Methods in Open Science

Statistical Inference

Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

WV VIEW

AmericaView
Empowering Earth Observation Education
AmericaView.org - Est. 2003

This module provides an introduction to the process and use of statistical inference to test hypotheses. Multiple courses could be devoted to the study and application of statistical methods. My goal here is just to provide a broad overview and focus on how key techniques are used by data scientists. Instead of focusing on the math, we will focus on what specific tests are used for, how they are conducted, how they are interpreted, and what assumptions must be met for the test to be valid or not misleading. Implementation of statistical tests in software is explored in the Python and R material. In Python, we will make use of SciPy. In R, we will explore a variety of packages.

Book Recommendation

https://statisticsbyjim.com/

Frost, J., 2020. *Hypothesis testing: An intuitive guide for making data driven decisions*. Statistics by Jim Publishing.

2

This text by Jim Frost is a great resource on hypothesis testing. It is a very practical text focused on statistical thinking and how to properly use and interpret statistical tests. If you have an interest in statistical inference and hypothesis testing, I highly recommend this book.

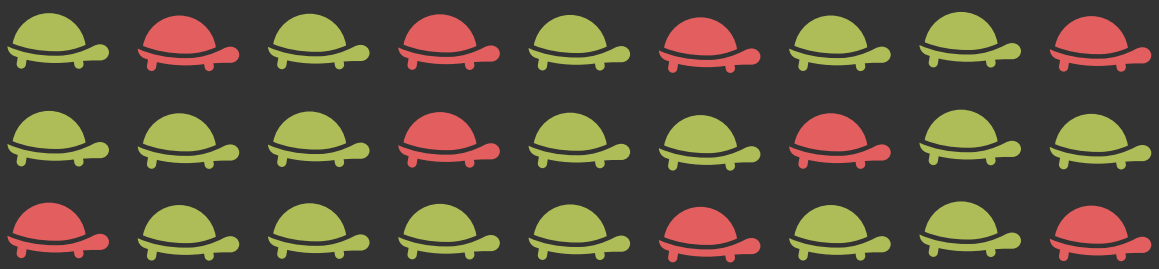
Statistical Inference

3

Let's now step through the process of statistical inference and explore why statistical inference is necessary.

Population vs. Samples

- Population = the entire set of things you are interested in
- Sample = a subset of the population that you measure/study



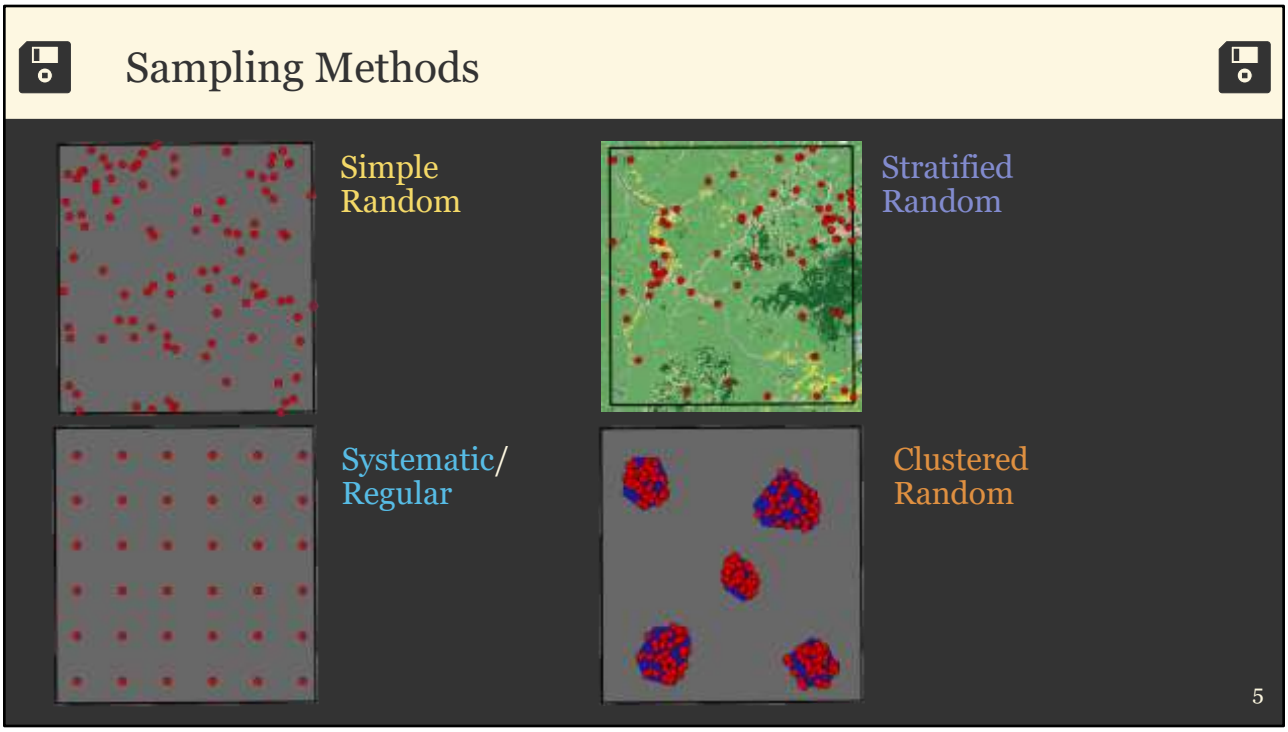
4

When performing a study, we are commonly interested in understanding something about a population. For example, in assessing the effectiveness of a new drug to treat high blood pressure, the population of interest would be all individuals that suffer from high blood pressure. However, it is not possible to administer the new drug to every individual with the disorder in order to assess if it effectively lowers blood pressure. As another example, we may be interested in assessing factors that increase the susceptibility of a specific species of tree occurring in a region to a specific disease or blight. In this case, the population would be every tree of that specific species within the defined geographic extent. Again, not all of these individual trees could be studied.

Since we cannot observe or measure every individual in the population of interest, it is necessary to select a subset of the population to study. This subset is known as a sample. Statistical methods grew out of the need to study a population using a limited sample from a larger population. Thus, the field of statistics grew from a very practical problem: we want to determine whether some type of treatment, such as a new drug, will cause an effect, such as lower blood pressure, in a specific population. Since it is not possible to study every individual in the entire population, we need to be able to make an inference with an estimated level of confidence from a subset of the population.

One of the key considerations in statistics is how well the sample represents

the population. If the sample is biased, skewed, or non-representative of the entire population, then the findings obtained from studying the sample may not hold or be misleading for the entire population, which defeats the purpose of performing a statistical test.



How can we collect a sample from a larger population that is representative of the entire population and is not biased? This is generally accomplished using some form of randomized approach. For example, if your goal is to assess the quality of products being produced at a factory, it would not be possible to test every product created. Instead, you would need to sample products. If you tested the first 10 products produced each day, this may not be representative of the true quality. Maybe mistakes are more common later in the day as the equipment heats up or the workers become tired. So, it would make more sense to randomly select products throughout the entire workday.

There are different sampling routines that can be applied to generate a sample from a larger population. Simple random sampling is purely random. For example, if the goal was to collect a sample from all the students in a large lecture hall with 250 seats, you could place each student's name in a hat then randomly draw a desired number of names. Systematic sampling or regular sampling is collecting data on a regular interval. For example, the student sitting in every 10^{th} seat could be selected.

A stratified random sample means to stratify based on some grouping variable. For example, if the lecture course was attended by students from all academic years, you could randomly select 15 students from each academic year by placing names in four separate hats, one for each academic year, then drawing

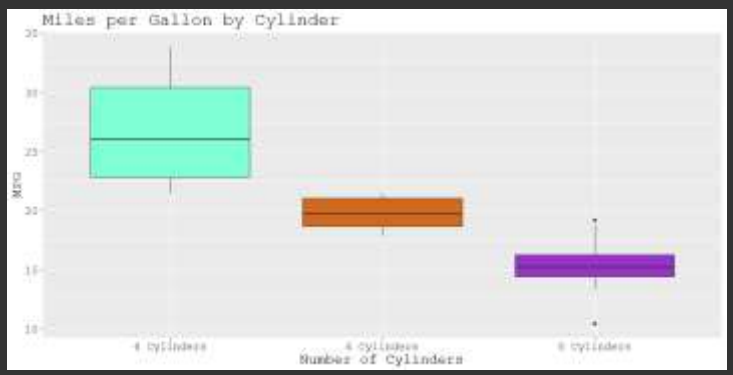
the desired number of names from each hat.

It is not always possible or practical to collect a purely random sample. For example, if your goal is to study all the forest stands in a state, randomly selecting stands of forest across a full state would mean that you would have to travel to, collect data in, and obtain permission to access all of these sites. This may simply be too time intensive, expensive, and/or infeasible. Some areas my be too remote or rugged to access. Some sites may be on private land on which the owner will not permit access. In such situations, it is common to employ clustered random sampling in which locations can be determined beforehand that can be accessed then random point locations can be selected within these clusters. Although not optimal, this does provide a more practical means to collect data.

In the real world, collecting a truly random and unbiased sample from a population may be very difficult due to time, cost, access, or other practical considerations. Thus, it is important to clearly document any limitations in the sampling methods so that the findings are clear and reproduceable. Researchers and scientists struggle to obtain a truly unbiased sample because the world is inherently complex. This is one reason we have laboratories, as they offer more controlled environments to perform experiments and control variables. Coming to grips with uncertainty and limitations is a key part of doing science and data analysis.

Statistical Tests

- You want to know something about a population
- However, you can't measure every member of a population
- So, you take a representative sample
- Measures made on the sample are used to test some hypothesis about a certain characteristic of the population with a certain level of confidence



Statistics may allow us to infer something about the population from a sample drawn from it.

6

The general workflow of performing a statistical test is to first determine what effect you want to assess within a defined population. It is key that your population is clearly defined. Once this population is defined, it is likely that you will not be able to measure every individual, so you must collect a subset of individuals to study. This could be a subset of people, products, turtles, images, counties, etc. The sampling unit and defined population will vary based on the problem of interest. In order to draw conclusions about the larger population from your sample, it is important that the sample is unbiased and representative of the population. Again, there are some practical limitations to being able to collect a large number of randomized samples, so methods and limitations should be clearly documented.

You can then measure each individual in the unbiased sample and use an appropriate statistical test to assess for an effect. If the sample is unbiased, you can then make an inference about the population with a certain level of confidence.

For example, your hypothesis may be that more engine cylinders will result in lower fuel efficiency. So, in this case your entire population is all vehicle models. You would then need to draw a sample of vehicles that is unbiased and representative of all vehicles. Next, you would need to measure or look up reported fuel efficiencies. Once these data are collected, you can then perform

a statistical test to compare the means of multiple groups (e.g., a One-Way ANOVA). Based on the results of this test, you can make an inference with a reported level of confidence about the impact of number of cylinders on fuel efficiency for the entire population based on your sample.

In short, the point of statistical inference is to allow us to infer something about the population from a sample drawn from it.

Null and Alternative Hypothesis

- Null Hypothesis (H_0) = there is no statistical significance or difference (no effect exists)
- Alternative Hypothesis (H_a) = there is statistical significance or difference (effect exists)

- If the statistical test does not show statistical significance, we fail to reject the null hypothesis.
- This is different than proving the alternative hypothesis is true!

"Innocent until proven guilty"

- Enough evidence (i.e., statistical significance) must be presented to take action (i.e., accept the alternative hypothesis) or punish the defendant
- You don't prove a defendant's innocence; you fail to provide enough evidence to reject their innocence

7

The language used in hypothesis testing can be a bit confusing, so we will take some time to step through key terminology and practices here.

Specific hypothesis tests are used to assess two mutually exclusive, or non-overlapping, hypotheses. The null hypothesis states that there is no effect or statistical differences. For the fuel efficiency example, the null hypothesis would be that there is no difference in mean fuel efficiency between vehicles with different numbers of cylinders. In contrast, the alternative hypothesis suggests that there is statistical significance or difference. In our example, the alternative hypothesis would be that not all means of groups of vehicles with the same number of cylinders are the same. The null hypothesis suggests that no effect exists (the number of cylinders does not impact fuel efficiency). In contrast, the alternative hypothesis suggest that an effect does exists (the number of cylinders does impact fuel efficiency).

Another important point is that we are attempting to infer something about the population from the sample. So, the hypothesis test is meant to assess for an effect in a population given the data available as a sample with a certain level of confidence. We will discuss what factors impact the level of confidence in later slides.

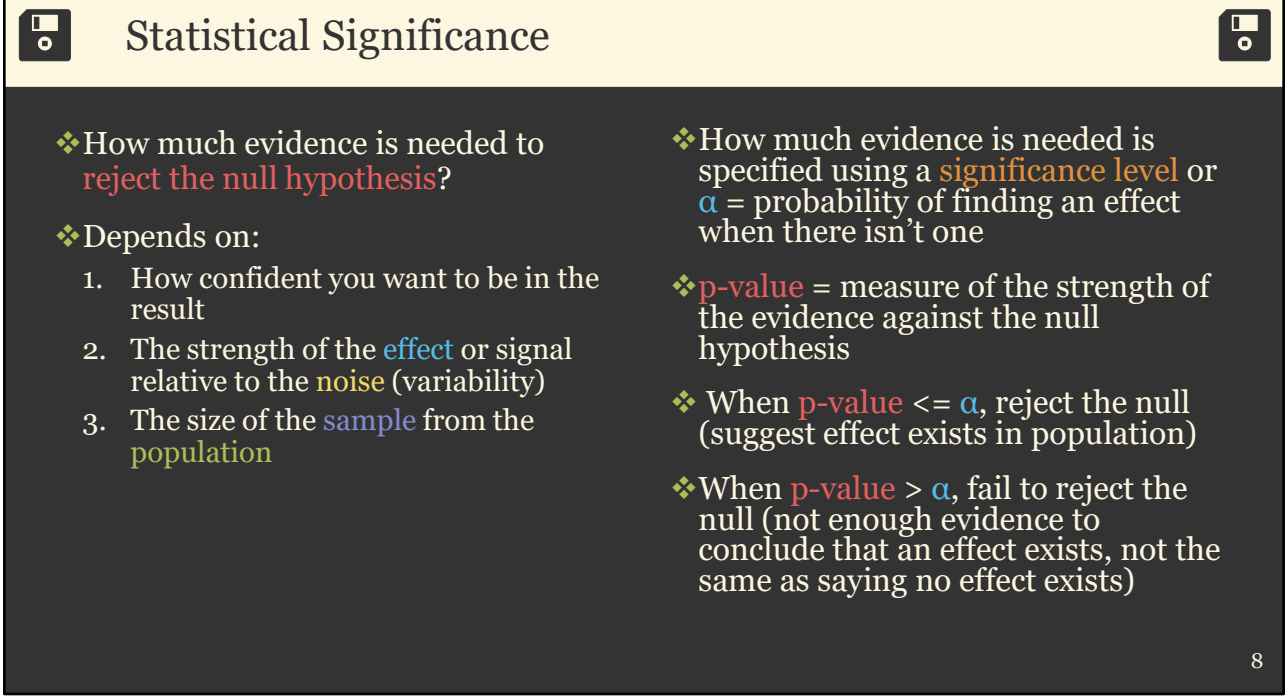
A valid statistical test yields one of two results: (1) if the result is not

statistically significant, we fail to reject the null hypothesis; (2) if the result is statistically significant, we reject the null hypothesis. Note that the result is stated relative to the null hypothesis as opposed to the alternative hypothesis. This is because statistical inference is designed to differentiate these potential outcomes specifically. In other words, statistical tests are designed to test whether there is enough evidence to reject the null hypothesis as opposed to prove the alternative hypothesis.

Although this language may seem overly complicated or that we are splitting hairs, this is very important. Rejecting the null hypothesis in favor of the alternative hypothesis is different from proving that the alternative hypothesis is true. We cannot prove an alternative hypothesis. This is one reason why it is generally frowned upon to use language such as “our study shows” or “our research proves” in scientific reports and articles. This goes against the core concepts of statistical inference and the scientific method more generally. Instead, we tend to use language such as “our study suggests” or “our research supports.”

I like to think of hypothesis tests in the context of legal verdicts. In a court case, the defendant is innocent until proven guilty. Enough evidence must be presented in order to reject the null hypothesis of innocence and accept the alternative hypothesis of guilt. Note that this is different from proving a defendant’s innocence. Part of the distinction here is that accepting the alternative hypothesis implies that some action should be taken to punish the defendant. In other words, the severity of rejecting the null hypothesis or favoring the alternative hypothesis of guilt carries more weight than failing to reject the null hypothesis of innocence. This imbalance in the severity of the action resulting from the decision is important.

If you find this terminology confusing, you are not alone. I have also struggled with this. However, it is important that we report the results of statistical inference correctly so that they are not misleading or inaccurate. This is part of the reason for the need for such specific language.



How much evidence is needed to reject the null hypothesis? This depends on several factors. First, being more confident in your conclusion to reject the null hypothesis requires more evidence. A stronger effect is generally easier to detect and requires less evidence. For example, if fuel efficiency varies greatly based on number of cylinders, then it will be easier to detect. In contrast, if the differences are more subtle, then this effect may be more difficult to detect. This is also impacted by the inherent variability or noise in the data. More variability in the data generally means that an effect is harder to detect.

The concept of effect or signal and noise or variability is often equated to the concept of signal-to-noise ratio. Is there enough of a signal that it can be distinguished from the random background noise or variability?

The sample size also has an effect. It is generally easier to detect an effect if one exists within a larger sample size. In fact, increasing the sample size generally increases the likelihood of detecting an effect even if the signal is not strong and/or there is a lot of noise. If the sample size is very large, the test may yield a statistically significant result that is not practically meaningful.

The amount of evidence needed to reject the null hypothesis depends on how confident you want or need to be in the result. This is generally defined based on a significance level or alpha, which is the probability of finding an effect

when there isn't one. For example, if your confidence level is set to 5%, a common choice, then there is a 5% chance of incorrectly finding an effect when one is not actually present. The acceptable level of finding an effect when there isn't one will depend on the problem at hand. Generally, higher levels of confidence may require larger sample sizes, especially if there is a lot of noise or variability in the data or the effect or signal is small relative to the noise.

Statistical tests allow for the calculation of a p-value, which is a measure of the strength of the evidence against the null hypothesis. If the p-value is smaller than the significance level or alpha, then the evidence suggests to reject the null hypothesis. In other words, there is enough evidence to reject the null hypothesis given the desired level of confidence, the sample size, and the strength of the effect relative to the noise or variability in the data.

If the p-value is larger than the significance level, then the evidence suggests to fail to reject the null hypothesis. In other words, there is not enough evidence to reject the null hypothesis given the desired confidence level, the sample size, and the strength of the effect relative to the noise or variability in the data. Again, this is different from proving the alternative hypothesis correct.

Steps in Conducting a Hypothesis Test

1. Determine what effect you are interested in detecting
2. Define the population for which you want to make an inference about the proposed effect
3. Draw a representative sample from the population
4. Collect the required data for each individual in the sample
5. Perform a statistical test to obtain a test statistic
6. Compare obtained test statistic to a distribution
7. Obtain a p-value
8. Compare p-value to pre-determined α
9. Assess validity of test based on assumptions
10. Report findings

9

This slide outlines the general steps in conducting a hypothesis test. Before moving on, I want to make a few points.

First, your goal is to make an inference about a population from your sample that takes into account a desired level of confidence that is impacted by the sample size, strength of the effect, and inherent level of noise or variability. You are not making an inference about the sample. So, in order to reach this goal, you need to clearly define the effect you are investigating, in the form of mutually exclusive null and alternative hypotheses, and the population to which the inference will be made. Without these considerations clearly defined, your experiment or analysis may be flawed for fail to investigate the phenomenon of interest.

In order to make an inference about the population, it is important that the sample be unbiased and representative of the population, or the results may be misleading.

Many statistical tests require the calculation of a test statistic, such as a T-Value or F-Value, that is then compared to a distribution. Different tests use different distributions. The reason why distributions are used is that we need to be able to model the inherent variability in the population relative to the sample. Since we don't really know the distribution of the measures for the

entire population, they must be modeled based on a theoretical distribution.

Another important point is that the alpha or significance level should be determined prior to performing the test. If you obtain a p-value then decide to lower or raise the confidence level to obtain the desired result (i.e., reject the null hypothesis), then you are augmenting your experiment to fit your needs, which is unethical.

Statistical tests have assumptions that may invalidate the results if they are not met. We will talk about assumptions for specific tests later in this module. It is generally important to assess assumptions to make sure your results can be trusted.

Lastly, you should report your results using statistically accurate terminology. For example, for the fuel efficiency example if you fail to reject the null hypothesis you could say: “A p-value of 0.41 was obtained, which suggest failure to reject the null hypothesis that the means are not different between vehicles with different numbers of engine cylinders at a 0.05 significance level.” If you reject the null hypothesis, you could state: “A p-value of 0.003 was obtained, which suggest a rejection of the null hypothesis that the means are not different between vehicles with different numbers of cylinders at a 0.05 confidence level in favor of the alternative hypothesis that at least two of the means are different.”

Confidence Intervals

- Goal is to estimate a population parameter (e.g., population mean)
  - Population parameter is unknown
- Draw representative sample from the population to obtain a sample statistic (e.g., the sample mean)
- The sample statistic will not exactly approximate the population parameter
- Confidence Interval (CI) = range of values that the population parameter will likely fall within based on the sample statistic, sampling error, and a defined level of confidence
- Confidence Interval depends on:
  - Sample size
  - Variability
  - Level of confidence desired

10

There is a relationship between the confidence interval and significance level or alpha. A 95% or 0.95 confidence interval is equivalent to an alpha of 0.05. In other words, the confidence interval is equal to 1 minus alpha.

In terms of statistical significance, if the p-value is less than the pre-defined alpha level then you reject the null hypothesis. Similarly, if the range of values at the desired confidence level does not include the value associated with the null hypothesis, then you reject the null hypothesis. For example, if you are assessing if two means are different, the confidence interval must exclude a value of zero (i.e., no difference) in order to reject the null hypothesis at the given significance level or alpha.

Confidence Intervals

For a statistical test:

$$CI = 1 - \alpha$$

If test is statistically significant:

- p-value is smaller than the α = reject the null hypothesis
- CI excludes the value associated with the null hypothesis

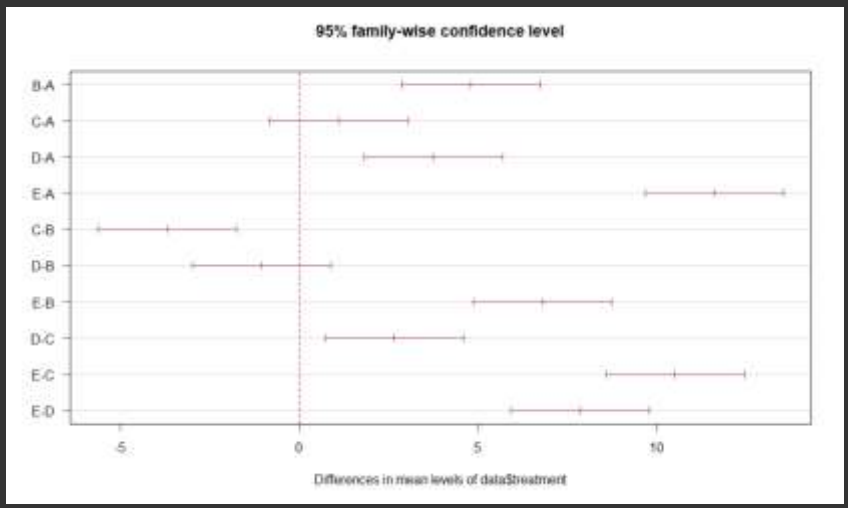
11

There is a relationship between the confidence interval and significance level or alpha. A 95% or 0.95 confidence interval is equivalent to an alpha of 0.05. In other words, the confidence interval is equal to 1 minus alpha.

In terms of statistical significance, if the p-value is less than the pre-defined alpha level then you reject the null hypothesis. Similarly, if the range of values at the desired confidence level does not include the value associated with the null hypothesis, then you reject the null hypothesis. For example, if you are assessing if two means are different, the confidence interval must exclude a value of zero (i.e., no difference) in order to reject the null hypothesis at the given significance level or alpha.

Graphical Explanations

- Example from: https://r-graph-gallery.com/84-tukey-test.html



12

Here is an example taken from the provided website in which pairs of groups are being compared to determine if their means are different. This is a post-hoc pairwise test, which we will discuss later in the module. At the 95% confidence level, if the range of values includes zero then we fail to reject the null hypothesis that the difference between the group means is zero. So, in this example, you would fail to reject the null hypothesis for Group C and A and Group D and B since the confidence interval contains zero. For all other pairs, you would reject the null hypothesis.

Concerns with the *p*-value

- Only statistically significant results published
- Pressure to publish
- "p-hacking"
- Cherry picking results
- Experiments not reproduceable
- No incentive to replicate experiments
- Selecting a significance level
- Oversimplifying results

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Wasserstein, R.L. and Lazar, N.A., 2016. The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), pp.129-133.

13

It is common in scientific studies to set a significance level, such as 0.05, as a hard cutoff to determine whether the results are statistically significant. However, this practice has come under scrutiny.

Academic journals generally aim to publish results that are statistically significant. Since many researchers must publish their work in order to keep their jobs, get promotions, and be able to competitively pursue grant funding, this leads to a desire and need to obtain statistically significant results. Researchers under pressure may be tempted to cherry pick results that are statistically significant or allow for obtaining a statistically significant p-value. This is known as p-hacking.

Further, the use of a single p-value to determine whether an effect exits, or results are statistically significant, was not intended by the originators of statistical inference. Instead, this was only meant to serve as a single piece of evidence. Repeated experimentation and replication of other researchers' experiments are meant to further support, or potentially refute, results. Unfortunately, there is no incentive for researchers to replicate their own work or replicate the experiments of other researchers. This is because researchers are encouraged to investigate new problems, and journals do not generally publish replications of existing experiments.

In short, there are some systematic issues in how modern research is conducted, how likely it is for research to be published, and how researchers and professionals are rewarded or evaluated. The phenomenon of p-hacking is one component of this issue.

I have several opinions relating to these issue; however, these are just my opinions and others may disagree. First, I feel that the statistical significance of results should not be considered when a decision is made to publish a paper. This leads to a bias in the literature and the state of knowledge. This bias is counter to the process of experimentation and statistical inference. More broadly, it is counter to the scientific method. Instead, the quality and soundness of the works and methods used should be of highest concern. I also think there should be more incentive to replicate results. The publishing of data and research code and the adoption of reproduceable scientific methods is key to replication and transparency.

In regards to p-values specifically, I feel that they do have value. However, I think the magnitude of the p-value is of more importance than whether it is higher or lower than some arbitrary cut-off. For example, a p-value of 0.00001 and a p-value of 0.045 both would yield a decision to reject the null hypothesis at an alpha of 0.05. However, the much lower p-value suggests a potentially much stronger effect than the one closer to the arbitrarily defined significance threshold. Obtained p-values are just one piece of evidence that can be considered in assessing data and reporting results. A well-designed, reproduceable experiment, associated, thoughtful evaluation and reporting are more important than if a single p-value is higher or lower than a threshold. Again, this is just my opinion.

Parametric vs. Nonparametric

- Parametric = assumes data from population follow a specified probability distribution
- Nonparametric = does not assume data from population follow a specified probability distribution

This does not imply that nonparametric tests don't have assumptions!

For Example:

T-Test
- Parametric test to compare the means of two groups
- Assumes a normal distribution of the measure of interest in population

Mann-Whitney Test
- Nonparametric test to compare the medians of two groups
- Does not assume a normal distribution of the measure of interest

14

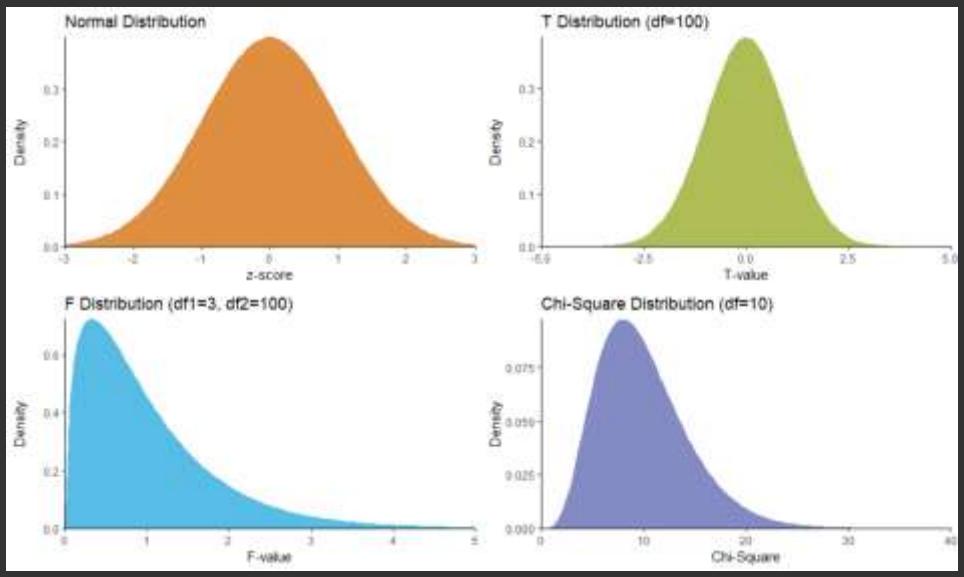
Before moving on to discuss some specific statistical tests, I want to comment on the difference between parametric and nonparametric tests. Parametric tests have distribution assumptions, such as assuming that the data are normally distributed. In contrast, nonparametric tests do not have distribution assumptions or do not assume that the data are normally distributed.

As a specific example, a T-Test is a parametric test that is used to compare the means of two groups. In this test, the variable of interest is assumed to be normally distributed in the population. In contrast, the nonparametric Mann-Whitney Test compares the medians of two groups, as an alternative measure of central tendency, and does not assume the measure of interest is normally distributed.

Practically, if the assumption of normality is not met then a parametric test may not be appropriate, or the results may be misleading. In such as case, the Mann-Whitney Test may be more appropriate.

There are a few important caveats. First, some tests are robust to the assumption of normality if the sample size is sufficiently large. We will discuss this for specific tests later in this module. Second, nonparametric tests are not assumption-free; they just do not have distribution assumptions.

Example Distributions



15

Different statistical tests use different distributions to compare the obtained sample statistic. For example, a T-Test uses the T-Distribution, the F-Test uses the F-Distribution, and the Chi-Square Test using the Chi-Square Distribution. Further, these distributions are not one specific distribution; instead, distributions vary based on some characteristics of the analysis, notably the degrees of freedom. For example, the T-Distribution varies based on the number of samples while the F-Distribution varies based on the number of samples and the number of different groups being compared. The Chi-Square distribution depends on the number of levels in each of the two categories being compared.

Once a test statistic is obtained, it must be compared to the appropriate distribution in order to determine the associated p-value, which is then compared to the pre-determined significance level or alpha.

Why is it necessary to use these theoretical distributions? Again, this arises from the fact that we do not know the true distribution or natural variability within the population. So, we must compare the results obtained from the sample to a theoretical distribution that represents a population. Using such distributions in this manner is one of the key characteristics of statistical inference. Some distribution assumptions are made in order to make an inference about a population from a sample when the population parameter is

unknown, or this arises from the inability to measure every member of the population.

Type I and Type II Error

	Test Rejects Null	Test Fails to Reject the Null
Null is True	False Positive	No Effect
Null is False	Effect Exists	False Negative

- False Positive = Type I Error = Suggests an effect exists when one does not
- False Negative = Type II Error = Suggests an effect does not exist when one does

- Probability of making a Type I Error = α = significance level
- Probability of making a Type II Error = β

16

When performing statistical tests, four outcomes are possible when comparing the results of the statistical test to the actual effect in the population. Again, we don't generally know the true effect within the population, so we describe the likelihood of making certain types of errors or obtaining incorrect results.

The following four outcomes are possible.

1. Suggest an effect is present when one is present
2. Suggest an effect is not present when one is not present
3. Suggest an effect exists when one does not
4. Suggest an effect does not exist when one does

Of these possible outcome, the first two would represent correct inferences and the second two would represent incorrect inferences about the population.

Suggesting that an effect exists when one does not is a False Positive or Type I Error. When performing a statistical test, the probability of making a Type I Error is equivalent to the significance level (i.e., alpha). So, at a 95% confidence level if the null hypothesis is rejected, we can say that there is less than a 5% chance that we rejected the null hypothesis when it was true or there was no effect.

A Type II Error is a False Negative: you suggest that an effect does not exist when one does. The probability of making this type of error is known as beta.

One important consideration is that the seriousness of making a Type I vs. making a Type II error is often not the same. For example, for a medical test that detects a specific cancer, a False Negative or Type II Error (i.e., suggesting the patient does not have cancer when they does) is likely more serious than a False Positive or Type I Error (i.e., suggesting the patient has cancer when they do not).

The relative severity of Type I and Type II errors may impact the significance level chosen and the interpretation of the hypothesis test. For example, if a False Negative is deemed to be more serious, you may choose to lower the significance level threshold. In other words, it would be better to be less confident in suggesting that a patient has cancer when they do in order to reduce the number of times a test yields a False Negative. There is often a trade off between Type I and Type II error.

Statistical Power

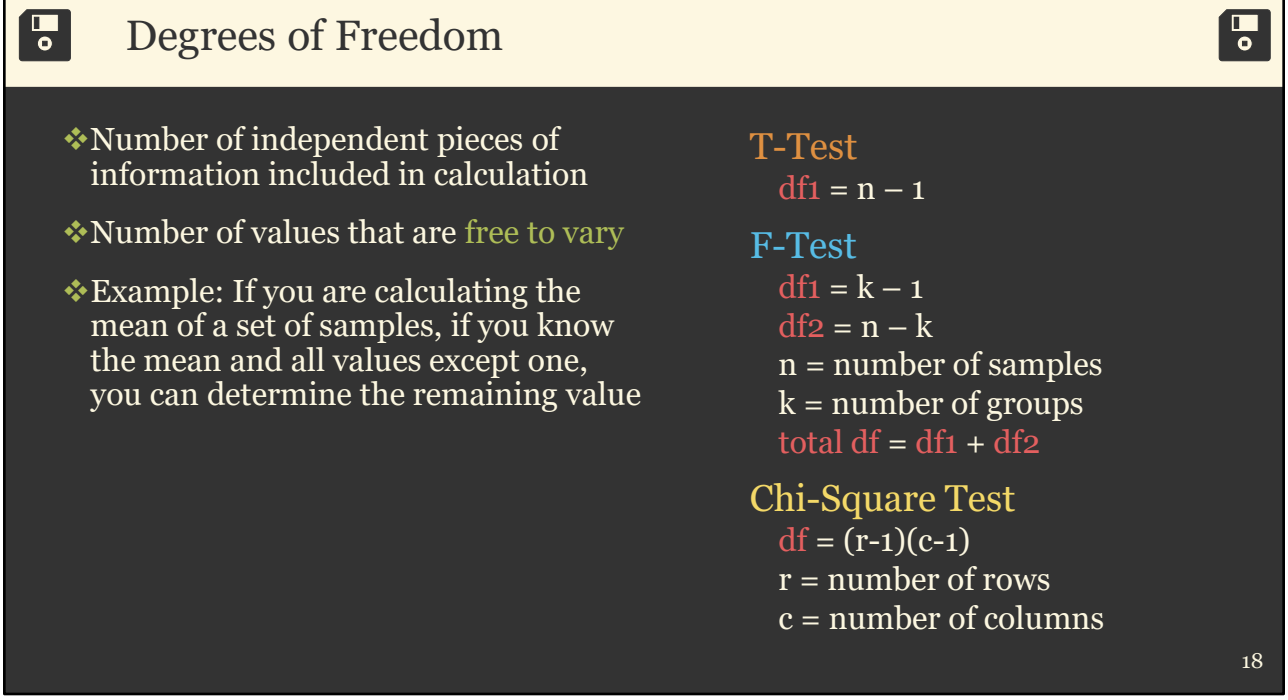
- Probability of correctly detecting an effect
- 1 – Type II Error Rate (β)
- The statistical power of a hypothesis test depends on:
  - Sample Size
  - Variability in the Population
  - Effect Size

- For a specific statistical test, given the anticipated effect size and desired significance level and power, what is the number of required samples?

17

Related to Type I and Type II error, statistical power is the probability of correctly detecting an effect, equivalent to one minus beta (i.e., 1 minus the Type II Error Rate). Statistical power is impacted by the sample size, the variability in the population, and the effect size. Generally, statistical power increases with larger sample sizes and larger effect sizes while decreasing with the variability in the population. In other words, we generally have a higher probability of correctly detecting an effect if the signal is large relative to the noise and our results are based on a large sample.

There are methods available to assess statistical power based on sample size, variability, and effect size. This generally requires an estimate of the effect size and variability. Again, we don't know the actual population variability. With assumptions of effect size and population variability, it is possible to estimate a required sample size to obtain the desired level of statistical power. This is one key component of designing an experiment. Having too small of a sample size can limit statistical power and make results less useful.



Lastly, I wanted to briefly discuss the concept of degrees of freedom since this comes up a lot when we are discussing statistical tests.

Degrees of freedom is the number of independent pieces of information included in a calculation. The key word here is independent. For example, if you know your mean test score in a course then you can determine a missing test score if you know all other exam scores. So, one of the test scores is not free to vary because only a specific test score would yield the obtained mean if all other test scores are held constant.

In short, degrees of freedom takes into account how many pieces of information can vary in an analysis and is an important component of many statistical tests. For example, degrees of freedom are important for determining the correct T-, F-, or Chi-Square distribution for a specific experiment. The T-Test degrees of freedom is the number of samples minus 1. Or, if we know the mean and all sample values, the remaining sample value is not free to vary. An F-Test has two degrees of freedom that are based on the number of groups being compared and the number of total samples. The degrees of freedom for a Chi-Square test depend on the number of levels in the two categories being compared.

The mathematics behind the concept of degrees of freedom are outside the

scope of this course but would be covered in a more advanced statistics course.

Summary of Key Points: Statistical Inference

- Statistic inference was developed to allow for the investigation of a population using a sample from the population.
- Sample must be randomized and unbiased.
- Statistical inference allows for hypothesis testing for mutually exclusive null and alternative hypotheses.
- Statistical tests yield a test statistic that is compared to a distribution for a p-value. The p-value is then compared to a threshold to determine if results are statistically significant relative to a pre-determined significance level.
- The use of p-values and comparison to a signifance threshold have come under scrutiny and some valid concerns have been raised.
- A confidence interval can be calculated based on the sample size, the effect size relative to the inference variability in the population, and the desired confidence level.
- Statistical inference errors: Type I and Type II error. Statistical power relates to the likelihood of making a Type II error.
- Parametric tests have distribution assumptions. If assumptions are not valid, the results may be incorrect or misleading.

Example Statistical Tests

20

Now that we have discussed the key concepts associated with statistical inference, we will move on to explore some common statistical tests. Again, our focus will be understanding what these tests are for, how they are implemented, how they are interpreted, and assumptions that must be met for the results to be valid and not misleading.

1-Sample T-Test

- Purpose = Infer whether a population mean equals a hypothesized mean
- H_0 (Population mean = hypothesized mean)
- H_1 (Population mean ≠ hypothesized mean)
- Assumptions
  1. Representative, random sample
  2. Continuous data
  3. Sample data normally distributed or more than 20 observations
  4. Observations are independent

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

Signal

Noise

$\bar{x}$ = sample mean
μ_0 = mean comparing sample to
s = standard deviation of sample
n = sample size

21

A 1-Sample T-Test is used to assess if the mean of a population is equal to a hypothesized mean. For example, you could assess whether the change in blood pressure for individuals with high blood pressure is equal to zero (no change) after taking a medication. However, the hypothesized mean does not need to be zero; it can be any value to which you want to compare the population mean.

The null hypothesis is that the population mean is equal to the hypothesized mean while the alternative hypothesis is that the population mean is not equal to the hypothesized mean. Note that the null and alternative hypotheses are stated relative to the population mean as opposed to the sample mean since our goal is to make an inference about the population from the sample.

The obtained T-Value is compared to a T-Distribution with the correct degrees of freedom (the number of samples minus one) in order to obtain a p-value. If the p-value is lower than the desired significance level or alpha, then the results suggest to reject the null hypothesis that the population mean is equal to the hypothesized mean in favor of the alternative hypothesis that the mean is not equal to the hypothesized mean. If the p-value is larger than alpha, we fail to reject the null hypothesis that the population mean is equal to the hypothesized mean.

The T-Test is a good example of the conceptualization of a statistical test as a tool to distinguish a signal from inherent noise. The numerator can the thought of as the signal; in this case, it is the difference between the sample mean and the hypothesized mean. The denominator can be thought of as the noise: it is the standard deviation divided by the number of samples. A larger standard deviation results in a larger noise component, which means that a stronger signal or effect must be present for it to be distinguished from the noise.

The T-Test is a parametric test with several assumptions. First, it assumes that the sample is representative of the population. Without an unbiased sample, making an inference about the population from the sample can be misleading. The quantity being investigated must be numeric or continuous as opposed to nominal or ordinal. It is also assumed that the quantity of interest is normally distributed unless the sample size is larger then 20. When the sample size is larger than 20, this assumption can be violated without invalidating the results.

2-Sample T-Test

- Purpose = Infer whether population means of two groups are different
- H_0 (Group 1 mean = Group 2 mean)
- H_1 (Group 1 mean ≠ Group 2 mean)

Assumptions

1. Representative, random sample
2. Continuous data
3. Sample data normally distributed or more than 20 observations
4. Observations are independent
5. Correct formula used depends on whether group variances are equal

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$ If variance of two groups is equal

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_1}{n_1}+\frac{s^2_2}{n_2}}}$$ If group variances are different

$\bar{x}$ = sample means
s = standard deviation of sample
s^2 = variance of sample
n = sample size

22

In contrast to a 1-Sample T-Test, a 2-Sample T-Test is used to compare the population means of two groups as opposed to the population mean of a single group to a hypothetical mean. The null hypothesis is that the population means of the two groups are equal while the alternative hypothesis is that the two population group means are not equal.

The test will yield a T-Value that is compared to the appropriate T-Distribution based on degrees of freedom. If the p-value is lower than alpha, you reject the null hypothesis that the population group means are equal in favor of the alternative hypothesis that they are not equal. If the p-value is larger than alpha, you fail to reject the null hypothesis that the population group means are equal. As an example, a T-Test could be used to assess whether the mean fuel efficiency of vehicles with 8-cylinder engines is different from that of vehicles with 4-cylinder engines.

The numerator or signal is simply the difference between the means. If the variance of the two groups is equal, a pooled measure of variance is used in the denominator to represent the noise. This is the top equation on the slide. If the group variances are different, than the bottom equation is used.

A 2-Sample T-Test has the same assumptions as a 1-Sample T-Test with the added concern that the correct formula must be selected based on if the

variance of the two groups is equal. The normality assumption is not an issue if more than 20 sample data points are collected.

Paired T-Test

- Purpose = Infer whether population means are different for two measurements collected on dependent samples
- H_0 (Mean difference between paired measures = 0)
- H_1 (Mean difference between paired measures ≠ 0)

Assumptions

1. Representative, random sample
2. Continuous data
3. Sample data normally distributed or more than 15 observations per group
4. Groups are independent
5. Sample pairs are dependent

23

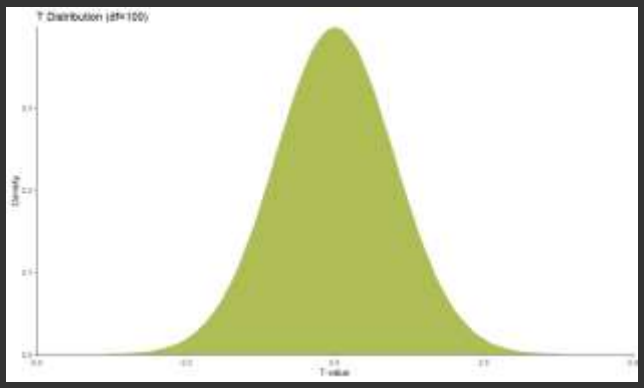
A Two-Sample T-Test is used when the samples across groups are independent. In contrast, a Paired T-Test is used when sample pairs are not independent. For example, a Paired T-Test would be appropriate when a measure is being compared for a set of patients before and after a drug is administered. In this case, a single patient will be represented as a dependent before-and-after sample pair.

T-Test Interpretation

- Use resulting T-Value and T-Distribution to determine p-value
- Compare p-value to α
- When p-value <= α, reject the null
- When p-value > α, fail to reject the null

Two-Tailed = Just assessing difference in means

One-Tailed = Assessing specific direction, larger or smaller mean



```
t = -7.0968, df = 33.012, p-value = 5.976e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.069117 -2.253890
sample estimates:
mean in group North Dakota         mean in group Utah
                  5.109584                   8.272087
```

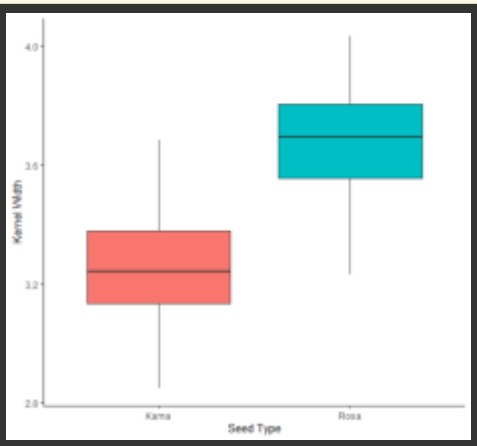
24

This slide highlights the interpretation of a T-Test. Note the use of the T-Distribution, which will vary based on sample size or degrees of freedom.

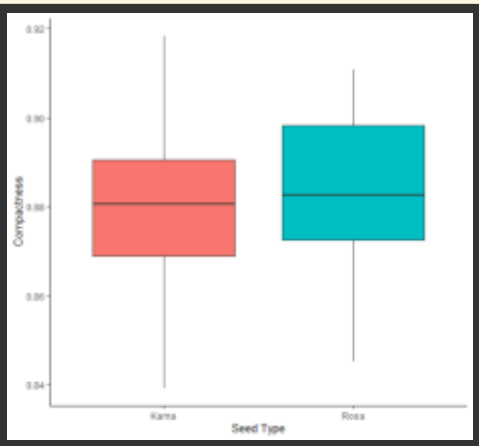
Another distinction that can be made is if you are assessing whether means are different or whether one group has a mean that is specifically larger or smaller than the other group. A two-tailed test, which is generally the default, is just assessing for difference; it is not assessing for the direction of the difference. In contrast, a one-tailed T-Test assesses whether the population mean of one group is specifically higher or lower than that of the other group.

The choice of one- vs. two-tailed tests depends on the question being investigated. Again, if you are only interested in assessing for a difference, a two-tailed test is appropriate. However, if you are assessing if one population group mean is higher or lower than the other, then a one-tailed test is appropriate. If you are in doubt, it is generally recommended to use a two-tailed test.

Example Data: Seeds




```
        Welch Two Sample t-test

data:  kernel by type
t = -14.097, df = 137.74, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4934688 -0.3720825
sample estimates:
mean in group 1 mean in group 2
       3.244629        3.677414
```



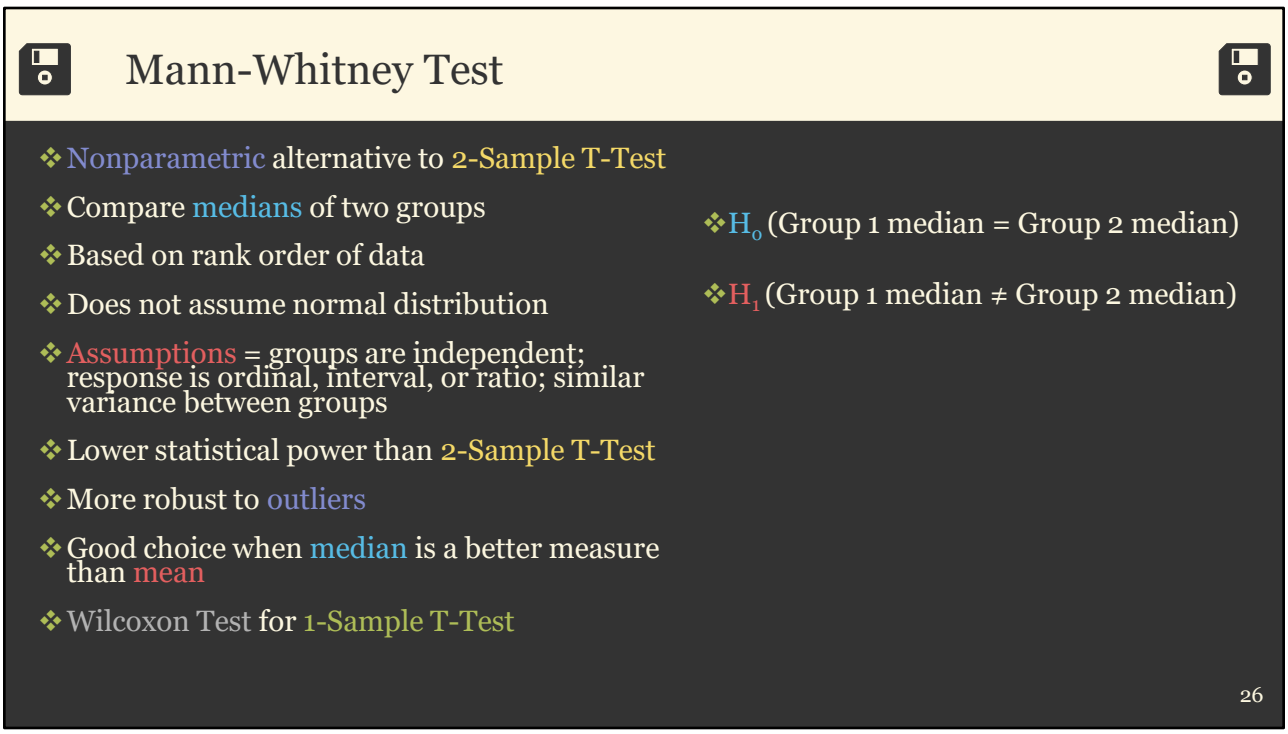

```
        Welch Two Sample t-test

data:  compactness by type
t = -1.2867, df = 117.74, p-value = 0.2003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.008744452  0.001850147
sample estimates:
mean in group 1 mean in group 2
      0.8800700       0.8835171
```

25

This slide represents the results for Two-Sample T-Tests. In the box plot and results on the left, the kernel width of the Kama and Rosa wheat varieties are compared while on the right the compactness of these varieties is compared. Only two of the three varieties can be compared using the T-Test since this test does not allow for comparing the means of more than two groups.

Based on the box plot for the kernel width comparison, it appears as if the mean kernel width of the two groups is different. This is confirmed by the statistical test in which a p-value of less than 2.2e-16 is obtained. This suggests that the null hypothesis should be rejected in favor of the alternative hypothesis that the kernel widths between the two groups are different. Note also that the confidence interval does not include zero, which again indicates that there is sufficient evidence to reject the null hypothesis.

For the comparison of compactness, we see less of a difference between the groups based on the box plots. This is confirmed by a p-value of 0.20, which suggests that we fail to reject the null hypothesis that the means are different. Also, note that the confidence interval includes zero.



The nonparametric alternative to the Two-Sampled T-Test is the Mann-Whitney Test while the alternative for the 1-Sample T-Test is the Wilcoxon Test. These tests rely on ranks and assess the population medians as opposed to means. They are nonparametric because they do not assume the quantity being compared is normally distributed.

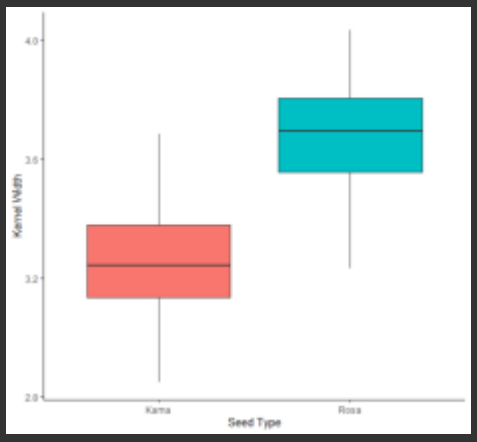
It is important to note that, although normality is not assumed, there are some assumptions. For example, the Mann-Whitney Test assumes that the groups are independent; the variable being compared is of the ordinal, interval, or ratio measurement level; and that the variance between groups is similar.

When might you chose to use these nonparametric alternatives as opposed to the parametric methods? First, you may choose the nonparametric options if you have a small sample size, the variable of interest is not normally distributed, or the distribution of the quantity of interest is far from normal. The nonparametric tests tend to be more robust to outliers, so this is another consideration. Lastly, these nonparametric tests may be a better option if the median is a better measure of central tendency than the mean for the specific dataset or problem of interest.

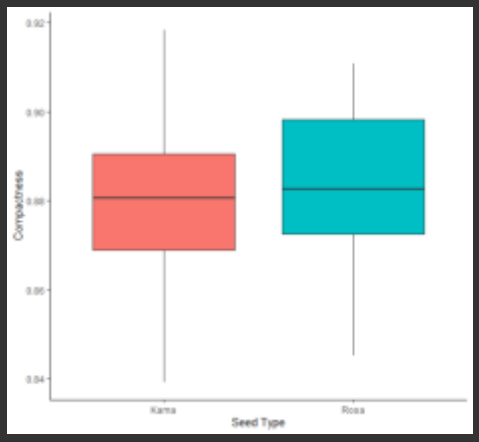
I often find it useful to calculate both the parametric and associated nonparametric tests then compare the results. If both results are statistically

significant or not statistically significant, then this can add further support to your results. If the results are different, then you may have to more thoroughly scrutinize the data to determine which test is most valid.

Example Data: Seeds




```
        Wilcoxon rank sum test with continuity correction

data:  wkernel by type
W = 234, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```



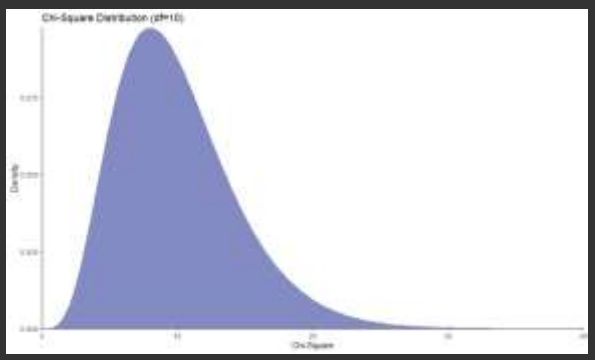

```
        Wilcoxon rank sum test with continuity correction

data:  compactness by type
W = 2140.5, p-value = 0.1978
alternative hypothesis: true location shift is not equal to 0
```

27

These graphs show the Mann-Whitney results for the kernel width and compactness comparisons. These results support the Two-Sample T-Test results: statistical significance is observed for kernel width but not for compactness.

Chi-Square Test

- Purpose = Infer whether there is a relationship between two categorical variables within a population of interest
- H_0 (No relationship between categorical variables)
- H_1 (Relationship between categorical variables does exist)



	A	B	C	D	E	F
1	#	#	#	#	#	#
2	#	#	#	#	#	#
3	#	#	#	#	#	#
4	#	#	#	#	#	#

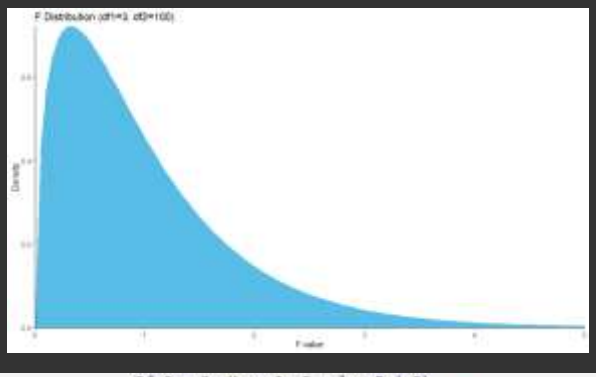
28

The Chi-Square test is used to infer if there is a relationship between two categorical variables. It relies on a Chi-Square distribution with the appropriate degrees of freedom. The degrees of freedom is based on the number of levels in each of the two categorical variables included in the test.

The null hypothesis is that no relationship exists between the categorical variables in the population. In other words, the label for Category 1 has no relationship or dependency with the label for Category 2. The alternative hypothesis is that there is a relationship between the categorical variables in the population.

One-Way ANOVA

- Purpose = Infer whether at least two population group means are different when there are three or more groups
- H_0 (All group means are equal)
- H_1 (Not all group means are equal)
- Assumptions
  1. Representative, random sample
  2. Independent groups
  3. Continuous dependent variable
  4. Categorical independent variable
  5. Sample data normally distributed or more than 15 or 20 observations per group
  6. Equal variance between groups (Welch's ANOVA)



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
STATE_NAME	7	3356	479.5	176.9	<2e-16	***
Residuals	481	1303	2.7			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

29

A Two-Sample T-Test can only be used to compare the means of two groups. If you have more than two groups, you can use One-Way Analysis of Variance (ANOVA). The null hypothesis for a One-Way ANOVA is that all population group means are equal. In contrast, the alternative hypothesis is that not all population group means are equal. Note that this test does not tell you which group means are different. A statistically significant result only suggests that at least two of the included groups have population means that are different.

ANOVA has several assumptions. First, the sample must be representative of the population in order to obtain results that are not misleading or can be extrapolated to the population. The different groups should be independent of each other or the values in one group should not impact the values in any other group. The value being compared between groups should be a numeric variable while the groups are defined by a nominal variable. This is a parametric test that assumes the numeric variable being compared is normally distributed; however, this assumption can be relaxed if more than 15 or 20 samples are included per group. It is also assumed that there is equal variance between the groups. However, if this is assumption is violated, the Welch's ANOVA can be used.

F-Table

Source	DF	SS	MS	F-Value	P-Value
Treatments	k-1	SST	MST = SST/(k-1)	F-Value = $\frac{MST}{MSE}$	p-value based on F-Distribution with appropriate DFs and F-Value
Error	n-k	SSE	MSE = SSE/(n-k)		
Total	n-k	SS			

```
                Df Sum Sq Mean Sq F value Pr(>F)
STATE_NAME       7   3356   479.5   176.9 <2e-16 ***
Residuals      481   1303     2.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- k = number of groups
- n = number of samples
- SST = Sum of square difference between group means and the global mean where each group difference is multiplied by the sample size in that group
- SSE = Sum of square differences between each sample and its group mean
- SS = Sum of square difference between each sample and global mean (SST + SSE)

$$\text{F-Value} = \frac{MST}{MSE}$$

Signal

Noise

30

This slide describes the F-Table and the calculation of the F-Value, which is used in the ANOVA test.

Similar to the T-Value, the F-Value can be thought of as a measure of signal-to-noise. The calculations rely on comparing the differences in group means to the global group means, as a measure of signal, and the within-class variability, as the noise, based on the difference between each sample and its group mean. So, even though the goal is to compare group means, the assessment is based on characterizing between-group and within-group variability. That is why this technique is known as analysis of variance.

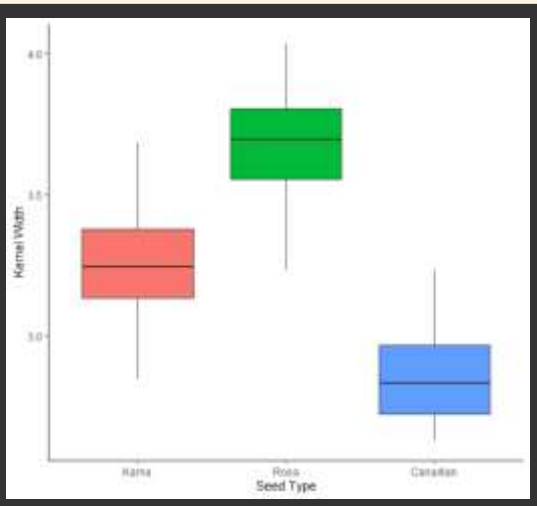
The difference between group means, or signal, is summarized using the sum of squares treatment. This is simply calculated as the square difference between group means and the global mean where each group difference is multiplied by the sample size in that group. This is then divided by the number of groups minus one to obtain the mean sum of squares treatment. The noise is characterized using the sum of squares error, which is calculated as sum of square differences between each sample and its group mean. The mean sum of squares error is then calculated by dividing the sum of squares error by the number of samples minus the number of groups. The F-Value is then calculated as the mean sum of squares treatment, signal, divided by the mean sum of squares error, noise. It should be noted that the terminology associated

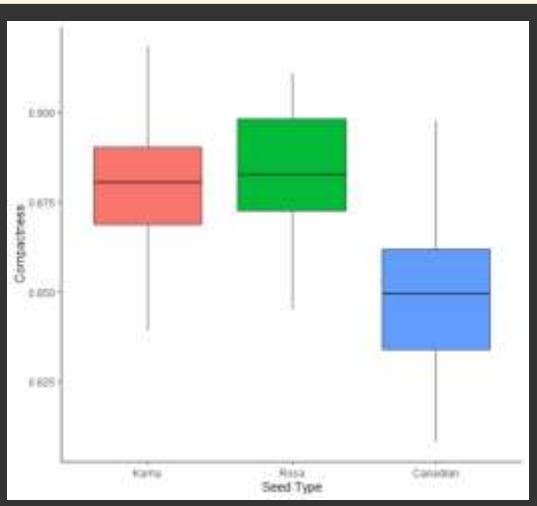
with the table is not used consistently, so you may see other labels for the treatment and error components of the variance.

In order to obtain a p-value, the F-Value is compared to an appropriate F-Distribution defined by two different degrees of freedom: the number of groups minus one and the number of samples minus the number of groups.

Another way to think about ANOVA is that the total amount of variance can be divided into two components: the variability between groups and the variability within groups. The signal, or variability between groups, must be large enough that it can be distinguished form the noise, the variability within groups.

Example Data: Seeds





	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	23.764	11.882	406.3	<2e-16 ***
Residuals	207	6.054	0.029		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	0.04936	0.024680	75.87	<2e-16 ***
Residuals	207	0.06734	0.000325		

31

These results expand upon the Two-Sample T-Test results described previously. I have added in the Canadian group since we can now compare more than two groups. Again, the box plot for the kernel width suggests that all three groups are different. In contrast, the box plots for compactness suggest that the Kama and Rosa varieties are not different while the Canadian variety is different from the other two.

As described in the table, both experiments suggest statistical significance, or we reject the null hypothesis in favor of the alternative hypothesis that at least two of the population group means are different for the measure of interest: kernel width or compactness.

Note that One-Way ANOVA does not indicate which means are different, only that at least two are different.

Before we move on, I wanted to discuss the degrees of freedom. The sum of squares treatment, which in this case is associated with the “type” row, is 2. This is because the degrees of freedom for the treatment is equal to the number of groups, in this case 3, minus 1. The degrees of freedom for the sum of squares error is equal to the number of samples, in this case 210 (70 per group), minus the number of groups (3).

Pair-Wise Comparison

- One-Way ANOVA does not tell you which means are different
- Multiple hypothesis tests increases chance of a false positive
- Family error rate limited by significant level
- Example: Tuckey's Method
  - Adjusted p-values
  - Confidence intervals

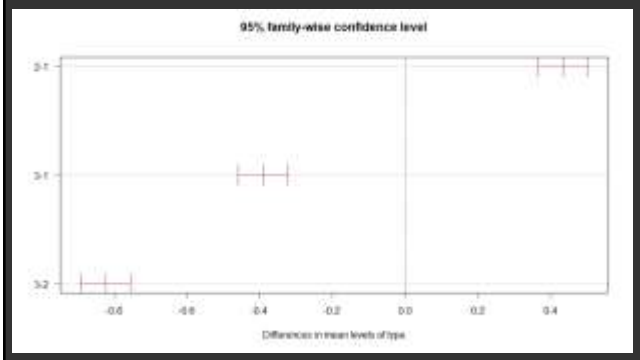
$STATE_NAME	diff	lwr	upr	p adj
Kansas-Colorado	5.79408322	4.909344528	6.6888219	0.0000000
Montana-Colorado	-1.04806783	-1.965074015	-0.1310616	0.0127131
Nebraska-Colorado	2.97853221	2.164487922	3.7925765	0.0000000
North Dakota-Colorado	-1.79668701	-2.667431873	-0.9259421	0.0000006
South Dakota-Colorado	0.87163656	-0.007538505	1.7508116	0.0538864
Utah-Colorado	1.42581618	0.584009224	2.5476231	0.0031044
Wyoming-Colorado	-0.96872817	-2.187076194	0.2496198	0.2356977
Montana-Kansas	-6.84215105	-7.671489522	-6.0128126	0.0000000
Nebraska-Kansas	-2.81555101	-3.529162961	-2.1019391	0.0000000
North Dakota-Kansas	-7.53077022	-8.375196232	-6.6863442	0.0000000
South Dakota-Kansas	-4.92244666	-5.709668399	-4.1352249	0.0000000
Utah-Kansas	-4.36826704	-5.419561536	-3.3169725	0.0000000
Wyoming-Kansas	-6.76281139	-7.916562587	-5.6090603	0.0000000
Nebraska-Montana	4.02660004	3.178936919	4.8742632	0.0000000
North Dakota-Montana	-0.68861917	-1.649007633	0.2717693	0.3639826
South Dakota-Montana	1.91970439	1.095205179	2.8302036	0.0000000
Utah-Montana	2.47388401	1.327362000	3.6204052	0.0000000
Wyoming-Montana	0.07933908	-1.161801728	1.3204818	0.9999995
North Dakota-Nebraska	-4.71521921	-5.577726274	-3.8527121	0.0000000
South Dakota-Nebraska	-2.10689565	-2.913481747	-1.3003095	0.0000000
Utah-Nebraska	-1.55271602	-2.615588096	-0.4898440	0.0003041
Wyoming-Nebraska	-3.94726038	-5.114389789	-2.7801310	0.0000000
South Dakota-North Dakota	2.60832357	1.683988946	3.5326582	0.0000000
Utah-North Dakota	3.16250319	2.004964184	4.3200422	0.0000000
Wyoming-North Dakota	0.76795883	-0.483967502	2.0198852	0.5734358
Utah-South Dakota	0.55417962	-0.562316593	1.6706738	0.0013743
Wyoming-South Dakota	-1.84036471	-3.053822760	-0.6269067	0.0001347
Wyoming-Utah	-2.39454496	-3.795823612	-0.9932664	0.0000078

32

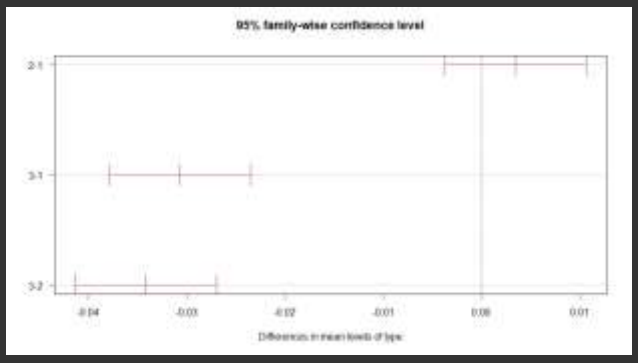
As already noted, if One-Way ANOVA suggests statistical significance, it doesn't indicate what groups are suggested to have different population means. Instead, it only tells you that at least two group means were suggested to be different. In order to determine what population group means are different, a post-hoc test can be performed. There are several options. Here, I demonstrate pair-wise comparison and Tuckey's Method.

There is one important issue here. Performing multiple hypothesis tests increases your likelihood of obtaining a false positive result or suggesting an effect exists when one does not (i.e., a Type I Error). So, the results must be adjusted to take this into account so that the family of hypothesis tests have a collective Type I Error rate that sums to the overall significance level, such as 0.05. How this is accomplished is beyond the scope of this course. However, it is important to be aware of this important issue.

Example Data: Seeds



Kernel Width



Compactness

33

These graphs highlight the results of pair-wise comparisons associated with the One-Way ANOVA results discussed previously. Specifically, the confidence intervals are shown for each pair of groups. If the confidence interval includes zero, then this suggests that the groups are not different. In contrast, if the confidence interval does not include zero, this suggests difference.

For the kernel width results, no confidence intervals include zero, which suggests that all groups are different. For the compactness, both Kama and Rosa are suggested to be different from Canadian. However, Kama and Rosa are not suggested to be different from each other. Generally, these results support the data summarization provided by the box plots presented above.

Kruskal-Wallis

- Nonparametric alternative to One-Way ANOVA
- H_0 (All group medians are equal)
- H_1 (All group medians are not equal)
- Generally lower statistical power than One-Way ANOVA
- Pair-wise comparison to compare groups if fail to reject null
- Assumptions
  - Representative, random samples
  - Independence between samples
  - Independence between groups
  - Dependent variable is ordinal, interval, or ratio
  - Independent variable is categorical
  - Similar variance between groups

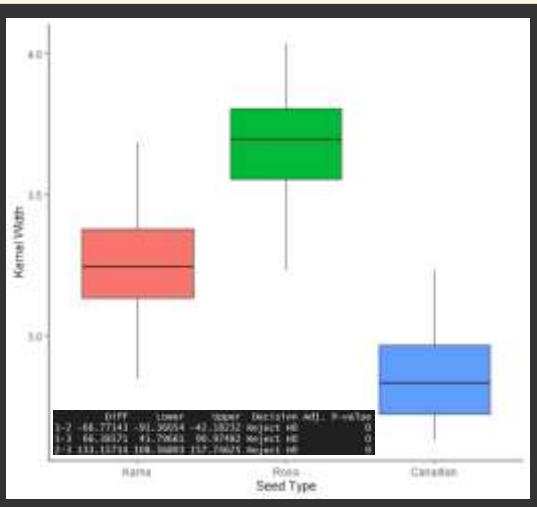
```
Kruskal-Wallis chi-squared = 367.65, df = 7, p-value < 2.2e-16
pairw.kw(hp_data2$temp, hp_data2$STATE_NAME, conf=0.95)
```

	Diff	Lower	Upper	Decision
Colorado-Kansas	-263.02917	-323.02863	-183.02971	Reject H0
Colorado-Montana	67.00007	-12.98254	146.83468	FTR H0
Kansas-Montana	520.09524	247.85552	393.93515	Reject H0
Colorado-Nebraska	-131.1539	-202.8432	-59.4646	Reject H0
Kansas-Nebraska	121.87527	59.02134	184.7292	Reject H0
Montana-Nebraska	-199.01997	-273.68004	-124.359	Reject H0
Colorado-North Dakota	102.26709	20.289	184.24638	Reject H0
Kansas-North Dakota	355.29686	280.92101	429.6727	Reject H0
Montana-North Dakota	36.48162	-50.18844	118.99128	FTR H0
Nebraska-North Dakota	233.42159	157.45318	309.38999	Reject H0
Colorado-South Dakota	-19.88958	-97.31758	57.55582	FTR H0
Kansas-South Dakota	233.14048	163.81111	302.48886	Reject H0
Montana-South Dakota	-87.70675	-167.94224	-7.55127	Reject H0
Nebraska-South Dakota	111.27922	40.23025	182.31618	Reject H0
North Dakota-South Dakota	-122.14857	-203.58248	-40.73428	Reject H0
Colorado-Utah	-52.97629	-151.78346	45.83088	FTR H0
Kansas-Utah	200.05287	107.45634	292.64941	Reject H0
Montana-Utah	-120.84296	-221.82633	-19.8584	Reject H0
Nebraska-Utah	78.1776	-15.7029	172.05811	FTR H0
North Dakota-Utah	-155.24998	-257.19838	-53.28958	Reject H0
South Dakota-Utah	-33.09561	-131.43484	65.24362	FTR H0
Colorado-Wyoming	57.00272	-50.30765	164.31309	FTR H0
Kansas-Wyoming	310.03158	208.45113	411.85203	Reject H0
Montana-Wyoming	-10.86335	-120.18133	98.45463	FTR H0
Nebraska-Wyoming	188.15662	85.36455	290.94868	Reject H0
North Dakota-Wyoming	-45.26497	-155.48003	64.95008	FTR H0
South Dakota-Wyoming	76.8854	-29.89627	183.76307	FTR H0
Utah-Wyoming	109.97901	-13.26753	233.22556	FTR H0

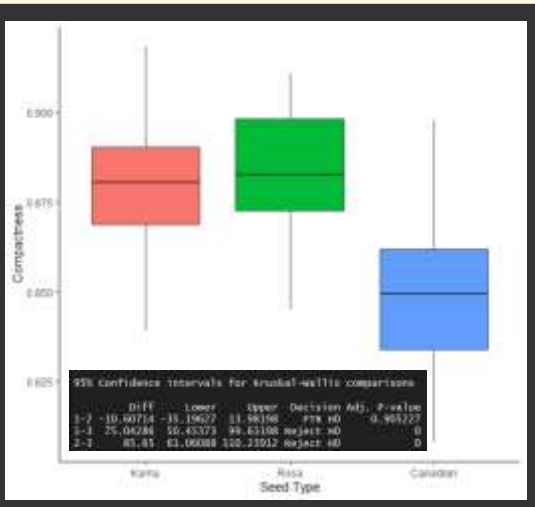
One-Way ANOVA is a parametric test. Its nonparametric equivalent is the Kruskal Wallis Tests, which is based on ranks and group medians. The null hypothesis is that all population group medians are equal while the alternative hypothesis is all population group medians are not equal (or, at least two population group medians are not equal).

Again, nonparametric tests still have assumptions, they just don't assume that the measure being compared between groups is normally distributed. The slide lists the assumptions of the Kruskal-Wallis Test.

Example Data: Seeds



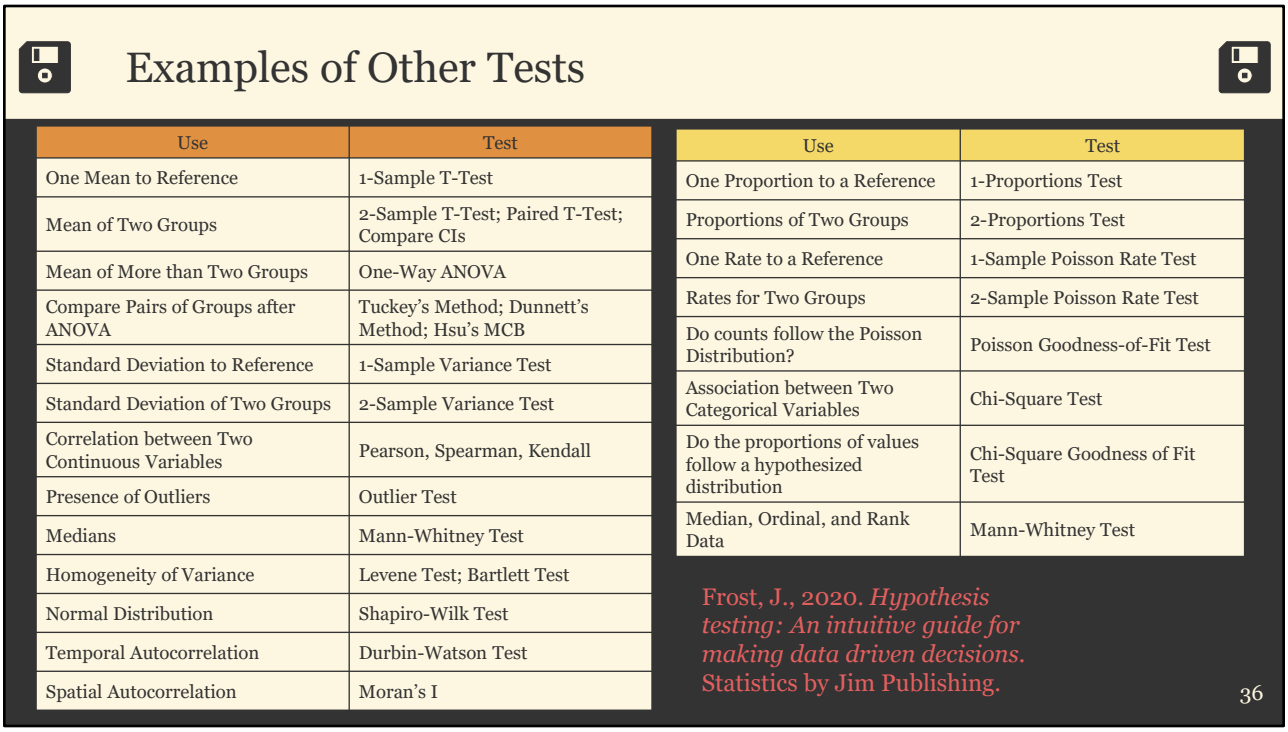

```
        Kruskal-Wallis rank sum test

data:  wkernel by typo
Kruskal-Wallis chi-squared = 168.07, df = 2, p-value < 2.2e-16
```




```
        Kruskal-Wallis rank sum test

data:  compactness by typo
Kruskal-Wallis chi-squared = 82.655, df = 2, p-value < 2.2e-16
```

35

This slide presents the Kruskal-Wallis results for the Seeds data. These results reinforce those obtained using One-Way ANOVA with statistical differences observed. Like One-Way ANOVA, this does not tell us which groups are suggested to be different. In order to determine this, nonparametric pair-wise comparisons can be used. We will not discuss those methods here, but they are interpreted similar to Tuckey's Method described above. Since Kruskal-Wallace is a nonparametric test, a nonparametric pair-wise test should be used as opposed to Tuckey's Method, which is parametric, if difference in means is suggested.



There are many other statistical tests designed for different use cases and to explore different hypotheses. The tables on this page were modified from Frost's text and provide some example tests and their associated uses.

We don't have space here to discuss additional tests. However, if you need to use other tests, I think you will find that the process of interpreting them is similar to the methods discussed here. First, you need to determine what test is most appropriate for assessing your hypothesis. You then need to prepare your data as input for the test, perform the test to obtain an index value, use the index value to obtain a p-value, and compare the p-value to a pre-defined significance level or alpha. Parametric tests will use a theoretical distribution based on degrees of freedom to obtain a p-value from the test statistic.

Statistical Inference via Bootstrapping

- Repeated random sampling with replacement
- Allows for creating multiple sub-samples from a single sample from the population
- Minimizes the need to use theoretical data distributions
- Can be computationally intensive

https://commons.wikimedia.org/wiki/File:JM_marbles_01.jpg

37

Although we will not focus on this topic here, I did want to mention that bootsrapping can be used as an alternative means to undertake statistical inference. This involves generating multiple samples by random sampling the original sample from the population with replacement. This allows for characterizing the distribution of the data and minimizes the need for using theoretical data distributions to compare the sample to. Traditional statistical methods were developed when such methods were too computationally intensive. However, such repeated subsampling is generally not a problem for modern computers.

In short, this is a powerful method to undertake statistical inference and offers and alternative to traditional methods.

Summary of Key Points

- Various statistical tests are available to assess different types of hypotheses.
- Tests can be conceptualized as assessing whether an effect or signal can be detected or differentiated from inherent noise.
- Parametric tests generally have nonparametric alternatives that can be used when distribution assumptions are violated.
- Assumptions of tests can be examined to assess whether results may be misleading.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.