ST 516: Foundations of Data Analytics

Point Estimates

YRBSS data

We're going to consider the population of all 13,583 high school students who participated in the 2013 Youth Risk Behavior Surveillance System (YRBSS)(for more info see:`www.cdc.gov/healthyyouth/data/yrbs/data.htm`). We took a simple random sample of 100 students from this population, a few rows which are shown on the next slide.

We will use this sample, to draw conclusions about the population of YRBSS participants. This is the practice of statistical inference in the broadest sense.

YRBSS data

ID	age	gender	grade	height	weight	helmet	active	lifting
5653	16	female	11	1.50	52.62	never	0	0
9437	17	male	11	1.78	74.84	rarely	7	5
2021	17	male	11	1.75	106.60	never	7	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2325	14	male	9	1.70	55.79	never	1	0

We would like to estimate two features of the high schoolers in YRBSS using the sample.

(1) What is the average height of the YRBSS high schoolers?
(2) On average, how many days per week are YRBSS high schoolers physically active?

These questions focus on means, but the principles of estimation apply to any other parameter too.

Point estimates

We want to estimate the *population mean* based on the sample. The most intuitive way to go about doing this is to simply take the *sample mean*.

That is, to estimate the average height of all YRBSS students, take the average height for the sample:

$$\overline{x}_{height} = \frac{1.50 + 1.78 + \cdots + 1.70}{100} = 1.697$$

The sample mean $\overline{x} = 1.697$ meters (5 feet, 6.8 inches) is called a **point estimate** of the population mean: if we can only choose one value to estimate the population mean, this is our best guess.

Variability in point estimates

Suppose we take a new sample of 100 people and recompute the mean; we will probably not get the exact same answer that we got using this sample. Estimates generally vary from one sample to another, and this **sampling variation** suggests our estimate may be close, but it will not be exactly equal to the parameter.

We described this variation last module using the sampling distribution for the sample average.

Recall: The sampling distribution for the sample mean is centered around the population mean, μ, and has standard deviation $\frac{\sigma}{\sqrt{n}}$.

Additionally if the CLT applies, the shape of the sampling distribution is approximately Normal.

Unbiased point estimates

It's reassuring that the sampling distribution for the sample mean is centered around the parameter we are trying to estimate.

On average, using the sample mean to estimate the population mean will give us the correct value (but of course in any particular experiment we won't get exactly the correct value).

If the sampling distribution of a point estimate is centered around the population parameter we are trying to estimate, we say the point estimate is **unbiased**. For example, the sample mean is an unbiased estimate of the population mean.

Variability in point estimates

The spread of the sampling distribution tells us about the variability of our point estimate.

If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very precise.

If it varies widely from one sample to another, then we should not expect our estimate to be very good.

The standard deviation of the sampling distribution for our estimate tells us how far the typical estimate is away from the actual population value. It describes the typical error of the point estimate, and for this reason we usually call this standard deviation the **standard error (SE)** of the estimate.

Variability of the sample mean

The standard error for the sample mean is,

$$SE_{\overline{x}} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation of the individual observations.

There is one subtle issue here: the population standard deviation is typically unknown.

We often just use the sample standard deviation s instead of σ.

This estimate of σ tends to be sufficiently good when the sample size is at least 30 and the population distribution is not strongly skewed. Next week we will see a method to account for this extra uncertainty in the standard error.

As sample size increases

A particularly desirable property for a point estimate, is that the spread of it's sampling distribution gets smaller, as the sample size increases.

In other words larger samples, lead to more precise estimates.

This is true for the sample mean, since the sample size appears in the denominator of the standard error.

Summary

A statistic used as a guess for a population parameter is called a point estimate.

The properties of a point estimate depend on it's sampling distribution:

- the mean of the sampling distribution tells us if our estimate is unbiased,
- the standard deviation of the sampling distribution tells us about the uncertainty in our estimate, and is called the standard error.

The sample mean is an estimate for the population mean. It is unbiased. We estimate its standard error with $s/\sqrt{n}$.