Point Estimates

- Your friend claims that Urban Dictionary tweets 5000 times per week.
    - We consider 5000 tweets to be a **point estimate** for the true average number of daily tweets.
- The true number is what we would get if we counted all of Urban Dictionary's tweets and divided by the total number of weeks (as of November 2020, the true average was closer to 8000!).
    - This is the **parameter** of interest.

Point Estimates

- When the parameter is a mean, it is denoted by μ (mu).
- The sample mean is denoted $\bar{x}$ (x-hat).
- Unless we collect responses from every case in the population, μ is unknown.
- We use $\bar{x}$ as our estimate of μ.

Sampling Distribution

Sample #	Observations	Mean
1	$x_{1,1}\ x_{1,2}\ \ldots x_{1,n}$	$\bar{x}_1$
2	$x_{2,1}\ x_{2,2}\ \ldots x_{2,n}$	$\bar{x}_2$
3	$x_{3,1}\ x_{3,2}\ \ldots x_{3,n}$	$\bar{x}_3$

Etc.

$\bar{x}$ will change each time we get a new sample. Therefore, when x is a random variable, $\bar{x}$ is also a random variable.

Error

- The difference between the point estimate and the population parameter is called the **error** in the estimate.
- Error consists of two aspects:
  1. sampling error
  2. bias.

Bias

- **Bias** is a *systematic* tendency to over- or under-estimate the population true value.
- E.g., Suppose we were taking a student poll asking about support for the Sac State football team.
- Depending on how we phrased the question, we might end up with very different estimates for the proportion of support.
- We try to minimize bias through thoughtful data collection procedures.

Sampling error

- **Sampling error** is how much an estimate tends to vary between samples.
- This is also referred to as *sampling uncertainty*.
    - In one sample, the estimate might be 1% above the true population value.
    - In another sample, the estimate might be 2% below the truth.
- Our goal is often to quantify this error.

Sampling Distribution

We can characterize the distribution of the sample statistic (the *sampling distribution*) as follows:

- The center is $\bar{x} = \mu$.
- The standard deviation of the sampling distribution (the **standard error**) is $s_{\bar{x}} = \sigma/\sqrt{n}$.
- Symmetric and bell-shaped.

Note that the sampling distribution is never actually observed!

Central Limit Theorem

When observations are independent and the sample size is sufficiently large ($n \geq 30$), the sample mean $\bar{x}$ will tend to follow a normal distribution with mean

$$\mu_{\bar{x}}$$

and standard error

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Example

Find the distribution of $\bar{x}$ given $\mu = 15$, $\sigma = 10$ and $n = 100$.

Example

Estimate how frequently the sample mean $\bar{x}$ should be within 2 units of the population mean, $\mu = 15$.

Central Limit Theorem in the Real World

- In a real-world setting, we almost never know the true population standard deviation.
- Instead, we use the *plug-in principle*:

$$SE_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

This estimate of the standard error tends to be a good approximation of the true standard error.

More About the CLT

- The sampling distribution is always centered at the true population mean μ.
- So $\bar{x}$ is an *unbiased* estimate of μ.
    - (As long as the data are independent!)

More About the CLT

- The variability decreases as the sample size n increases.
- Remember our formula for standard error!
- Estimates based on a larger sample are intuitively more likely to be accurate.