P value and F ratio[1]

One-way ANOVA compares three or more unmatched groups, based on the assumption that the populations are Gaussian or have a normal distribution.

P value

The P value tests the null hypothesis that data from all groups are drawn from populations with identical means. Therefore, the P value answers this question:

> If all the populations really have the same mean, what is the chance that random sampling would result in means as far apart (or more so) as observed in these samples?

If the overall P value is large, the data do not give you any reason to conclude that the means differ. Even if the population means were equal, it is possible to find sample means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have compelling evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean, only that ***at least*** one differs from the rest. Look at the results of post tests (as Tukey) to identify where the differences are.

F ratio and ANOVA table

The P value is computed from the F ratio which is computed from the ANOVA table.

ANOVA partitions the variability among all the values into:

1- one component that is due to variability among group means; and

2- another component that is due to variability within the groups (also called residual variation).

Variability within groups (within the columns) is quantified as the sum of squares of the differences between each value and its group mean. This is the residual sum-of-squares.

Variation among groups is quantified as the sum of the squares of the differences between the group means and the grand mean (the mean of all values in all groups). *Adjusted for the size* of each group, this becomes the sum-of-squares.

Each sum-of-squares is associated with a certain number of degrees of freedom (df, computed from number of subjects and number of groups), and the mean square (MS) is computed by dividing the sum-of-squares by the appropriate number of degrees of freedom. These can be thought of as variances. The square root of the mean square residual can be thought of as the pooled standard deviation.

The F ratio is the ratio of two mean square values. *If the null hypothesis is true, you expect F to have a value close to 1.0 most of the time.* A large F ratio means that the variation among group means is more than you'd expect to see by chance. *You'll see a large F ratio both when the null hypothesis is wrong (the data are not sampled from populations with the same mean) and when random sampling happened to end up with large values in some groups and small values in others.*

The P value is determined from the F ratio and the two values for degrees of freedom shown in the ANOVA table.

---

1 Extracted and modify from GraphPad Statistics Guide

What is a P value?

Suppose that you've collected data from two samples and the means are different. You want to know if the difference is statistical significant.

Observing different sample means is not enough to persuade you to conclude that the populations have different means. It is possible that the populations have the same mean and that the difference you observed between sample means occurred only by chance. There is no way you can ever be sure if the difference you observed reflects a true difference or if it simply occurred in the course of random sampling. **All you can do is calculate probabilities.**

The first step is to state the null hypothesis, that is all differences are due to random sampling.

The P value is a probability, with a value ranging from zero to one, that answers this question:

> In a sample of this size, if the populations really have the same mean, what is the probability of observing at least as large a difference between sample means as was, in fact, observed?

The most common misinterpretation of a P value

Many people misunderstand what a P value means. Let's assume that you compared two means and obtained a P value equal to 0.03.

Correct definitions of this P value:

There is a 3% chance of observing a difference as large as you observed even if the two population means are identical (the null hypothesis is true).

or

Random sampling from identical populations would lead to a difference smaller than you observed in 97% of experiments, and larger than you observed in 3% of experiments.

Wrong:

~~There is a 97% chance that the difference you observed reflects a real difference between populations, and a 3% chance that the difference is due to chance.~~

This latter statement is a common mistake. If you have a hard time understanding the difference between the correct and incorrect definitions, read this Bayesian perspective.

Kline[2] lists commonly believed fallacies about P values, which are:

Fallacy 1: **P value is the probability that the result was due to sampling error**

Why: *The P value is computed assuming the null hypothesis is true*. In other words, the P value is computed based on the assumption that the difference was due to sampling error. Therefore the P value cannot tell you the probability that the result is due to sampling error.

Fallacy 2: The P value Is the probability that the null hypothesis is true

No!. The P value is computed assuming that the null hypothesis is true, so cannot be the probability that it is true.

Fallacy 3: *1-P*, is the probability that the alternative hypothesis is true

If the P value is 0.03, it is very tempting to think: If there is only a 3% probability that my difference would have been caused by random chance, then there must be a 97% probability that it

2RB Kline, Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research, 2004, IBSN:1591471184

was caused by a real difference. But this is wrong!
What you can say is that *if the null hypothesis were true, then 97% of experiments would lead to a difference smaller than the one you observed, and 3% of experiments would lead to a difference as large or larger than the one you observed.*
Calculation of a P value is predicated on the assumption that the null hypothesis is correct. P values cannot tell you whether this assumption is correct. P value tells you how rarely you would observe a difference as larger or larger than the one you observed if the null hypothesis were true.
The question that the data scientist must answer is whether the result is so unlikely that the null hypothesis should be discarded.

Fallacy 4: *1-P,* is the probability that the results will hold up when the experiment is repeated
If the P value is 0.03, it is tempting to think that this means there is a 97% chance of getting 'similar' results on a repeated experiment. Not so.

Fallacy 5: A high P value proves that the null hypothesis is true.
No. A high P value means that *if the null hypothesis were true, it would not be surprising to observe the difference seen in this samples.* But that does not prove the null hypothesis is true.

Fallacy 6: The P value is the probability of rejecting the null hypothesis
You reject the null hypothesis (and deem the results statistically significant) when a P value from a particular experiment is less than the significance level α, which you (should have) set as before you collect the sample. So *if the null hypothesis is true, α is the probability of rejecting the null hypothesis.*
The P value and α are not the same. A P value is computed from each comparison, and is a measure of the strength of evidence. The significance level α is set once as part of the experimental design.

One-tail vs. two-tail P values

When comparing two groups, you must distinguish between one- and two-tail P values. Some books refer to one-sided and two-sided P values, which mean the same thing.

What does one-sided mean?
It is easiest to understand the distinction in context. So let's imagine that you are comparing the mean of two groups (with an unpaired t test). Both one- and two-tail P values are based on the same null hypothesis, that two populations really are the same and that an observed discrepancy between sample means is due to chance.

A two-tailed P value answers this question:

> Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart as (or further than) you observed in this experiment with either group having the larger mean?

To interpret a one-tail P value, you must predict which group will have the larger mean before collecting any data.

The one-tail P value answers this question:

> Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart as (or further than) observed in this experiment with the specified group having the larger mean?

If the observed difference went in the direction predicted by the experimental hypothesis, the one-tailed P value is half the two-tailed P value (with most, but not quite all, statistical tests).

When is it appropriate to use a one-sided P value?

A one-tailed test is appropriate when previous data, physical limitations, or common sense tells you that the difference, if any, can only go in one direction. You should only choose a one-tail P value when both of the following are true.

- You predicted which group will have the larger mean (or proportion) before you collected any data.

- If the other group had ended up with the larger mean – even if it is quite a bit larger – you would have attributed that difference to chance and called the difference 'not statistically significant'.

The issue in choosing between one- and two-tailed P values is not whether or not you expect a difference to exist. If you already knew whether or not there was a difference, there is no reason to collect the data. Rather, the issue is whether the direction of a difference (if there is one) can only go one way. You should only use a one-tailed P value when you can state with certainty (and before collecting any data) that in the overall populations there either is no difference or there is a difference in a specified direction. If your data end up showing a difference in the “wrong” direction, you should be willing to attribute that difference to random sampling without even considering the notion that the measured difference might reflect a true difference in the overall populations. If a difference in the “wrong” direction would intrigue you (even a little), you should calculate a two-tailed P value.

Recommendation

I recommend using only two-tailed P values for the following reasons:

- The relationship between P values and confidence intervals is more straightforward with two-tailed P values.

- Two-tailed P values are larger (more conservative). Since many experiments do not completely comply with all the assumptions on which the statistical calculations are based, many P values are smaller than they ought to be. Using the larger two-tailed P value partially corrects for this.

- Some tests compare three or more groups, which makes the concept of tails inappropriate (more precisely, the P value has more than two tails). A two-tailed P value is more consistent with P values reported by these tests.

- Choosing one-tailed P values can put you in awkward situations. If you decided to calculate a one-tailed P value, what would you do if you observed a large difference in the opposite direction to the experimental hypothesis? To be honest, you should state that the P value is large and you found “no significant difference.” But most people would find this hard. Instead, they’d be tempted to switch to a two-tailed P value, or stick with a one-tailed P value, but change the direction of the hypothesis. You avoid this temptation by choosing two-tailed P values in the first place.

When interpreting published P values, note whether they are calculated for one or two tails. If the author didn’t say, the result is somewhat ambiguous.

How to convert between one- and two-tail P values

The one-tail P value is half the two-tail P value.

The two-tail P value is twice the one-tail P value (assuming you correctly predicted the direction of the difference).

Advice: Use two-tail P values

If in doubt, choose a two-tail P value. Why?

- The relationship between P values and confidence intervals is easier to understand with two-tail P values.
- Some tests compare three or more groups, which makes the concept of tails inappropriate (more precisely, the P values have many tails). A two-tail P value is more consistent with the P values reported by these tests.
- Choosing a one-tail P value can pose a dilemma. What would you do if you chose to use a one-tail P value, observed a large difference between means, but the “wrong” group had the larger mean? In other words, the observed difference was in the opposite direction to your experimental hypothesis. To be rigorous, you must conclude that the difference is due to chance, even if the difference is huge. While tempting, it is not fair to switch to a two-tail P value or to reverse the direction of the experimental hypothesis. You avoid this situation by always using two-tail P values.

Advice: How to interpret a small P value

Before you interpret the P value

Before thinking about P values, you should:

- Review the science. If the study was not designed well, then the results probably won't be informative. It doesn't matter what the P value is.
- Review the assumptions of the analysis you chose to make sure you haven't violated any assumptions. We provide an analysis checklist for every analysis that Prism does. If you've violated the assumptions, the P value may not be meaningful.

Interpreting a small P value

A small P value means that the difference (correlation, association,...) you observed would happen rarely due to random sampling. There are three possibilities:

- **The null hypothesis of no difference is true, and a rare coincidence has occurred**. You may have just happened to get large values in one group and small values in the other, and the difference is entirely due to chance. How likely is this? The answer to that question, surprisingly, is not the P value. Rather, the answer depends on the scientific background of the experiment.
- **The null hypothesis is false. There truly is a difference (or correlation, or association...) that is large enough to be scientifically interesting.**
- **The null hypothesis is false. There truly is a difference (or correlation, or association...), but that difference is so small that it is scientifically boring.** The difference is real, but trivial.

Deciding between the last two possibilities is a matter of scientific judgment, and no statistical calculations will help you decide.

Using the confidence interval to interpret a small P value

If the P value is less than 0.05, then the 95% confidence interval will not contain zero (when comparing two means). To interpret the confidence interval in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that you consider to be scientifically important or scientifically trivial. This section assumes you are comparing two means with a t test, but it is straightforward to use these same ideas in other contexts.

There are three cases to consider:

- **The confidence interval only contains differences that are trivial.** Although you can be 95% sure that the true difference is not zero, you can also be 95% sure that the true difference between means is tiny and uninteresting. The treatment had an effect, but a small one.
- **The confidence interval only includes differences you would consider to be important.** Since even the low end of the confidence interval represents a difference large enough that you consider it to be scientifically important, you can conclude that there is a difference between treatment means and that the difference is large enough to be scientifically relevant.
- **The confidence interval ranges from a trivial to an important difference.** Since the confidence interval ranges from a difference that you think would be scientifically trivial to one you think would be important, you can't reach a strong conclusion. You can be 95% sure that the true difference is not zero, but you cannot conclude whether the size of that difference is scientifically trivial or important.

Advice: How to interpret a large P value

Before you interpret the P value

Before thinking about P values, you should:

- Assess the science. If the study was not designed well, then the results probably won't be informative. It doesn't matter what the P value is.
- Review the assumptions of the analysis you chose to make sure you haven't violated any assumptions. If you've violated the assumptions, the P value may not be meaningful.

Interpreting a large P value

If the P value is large, the data do not give you any reason to conclude that the overall means differ. Even if the true means were equal, you would not be surprised to find means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have convincing evidence that they differ.

Using the confidence interval to interpret a large P value

How large could the true difference really be? Because of random variation, the difference between the group means in this experiment is unlikely to be equal to the true difference between population means. There is no way to know what that true difference is. The uncertainty is expressed as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between the two means. When the P value is larger than 0.05, the 95% confidence interval will start with a negative number (representing a decrease) and go up to a positive number (representing an increase).

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference that would be scientifically important or scientifically trivial. There are two cases to consider:

- The confidence interval ranges from a decrease that you would consider to be trivial to an increase that you also consider to be trivial. Your conclusions is pretty solid. Either the treatment has no effect, or its effect is so small that it is considered unimportant. This is an informative negative experiment.
- One or both ends of the confidence interval include changes you would consider to be scientifically important. You cannot make a strong conclusion. With 95% confidence you can say that either the difference is zero, not zero but is scientifically trivial, or large enough to be scientifically important. In other words, your data really don't lead to any solid conclusions.

Calculations with Excel

If you want to compute P values using Excel, use these functions:

P value from F:	FDIST (F, DFn, DFd)
P value from t (two tailed):	TDIST (t, df, 2) (The third argument, 2, specifies a two-tail P value.)
P value from ChiSquare:	CHIDIST (ChiSquare, DF)
P value from z (two tailed):	TDIST(z, 10000,2) (With a huge number of degrees of freedom, t and z are identical.)
or	2*(1.0-NORMSDIST(z))

Hypothesis testing and statistical significance

"Statistically significant". That phrase is commonly misunderstood. Before analyzing data and presenting statistical results, make sure you understand what statistical 'significance' means and doesn't mean.

Statistical hypothesis testing
Much of statistical reasoning was developed in the context of quality control where you need a definite yes or no answer from every analysis. Do you accept or reject the batch? The logic used to obtain the answer is called hypothesis testing.

First, define a threshold P value before you do the experiment. Ideally, you should set this value based on the relative consequences of missing a true difference or falsely finding a difference. In practice, the threshold value (called alpha) is almost always set to 0.05 (an arbitrary value that has been widely adopted).

Next, define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same mean.

Now, perform the appropriate statistical test to compute the P value.

• If the P value is less than the threshold, state that you “reject the null hypothesis” and that the difference is “statistically significant”.

• If the P value is greater than the threshold, state that you “do not reject the null hypothesis” and that the difference is “not statistically significant”. Y*ou cannot conclude that the null hypothesis is true. All you can do is conclude that you don't have sufficient evidence to reject the null hypothesis.*

Extremely significant?
Once you have set a threshold significance level (usually 0.05), every result leads to a conclusion of either "statistically significant" or not "statistically significant". Some statisticians feel very strongly that the only acceptable conclusion is significant or 'not significant', and oppose use of adjectives or asterisks to describe values levels of statistical significance.

Many scientists are not so rigid, and so prefer to use adjectives such as “very significant” or “extremely significant”.

P value	Wording	Summary
< 0.0001	Extremely significant	****
0.0001 to 0.001	Extremely significant	***
0.001 to 0.01	Very significant	**
0.01 to 0.05	Significant	*
≥ 0.05	Not significant	ns

Advice: Avoid the concept of 'statistical significance' when possible

The term "significant" is seductive and easy to misinterpret, because the statistical use of the word has a meaning entirely distinct from its usual meaning. Moreover, a result that is not statistically significant (in the first experiment) may turn out to be very important.

Using the conventional definition with alpha=0.05, a result is said to be statistically significant when a difference that large (or larger) would occur less than 5% of the time if the populations were, in fact, identical.

The entire construct of 'hypothesis testing' leading to a conclusion that a result is or is not 'statistically significant' makes sense in situations where you must make a firm decision based on the results of one P value. While this situation occurs in quality control and maybe with clinical trials, it rarely occurs with basic research. If you do not need to make a decision based on one P value, then there is no need to declare a result "statistically significant" or not. Simply report the P value as a number, without using the term 'statistically significant'. Or consider simply reporting the confidence interval, without a P value.