Basic Statistics Crash Course

Suppose you are interested in knowing how many Netflix shows the average UW–Madison student has seen, so you conduct a survey of a simple random sample of students. You obtain the following data:

	Number of Shows Watched
Student 1	3
Student 2	4
Student 3	3
Student 4	10
Student 5	6
Student 6	5
Student 7	2
Student 8	0
Student 9	3

Five statistics will help you make sense of these data: the **mean**, the **mode**, the **sample variance**, the **sample standard error**, and the **95% confidence interval**.

Mode

The *mode* is the most common response. What is the mode of this data?

Why do we want to know the most common response?

Mean

The *mean* is the average response in the sample. In statistics, we often use $\bar{x}$ (pronounced "x-bar") or μ (the Greek letter mu) to refer to the mean.

Why do we want to know the mean?

We calculate the mean by adding up all of our responses and dividing by the total number of responses. Mathematically, this looks like:

$$\bar{x} = \frac{\sum x_i}{n}$$

But again, all we're doing is adding up our responses and dividing by the total number of responses.

What is the mean of this data, i.e. how many Netflix shows have students in our sample seen on average?

Sample Variance

The *sample variance* is a measure of how spread out our data are. Put differently, how much do they differ, on average, from the mean? In statistics, we use σ (the Greek letter sigma) to refer to the variance.[1]

Why do we want to know the sample variance?

We calculate the sample variance by:

1. calculating the mean
2. subtracting the mean from each response
3. squaring each of these numbers
4. adding these numbers up
5. dividing them by the number of responses minus 1

Mathematically, this looks like:

$$\sigma = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

But again, all we're doing is subtracting the mean from each response, squaring those numbers, adding them together, and dividing by n-1.

What is the sample variance of this data, i.e. how spread out are our responses?

[1] You may have seen slightly different notation if you took statistics in a math or economics department. We'll use σ in this class.

Sample Standard Error

The *sample standard error* is related to the variance and tells us, on average, how far off our observations are from the true mean in the overall population. Put differently, how accurate are our results, on average? In statistics, we use SE to refer to the standard error.

Why do we want to know the sample standard error?

We calculate the sample standard error by:

1. calculating the variance
2. dividing it by n
3. taking the square root of that number

Mathematically, this looks like:

$$\text{SE} = \sqrt{\frac{\sigma}{n}}$$

What is the sample standard error of this data, i.e. how far off are we from the true average number of Netflix shows that UW–Madison students have seen?

95% Confidence Interval

A *confidence interval* is a range of numbers. The true average response (mean) in the population will fall somewhere in that range. We can calculate different confidence intervals depending on how confident we want to be: in political science, we usually want to find a range of numbers that we are 95% confident includes the true mean.

Why is it important to know the confidence interval?

We calculate the 95% confidence interval by:

1. calculating the mean
2. calculating the standard error
3. multiplying the standard error by a z-statistic associated with how confident we want to be. For this class, you don't need to know where the z-statistic comes from. Just know that for 95% confidence interval, it's about 1.96.

4. adding the number from #3 to the mean to get the upper bound of our confidence interval
5. subtracting the number from #3 from the mean to get the lower bound of our confidence interval

Mathematically, this looks like:

$$\texttt{Confidence Interval} = \bar{x} \pm z * SE$$

But again, all we're doing is multiplying the standard error by 1.96 and then adding and subtracting that number to and from the mean.

What is the 95% confidence interval for this data, i.e. in what range of numbers are we 95% confident that the true average number of Netflix shows seen by UW–Madison students falls?