

Inference in Practice

Diana Mindrila, Ph.D.

Phoebe Balentyne, M.Ed.

Based on Chapter 16 of *The Basic Practice of Statistics* (6th ed.)

Concepts:

- Conditions for Inference in Practice
- Cautions About Confidence Intervals
- Cautions About Significance Tests
- Planning Studies: Sample Size for Confidence Intervals
- Planning Studies: The Power of the Statistical Test

Objectives:

- Describe the conditions necessary for inference.
- Describe cautions about confidence intervals.
- Describe cautions about significance tests.
- Calculate the sample size for a desired margin of error in a confidence interval.
- Define Type I and Type II errors.
- Define the power of a significant test.

References:

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

z Procedures

- Statistical inference involves the use of information collected from a sample to draw conclusions about the population.
- Two examples of types of inferences will be discussed.
- When the “simple conditions” are true (the data are an SRS, the population has a Normal distribution, and the standard deviation σ of the population is known), a confidence interval for the mean μ is:

$$CI = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- The purpose of this first procedure is to estimate the population mean based on the sample mean by computing a confidence interval.
- To test a hypothesis $H_0: \mu = \mu_0$, use the one-sample z statistic:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- This second procedure is a test of statistical significance for the difference between the sample mean and a given value, which is often the population mean.
- These are called **z procedures** because they both involve a one-sample z statistic and use the standard Normal distribution.
- These computation procedures should not be used if a variable does not have a distribution that is close to Normal.

Conditions for Inference in Practice

Any confidence interval of a significance test can be trusted only under specific conditions.

These are some guidelines:

Where did the data come from?

When researchers use statistical inference, they are acting as if the data are a random sample or come from a randomized comparative experiment.

- If the data do not come from a random sample or randomized comparative experiment, conclusions may be challenged.
- Practical problems such as nonresponse or dropouts from an experiment can hinder inference.
- Different methods are needed for different designs.
- There is no cure for fundamental flaws like voluntary response.

What is the shape of the population distribution?

Many of the basic methods of inference are designed for Normal populations.

- Any inference procedure based on sample statistics like the sample mean that are not resistant to outliers can be strongly influenced by a few extreme observations.

Cautions About Confidence Intervals

- A sampling distribution shows how a statistic varies in repeated random sampling.
- The variation causes random sampling error because the statistic missed the true parameter by a random amount.
- No other source of variation or bias in the sample data influences the sampling distribution.

The margin of error in a confidence interval covers only random sampling errors. Practical difficulties such as undercoverage and nonresponse are often more serious than random sampling error. The margin of error does not take such difficulties into account.

- When estimating the population mean, the margin of error must be computed prior to obtaining the confidence interval.
- The mean will be slightly different for each random sample for which data is collected.
- This is called sampling error, and the margin of error represents the amount of error that may be induced in the estimation process due to random sampling.

Sample Size for Confidence Intervals

It is important to plan for inferences when planning a sample or an experiment. The number of observations is a critical part of planning the study.

- When estimating confidence intervals, the larger the sample, the smaller the margin of error.
- However, selecting a large random sample is difficult and expensive.
- Therefore, when planning a study, researchers must decide how many individuals should be included in the sample in order to have a reasonable degree of estimation confidence and a small margin of error.
- First, researchers must determine the specific confidence level they desire (for example: 95%) and a margin of error that would be acceptable.

The margin of error ME of the confidence interval for the population mean μ is:

$$ME = z^* \frac{\sigma}{\sqrt{n}}$$

If this equation is solved for n to find out the number of participants needed, the following equation is obtained:

$$n = \left(\frac{z^* \sigma}{ME} \right)^2$$

Example:

CI (95%) $\Rightarrow z^*=1.96$

Margin of Error=3

Population Standard Deviation=15

$$n = \left(\frac{1.96 \cdot 15}{3} \right)^2 = 96.04 \approx 96$$

- To estimate the population mean with 95% confidence and a margin of error of 3, the sample must include at least 96 individuals.

Cautions About Significance Tests

Significance tests are widely used in most areas of statistical work. Some points to keep in mind when using or interpreting significance tests are:

How small a P is convincing?

The purpose of a test of significance is to describe the degree of evidence provided by the sample against the null hypothesis. How small a P -value is convincing evidence against the null hypothesis depends mainly on two circumstances:

- If H_0 represents an assumption that has been believed for years, strong evidence (a small P) will be needed.
- If rejecting H_0 means making a costly changeover, strong evidence is required.

Significance Depends on the Alternative Hypothesis

The P -value for a one-sided test is one-half the P -value for the two-sided test of the same null hypothesis based on the same data.

- The evidence against the null hypothesis is stronger when the alternative is one-sided because it is based on the data *plus* information about the direction of possible deviations from the null.
- If this added information is not available, always use a two-sided alternative hypothesis.

Sample Size Affects Statistical Significance

- ❖ Because large random samples have small chance variation, very small population effects can be highly significant if the sample is large.
- ❖ Because small random samples have a lot of chance variation, even large population effects can fail to be significant if the sample is small.

The Power of a Statistical Test

Sometimes the conclusions drawn from a significance test can be wrong. There are two types of mistakes that can be made.

If H_0 is rejected when H_0 is true, a **Type I error** has been committed.

If H_0 is accepted when H_0 is false, a **Type II error** has been committed.

		Truth about the null hypothesis	
		H_0 true	H_0 false
Decision	Reject H_0	Type I error (α)	<i>Correct conclusion (Power)</i>
	Accept H_0	<i>Correct conclusion</i>	Type II error (β)

- When conducting a test of significance, researchers must use the evidence collected from the sample to make a decision about the null hypothesis. When these conclusions are wrong, this is called Type I or Type II error.
- There are four possible scenarios when a decision is made. These possible scenarios are pictured in the above image.
 - The null hypothesis can either be true or false.
 - The null hypothesis can either be accepted or rejected.
 - ❖ If the null hypothesis was accepted when it is true, this was the correct decision.
 - ❖ If the null hypothesis was rejected when it was false, this was the correct decision. (The ability of a statistical test to reject the null hypothesis when it is, in fact, false, is also called the statistical power of a test).
 - ❖ If the null hypothesis is true, but it is rejected, this is called a **Type I error**.
 - ❖ If the null hypothesis is false, but it is accepted, this is called a **Type II error**.
- When researchers try to reduce the probability of making a Type I error, they automatically increase the probability of making a Type II error.
- Depending on the type of research being conducted, researchers must decide which type of error is less detrimental to the study.
- Generally, consequences are not as severe if a null hypothesis that is true is rejected than if a null hypothesis that is false is accepted.

The Power of a Statistical Test

The probability of a Type I error is the probability of rejecting H_0 when it is really true. This is exactly the significance level of the test.

Significance and Type I Error

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true. Consider the consequences of a Type I error before choosing a significance level.

A significance test makes a Type II error when it accepts a null hypothesis that really is false. There are many values of the parameter that satisfy the alternative hypothesis, so researchers must concentrate on one value. Researchers can calculate the probability that a test *does* reject H_0 when an alternative is true. This probability is called the **power** of the test against that specific alternative.

The **power** of a test against a specific alternative is the probability that the test will reject H_0 at a chosen significance level α when the null hypothesis is false.

The Power of a Statistical Test

How large of a sample should be taken when a significance test is planned? The answer depends on what alternative values of the parameter are important to detect.

Summary of influences on how many observations are needed:

- If a smaller significance level is required (such as 1% rather than 5%), then a larger sample must be taken. A smaller significance level requires stronger evidence to reject the null hypothesis.
- If higher power is required (such as 99% rather than 90%), a larger sample is needed. Higher power gives a better chance of detecting a difference when it is really there.
- At any significance level and desired power, detecting a small difference requires a larger sample than detecting a large difference.