Statistical inference

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

There are **two forms of estimation:**

- Point estimation (maximally likely value for parameter)
- Interval estimation (also called confidence interval for parameter)

This chapter introduces estimation. The following chapter introduced NHTS.

Both estimation and NHTS are used to infer parameters. A **parameter** is a statistical constant that describes a feature about a phenomena, population, pmf, or pdf.

Examples of parameters include:

- Binomial 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")

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

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

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

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

Estimating μ with confidence

Sampling distribution of the mean

Although point estimate $\bar{x}$ is a valuable reflections of parameter μ, it provides no information about the precision of the estimate. 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 too much detail, the SDM reveals that:

- $\bar{x}$ is an unbiased estimate of μ;
- the SDM tends to be normal (Gaussian) when the population is normal or when the sample is adequately large;
- the standard deviation of the SDM is equal to $\sigma/\sqrt{n}$. This statistic—which is 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 a reflection of the precision of $\bar{x}$ as an estimate of μ:

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

Note that this formula is based on σ and <u>not</u> on sample standard deviation *s*. Recall that σ is NOT calculated from the data and is derived from an external source. Also note that the *SEM* is inversely proportion to the square root of *n*.

Numerical example. Suppose 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**:

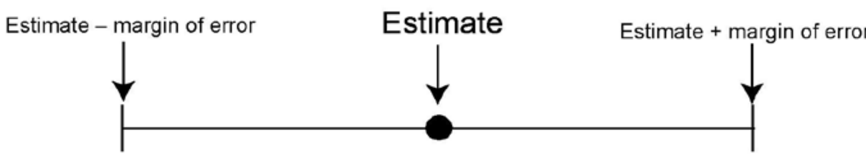


Fig: confidence-interval.ai

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

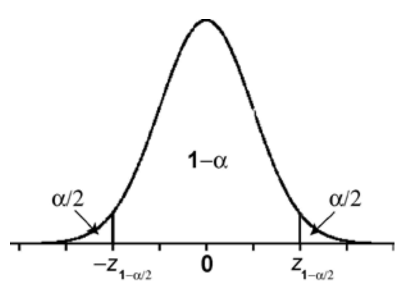
Note: The margin of error is the plus-or-minus wiggle-room drawn around the point estimate; it is equal to half the confidence interval length.

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:

$(1-\alpha)100\%$	α	$z_{1-\alpha/2}$
90%	.10	1.64
95%	.05	1.96
99%	.01	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 (*m*) is the "plus or minus" value surrounding the estimate. In this case *m* = ±9.27.
- We use these confidence interval to address potential locations of the population mean μ, NOT the 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 margin of error $m = z_{1-\alpha/2}^{2} \frac{\sigma}{\sqrt{n}}$. Solving this equation for *n* derives,

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

We always round results from this formula up to the next integer to ensure that we have a margin of error 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 $n = z_{1-\varepsilon/2}^{2} \frac{\sigma^{2}}{m^{2}} = 1.96^{2} \cdot \frac{15^{2}}{5^{2}}$ = 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.

Again, doubling the precision requires quadrupling the sample size.

Estimating *p* with confidence

Sampling distribution of the proportion

Estimating parameter *p* is analogous to estimating parameter 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") represents 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, $\hat{p}$ = 17 / 57 = 0.2982. We ask, How precise is $\hat{p}$ as are reflection of *p*? How much can we expect any given $\hat{p}$ to vary from *p*?

In samples that are large, the sampling distribution of $\hat{p}$ is 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 parameter *p*.

Confidence interval for p

This approach should be used only in samples that are large.[a] Use this rule to determine if the sample is large enough: if $npq \geq 5$ → proceed with this method. (Call this "the *npq* rule").

An approximate (1−α)100% CI for *p* is given by

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

where the estimated $SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Numerical example. An SRS of 57 individuals reveals 17 smokers. Therefore, $\hat{p}$ = 17 / 57 = 0.2982, $\hat{q}$ = 1 – 0.2982 = 0.7018 ad $n\hat{p}\hat{q} = (.2982)(.7018)(57) =$ 11.9. Thus, the sample is large to proceed with the above formula. The estimated $SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{.2982 \cdot .7018}{57}}$ = 0.06059 and the 95% CI for *p* = .2982 ± (1.96)(.06059) = .2982 ± .1188 = (.1794, .4170). Thus, the population prevalence is between 18% and 42% with 95% confidence.

Estimation of a proportion (step-by-step summary)

Step 1. Review the research question and identify the parameter. Read the research question. Verify that we have a single sample that addresses a binomial proportion (*p*).

Step 2. Point estimate. Calculate the sample proportion ($\hat{p}$) as the point estimate of the parameter.

Step 3. Confidence interval. Determine whether the *z* (normal approximation) formula can be used with the "*npq* rule." If so, determine the *z* percentile for the given level of confidence (table) and the standard error of the proportion $SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}}$. Apply the formula $\hat{p} \pm (z_{1-\alpha/2})(SEP)$.

Step 4. Interpret the results. In plain language report what proportion and the variable it address. Report the confidence interval; being clear about what population is being addressed. Reported results should be rounds as appropriate to the reader.

Illustration

Of 2673 people surveyed, 170 have risk factor X. We want to determine the population prevalence of the risk factor with 95% confidence.

Step 1. Prevalence is the proportion of individuals with a binary trait. Therefore we wish to estimate parameter *p*.

Step 2. $\hat{p}$ = 170 / 2673 = .06360 = 6.4%.

Step 3. $n\hat{p}\hat{q} = 2673(.0636)(1-.0636) = 159$ → z method OK.

$$SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(.0636)(1-.0636)}{2673}} = .00472$$

The 95% CI for p = $\hat{p} \pm (z_{1-\alpha/2})(SEP)$ = 0.636 ± 1.96 · .00472 = .0636 ± .0093 = (.0543, .0729) = (5.4%, 7.3%)

Step 4. The prevalence in the sample was 6.4%. The prevalence in the population is between 5.4% and 7.3% with 95% confidence.

Sample size requirement for estimating p with confidence

In planning a study, we want to collect enough data to estimate *p* with adequate precision. Earlier in the chapter we determined the sample size requirements to estimate μ with confidence. We apply a similar method to determine the 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*,

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

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 roughly 30%

Solution: 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.

How large a sample is needed to shrink the margin of error to 0.03?

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 study 897 individuals.