5: Introduction to Estimation

Table of Contents

Acronyms and symbols

$\hat{q}$	complement of sample proportion
$\bar{x}$	sample mean
$\hat{p}$	sample proportion
$1 - \alpha$	confidence level
CI	confidence interval
LCL	lower confidence limit
m	margin of error
n	sample size
NHTS	Null hypothesis tests of significance
p	binomial (“population proportion”) parameter
s	sample standard deviation
SDM	sampling distribution of mean (hypothetical probability model)
SEM	standard error of the mean
SEP	standard error of the proportion
UCL	upper confidence limit
α	alpha level
μ	expected value (“population mean”) parameter
σ	standard deviation parameter

Statistical inference

Statistical inference is the act of generalizing from the data ("sample") to larger phenomenon ("population") with calculated degree of certainty. The act of **generalizing** and deriving statistical judgements *is* the process of inference.

Note: There is a distinction between *causal inference* and *statistical inference*. Here we consider only *statistical inference*.

The **two common forms of statistical inference** are:

- Estimation
- Null hypothesis tests of significance (NHTS)

Both estimation and NHTS are used to infer parameters. A **parameter** is a statistical constant that describes something about the population or phenomena being studied.

Examples of parameters include:

- Probability of "success" p (also called the population proportion)
- Expected value μ (also called the population mean)
- Standard deviation σ (also called the population standard deviation).

Notice how each of the above parameters describes probability mass or density functions we have studied.

Point estimates are single points that infer parameters directly. For example,

- sample proportion $\hat{p}$ ("p hat") is the point estimate of p
- sample mean $\bar{x}$ ("x bar") is the point estimate of μ
- sample standard deviation s is the point estimate of σ

Notice the use of different symbols to distinguish estimates and parameters. More importantly, point estimates and parameters are fundamentally different.

- Point estimate are calculated from the data; parameter are <u>not</u>.
- Point estimate vary from study to study; parameter are <u>not</u>.
- Point estimates are random variables: parameters are constants.

Estimating μ with confidence

Sampling distribution of the mean

Although point estimates are valuable reflections of parameters, they do not provide information about sample to sample variability. For example, we ask How precise is $\bar{x}$ as estimate of μ? How much can we expect any given $\bar{x}$ to vary from μ?

The variability of $\bar{x}$ as the point estimate of μ starts by considering a hypothetical distribution called the **sampling distribution of a mean** (**SDM** for short). Understanding the SDM is difficult because it is based on a thought experiment that doesn't occur in actuality, being a hypothetical distribution based on mathematical laws and probabilities. The SDM *imagines* what would happen if we took repeated samples of the same size from the same (or similar) populations done under the identical conditions. From this hypothetical experiment we "build" a pmf or pdf that is used to determine probabilities for various hypothetical outcomes.

Without going into detail, the SDM reveals that:

- $\bar{x}$ as an unbiased point estimate of μ
- the SDM mimics a normal pdf when the sample is large or "adequate"
- the standard deviation of the SDM is equal to $\sigma/\sqrt{n}$

This last statistic—sometimes called the **standard error of the mean (*SEM*)—**predicts how closely the $\bar{x}$s in the SDM are likely to cluster around the value of μ and is thus a reflection of the precision of $\bar{x}$ as an estimate of μ.

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

This particular version of the *SEM* is based on σ and <u>not</u> on sample standard deviation *s*. Keep in mind that σ is NOT calculated from the data and must therefore be derived from some other source. Also note that the *SEM* is inversely proportion to the square root of *n*

Numerical example. Suppose we have a measurement that has σ = 10.

- A sample of $n = 1$ for this variable derives *SEM* = $\sigma/\sqrt{n}$ = 10 / √1 = 10
- A sample of $n = 4$ derives *SEM* = $\sigma/\sqrt{n}$ = 10 / √4 = 5
- A sample of $n = 16$ derives *SEM* = $\sigma/\sqrt{n}$ = 10 / √16 = 2.5

Each time we quadruple *n*, the *SEM* is cut in half. This is called the **square root law—** the precision of the mean is inversely proportional to the square root of the sample size.

Confidence Interval for μ when σ is known before hand

To gain further insight into μ, we surround the point estimate with a **margin of error**:

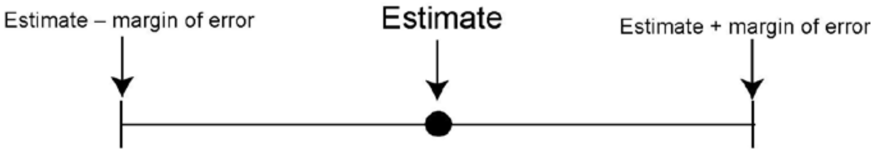


Fig: confidence-interval.ai

This forms a **confidence interval (CI)**. The lower end of the confidence interval is called the **lower confidence limit (LCL)**. The upper end is the **upper confidence limit (UCL)**.

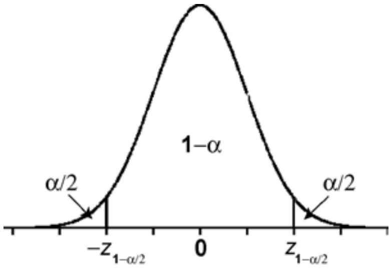
Note: The margin of error is the plus-or-minus wiggle-room drawn around the point estimate; it is equal to half the confidence interval length.

Let (1−α)100% represent the **confidence level** of a confidence interval. The α ("alpha") level represents the "lack of confidence" and is the chance the researcher is willing to take in *not* capturing the value of the parameter.

A (1−α)100% CI for μ is given by:

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

The $z_{1\text{-}\alpha/2}$ in this formula is the z quantile association with a 1 – α level of confidence. The reason we use $z_{1\text{-}\alpha/2}$ instead of $z_{1\text{-}\alpha}$ in this formula is because the random error (imprecision) is split between underestimates (left tail of the SDM) and overestimates (right tail of the SDM). The confidence level 1−α area lies between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$:



normal_alpha_over_2.ai

You may use the z/t table on the *StatPrimer* website to determine z quantiles for various levels of confidence. Here are the common levels of confidence and their associated alpha levels and *z* quantiles:

(1−α)100%	α	$z_{1\text{-}\alpha/2}$
90%	.10	$z_{1-.10/2} = z_{.95} = 1.64$
95%	.05	$z_{1-.05/2} = z_{.975} = 1.96$
99%	.01	$z_{1-.01/2} = z_{.995} = 2.58$

Numerical example, 90% CI for μ. Suppose we have a sample of *n* = 10 with *SEM* = 4.30 and $\bar{x}$ = 29.0. The *z* quantile for 10% confidence is $z_{1-.10/2} = z_{.95} = 1.64$ and the 90% CI for μ = 29.0 ± (1.64)(4.30) = 29.0 ± 7.1 = (21.9, 36.1). We use this inference to address population mean μ and NOT about sample mean $\bar{x}$. Note that the margin of error for this estimate is ±7.1.

Numerical example, 95% CI for μ. The *z* quantile for 95% confidence is $z_{1-.05/2} = z_{.975} = 1.96$. The 95% CI for μ = 29.0 ± (1.96)(4.30) = 29.0 ± 8.4 = (20.6, 37.4). Note that the margin of error for this estimate is ±8.4.

Numerical example, 99% CI for μ. Using the same data, α = .01 for 99% confidence and the 99% CI for μ = 29.0 ± (2.58)(4.30) = 29.0 ± 11.1 = (17.9, 40.1). Note that the margin of error for this estimate is ±11.1.

Here are confidence interval lengths (UCL – LCL) 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 and margin of error. You can achieve a smaller margin of error if you are willing to pay the price of less confidence. Therefore, as Dr. Evil might say, 95% is “pretty standard.”

Numerical example. Suppose a population has σ = 15 (not calculated, but known ahead of time) and unknown mean μ. We take a random sample of 10 observations from this population and observe the following values: {21, 42, 5, 11, 30, 50, 28, 27, 24, 52}. Based on these 10 observations, $\bar{x}$ = 29.0, *SEM* = 15/√10 = 4.73 and a 95% CI for μ = 29.0 ± (1.96)(4.73) = 29.0 ± 9.27 = (19.73, 38.27).

Interpretation notes:

- The margin of error is the “plus or minus” value of 9.27.
- We use these confidence limits to address population mean μ and NOT about sample mean $\bar{x}$.

Sample Size Requirements for estimating µ with confidence

One of the questions we often faces is “How much data should be collected?” Collecting too much data is a waste of time and money. Also, by collecting fewer data points we can devote more time and energy into making these measurements accuracy. However, collecting too little data renders our estimate too imprecise to be useful.

To address the question of sample size requirements, let *m* represent the desired **margin of error** of an estimate. This is equivalent to half the ultimate confidence interval length. Note that $m = z^2_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. Solving this equation for *n* derives a required sample size of

$$n = z^2_{1-\alpha/2}\frac{\sigma^2}{m^2}$$

Always round results from this formula up to ensure that you have a margin of error of no greater than *m*.

Note that to determine the sample size requirements for estimating µ with a given level of confidence requires specification of the z quantile based on the desired level of confidence ($z_{1-\alpha/2}$), population standard deviation (σ), and desired margin of error (*m*).

Numerical examples. Suppose we have a variable with standard deviation σ = 15 and want to estimate µ with 95% confidence.

The samples size required to achieve a margin of error of 5 $n = z^2_{1-\varepsilon/2}\frac{\sigma^2}{m^2} = 1.96^2 \cdot \frac{15^2}{5^2}$ = 36.

The samples size required to achieve a margin of error of 2.5 is $n = 1.96^2 \cdot \frac{15^2}{2.5^2}$ = 144.

Greater precision requires a larger sample size.

Estimating p with confidence

Sampling distribution of the proportion

Estimating probability of success *p* is analogous to estimating expected value μ. However, instead of using $\bar{x}$ as an unbiased point estimate of μ, we use $\hat{p}$ as an unbiased estimate of *p*.

The symbol $\hat{p}$ ("p-hat") is used to represent the **sample proportion**:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

For example, if we find 17 smokers in an SRS of 57 individuals, the sample proportion is $\hat{p}$ = 17 / 57 = 0.2982. We ask, How precise is $\hat{p}$ as are reflection of parameter *p*? How much can we expect any given $\hat{p}$ to vary from *p*?

In samples that are large, the sampling distribution of $\hat{p}$ will be approximately normal with a mean of *p* and standard error of the proportion $SEP = \sqrt{\frac{pq}{n}}$ where *q* = 1 – *p*. The *SEP* quantifies the precision of the sample proportion as an estimate of the binomial parameter and population proportion.

Confidence interval for p

In large samples, an approximate (1−α)100% CI for *p* is given by

$$\hat{p} \pm (z_{1-\alpha/2})(SEP)$$

Since *p* is not known, we cannot derive SEP directly and instead use this approximation:

$$\text{The estimated } SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

The above CI formula should be used only in samples in large sample in which *npq* ≥ 5 ("the *npq* rule"). A more precise formula that may be applied to smaller samples is provided in later chapters.

Numerical example. In an SRS of 57 individuals with 17 smokers, $\hat{p}$ = 17 / 57 = 0.2982, q-hat = 1 – 0.2982 = 0.7018. Therefore $n\hat{p}\hat{q} = (.2982)(.7018)(57) =$ 11.9 and the sample is sufficiently large to proceed with the above formula. The estimated $SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{.2982 \cdot .7018}{57}}$ =0.0606 and the 95% CI for *p* = .2982 ± (1.96)(.0606) = .2982 ± .1188 = (.1794, .4170).

Sample size requirement for estimating p with confidence

n planning a study, we want to collect enough data to estimate population proportion *p* with adequate precision. In an earlier chapter we had determined the sample size requirements to estimate µ with confidence. We apply a similar method in determining sample size requirements to estimate *p*.

Let *m* represent the margin of error. This provides the "wiggle room" around $\hat{p}$ for our confidence interval and is equal to half the confidence interval length. To achieve margin of error *m*, study

$$n = \frac{z^2_{1-\frac{\alpha}{2}} p^* q^*}{m^2}$$

where p^* represent the an educated guess for the proportion and $q^* = 1 - p^*$.

When no reasonable guess of *p* is available, use *p** = 0.50 to provide a "worst-case scenario" sample size that will provide more than enough data.

Numeric example: We want to sample a population and calculate a 95% confidence for the prevalence of smoking. How large a sample is needed to achieve a margin of error of 0.05 if we assume the prevalence of smoking is about 30%? How large a sample is needed to shrink the margin of error to 0.03?

Solutions:

To achieve a margin of error of 0.05, $n = \frac{z^2_{1-\frac{\alpha}{2}} p^* q^*}{m^2} = \frac{1.96^2 \cdot 0.30 \cdot 0.70}{0.05^2}$ = 322.7. Round this up to 323 to ensure adequate precision.

To achieve a margin of error of 0.05, $n = \frac{1.96^2 \cdot 0.30 \cdot 0.70}{0.03^2}$. = 896.4, so use 897 individuals.