1. Why do we use statistics in research?

You may already be familiar with "descriptive statistics", values such as mean, median, mode, interquartile range, etc. These values can tell us the central tendency of our data (i.e., where the bulk of our data points are) or the variability of our data (i.e., how close together our data points are). Therefore, descriptive statistics allow us to understand our data, but we cannot use it to determine how variables are interacting with each other or make conclusions about how our experimental sample may represent the larger population. To do this, we need to use "inferential statistics".

2. Framework for statistical analysis

2.1 *Hypotheses:*

When you perform a statistical test, you always have two hypotheses in mind, the null (H_0) and alternative (H_a) hypothesis.

- The null hypothesis assumes the "status quo" between your two populations - that is there is equality between them. In other words, it assumes there is no difference or change between what you are comparing.
    - Example: There is no difference in ice cream consumption between people older than 18 and those under 18.
- The alternative hypothesis assumes that there is inequality between your two populations. This is the question you actually want to test.
    - Example: There is a difference in ice cream consumption between people older than 18 and those under 18.

Note: for JEI manuscripts, you should not present your null and alternative hypotheses. Only the alternative hypothesis, in the larger context of your research, should be presented.

2.2 *P-values:*

2.2.1 What is a *p*-value?

It is a value between 0 and 1 that provides a measurement of probability, assuming that the null hypothesis is true.

2.2.2 What does it mean?

There are three main ways we can think about a *p*-value and its meaning. They are:

1. The probability we observe a test statistic as extreme or more extreme than the one observed
2. Probability that we observe data in our population that is at least as extreme as what we observed
3. If the experiment was repeated, the probability you would observe results as extreme by chance

2.3 *How do we decide what is significant?*

A *p*-value in itself does not tell us if data is significant or not. We need to set a significance level to help us decide whether our hypothesis is supported. Your significance level should always be set prior to collecting data and running any statistical analysis.

The significance level is determined by "α" (alpha), which represents type I error. When you reject a null hypothesis that is true (i.e., you get a false positive result), this is a type I error. You can set your α-level to any value, but typically 0.05 is used, which means that we are okay with getting false positive results 5% of the time. Another way to think of this is saying that you are 95% confident that the results you saw are true (i.e., you have a 95% confidence interval).

2.4 *Assumptions to perform inferential statistics*

When you perform any of the following statistical tests, your data should meet the assumptions listed below. Failure to meet these assumptions means that the results of any test you perform may not be accurate.

1. Your data is normally distributed (and groups are of approximately equal size)
2. Your sample is representative of the larger population
3. Observations are independent of each other

If these assumptions are not met, then it means you cannot trust your *p*-value or confidence interval with the t-tests, or one-way ANOVA described below. In these situations, it is better to use descriptive statistics to make comparisons with your data or use a non-parametric test (described here).

2.4.1 Checking for normal distribution of data

We recommend checking whether your data follows a normal distribution in two ways. The first is to calculate the mean, median and mode for your data. If it is normally distributed, then these values should be equal. The second way is to make a histogram of your data and look at the shape. Normally distributed data should follow a "bell curve" shape (see panel (b) below). The bin size (i.e., how wide your groups are on the x-axis) for your histogram can have an impact on what shape it takes, so it is worthwhile to try a few different bin sizes and see what the overall appearance of the data is.

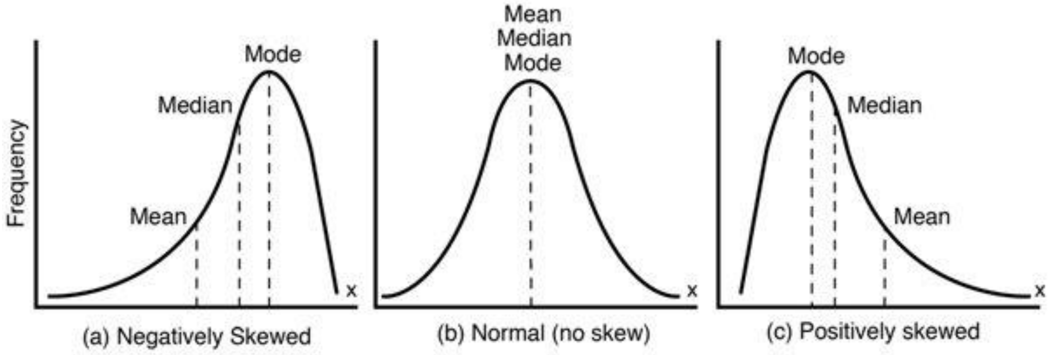


Figure from Faria Marco TC, et al. "Acoustic Emission Tests on the Analysis of Cracked Shafts of Different Crack Depths." 23rd ABCM International Congress of Mechanical Engineering, Dec. 2015, doi:10.20906/CPS/COB-2015-1434

2.5 *Types of Data*

There are two types of data that can be collected. Your data will either be "categorical" or "continuous". The type of data you have, in addition to the type of comparison you want to make, will dictate what statistical test is most appropriate to use.

2.5.1 Categorical Data

This type of data has a finite number of categories or distinct groups, and you record the frequency within each category/group.

For example, asking students what their favorite subject at school is.

2.5.2 Continuous Data

This is numerical data with an infinite number of values between any two values.

For example, a score on an exam can be any possible value between 0 and 100 points.

3. Common statistical tests for JEI work:

Below we give the common types of statistical tests seen in JEI manuscripts. We provide an overview of when each test is appropriate and how to perform the test using QuickCalcs from GraphPad. At the very end, we provide a practice data set for each type of test. We recommend you verify that you can achieve the same outputs we did using the provided data to ensure you are properly executing a specific test before analyzing your own data.

3.1 *t-Tests*

t-tests (or Student's t-test) is probably the most widely used statistical test on continuous data. It is used to compare the mean of a population to a known mean (one-sample t-test) or to the mean of another experimental group (two-sample t-test). It is much more likely that you will need a two-sample t-test, but we will still discuss when it is appropriate to use a one-sample t-test.

Note: two-sample t-tests should not be used multiple times to compare multiple experimental groups to a single control group. Instead, a one-way ANOVA should be used. An explanation on why multiple t-tests is bad is can be found here.

For all t-tests you can perform a one-tailed or two-tailed test. We recommend at JEI that you always perform a two-tailed test. If you have reason to believe that the effect you will see should only happen in one direction, then it may be appropriate to perform a one-tailed test. The decision to do a one vs. two-tailed test should be made before any data is collected.

Example Hypotheses:
One-tailed: People will eat more ice cream than the previous year
Two-tailed: People will eat a different amount of ice cream than the previous year

3.1.1 One Sample t-test

A one-sample t-test is used when you want to compare the mean of your experimental data to the known mean of a population. In this case, the only data you will have collected are your experimental values.

Online Calculator

You will go to "continuous data" and select "one-sample t-test". There are many different options on how you can input your data. The best way to limit any error on your end is to directly input your raw data values instead of using mean, SEM (or SD), and N values you previously calculated. After entering your data, set the hypothetical mean value (which represents the known population mean). Once you have done this, hit "calculate now".

3.1.2 Two-Sample t-test

A two-sample t-test is used when you want to compare the mean of two groups to each other and you have collected the data for both groups.

There are two types of a two-sample t-test, paired or unpaired. In most cases, you will perform an unpaired t-test. A paired t-test should be used if you measured the same group/subjects at different times.

Online Calculator

You will go to "continuous data" and select "t-test to compare two means". There are many different options on how you can input your data. The best way to limit any error on your end is to directly input your raw data values instead of using mean, SEM (or SD), and N values you previously calculated. After entering your data, you must choose the test you want to perform. In most cases this will be an unpaired t-test. If your data meets the criteria for a paired t-test as described above, then you should choose that option. Once you have done this, hit "calculate now".

Outputs:
The output of a t-test is a *p*-value. The *p*-value is interpreted in relation to the significance level (α) you set earlier. This is why it is important you set your significance level prior to doing any statistical analysis otherwise you could decide to change it based on the *p*-values of your test which isn't ethical!

If your *p*-value is less than or equal to your significance level, then you would state that you saw a significant difference between your two groups (two-sample) or between your group and the hypothetical mean (one-sample).

If you *p*-value is greater than your significance level, then you would state that no significant difference was seen.

3.2 *ANOVA*

ANOVA stands for "Analysis of Variance". There are multiple types of ANOVA tests, but we will focus on the one-way ANOVA test as it is the most commonly used. A one-way ANOVA is used (on continuous data) when you have three or more groups (independent variable) for a single type of measurement (dependent variable) and want to know if there is an interaction between the variables. Think of it like adding one or more groups to data you would use a two-sample t-test on.

Example Hypothesis:
You want to know if ice cream flavor (vanilla, chocolate, cookies and cream, and Neapolitan) has any effect on the amount consumed.

Outputs:
When you do an ANOVA there are two main values that are calculated.

The first is the F-value which is similar to a t-value you can get with t-tests. The F-value is used in combination with the degrees of freedom to calculate the *p*-value. Most online tools will automatically report the *p*-value, but for some you may have to look up a F-distribution table to get the *p*-value.

As before, the *p*-value tells you whether your data is significantly different from the assumption of the null hypothesis (in this case the null hypothesis is always that the mean of each group is equal to each other). So, the *p*-value calculated from the F-value tells you that there is a difference among your groups, but this is an overall conclusion.

If you get a significant result from your ANOVA and want to know if there is a difference between two specific groups then you will need to perform additional tests, which are called post-hoc tests. The online tool we recommend below will automatically perform multiple comparisons with a Tukey's HSD post-hoc test.

Online Calculator
Note: This test uses an online resource from Vassar Stats instead of QuickCalcs. The number of samples represents the number of groups you have (you must have at least 2 but cannot have more

than 5 with this tool). Click "Independent Samples" - it is very unlikely that you will have correlated samples for your analysis. Hit "Calculate" at the bottom to get your results.

3.3 Chi-Square (χ^2)
Note: There are two different types of Chi-Square tests, we only describe the Chi-Square test for goodness of fit below.

A Chi-Square test is used when you want to look at the difference between categorical observed and expected data to determine if the difference is due to chance or due to a relationship between the variables. For example, you could use a Chi-Square test to determine if all sides of a die are as likely to be rolled as the other.

Example Hypothesis:
You want to know if people order ice cream in cups, waffle cones, and wafer/cake cones in equal amounts.

Online Calculator
You will go to "categorical data" and select "Chi-square". Compare observed and expected frequencies". Choose how you will enter your expected values. When entering observed values into the data table, make sure you provide the actual number observed and not a percent or fraction. Once you have done this, hit "calculate now".

4. Correcting for Multiple Comparisons

4.1 *What are multiple comparisons*
Similar to increasing the sample size for your experiment, the more tests you run, the greater chance that you will find a result with a *p*-value below your given threshold. The probability you get at least one "significant" result is defined by $100(1.00\text{-}0.95^N)$, where N is the number of comparisons being made and assuming $\alpha = 0.05$. Using this equation, if you did 10 separate t-tests the probability you get at least one positive result (that may or may not actually be significant) is approximately 40%. This is a little high for us to be comfortable with conclusions drawn from the data!

4.2 *Considerations when doing multiple comparisons*
The most important thing to keep in mind when doing multiple comparisons is that you should have pre-planned the multiple comparisons you want to do before you collect any data. This helps prevent you from "p-hacking" or looking for significant results where you can find them. It is important to note that negative results (a significant difference is not detected) in science are just as important as positive results, so do not be discouraged if you do an experiment and nothing turns out significant!

4.3 *Tests for correcting multiple comparisons*
There are several ways that we can account for multiple tests and the associated α level inflation. We outline the two most common ways to correct for multiple comparisons and ensure you aren't seeing false positive results.

4.3.1 Bonferroni
A Bonferroni correction is usually the easiest (and most stringent) way to correct for multiple comparisons. To perform a Bonferroni correction, you will take your α-level and divide it by the total number of tests you performed to get an "adjusted α-level". So, if you perform 10 tests with $\alpha = 0.05$, your new significance level would become 0.005 (0.05/10). This means, in order for a result to be considered statistically significant it needs to have a *p*-value of 0.005 or less.

4.3.2 Tukey HSD post-hoc test

This correction can only be used to correct for multiple comparisons when performing an ANOVA. We won't delve into the details of how this test works, but if you are interested in learning more about it check out this page. The online calculator we linked to, along with many others, will automatically perform this test when you input data for a one-way ANOVA.

4.4 *How to acknowledge a post-hoc test in your results:*

If you use a post hoc test, it is important to remember that you used it in combination with another statistical test. This means when you are talking about your statistics you should phrase it as "a two-sample t-test with a Bonferroni correction for multiple comparisons was performed. An adjusted α-level of 0.005 was taken as significant."

Example Problems

The following tables will be needed to complete the practice problems given below.

Year	Per Capita Consumption (lbs)	Year	Per Capita Consumption (lbs)
2000	16.1	2010	14.0
2001	15.8	2011	13.2
2002	16.2	2012	13.2
2003	15.9	2013	13.1
2004	14.6	2014	12.5
2005	15.1	2015	12.9
2006	15.3	2016	12.9
2007	14.8	2017	12.3
2008	14.2	2018	12.0
2009	13.9	2019	12.1

Table 1: Average per capita consumption (lbs) of ice cream in the United States from 2000-2019. Data from Economic Research Service: USDA Foreign Agricultural Service

Person	Consumption (lbs)
1	10.0
2	12.6
3	14.2
4	13.4
5	14.7
6	12.0
7	7.8
8	8.0
9	9.6
10	10.6

Table 2: 2020 Ice Cream Consumption (lbs)

Flavor	Number of People
Chocolate	35
Cookies and Cream	14
Neapolitan	11
Vanilla	40
TOTAL	**100**

Table 3: Count of ice cream flavors ordered over two hours.

1. **You are interested in knowing if there was any difference between the average amount of ice cream consumed from 2000-09 vs. 2010-19 (Table 1).**

2. **Now, you want to look at 5-year groups (2000-04, 2005-09, 2010-14, 2015-19 Table 1). Is there any difference across all groups and/or between specific groups?**

3. **You surveyed 10 individuals about their ice cream consumption during 2020 (Table 2). You want to know if there is a difference from the average between 2000-2019 (Table 1).**

4. **Your friend owns an ice cream store that sells four flavors. They provide you with the total number of people that ordered each flavor over the course of two hours. You are interested in knowing whether each flavor is ordered equally.**

ANSWERS

1. **You are interested in knowing if there was any difference between the average amount of ice cream consumed from 2000-09 vs. 2010-19 (Table 1).**

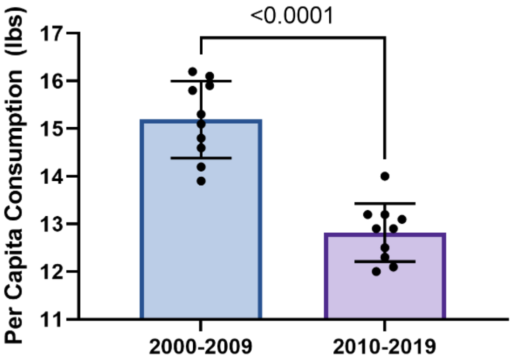
For this scenario you would use a two-sample t-test since you are interested in comparing two means.

Mean ± SD
2000-09: 15.19 ± 0.81
2010-19: 12.81 ± 0.61

p-value < 0.0001

Since the *p*-value is less than 0.05, we would consider this a significant result and reject the null hypothesis that the average consumption of ice cream between the two decades is equal. This supports that there is a difference in ice cream consumption.

In the context of a JEI manuscript, you could say something like: Ice cream consumption significantly decreased in 2010-19 compared to 2000-09 ($p < 0.0001$).



Notice, since we have multiple data points for each decade, we have included error bars that represent the standard deviation of our data.

2. **Now, you want to look at 5-year groups (2000-04, 2005-09, 2010-14, 2015-19 Table 1). Is there any difference across all groups and/or between specific groups?**

Since we have more than two groups, we need to use a one-way ANOVA here.

F = 34.69
df = 3
$p < 0.001$

Since the *p*-value is less than 0.05 we would consider this a significant result. With an ANOVA, this means that we are rejecting the null hypothesis that the mean of each group is equal to each other. This tells us that our "factor" (in this case the year) has an effect on ice cream consumption.

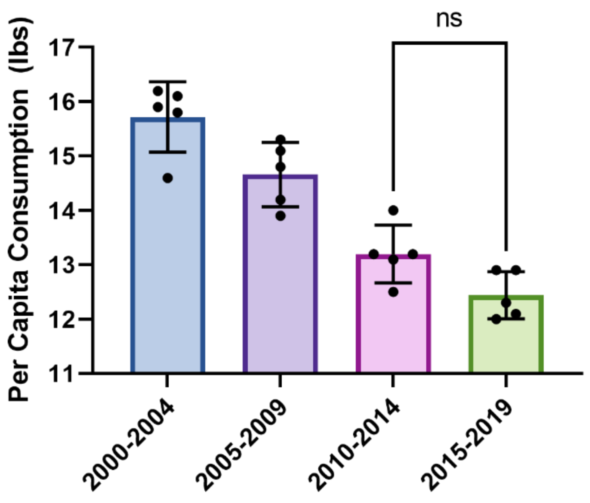
In the context of a JEI manuscript, you could say something like: The five-year period looked at does have an effect on the amount of ice cream consumed ($p < 0.001$).

If we want to know if there is a significant difference between any of the groupings, then we need to do a Tukey HSD post-hoc test (if you use the online calculator above, this is automatically done for you). When we do this, we get the following results:

2000-04 vs 2005-09: p =0.0377
2000-04 vs 2010-14: $p < 0.0001$
2000-04 vs 2015-19: $p < 0.0001$

2005-09 vs 2010-14: $p = 0.0038$
2005-09 vs 2015-19: $p < 0.0001$
2010-14 vs 2015-19: $p = 0.1781$

From these results we can see every comparison, except for 2010-14 vs. 2015-19 is significant.



Since we performed multiple tests and most of the results were significant it is simplest to note the results that were not significant on our graph.

3. You surveyed 10 individuals about their ice cream consumption during 2020 (Table 2). You want to know if there is a difference from the average between 2000-2019 (Table 1).

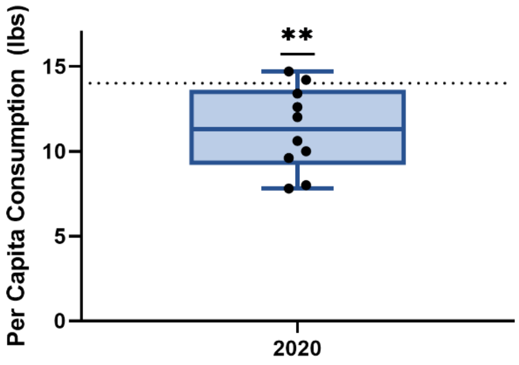
You would use a one-sample t-test here. The first thing you need to do is find your hypothetical mean which in this case is the average from 2000-2019.

Hypothetical mean = 14.01 ± 1.4
2020 mean = 11.29 ± 2.47

p-value = 0.0069

Since the *p*-value is less than 0.05, we would reject the null hypothesis and state that there is a significant difference in the amount of ice cream consumed in 2020 compared to the previous 20 years.

In the context of a JEI manuscript, you could say something like: The amount of ice cream consumed in 2020 is significantly less than the average consumed from 2000-19 (*p* =0.0069).



Since the only data we collected was 2020 ice cream consumption, a box and whisker plot is a nice way to show our data. This allows people to see the "spread" of our data along with the individual data points. We have also included a line at the hypothetical mean.

4. **Your friend owns an ice cream store that sells four flavors. They provide you with the total number of people that ordered each flavor over the course of two hours. You are interested in knowing whether each flavor is ordered equally.**

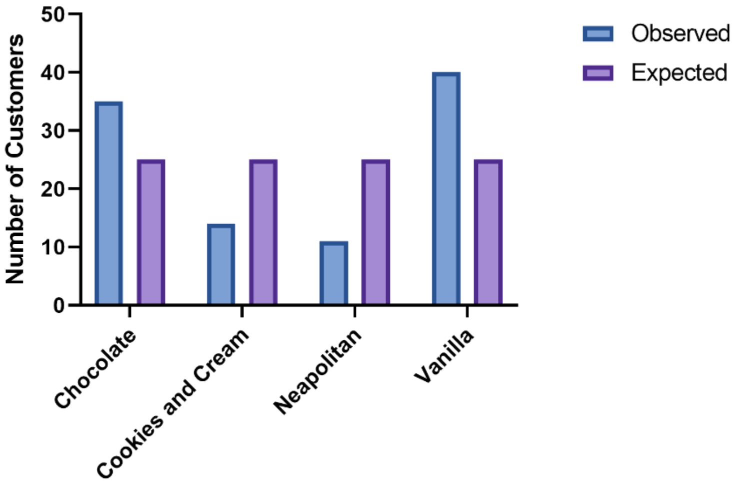
For this scenario you would use a Chi Square (χ^2) test. Since the assumption is that each flavor would be ordered in equal proportions, you would get an expected value of 25 for each flavor.

$\chi^2 = 25.68$
df = 3
$p < 0.0001$

Once again, the *p*-value is less than 0.05, so we would consider this a significant result that supports our observed values being different from the expected values.

In the context of a JEI manuscript, you could say something like: Ice cream flavors were not ordered in equal amounts over the two-hour observation period ($p < 0.0001$).

Generally, you would present this data in a table like Table 3 above and state in your results and methods what your expected value is. If you wanted to include a graph, then you could do something similar to what is below.


Number of Customers
Chocolate
Cookies and Cream
Neapolitan
Vanilla
Observed
Expected