5: Introduction to Estimation

Parameters and statistics

Statistical inference is the act of generalizing from a sample to a population with calculated degree of certainty. The two forms of statistical inference are **estimation** and **hypothesis testing**. This chapter introduces estimation. The next chapter introduces hypothesis testing.

A statistical **population** represents the set of all possible values for a variable. In practice, we do not study the entire population. Instead, we use data in a **sample** to shed light on the wider population. The process of **generalizing** from the sample to the population *is* statistical inference.

The term **parameter** is used to refer to a numerical characteristic of a population. Examples of parameters include the population mean (μ) and the population standard deviation (σ). A numerical characteristic of the sample is a **statistic**. In this chapter we introduce a particular type of statistic called an **estimate**. The sample mean $\bar{x}$ is the natural estimator of population mean μ. Sample standard deviation *s* is the natural estimator of population standard deviation σ.

Different **symbols** are used to denote parameters and estimates. (e.g., μ versus $\bar{x}$). The parameter is a fixed constant. In contrast, the estimator varies from sample to sample. Other differences are summarized:

	Parameters	Estimators
Source	Population	Sample
Value known?	No	Yes (calculate)
Notation	Greek (μ)	Roman ($\bar{x}$)
Vary from sample to sample	No	Yes
Error-prone	No	Yes

Sampling distribution of a mean (SDM)

If we had the opportunity to take repeated samples from the same population, samples means ($\bar{x}$ s) would vary from sample to sample and form a **sampling distribution means (SDM)**. The SDM is used to help us understand the random behavior of a sample mean. You must use your *imagination* to understand this concept.

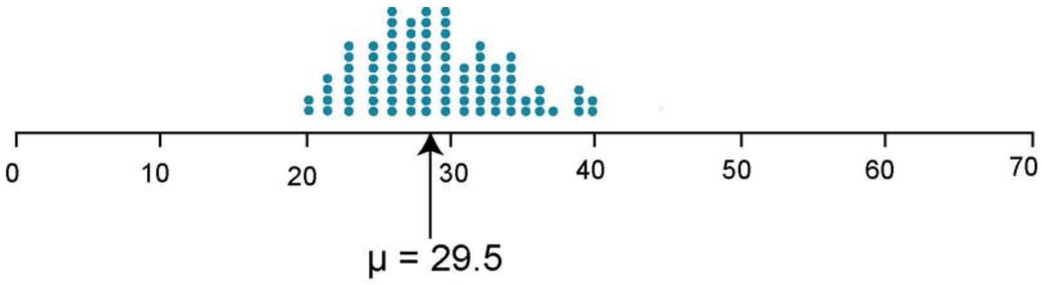
Suppose you want to estimate population mean μ. You sample the population randomly and repeatedly, each time using the same sample size *n*. From each of these repeated samples you calculate independent sample means. The first thing you note is that each of the $\bar{x}$ s differs. The sample means varies, and if your sample is a random sample, it varies randomly. This is not surprising, so you clarify to yourself, "any given sample mean is just an estimate."

Imagine taking all possible samples of size *n* from a population. You then calculate the mean of each of these samples and arrange them to form a distribution. This is the SDM for the variable based on *n*. Of course you would never do this in practice! It is a hypothetical model that allows us to gain an appreciation of the nature of $\bar{x}$. Ultimately, it will allow us to make predictions about the value of population parameter μ.

Let's run a **simulation experiment**. Our simulation will be based on sampling a population of *N* = 600 age values. If you are interested, data for the population is stored in **`populati.sav`**. The population mean age μ = 29.5. The population standard deviation σ = 13.6.

Keep your imagination active.

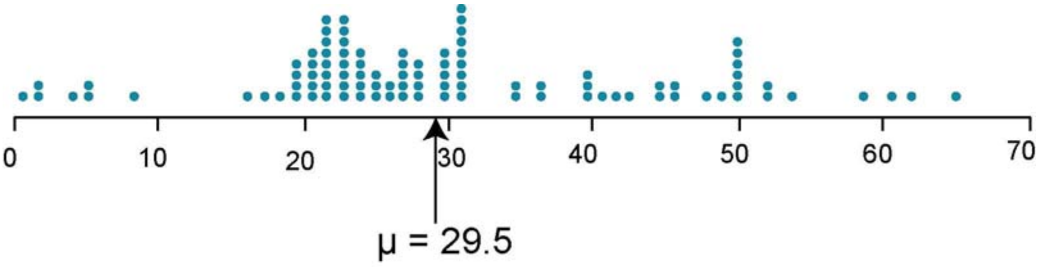
Imagine taking repeated samples, each of *n* = 10, from `populati.sav`. Do this 100 times. In one experiment, it just so happened that the first $\bar{x}$ was 36.4, the second $\bar{x}$ was 30.2, and the third $\bar{x}$ was 24.6. The experiment took an additional 97 SRSs, calculated the means, and plotted them to see their distribution. Here is what the plot looked like:



SDM_n=10.ai

Some of the $\bar{x}$ s fell above the true value of μ, and some fell below. The average of the $\bar{x}$ s was μ: the mean of the SDM is μ. This makes $\bar{x}$ an **unbiased estimator** of μ

How does this SDM compare with the distribution of single observations? I randomly sampled 100 individuals from the population and got the following distribution:



SDM_n=1.ai

This distribution is also centered on μ. However, it is more spread-out than the SDM. **Averages are less variable than individual observations**.

A third simulation experiment was carried out. Now n = 30 for each sample. Here is the distribution of $\bar{x}$s (each based on n = 30).

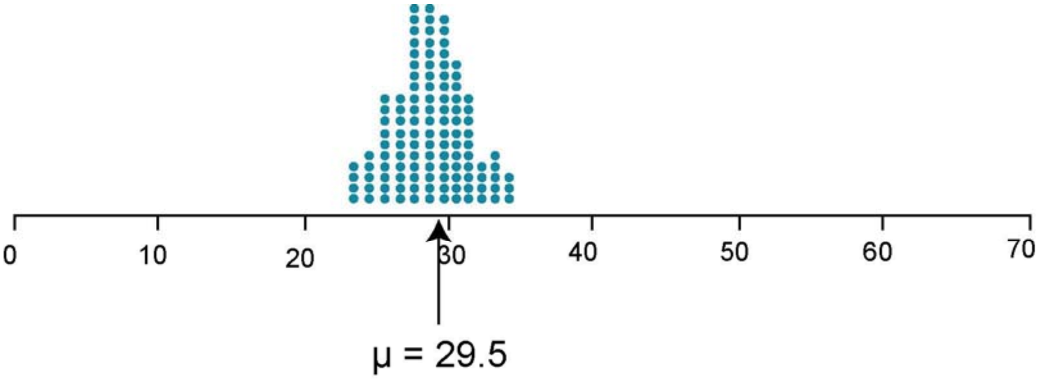
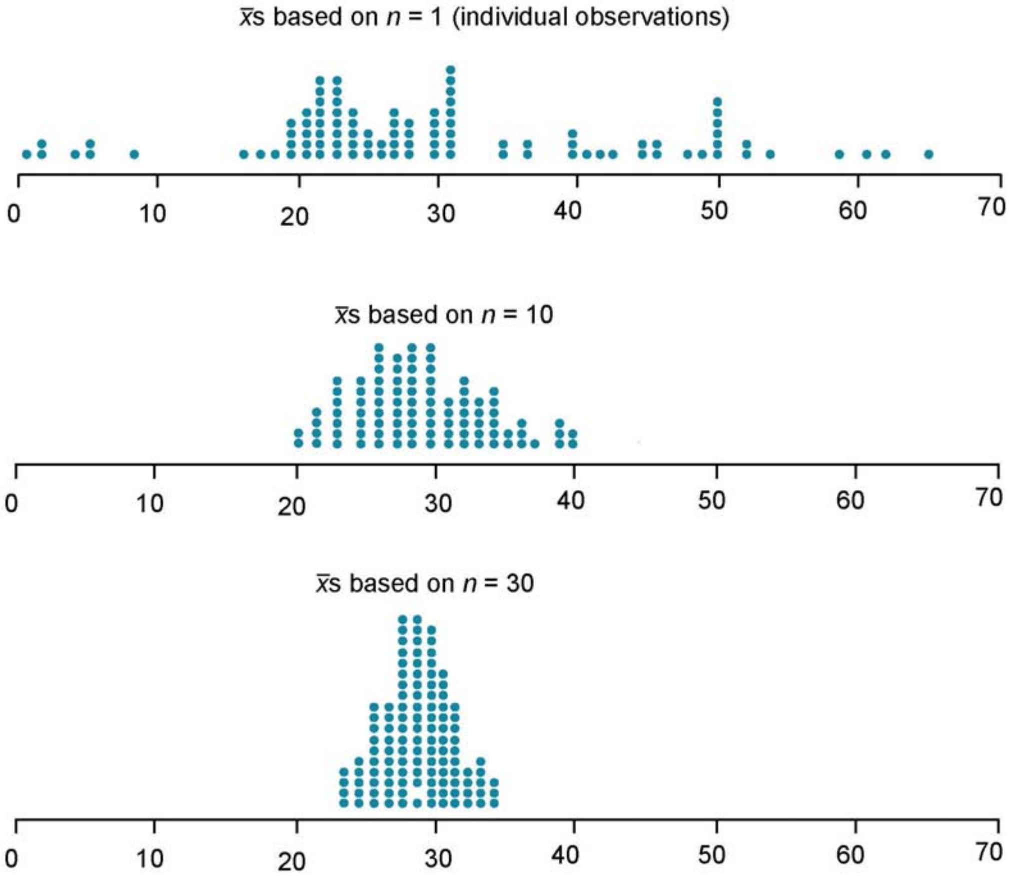


Fig: SDM_n=30.ai

This distribution is also centered on μ and has even less spread than the distribution of $\bar{x}$s based on n = 10.

Here are the experimental results, one plotted on top of the other:


$\bar{x}$s based on n = 1 (individual observations)

$\bar{x}$s based on n = 10

$\bar{x}$s based on n = 30

Fig: SDM_all.ai

With increasing sample size, the SDMs become increasingly Normal. This is due to something called the **Central Limit Theorem (CLT)**. The CLT states that SDMs tend toward Normality as *n* gets large. This justifies use of procedures based on the Normal distribution even when the underlying populations is not Normal. This CLT phenomenon becomes increasing strong with large samples.

Our experiments demonstrate:

1. The sample mean $\bar{x}$ is an unbiased estimate of μ.
2. Averages are less variable than individual observations.
3. The SDM becomes increasingly Normal as *n* increases. (the Central Limit Theorem).

Standard error of the mean

How much can we expect any given sample mean to vary from μ. A way to quantify this variability is by determining the standard deviation of the SDM. This standard deviation is called the **standard error of the mean** (***SEM***).

Large sample sizes produce $\bar{x}$s that closely cluster around the true value of μ. When individual values have standard deviation σ, sample mean $\bar{x}$ based on *n* has deviation (error) σ / √*n*:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

The *SEM* is *inversely* proportion to the square root of *n*. We can call this "the square root law."

Illustrative example. The `AGE` variable in `populati.sav` has standard deviation σ = 13.586.

- A sample of *n* = 1 from this population has *SEM* = 13.586 / √1 = 13.586.
- A sample of *n* = 10 from this population has *SEM* = 13.586 / √10 = 4.296 .
- A sample of *n* = 30 from this population has *SEM* = 13.586 / √30 = 2.480.

As *n* increases, the *SEM* decreases.

Confidence Interval for μ (σ known)

Sample mean $\bar{x}$ is the point estimator of population mean μ. To gain insight into its precision, we surround the point estimate with a **margin of error**:

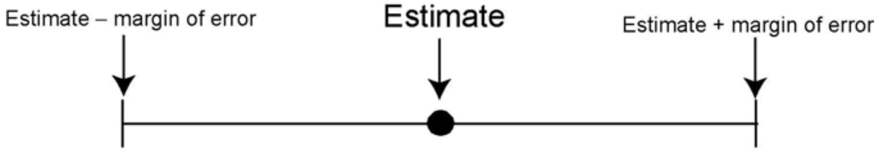


Fig: confidence-interval.ai

This is a confidence interval. The lower end of the confidence interval is the **lower confidence limit (LCL)**. The upper end is the **upper confidence limit (UCL)**. The length of the confidence interval (UCL – LCL) quantifies the precision of the estimate.

A **95% confidence interval for μ** is given by

$$\bar{x} \pm (1.96)(SEM)$$

where $SEM = \sigma / \sqrt{n}$.

Illustrative example. A population has standard deviation σ = 13.586 and unknown mean μ. We take a random sample of 10 observations from this population and observe the following age values: {21, 42, 5, 11, 30, 50, 28, 27, 24, 52}. Based on these 10 observations, $\bar{x}$ = 29.0. We want to estimate population mean μ with 95% confidence. **Solution**: The standard error of the mean *SEM* is 13.586 / √10 = 4.3. The 95% confidence interval for μ = 29.0 ± (1.96)(4.30) = 29.0 ± 8.4 = (20.6, 37.4). We have 95% confidence that the true value of population mean μ will be captured by this interval.

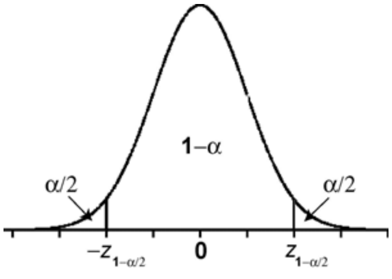
We can calculate a confidence interval at almost any level of confidence. Let **α** represent the chance we are willing to take in *not* capturing μ. This is our "lack of confidence."

Lack of Confidence α	Confidence (1−α)100%
.01	(1−.01)100% = 99%
.05	(1−.05)100% = 95%
.10	(1−.10)100% = 90%

A **(1−α)100% confidence interval for μ** is now given by:

$$\bar{x} \pm (z_{1-\alpha/2})(SEM)$$

The reason we use $z_{1-\alpha/2}$ instead of $z_{1-\alpha}$ in this formula is so we have α in the combined tails of the sampling distribution of means and 1−α area between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$:



normal_alpha_over_2.ai

Here are common confidence levels:

(1−α)100%	α	$z_{1-\alpha/2}$
90%	.10	$z_{1-.10/2} = z_{.95} = 1.645$
95%	.05	$z_{1-.05/2} = z_{.975} = 1.96$
99%	.01	$z_{1-.01/2} = z_{.995} = 2.58$

Illustrative example, 90% confidence interval. Recall that the sample of 10 ages used previously for illustration has *SEM* = 4.30 and $\bar{x}$ = 29.0. The critical value of 10% confidence is $z_{1-.10/2} = z_{.95}$ = 1.645. The 90% confidence interval for μ = 29.0 ± (1.645)(4.30) = 29.0 ± 7.1 = (21.9, 36.1).

Illustrative example, 95% confidence interval. The critical z value for 95% confidence is $z_{1-.05/2} = z_{.975}$ = 1.96. The 95% confidence interval for μ = 29.0 ± (1.96)(4.30) = 29.0 ± 8.4 = (20.6, 37.4).

Illustrative example, 99% confidence interval. Using the same data, α = .01 for 99% confidence. The 99% confidence interval for μ = 29.0 ± (2.58)(4.30) = 29.0 ± 11.1 = (17.9, 40.1).

Here are confidence interval lengths of the three intervals just calculated:

Confidence Level	Confidence Interval	Confidence Interval Length
90%	(21.9, 36.1)	36.1 − 21.9 = 14.2
95%	(20.6, 37.4)	37.4 − 20.6 = 16.8
99%	(17.9, 40.1)	40.1 − 17.9 = 22.2

The confidence interval length grows as the level of confidence increases from 90% to 95% to 99%.This is because there is a trade-off between the confidence level and margin of error. To obtain a smaller margin of error, you must be willing to accept lower confidence.

Sample Size Requirements

One of the questions a statistician often faces is “How much data should be collected?” Collecting too much data is a waste of time, and collecting too little data renders an estimate too imprecise to be useful.

To address the question of sample size requirements, let *d* represent the **margin of error** of a confidence interval. This is half the confidence interval length and, for a 95% confidence interval for μ from a Normal population with known standard deviation is

$$d = 2\frac{\sigma}{\sqrt{n}}$$

Solve this equation for n to get the sample size requirement to achieve margin of error *d* for the 95% confidence interval:

$$n = \frac{4\sigma^2}{d^2}$$

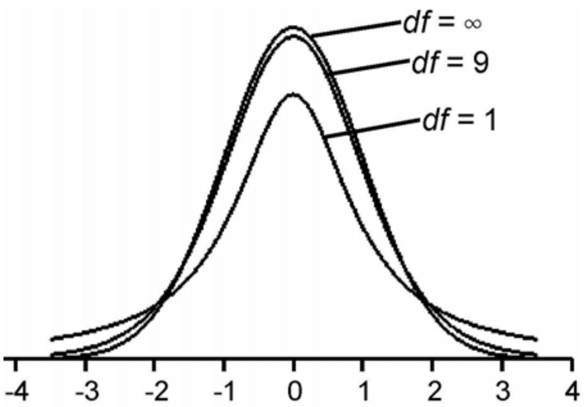
Illustrative examples. Suppose we have a variable with standard deviation σ = 15. The samples size required to achieve a margin of error of 5 is $n = \frac{4\sigma^2}{d^2} = \frac{4 \cdot 15^2}{5^2}$ = 36. The samples size required to achieve a margin of error of 2.5 is $n = \frac{4\sigma^2}{d^2} = \frac{4 \cdot 15^2}{2.5^2}$ = 144.

Student's *t* Distributions

Methods discussed so far have depended on knowing population standard deviation σ from information provided prior to the study. In practice, however, this is seldom the case. In such instances, we use *s* as our estimate σ and use a modification of the *z* distribution s known as Student's *t* distribution as part of the estimate.

Student's *t* distributions are a family of probability models that resemble the *z* distribution. Like the *z* distribution, *t* distributions are bell-shaped and centered on 0. However, *t* distributions have broader tails than *z* distributions.

Keep in mind that there is more than one *t* distribution. Each member of the *t* distribution family is identified by a parameter known as its **degree of freedom** (*df*). *T* distributions with *df* = 1, *df* = 9, and *df* = infinity are pictured below.



Notice that *t* distributions become increasing Normal as the degrees of freedom rises. Also notice that *t* distributions with few degrees of freedom have relatively broad tails. This compensates for the additional uncertainty introduced by using *s* instead of σ in the course of our inference.

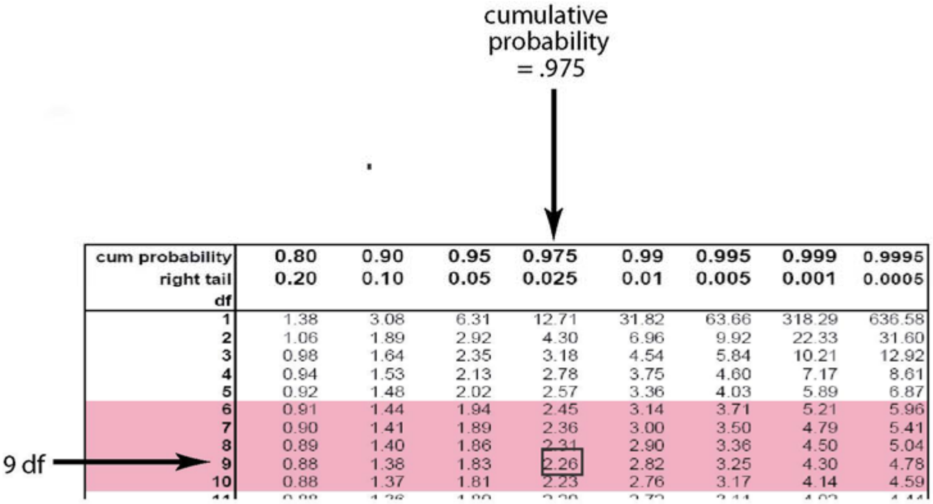
A *t* distribution with infinite degrees of freedom *is* a standard Normal distribution. For all intensive purposes, it makes little difference whether we use a *t* or *z* once that sample size gets over 1000.

Notation: Let t_{df} denote a *t* distribution with *df* degrees of freedom. For example, t_9 will denote a *t* distribution with 9 degrees of freedom.

t Table

We use a *t* table to determine the critical values *t* scores on *t* distributions. T table differ from *z* tables in the way they are set up.

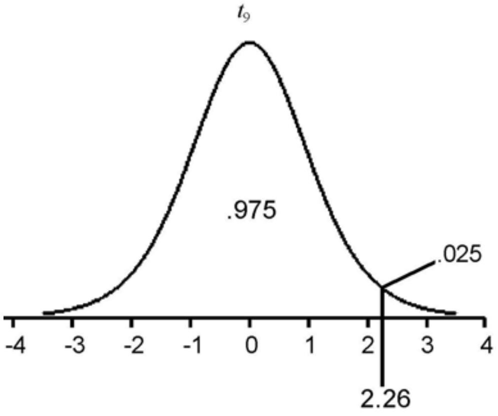
Each row in the *t* table contains *t* scores for a *t* distribution with a given number of degrees of freedom. The cumulative probability for each landmark and the upper tail probability are given in the heading row of the table. The bottom row of table provides confidence levels, or the area under the curve between ±*t* scores. Here's a portion of our *t* table:


cumulative probability = .975
9 df

cum probability	0.80	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
right tail	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
df								
1	1.38	3.08	6.31	12.71	31.82	63.66	318.29	636.58
2	1.06	1.89	2.92	4.30	6.96	9.92	22.33	31.60
3	0.98	1.64	2.35	3.18	4.54	5.84	10.21	12.92
4	0.94	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	0.92	1.48	2.02	2.57	3.36	4.03	5.89	6.87
6	0.91	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.90	1.41	1.89	2.36	3.00	3.50	4.79	5.41
8	0.89	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	0.88	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	0.88	1.37	1.81	2.23	2.76	3.17	4.14	4.59

t-table figure.ai

The *t* score for the 97.5th percentile (cumulative probability) for a t_9 is indicated in the table. This is what the *t* score looks like on the *t* density curve:



To aid in the process of getting probabilities from *t* scores, we will use the notation $t_{\mathrm{df},p}$ to denote the *t* score with *df* degrees of freedom and cumulative probability *p*. For example, $t_{9,\ .975}$ = 2.26, as shown above. Notice that $t_{9,\ .975}$ has a right-tail probability of .025.

Confidence Interval for μ (*t* procedure)

We want to estimate μ based on a SRS of size *n* from the population. The population is assumed to be Normal or the sample is large enough to allow us to assume that the sampling distribution of x-bar approaches Normality (Central Limit Theorem).

Since population standard deviation σ is not know, we use sample standard deviation *s* to calculate the **estimated standard error of the mean** (*sem*):

$$sem = \frac{s}{\sqrt{n}}$$

A (1 − α)100% confidence interval for μ is given by:

$$\bar{x} \pm (t_{n-1,1-\alpha/2})(sem)$$

where $\bar{x}$ represents the sample mean, $t_{n\text{-}1,1-\alpha/2}$ represents a *t* score with $n-1$ df and cumulative probability 1−(α/2), and *sem* represents the estimated standard error of the mean, above.

Illustrative Example (`%ideal.sav`). We measure body weight in a sample 18 diabetics. Each individual's body weights is divided by their ideal body weight and multiplied by 100, so that the data represent percentage of ideal body weight. For example, a score of 100 represent 100% of ideal body weight. Data are {107, 119, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 100, 125, 114, 95, 117} (Source: Pagano & Gauvreau, 1993, p. 208; Saudek et al., 1989). We want to estimate population mean μ with 95% confidence.

Solution: The first step is to calculate the sample mean and standard deviation: $\bar{x}$ = 112.778 and *s* = 14.424 (calculations not shown). Sample size *n* = 18, so *sem* = *s* / √*n* = 14.424 / √18 = 3.400. For 95% confidence, α = .05, and $t_{n\text{-}1,1-\alpha/2} = t_{18\text{-}1,1\text{-}.05/2} = t_{17,.975}$ = 2.11 (from *t* table). The 95% CI for μ is 112.78 ± (2.11)(3.400) = 112.78 ± 7.17 = (105.61, 119.95).

Comparison of confidence intervals based on the *Z* and *t* distribution. The confidence intervals of μ presented in this chapter both have the general form: estimate ± margin of error. In both instances, the estimate is provided by $\bar{x}$. Also, in both instances the margin of error = critical value × standard error. When σ is known, the critical value is $z_{1-\alpha/2}$ and the standard error is σ / √*n*. When σ is not know, the critical value is $t_{n\text{-}1,1-\alpha/2}$ and the standard error is *s* / √*n*

Vocabulary

Sampling distribution of the mean: the hypothetical distribution of sample means that would occur from repeated independent samples of size *n* from the population.

Central Limit Theorem: an axiom that states that the sampling distribution of the mean will tend toward normality when n is large.

Law of large numbers: the law of large numbers states that the larger the sample, the more likely it is to represent the population from which it was drawn -- specifically, the more likely it is that the sample mean will be close to the population mean.

Standard error of the mean (SEM or sem): a statistic that indicates how greatly a particular sample mean is likely to differ from the mean of the population.

Margin of error (*d*): the plus-or-minus wiggle-room drawn around the estimate in order to locate the location of the parameter; half the confidence interval width.