Print

Summary: Decision Making with Data

To go to the video where you learned a task or concept, select a link.

Introduction to Statistical Inference

The methods of inference essentially break down into two basic activities: **interval estimation** and **hypothesis testing**.

If you're asking the question, "What is the unknown value?", then you would compute an **interval estimate** of the value. If you're asking the question, "Is the unknown value greater than X?", then you would conduct a **hypothesis test**.

What Is a Confidence Interval?

Perhaps the most common type of interval estimate is a **confidence interval**. A confidence interval is a range of values that reflects both an approximation of a population attribute, such as the mean, and the amount of uncertainty attributable to sampling.

A common example is political polling results, which are often reported with a value referred to as the **margin of error**. The margin of error is a measure of the uncertainty in the polling results.

For example, suppose a political poll of 1,000 likely voters finds that 45% of the surveyed voters support a specific candidate. For a poll of 1,000 likely voters, the margin of error is ± 3%.

From this you can infer, with 95% confidence, that the true proportion of likely voters who support the candidate is between 0.42 and 0.48, or between 42% and 48%.

The term **confidence** refers to the statistical confidence that the calculated interval includes the true proportion.

A Practical Example

The speed of light, in a vacuum, is 186,282 miles per second (or 299,792 kilometers per second).

Albert A. Michelson of the US Navy was a scientist known for his work in estimating the speed of light. In June 1879, he attempted to determine the speed of light using the limited tools that were available to him at the time. We use Michelson's data as a case study throughout this module.

Estimating a Mean

Michelson was trying to measure an unknown **parameter**, the constant speed of light. Michelson planned to use a **sample statistic** to draw a conclusion about an unknown **population parameter.**

The mean of a sample is called a **point estimate**.

Point estimates, for samples drawn under the same conditions, are different. This idea, that different subsets of observations yield different estimates, is known as **sampling variation.**

A margin of error quantifies the uncertainty in our estimate of the true value. Adding a margin of error produces a **confidence interval**.

A confidence interval is an estimate that accounts for the likely range of the sampling error. Or, put another way, a confidence interval provides **a range of plausible values for the theoretical or true population parameter**.

Visualizing Sampling Variation

Interval estimation is rooted in the simple idea that random samples vary from one another.

This variability produces **sampling error.** Sampling error is the imprecision arising from using samples rather than an entire population. This is illustrated using the Sampling Distribution of Sample Means script.

After thousands of repetitions, the distribution of all sample means, also called the **sampling distribution of the mean**, approaches the familiar symmetric mound shape. It is centered at the known population mean.

Constructing Confidence Intervals

A confidence interval for the mean starts with the point estimate, and then you add a margin of error around this point estimate. To construct the margin of error, you need

- the standard error,
- the confidence level (1-α),
- and a critical value from the *t* distribution.

The standard error is an estimate of the standard deviation of sample means. The confidence level is specified by you, the analyst. It's typically set at 95%, but you can change this.

For any confidence interval, there is some risk that the interval doesn't capture the true parameter value. This risk is denoted as alpha (α). For a 95% confidence interval, the alpha risk is 5%.

Sample means follow the *t* distribution instead of the normal distribution. The *t* distribution is a sampling distribution that is centered at zero, like the standard normal distribution, and it has a mounded shape. But its shape changes slightly based on the sample size.

To construct a 95% confidence interval for the mean, you need to find values of the *t* distribution that encompass the central 95% of the distribution.

Prediction Intervals

You use a confidence interval to estimate a population parameter. If you want to estimate the next value, or the next set of values, you would construct a **prediction interval**.

A prediction interval is centered at the point estimate, exactly like a confidence interval. However, the upper and lower bounds are pushed farther from that point estimate to capture the additional uncertainty.

Tolerance Intervals

With a **tolerance interval**, you identify a range of values for the population that captures a specified proportion of observations, with a certain degree of confidence.

Tolerance intervals enable you to answer this question: Given the data that I have collected and my current process, what is the range of variation that I can expect to see in the process characteristic?

To construct a tolerance interval, you specify the confidence level (1- α). You also specify p, the proportion of the population that you want to include in the interval.

When your data are approximately normally distributed, you can construct a **parametric tolerance interval**. When your distribution is not normal, you can construct a **nonparametric tolerance interval** instead. Nonparametric methods do not rely on assumptions about the underlying distribution of the data.

Comparing Interval Estimates

Suppose that you are studying a continuous characteristic, like velocity.

- If your question is about the **average** velocity, you should use a confidence interval.
- If your question is about the **next velocity measurement**, or **set of measurements**, you should use a prediction interval.

And, if your question is about the **population** (in this case all of the potential velocity measurements), you should use a tolerance interval.

Introduction to Statistical Testing

In hypothesis testing, and in statistical testing in general, you need to make a decision based on sample data, without having full knowledge about a population.

Statistical Decision Making

Although you can use interval estimates to draw conclusions about the parameter that you're estimating, the hypothesis testing protocol provides a more formal framework for statistical decision making.

The goal of hypothesis testing is to make a decision based on incomplete information (a sample). As a result, hypothesis testing addresses two potential issues: inconclusive results and erroneous decisions. These two concerns underlie two important concepts in hypothesis testing: the null hypothesis and the *p*-value (or probability value).

Understanding the Null and Alternative Hypothesis

The **null hypothesis** is the default decision. You assume that the null is true unless your data tell you otherwise. The **alternative hypothesis** is the motivation to collect and analyze the data. You represent the null and alternative hypotheses as statements about the value of population

parameters.

In hypothesis testing, you start by assuming that the null hypothesis is true, and you use sample data to make decisions about the null hypothesis.

If you reject the null hypothesis, you accept the alternative. If you fail to reject the null hypothesis, you conclude that the null hypothesis ***might be*** true.

When you fail to reject the null hypothesis, you are simply saying that you don't have evidence, based on the available data, that the null hypothesis is not true.

Sampling Distribution under the Null

In hypothesis testing, you make decisions based on one sample. But different samples have slightly different means. You have to make a decision about the population in light of this sampling variation, using the one sample that you have.

If you assume that the null hypothesis is true, as the sample size *n* increases, the distribution of the sample mean approaches a normal distribution with a mean of mu (the hypothesized value) and a standard deviation of sigma over the square root of n.

Although you can't be certain in advance about an individual sample mean, you have a good idea how close it will be to the true population mean based on ordinary sampling variation. Means within the range of ordinary sampling variation are consistent with the null hypothesis. But if a sample mean is extreme relative to the distribution of sample means, you start to question whether the null hypothesis is correct.

The *p*-Value and Statistical Significance

In a statistical test, you ask, "How likely is our result under the null hypothesis?" and you set a boundary beyond which the sample is so implausible that the null hypothesis is simply not credible.

In statistical tests, you use a **probability value**, or ***p*-value**, to guide your decision making. A *p*-value is a measure of the strength of the evidence against the null hypothesis. The smaller the *p*-value, the stronger the evidence against the null. If you have a high *p*-value, you have little evidence against the null.

Sample results with *p*-values smaller than a chosen value, called the **significance level**, are referred to as **statistically significant**. The significance level, alpha, is often set at 0.05. However, this is not a hard and fast rule. Significant results are surprising or unusual enough to warrant attention, because they are unlikely to have occurred just due to random sampling variation.

Having significant results does ***not*** mean that you are correct in rejecting the null hypothesis! And a low *p*-value does not signify that the result is meaningful or important from a practical perspective.

Conducting a One-Sample *t* Test

A test for one mean is called a one-sample *t* test.

In a **two-tailed test**, you would reject the null hypothesis if the observed sample mean is much larger than, or much smaller than, the null hypothesis.

If your alternative hypothesis is that the mean is greater than some value, or less than some value, you'd conduct a one-tailed test.

For a one-sample *t* test, the test statistic is the *t* ratio. The *t* ratio measures the distance between the sample mean and the null hypothesized value, in units of standard error.

Equivalence Testing

If your goal is to provide evidence that the deviation from the null hypothesis is ***not too great***, then you would use an **equivalence test**.

In equivalence testing, you specify a practically acceptable interval around the hypothesized value, and then test to see whether the sample mean is outside this interval.

To do this, you conduct two one-sided *t* tests. As a result, this is known as the TOST (two one- sided *t* tests) approach. If you reject both of these tests, at some level of alpha, then you can conclude that the mean is "practically" equivalent to the hypothesized value.

Comparing Two Means

You want might want to use statistical tests to compare the means of two populations or processes. In such cases, the question is not about the mean of one population, but the **difference between the means of two populations**.

To compare two population means based on **independent** samples, you use a **two-sample *t* test**. There are two versions of this test: the **pooled two-sample *t* test** and the **unequal variances, or unpooled, two-sample *t* test**.

Like the one-sample *t* test, the test statistic for these two-sample tests is the *t* ratio.

Two-Sample *t* Tests

In both a **pooled two-sample *t* test** and an **unpooled two-sample *t* test**, the null hypothesis is that the difference between the means is zero. The alternative hypothesis is that the difference is not equal to zero.

The curve is similar to what you saw in the one-sample *t* test. It represents the distribution of the differences between the sample means, **under the null hypothesis that the average difference between means is zero.**

Three *p*-values are reported: one for the two-tailed test, and two for the potential one-tailed tests.

You can also use a confidence interval to evaluate your null hypothesis. The 95% confidence interval for the difference in means is the likely range of values for the average difference. If the hypothesized value, zero, doesn't fall within this interval, you have additional evidence that there is a significant difference.

Unequal Variance Tests

You use a two-sample *t* test to compare two independent means.

However, instead of comparing the centering of two distributions, you might want to compare the spread or dispersion of these distributions. To do this, you would perform a **hypothesis test for unequal variances**.

In an unequal variances test, your null hypothesis is that variances for the two groups are equal. Your alternative hypothesis is simply that they are not equal.

Paired Observations

A two-sample *t* test is used to compare population means based on samples drawn independently from two populations.

But suppose your data form a pair of measurements for an item? In this situation, instead of performing a two-sample *t* test to compare the means, you'd use a **paired *t* test**.

What if you mistakenly analyze paired data using a pooled two-sample *t* test instead? When you ignore the pairing, all of the variability between the pairs is included in the estimate of the standard deviation used to calculate the test statistic. When the data are matched pairs, the two-sample *t* test is generally less sensitive in detecting a true mean difference than the paired *t* test.

Comparing More Than Two Means

When you want to conduct a hypothesis test to compare more than two population means, one popular method is o**ne-way analysis of variance**, or **ANOVA** for short.

In ANOVA, the populations of interest are defined by the levels of a nominal variable, or factor.

The ANOVA method starts with the null hypothesis that all groups have the same mean value. The alternative hypothesis is that ***not*** all of the means are equal. The alternative is often stated this way: At least two of the means are not equal.

To determine whether there is a significant difference between the means, ANOVA analyzes two components of variation:

- the variation due to the different **levels**, or in this case, the different trials
- the pure **random variation**

One-Way ANOVA (Analysis of Variance)

ANOVA is based on a test statistic, called the F-Ratio. The F-Ratio is a ratio of two variances: the variance between the groups and the random error variance.

If the null hypothesis is true, the F-Ratio is close to 1.0. However, if the null is not true, the F-Ratio is greater than 1.0.

The F-Ratio follows the F distribution. The shape of the *F* distribution is based on the degrees of freedom for the mean square model (in the numerator of the F-Ratio) and the mean square error (in the denominator).

In ANOVA, the *p*-value measures how extreme an F-Ratio is relative to the *F* distribution. This *p*-value, and all of the computations that go into computing the F-Ratio, are reported in the Analysis of Variance table.

Multiple Comparisons

When it is unclear which means are different in an ANOVA situation, you can use a **multiple comparison procedure**.

The most basic type of multiple comparison analysis, the **Each Pair, Student's *t*** method, compares all possible pairs of means using pooled two-sample *t* tests at the 0.05 level (or whatever level you specify).

The more tests you perform, the higher the probability of making a false discovery. There are other multiple comparison techniques that are designed to minimize the false discovery rate, such as Tukey's HSD.

Instead of comparing group means to one another, you might be interested in identifying group means that are significantly different from the overall mean. If this is your analysis goal, you can use a technique called **analysis of means**, or ANOM.

Revisiting Statistical Versus Practical Significance

It is easy to fall into a habit of thinking that a small *p*-value indicates an important finding, but that is not always the case. Statistical significance is not the final word in decision making, despite its critical importance in statistical testing.

You, the analyst, determine whether the results that you observe are of practical importance, based on the context, the nature of the problem, and the available evidence.

Summary of Hypothesis Testing for Continuous Data

Statistical testing is a broad and deep topic, and there are a lot of concepts and methods that we did not have an opportunity to discuss. If you regularly use data for decision making, you might want to broaden your knowledge of statistical testing.

The methods that you use are determined largely by your data. If you encounter anything unusual or unexpected, you might want to consult with an expert or do some research on your own.

Introduction to Sample Size and Power

For interval estimation, smaller margins of error, which lead to more precise estimates, are obtained by increasing the sample size. For hypothesis testing, the ability of your test to reject a false null hypothesis is known as **statistical power**, or simply **power**. Increasing the sample size increases the power of your test.

The sample size computations used depend on your data, the context of your problem, and the analyses that you will run.

Determining the appropriate sample size is something that is done ***prior*** to collecting and analyzing data.

Sample Size for a Confidence Interval for the Mean

To construct the margin of error, you need three values:

- the confidence level (1-α)
- the standard error
- a critical value from the *t* distribution

The sample size influences both the critical t value and the standard error. If the precision of your interval estimate is of practical importance to you, you should determine the sample size ***prior*** to collecting the data.

Outcomes of Statistical Tests

When you make a decision based on the results of a hypothesis test, there are four possible outcomes.

When you commit a **false positive**, the null hypothesis is true, but you erroneously reject it. This is called a **type I error**. The probability of committing a type I error is **α**. This is the significance level of your test.

A **true negative** is when the null hypothesis is true and you do not reject it. You make the correct decision.

In a **false negative**, the null hypothesis is false, but you fail to reject it. This is called a **type II error**. The probability of a type II error is denoted by **β**.

For a **true positive**, the null hypothesis is false, and it is correctly rejected as false. You make the correct decision. The true positive rate is also known as the test's **power**. The probability of a true positive, or power, is 1- β.

Statistical Power

Power represents the ability of a hypothesis test to detect true differences between null and alternative hypothesized values.

There are several factors that influence the power of a test:

- the significance level, *α*
- the sample size, *n*
- the variability of the population
- the size of the difference that you want to detect

Exploring Sample Size and Power

The test statistic, the *t* ratio, is a ratio of the **signal**, the difference between the observed and hypothesized value, and the **noise** in the data, the standard error.

By changing the sample size or by increasing the difference between the observed and hypothesized means, you are more likely to reject the null hypothesis.

Power is the ability to correctly reject a false null hypothesis. This is largely based on your sample size. The power of your statistical test is also related to the critical difference that you need to detect, the variability in the population, and the significance level that you choose.

Calculating the Sample Size for One-Sample *t* Tests

When you are planning for data collection or designing a study, it's important to make sure that your test has enough sensitivity, or power, to detect differences that you consider important.

To calculate the sample size required to detect a critical difference, you need the following information:

- alpha, the significance level for your test
- an estimate of the standard deviation

- the difference between the sample mean and the hypothesized value that you need to be able to detect

You also need to specify the desired power of the test. This is typically set at 0.90 or 0.95.

Calculating the Sample Size for Two-Sample *t* Tests

When you calculate the sample size for a two-sample *t* test, the value provided is the total sample size between the two groups. You split this between the two groups.

You can also determine the sample size to detect a specified difference between any two means for one-way analysis of variance, or ANOVA.

Summary of Sample Size and Power

Statistical inference is about making sound decisions from your data.

Sampling comes at a cost. When planning for data collection, you need to consider the practical implication of collecting larger samples.

The sample size computations depend on the types of data that you will collect, your study design, and the analyses that you will run. Prior to data collection, consult with an expert for guidance on determining how much data you will need.