Hypothesis Testing and Confidence Intervals

From sample evidence to careful population conclusions

Ratiomera Statistics

Table of contents

Course-Page Summary

Inference uses a sample to learn about a population while keeping uncertainty visible. Begin with the research question and design, name the population quantity, and only then choose a confidence interval or hypothesis test that matches the data structure.

The Core Reasoning

Table 1: Reliable inference connects the research question, design, calculation, and interpretation.

Step	Question to ask	Main idea
1. Define	What population quantity is being studied?	Write the parameter and the null and alternative hypotheses before calculating
2. Match	Are the observations independent, paired, or categorical counts?	The design determines the standard error and reference distribution
3. Calculate	How far is the estimate from the null value in standard-error units?	A test statistic standardizes the observed discrepancy
4. Evaluate	How unusual would this statistic be if the null model were true?	The p-value is a conditional probability under the null model
5. Conclude	What does the evidence support at the chosen level?	State the population claim, uncertainty, direction, and study limits

For a sample mean with an estimated population standard deviation, the one-sample statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

with $n - 1$ degrees of freedom. A confidence interval uses the same estimate and standard error:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}.$$

The interval gives a range of population means compatible with the data and model at the chosen confidence level. Its repeated-sampling meaning is about the method: across many comparable samples, the stated proportion of intervals would contain the fixed population parameter.

Choose the Procedure From the Design

Table 2: The same-looking numbers can require different procedures when their dependence structure differs.

Data structure	Question	Procedure used here
One quantitative sample	Does one population mean differ from a reference value?	One-sample t procedure
Two independent quantitative groups	Do two population means differ?	Independent-samples t procedure
Two measurements on the same cases	Does the population mean of the within-case differences differ from zero?	Paired t procedure
Two categorical variables	Are the variables independent in the population?	Chi-square test of independence

Interpretation Checklist

- A p-value is the probability of data at least this incompatible with the null model, assuming that model and its conditions. It is not the probability that the null hypothesis is true.
- A decision to reject is not proof, and a decision not to reject is not proof of equality.
- Statistical significance does not state effect size, practical importance, measurement quality, or causality.
- Type I error means rejecting a true null hypothesis. Type II error means failing to reject a false null hypothesis. Power is the probability of rejecting the null when the specified alternative is true.
- A larger sample can reduce standard error and increase power, but it does not repair biased sampling, poor measurement, dependence ignored by the model, or a weak design.

Topic 3 supplies the inferential language used in every later topic. Correlation, regression, partial correlation, multiple regression, and ANOVA change the parameter and model, but they retain the same discipline: define the question, respect the design, quantify uncertainty, and keep the conclusion within scope.

Expanded Reference

Purpose and foundations

Statistical inference uses information from a sample to learn about a population while acknowledging sampling uncertainty. A **population** is the full group of cases addressed by the research question. A **sample** is the subset that was observed. A **parameter** is a numerical feature of the population, such as a population mean μ. A **statistic** is the corresponding quantity calculated from the sample, such as the sample mean $\bar{x}$. The statistic is known after the data are collected; the parameter usually remains unknown.

If a new random sample were drawn from the same population, its statistic would usually differ. The hypothetical distribution of that statistic across repeated samples is its **sampling distribution**. Its standard deviation is called the **standard error**. Standard error measures sample-to-sample variability of an estimate. It does not measure the spread of individual observations, which is the role of an ordinary standard deviation.

Inference depends on more than a formula. The sample must connect credibly to the population, the observations must fit the dependence assumptions of the method, and the measurement must represent the intended variable. A small standard error cannot correct selection bias, poor measurement, or an unsuitable design. Begin with the research question and study design, then inspect the descriptive statistics, and only then choose an inferential procedure.

Element	Sample language	Population language
Center	Sample mean $\bar{x}$	Population mean μ
Proportion	Sample proportion $\hat{p}$	Population proportion p
Variability of scores	Sample standard deviation s	Population standard deviation σ
Uncertainty of an estimate	Estimated standard error	Sampling-distribution standard deviation

Core ideas

A confidence interval gives a range of parameter values compatible with the estimate and its sampling uncertainty under the model. The confidence level describes the long-run performance of the procedure. If the same sampling and interval method were repeated many times, a stated proportion of the resulting intervals would contain the fixed population parameter. After one interval has been calculated, the parameter is not moving among values; the interval is the result that varied through sampling.

A hypothesis test begins with a **null hypothesis** H_0, a precise reference claim about a population parameter. An **alternative hypothesis** H_1 states the direction or difference of substantive interest. A test statistic measures how far the observed estimate lies from the null value in standard-error units. The **p-value** is the probability, assuming the null hypothesis and all model conditions, of obtaining a test statistic at least as incompatible with the null as the observed one. It is not the probability that the null hypothesis is true.

The significance level α is a decision threshold selected before looking at the result. If the p-value is at most α, the result is called statistically significant and H_0 is rejected. If the p-value is greater than α, the analysis fails to reject H_0. Failing to reject is not proof that there is no effect. The data may be imprecise, the true effect may be small, or the design may have limited power.

Reality and decision	Do not reject H_0	Reject H_0
H_0 is true	Correct retention decision	Type I error, with probability controlled by α
H_0 is false	Type II error, denoted β	Correct detection, with probability called power $1-\beta$

Power is the probability that a test rejects H_0 when a specified alternative is true. It increases when the true effect is larger, scores are less variable, the sample is larger, or the significance rule is made less strict. These influences involve tradeoffs. Planning therefore requires a substantively meaningful effect size and a defensible design, not a search for significance after data collection.

Procedure choice follows the structure of the research question. A one-sample mean procedure compares one group with a reference value. An independent-groups procedure compares separate groups. A paired procedure analyzes linked measurements, such as the same participants before and after an intervention, by working with within-pair differences. A chi-square procedure for a contingency table compares observed categorical counts with counts expected under the null model. In every case, the unit of analysis and dependence structure must be stated.

Formula guide

For an independent random sample, the estimated standard error of a sample mean is the sample standard deviation divided by the square root of the sample size:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

The square root explains why uncertainty decreases more slowly than sample size grows. Multiplying n by four halves this standard error when the variability stays the same.

A confidence interval combines an estimate $\hat{\theta}$ with its standard error and a critical value c chosen for the confidence level:

$$\hat{\theta} \pm c \cdot SE\left(\hat{\theta}\right)$$

If the population standard deviation σ is known and the stated normal model applies, use the standard-normal reference:

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}.$$

The alternative determines the reference area. A two-sided alternative uses both tails beyond $|z|$ or $|t|$. A one-sided alternative uses the prespecified directional tail. Choosing the direction after seeing the result does not constitute a prespecified one-sided test.

For a one-sample mean with an estimated population standard deviation, the interval uses a critical value from a t distribution with $n-1$ degrees of freedom:

$$\bar{x} \pm t_{1-\alpha/2,\,n-1}\frac{s}{\sqrt{n}}$$

The corresponding one-sample test statistic compares the observed sample mean with the null value μ_0:

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

The numerator is the observed difference from the null claim. The denominator translates that difference into standard-error units. For paired measurements, first calculate one difference d_i per pair and then apply the same one-sample reasoning to the mean difference $\bar{d}$:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

This preserves the pairing. Treating the measurements as unrelated would discard information about which two observations belong together.

For two independent samples under the equal-population-variance model taught here, first pool the two sample variances:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Then calculate

$$SE(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, \qquad df = n_1 + n_2 - 2.$$

The matching two-sided interval replaces the numerator with

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2,\, n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The equal-variance condition, procedure name, and calculation must agree. Paired data require the difference-score procedure instead.

For a one-sample known-σ planning question, define the standardized population difference

$$\delta = \frac{\mu - \mu_0}{\sigma}, \qquad \text{Power} = 1 - \beta.$$

In the one-sided z planning model used in the supplied material, the sample size required for significance level α and target power $1 - \beta$ is

$$n = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\delta}\right)^2.$$

Round the result up. This formula belongs to that stated model and is not a universal sample-size rule. Power increases with sample size and effect magnitude, and decreases when a stricter significance level moves the rejection boundary farther into the tail.

For two categorical variables, the expected count under independence in row i and column j is

$$m_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}.$$

The chi-square statistic and degrees of freedom are

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - m_{ij}\right)^2}{m_{ij}}, \qquad df = (k-1)(l-1).$$

For a two-by-two table, the magnitude of the phi coefficient is

$$|\phi| = \sqrt{\frac{\chi^2}{n}}.$$

The approximation used in this learning sequence requires a simple random sample and expected counts greater than 5 in every cell. A large χ^2 counts against independence; it is not evidence for independence.

Question structure	Course procedure	Quantity analyzed
One sample versus a reference	One-sample z or t	One sample mean
Two separate groups	Pooled independent-samples t	Difference between group means
Two linked measurements	Paired t	Mean of within-pair differences
Independent groups with a rank-based question	Wilcoxon rank-sum procedure	Relative ranks across groups
Paired observations with a rank-based question	Wilcoxon signed-rank procedure	Signed ranks of paired differences
Two categorical variables	Chi-square independence procedure	Observed versus expected cell counts

Reading the explanatory figure

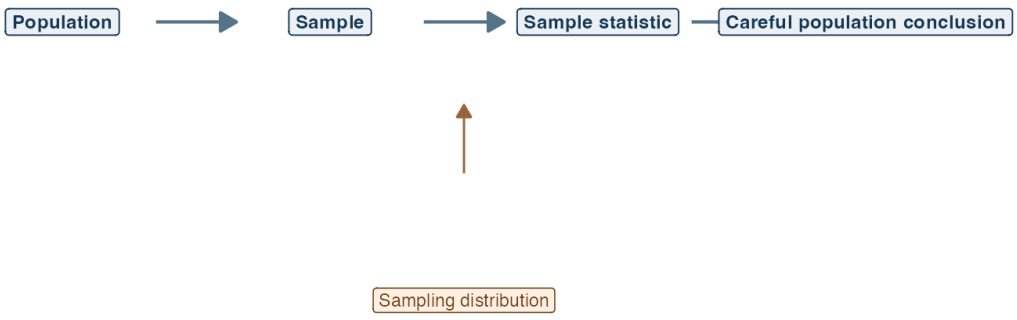

Figure 1: A horizontal flow moves from population to sample to sample statistic and then to a careful population conclusion, with a sampling distribution feeding the statistic.

Read the main line from left to right. The population is the target of the research question. The sample is the part that becomes observable. A sample statistic condenses relevant evidence, such as a sample mean, difference, proportion, or association. The final box is deliberately labeled a careful conclusion because the return from sample to population is never automatic.

The arrow rising from the sampling distribution is the bridge supplied by probability. It represents how the statistic would vary across repeated samples under stated assumptions. A confidence interval uses this variability to show precision. A test compares the observed statistic with the sampling behavior expected under H_0. The figure does not imply that a large sample guarantees

generalizability. Sampling method, measurement, missing data, dependence, and study design still determine which population conclusion is warranted.

The separation between “sample statistic” and “population conclusion” is a useful pause point. Before crossing it, ask whether the standard error reflects the actual design, whether the procedure’s assumptions are plausible, and whether the wording of the conclusion matches what was tested. A result can support an association or difference without establishing a causal effect.

Interpretation checklist

State the population, sample, parameter, statistic, and unit of analysis. Describe how cases entered the sample. Inspect the data and identify missingness or unusual observations. Match the procedure to the outcome scale and to independent, paired, or categorical data. State H_0 and H_1 in words and symbols. Report the estimate and confidence interval alongside the test statistic, degrees of freedom when applicable, p-value, and a contextual interpretation.

Avoid converting the p-value into a probability that a hypothesis is true. Do not use statistical significance as a synonym for practical importance. Compare the size and uncertainty of the estimate with the research question. When a result is not significant, discuss its interval and precision instead of declaring that groups are equal. When several tests are conducted, recognize that the chance of at least one Type I error can increase and use a planned multiplicity approach when required.

How this topic connects

Probability supplied the sampling distributions that make confidence intervals and tests possible. The same inferential pattern now accompanies every later coefficient. A correlation has a standard error and test. A regression slope has an estimate, interval, and p-value. A partial correlation and each multiple-regression coefficient are interpreted conditionally. Analysis of variance uses an F statistic to compare model-related and residual variability.

Inference is therefore not a separate ritual attached to the end of an analysis. It is a disciplined bridge from descriptive sample evidence to a bounded population claim. Keeping the estimate, uncertainty, design, and substantive meaning together makes that bridge trustworthy.

Document ID: `topic-03-hypothesis-testing-summary-en`

Notes 1 of 3

Notes 2 of 3

Notes 3 of 3