean of 5.19 is 0.81
$\rightarrow$ observations are 0.81 away from the mean, on average

Dispersion

Standard Deviation is a measure of spread. We can use it to compare distributions. Both of these have $\mu = 100$

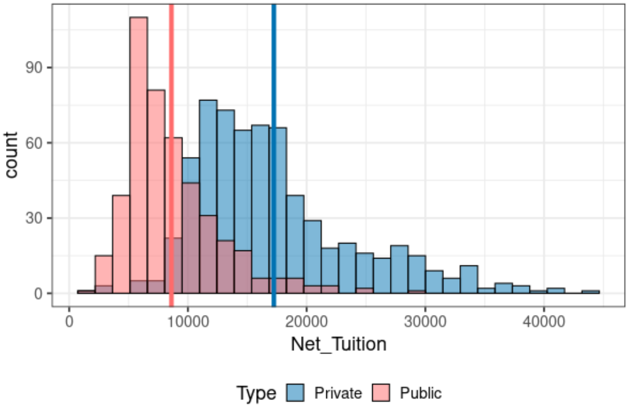


Better Centers?



$$\overline{x} = \$13713,\ \hat{\sigma} = \$7208$$

Better Centers?



$$\overline{x}_{public} = \$8615,\ \hat{\sigma}_{public} = \$3957$$
$$\overline{x}_{private} = \$17244,\ \hat{\sigma}_{private} = \$6829$$

Std. dev's. of both groups are lower than the overall std. dev. of when they were combined (group observations are closer to their own means)

Main Takeaways

Variance and standard deviation are metrics of dispersion and variability
Tell us how far things are from mean
Identify outliers
Allows us to see uncertainty based on a point estimate
Allows us to compare different centers to see if they offer improvement