8.1: Estimating a Population Mean

Recall: A *parameter* is a numerical summary of a population; a *statistic* is a numerical summary of a sample. (For example, the population mean and population standard deviation are parameters; the sample mean and sample standard deviation are statistics.)

Definition: A *point estimate* is the value of a statistic that estimates the value of a parameter.

Because it is usually unrealistic to measure or observe the entire population of interest, we use samples to gain information about the population. It seems reasonable to use a sample statistic to estimate a population parameter. However, we would not expect the sample statistic to exactly match the population parameter. How close should we expect them to be?

Confidence intervals:

Definition: A confidence interval (CI) for an unknown parameter is an interval of numbers generated by a point estimate for that parameter.

Definition: The *confidence level* (usually given as a percentage) represents how confident we are that the confidence interval contains the parameter.

If a large number of samples is obtained, and a separate point estimate and confidence interval are generated from each sample, then a 95% confidence level indicates that 95% of all these confidence intervals contain the population parameter.

A confidence interval is obtained by placing a *margin of error* on either side of the point estimate of the parameter.

In other words, the confidence interval consists of: Point estimate $\pm$ margin of error

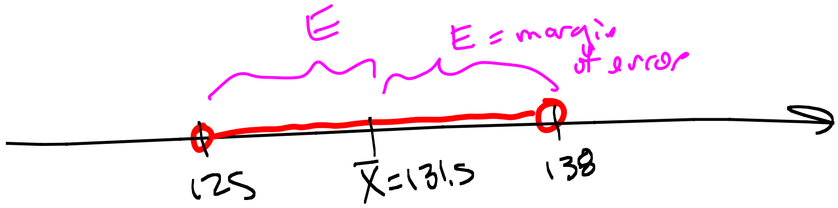
Point estimates for mean and standard deviation:

The point estimate of the population mean μ is the sample mean $\overline{x}$.
The point estimate of the population standard deviation σ is the sample standard deviation *s*.

So, for every sample, the sample mean will be in the center of the confidence interval. If we use *E* to indicate the margin of error, the confidence interval is $\overline{x} \pm E$, or $(\overline{x} - E, \overline{x} + E)$.

Simulations:
http://rpsychologist.com/d3/CI/
(Created by Kristoffer Magnussen; who permits use via Creative Commons License)

http://onlinestatbook.com/stat_sim/conf_interval/index.html
(Rice Virtual Lab in Statistics; public domain resource partially funded by the National Science Foundation; creation led by David Lane of Rice University)



Example 1: Suppose (125, 138) is the 95% confidence interval for μ generated by a sample. Find the sample mean *x* and the margin of error *E*.

$\bar{x}$ is in the middle: $\frac{125+138}{2} = 131.5 = \bar{x}$

Margin of error = distance from $\bar{x}$ to edge: $138 - 131.5 = 6.5$
or $131.5 - 125 = 6.5$

$\boxed{\text{Margin of Error: } E = 6.5 \quad \text{sample mean } \bar{x} = 131.5}$

Recall: The standard deviation of the sampling distribution of the sample means is called the *standard error*. It is calculated by dividing the population standard deviation by the square root of the sample size.

$$\text{Standard error: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Because the margin of error on each side of *x* will be the same, we should be able to write the confidence interval as $(\bar{x} - z_c\sigma_{\bar{x}}, \bar{x} + z_c\sigma_{\bar{x}})$, where $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of the sample means, and z_c is a multiplier that tells us how many standard deviations (of the sampling distribution of the sample means) lie between the sample mean $\bar{x}$ and the edge of the confidence interval. We call this z_c the *critical value* for a *z*-score in the sampling distribution of the sample means.

From the Empirical Rule, for bell-shaped distributions, about 95% of the observations will lie within 2 standard deviations of the mean.

Therefore, for samples of a given size, about 95% of the samples will lie within 2 standard errors of the mean. In other words, for the 95% confidence interval, the critical value z_c is 2. (approximate value)

> 95% Confidence Interval
>
> For a normally distributed variable with population standard deviation σ, using samples of size *n*, the 95% confidence interval for the population mean μ is
>
> $$(\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}),$$
>
> $$\text{where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$
>
> Note: If the variable is not normally distributed, this still applies as long as the sample is sufficiently large, generally for $n \geq 30$.

Example 2: Suppose the population standard deviation for a certain plant species' height is 4.3 cm. A sample of 42 plants of this species resulted in a mean height of 39.6 cm. Determine the 95% confidence interval for the plant species' height.

$\sigma = 4.3$ cm

$n = 42$

$\bar{x} = 39.6$ cm

std. error: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4.3}{\sqrt{42}} \simeq 0.6635$

Upper bound for 95% CI: $\bar{x} + 2\sigma_{\bar{x}} = 39.6 + 2(0.6635)$
(CI = Confidence Interval)
$= 40.927$

Lower bound for 95% CI: $\bar{x} - 2\sigma_{\bar{x}} = 39.6 - 2(0.6635)$
$= 38.273$

$\boxed{\text{95\% CI is } (38.27, 40.93)}$

In words, * sort of *
We're 95% confident the actual population mean is between 38.27 and 40.93.