STATISTICS SURVIVAL SERIES

Essential Dos and Don’ts of Statistics

GraphPad

The Importance of Statistics in Science

As a scientist, research is the foundation of your work. You begin with insightful questions, develop hypotheses, and then collect and analyze data from which you make your conclusions. This allows you to solve meaningful problems in the world.

This research process, however, isn't always that simple. No matter how carefully you perform your experiments, research always contains uncertainty and error. That is where statistics comes in. Statistics help you quantify uncertainty in your research. This provides credibility to your process and conclusions. That's why it's essential in every area of science.

> No matter how carefully you perform your experiments, research always contains uncertainty and error. That is where statistics comes

However, being a scientist doesn't make you a statistician. The truth is many scientists have only taken introductory stats courses. As a result, analyzing data from even relatively simple experiments can be overwhelming. Even worse, when done incorrectly, your interpretations of the results may be invalid.

The following guide is designed to help. It outlines ways you may be inadvertently—and negatively—impacting your research. From skipping important questions as part of your experimental design, to influencing the data collection process, review these essential dos and dont's and improve the accuracy of your research.

CHAPTER 2

Don’t Start Without a Plan

One of the most common mistakes people make in research is collecting a bunch of data without having thought through what questions they are trying to answer, what specific hypotheses they want to test, and what statistical tests they can use to test these hypotheses.

Analyzing data requires many important decisions: Parametric or nonparametric test? Eliminate outliers or not? Transform the data first? Normalize to external control values? Adjust for covariates? Use weighting factors in regression?

All these decisions (and more) should be part of your experimental design.

When decisions about statistical analysis are made after inspecting the data, it is easy for statistical analysis to become a high-tech Ouija board—a method to produce results you hoped to see, as opposed to an objective method of analyzing data. The new name for this is p-hacking, and we will get into that in much more detail in the next chapter.

However, having a plan does more than just help you avoid making critical mistakes in the analysis it can also save you time. Imagine after weeks of data collection, you examine your data and realize you should have added another variable or measured something differently. Changing the way the data is analyzed after you’ve seen the results of your work can and will impact the validity of the statistical results. Many researchers have fallen into this trap.

☑ **DO**

Before you start data collection, have an experimental design planned out. This includes how you plan to collect data, parameters to your data collection, as well as an idea of how you plan on analyzing the data.

CHAPTER 3

Don’t Be a P-Hacker

We must address p-hacking.

The act of p-hacking is a pervasive problem, and occurs when you influence the data collection process or statistical analyses performed in order to produce a statistically significant result, whether you mean to or not.

This is important: Statistical results can only be interpreted at face value when every choice in data analysis was performed exactly as planned and documented as part of the experimental design.

Unfortunately, p-hacking occurs quite frequently, and often you may not even realize it’s happening.

Imagine this scenario: You develop a hypothesis, define test parameters, and then collect and analyze the data. You’ve decided to use the traditional value for your significance level of α=0.05 for your statistical tests (i.e. P values less than 0.05 are considered “statistically significant”). The results you obtain aren’t statistically significant but show a difference or trend in the direction you expected, so you collect some more data and reanalyze. Or perhaps you try a different way to analyze the data: remove a few outliers; transform to logarithms; try a nonparametric test; redefine the outcome by normalizing; use a method to compare one variable while adjusting for differences in another—the list of possibilities is endless. The point is that you keep trying until you obtain a statistically significant result.

The results from data collected this way cannot be interpreted at face value. Even if there really is no difference (or no effect), the chance of finding a “statistically significant” result purely by chance after this process exceeds 5%.

The problem is that you introduce bias when you choose to collect more data (or analyze the data differently) only when the P value is greater than 0.05. If the P value was less than 0.05 in the first analysis, it might be larger than 0.05 after collecting more data or using an alternative analysis. But you would never see this if you only collected more data or tried different data analysis strategies when the first P value was greater than 0.05.

> Unfortunately, **p-hacking occurs quite frequently,** and often you may not even realize it’s happening.

Important concepts: Hypotheses and P values

Most statistical tests work by generating not one, but two hypotheses: the Null Hypothesis and the Alternative hypothesis. Before you perform an experiment and record your observations, you should understand these terms:

- **Alternative Hypothesis (H_a):** the hypothesis that the observations are due to some real effect
- **Null Hypothesis (H_0):** the hypothesis that the observations are due to random chance.
- **Significance level (α):** the probability of rejecting H_0 when it's actually true

What is a P value?

Generally, the reason you perform an experiment is because you're interested in H_a (for example, that a treatment improves outcomes or that two groups of measurements have different means). However, what you test is H_0 (that the treatment has no effect, or that the two groups have the same means. The P value is the probability of obtaining an outcome at least as extreme as the outcome you observed if H_0 were true.

What does "statistically significant" mean?

Once you've calculated a P value, you can test for statistical significance. If your P value is smaller than α, it's unlikely that you would have obtained your results if H_0 were true. Therefore, you can **reject** H_0, stating that the effect is "statistically significant."

Three Types of P-Hacking

1 Changing the values analyzed.

The first kind of P-hacking involves changing the actual values analyzed. Examples include ad hoc sample size selection, switching to an alternate control group (if you don't like the first results and your experiment involved two or more control groups), trying various combinations of independent variables to include in a multiple regression (whether the selection is manual or automatic), trying analyses with and without outliers, and analyzing various subgroups of the data.

2 Reanalyzing a single data set with different statistical tests.

Examples: Trying a parametric and then a nonparametric test. Analyzing the raw data, then analyzing the logarithms of the data.

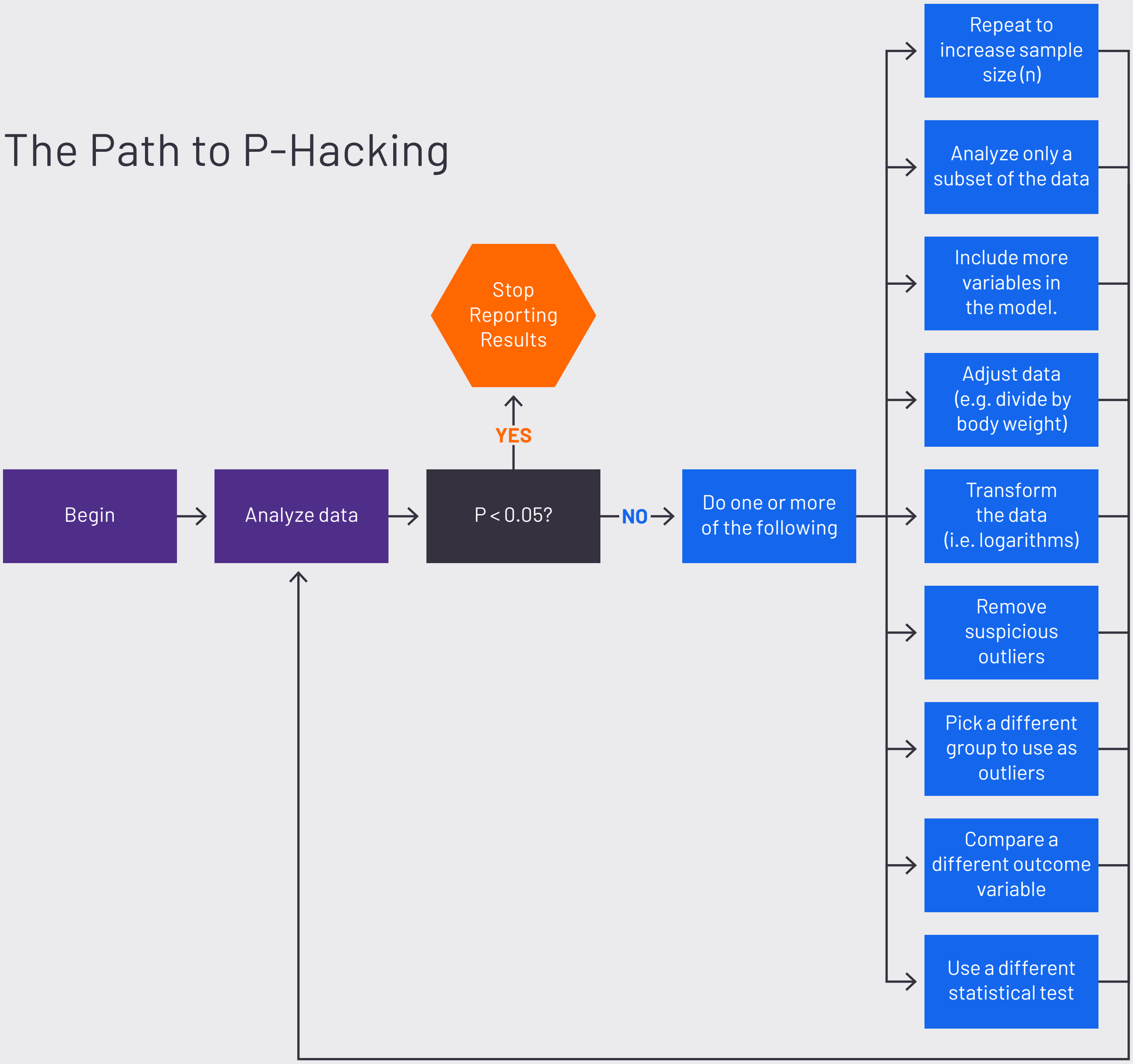
3 Inadvertently choosing the wrong analysis.

This happens when researchers performed a reasonable analysis given their assumptions and their data, but would have done other analyses that were just as reasonable had the data turned out differently.

☑ DO

The bottom line is that exploring your data can be a very useful way to generate hypotheses and make preliminary conclusions. When you do so, make sure all such analyses are clearly labeled, and then retested with new data.

The Path to P-Hacking


Begin
Analyze data
P < 0.05?
YES
Stop Reporting Results
NO
Do one or more of the following
Repeat to increase sample size (n)
Analyze only a subset of the data
Include more variables in the model.
Adjust data (e.g. divide by body weight)
Transform the data (i.e. logarithms)
Remove suspicious outliers
Pick a different group to use as outliers
Compare a different outcome variable
Use a different statistical test

CHAPTER 4

Don't Add Subjects Until You Hit Significance

Adding subjects until you hit significance may be tempting, but it is also misleading.

Here's a common scenario. Rather than choosing a sample size before beginning a study, you simply repeat the statistical analyses as you collect more data, and then:

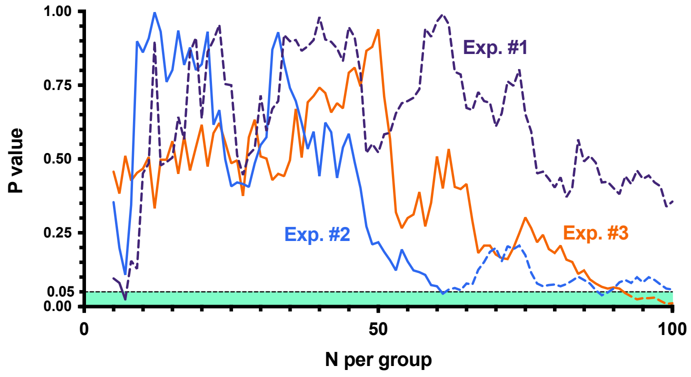
- If the result is not statistically significant, collect some more data, and reanalyze
- If the result is statistically significant, stop the study

The problem with this approach is that you'll keep going if you don't like the result, but stop if you do like the result. The consequence is that the chance of obtaining a "significant" result if the null hypothesis were true is a lot higher than 5%.

Simulations to Demonstrate the Problem

The graph below illustrates this point via simulation. We began by simulating two groups of data by drawing values from a Gaussian distribution (mean=40, SD=15, but these values are arbitrary). Both groups were simulated using exactly the same distribution, and so have the exact same true mean value. We picked N=5 in each group and computed an unpaired t test (comparing the means of two groups) and recorded the P value. Then we added one subject to each group (so N=6) and recomputed the t test and P value. We repeated this until N=100 in each group. Then we repeated the entire simulation three times. Because these simulations were done comparing two groups with identical population means, any "statistically significant" result we obtain must be a coincidence — a Type I error.

The graph plots P value on the Y axis vs. sample size (per group) on the X axis. The green shaded area at the bottom of the graph shows P values less than 0.05, so deemed "statistically significant".



Experiment 1 (purple) reached a P value less than 0.05 when N=7, but the P value is higher than 0.05 for all other sample sizes. Experiment 2 (blue) reached a P value less than 0.05 when N=61 and also at N=88 and 89. Experiment 3 (orange) curve hit a P value less than 0.05 when N=92 and remained lower than this value until N=100.

If we followed the sequential approach, we would have declared the results in all three experiments to be "statistically significant". We would have stopped when N=7 in the first (purple) experiment, so would never have seen the dotted parts of its curve. We would have stopped the second (blue) experiment when N=61, and the third (orange) experiment when N=92. In all three cases, we would have declared the results to be "statistically significant".

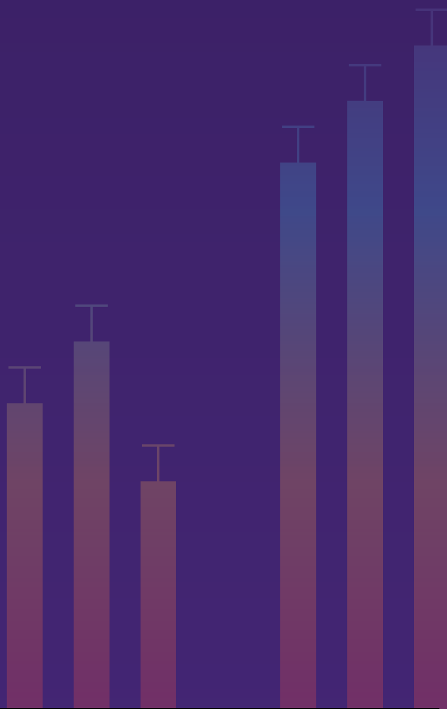
Since these simulations were created for values where the true mean in both groups was identical, any declaration of "statistical significance" is a Type I error. If the null hypothesis is true (the two population means are identical) we expect to see this kind of Type I error in 5% of experiments (if we use the traditional definition of α=0.05 so P values less than 0.05 are declared to be significant).

Carried out long enough, this kind of sequential approach will always result in a Type I error. In other words, if you extended any experiment long enough (infinite N), they would all eventually reach statistical significance. Of course, in some cases you would eventually give up even without "statistical significance". But this sequential approach will produce "significant" results in far more than 5% of experiments, even if the null hypothesis were true, and so this approach is invalid.

> Since these simulations were created for values where the **true mean in both groups was identical,** any declaration of “statistical significance” is a **Type I error.**

☑ **DO**

It is important that you choose a sample size and stick with it. You’ll fool yourself if you stop when you like the results, but keep going when you don’t. The alternative is using specialized sequential or adaptive methods that take into account the fact that you analyze the data as you go. To learn more about these techniques, research ‘sequential’ or ‘adaptive’ methods in advanced statistics books.



Statistical Analysis with GraphPad Prism

GraphPad Prism is the world's leading data analysis and graphing solution purpose-built for scientific research.

750,000 of the world's leading scientists use Prism to save time performing statistical analyses, make more accurate analysis choices, and elegantly graph and present their scientific research.

Download a free trial today—no credit cards, no commitments— and you will be on your way to sharing your research with the world!

www.graphpad.com

FOR MAC AND WINDOWS