**Overusing the Term “Statistically Significant” Makes You Look Clueless**

**A primer on interpreting other people’s hypothesis tests**

If you’re in the market for a new tongue-twister, try this paraphrase of a [classic](http://www.stat.columbia.edu/~gelman/research/published/signif4.pdf):

“The difference between statistically significant and statistically non-significant is not necessarily significant.”

As a [recovering statistician](https://www.linkedin.com/in/kozyrkov/), I have the pleasure of knowing many [data experts](http://bit.ly/quaesita_universe) and the displeasure of meeting a lot of [posers](http://bit.ly/quaesita_charlatan). Though the people I bump into are hardly a [random sample](https://bit.ly/quaesita_srstrees1), I have noticed a [correlation](http://bit.ly/quaesita_correlation) between throwing around the term “*statistically significant*” and having no idea what it means. Data experts hardly ever use it in casual conversation. Why is that?

Genuine data experts hardly ever shove the term “statistically significant” into casual conversation.

In short, because the term “*statistically significant*” isn’t especially significant from the perspective of *other people*. It matters ***who*** set up the calculation… and if it wasn’t you, the results likely aren’t for you either. Experts understand this, so they don’t foist their statistical significance on innocent bystanders.

There’s a gap between what the term really means and how charlatans (mis)use it to trick you. I tackled this in detail in an [earlier article](http://bit.ly/quaesita_insignif) and I’ve also added a quick summary to the bottom of this page.\* In this article, I’ll illustrate the process of statistical decision-making with an example and help immunize you against charlatans by giving you a primer on **interpreting other people’s hypothesis test results**.

**An insignificant example**

Imagine you have a friend who makes blue ribbons. Your friend also happens to be a skilled [decision-maker](http://bit.ly/quaesita_di) who understands how to use the classical hypothesis testing framework, which I’ve explained [here](http://bit.ly/quaesita_damnedlies) for those who need a refresher. I’ll assume that you know what a *default action* is and how the process works. If you don’t, I recommend a small detour before continuing here:

**[Never start with a hypothesis](https://towardsdatascience.com/hypothesis-testing-decoded-for-movers-and-shakers-bfc2bc34da41" \t "_blank)**

**[Lies, damned lies, and STAT101](https://towardsdatascience.com/hypothesis-testing-decoded-for-movers-and-shakers-bfc2bc34da41" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/hypothesis-testing-decoded-for-movers-and-shakers-bfc2bc34da41" \t "_blank)

Now let’s apply this framework, giving you a taste of the reasoning.

**Default action**

By [default](http://bit.ly/quaesita_damnedlies), your friend will switch to a cheaper supplier of blue dye as long as there’s no human-detectable visual difference ([operationalized](http://bit.ly/quaesita_opera) to include a tolerance for the fallibility of human sense) from the blue ribbons they’ve already made.

**Setting selection**

Before — before, not after (!!) — doing a hypothesis test, your friend must pick the quality at which the decision will be made by adjusting some settings, including [significance level, power/sample size, and assumptions](http://bit.ly/quaesita_statistics). I wish more people appreciated just how many knobs and dials statistics has.

**Hypothesis testing**

Then your friend collects [data](https://bit.ly/quaesita_srstrees1), [magic happens](http://bit.ly/quaesita_stc013)\*\*, and they end up with a [p-value](http://bit.ly/quaesita_pesky) summarizing whether they should cheapen up (*default action*) or stick with the expensive blue dye (*alternative action*).

**Conclusion**

At the end of the statistical odyssey, your friend rejects the [null hypothesis](http://bit.ly/quaesita_fisher), and stays loyal to the more expensive blue dye. It was statistically significant! That’s what your friend concluded… but what should *you* conclude?

**Interpretation — From their conclusion to yours**

**Was there a big difference?**

Should you conclude that there was a ***big difference*** between the dyes?

Not necessarily. Without knowing how your friend [operationalized](https://bit.ly/quaesita_metricdesign) *“human-detectable visual difference”* you won’t know if the difference that they now believe in is something you’d call big.

At this point in the angry statistician rant, we often get the misplaced fuss about “effect sizes” (which I’ll cover in another post). For now, all I’ll say is that those discussions often miss the point, since effect sizes should already be built into the hypotheses whenever we’re dealing with proper statistical decision-making.

**Was it an important finding?**

Was it an ***important*** finding? Not necessarily. It’s only important to folks who care about blue dyes **and** what *your friend* chooses to do (a list of dramatis personae that surely includes your friend, who took a [statistical](http://bit.ly/quaesita_statvocab) approach because the blue dye question was personally or professionally important to them).

Why would it be *important* to anyone else? Why indeed. It might not be.

Statistical summaries essentially boil down to a statement about surprise. As in, “this surprises me.”

Sometimes, though, the very fact that particular individuals were surprised enough to change their minds about something can be important news for the rest of us. For example, you might consider a result published in a medical journal *important* because the scientists who wrote it changed their minds about how a medication works. Sure, you have no idea what they actually did in that lab of theirs (even if they summarized it for you in their paper, you still lack a bunch of context... and if you’re not an expert, you won’t grokk the nuances of the [assumptions](http://bit.ly/quaesita_saddest) they made), but if [you choose to trust them](http://bit.ly/quaesita_scientists), you might take an interest in whatever *they* find important and surprising. Is the finding ***true***, though? Unfortunately, if all you have is a summary, the only conclusion you’re allowed to make is this: *“The people I trust are convinced, therefore I’m convinced.”* If you’re not going to do the research yourself, all you can do is trust in your chosen experts and parrot their conclusions, or waft through life doing your best to avoid forming opinions about anything at all. But I digress — if you want more on this line of reasoning, you’ll find it [here](http://bit.ly/quaesita_scientists):

**[Why do we trust scientists?](https://kozyrkov.medium.com/why-do-we-trust-scientists-98c24e3b9f0e" \t "_blank)**

**[Now’s a good time to rethink our assumptions about fact and fiction](https://kozyrkov.medium.com/why-do-we-trust-scientists-98c24e3b9f0e" \t "_blank)**

[kozyrkov.medium.com](https://kozyrkov.medium.com/why-do-we-trust-scientists-98c24e3b9f0e" \t "_blank)

Let’s get back to those blue ribbons, shall we?

**Is it meaningful?**

Everyone except your friend may consider the whole question of getting the perfect blue dye for ribbons to be trivial. So, is your friend’s finding meaningful?

Unless you know (and buy into) the entire set of [assumptions](http://bit.ly/quaesita_saddest), [risk settings](http://bit.ly/quaesita_statistics), and [decision framing](http://bit.ly/quaesita_damnedlies), the only *meaning* here is that you know which dye your friend will be using next week.

Does it *mean* the other dye from the cheaper supplier is of lower quality? Certainly not.

Your friend’s decision setup had nothing to do with the question of dye *quality*, so their conclusions don’t cover that topic. And even if quality were part of the question, it’s important to check the [operationalization](https://bit.ly/quaesita_metricdesign). Your notion of dye quality might be different from your friend’s.

Unfortunately, the leap from your friend’s question (*“Should I abort my planned move to Supplier B’s cheaper blue dye?”*) to a gossip blogger’s hot take (*“Supplier B makes bad quality dye, according to science!!!”*) is exactly the kind of feebleminded data illiteracy that makes us statisticians want to punch something.

It's exactly the kind of feebleminded data illiteracy that makes us statisticians want to punch something.

**Should you avoid the cheaper dye?**

How about if you’re in the blue ribbon business too — should you be convinced that the cheaper dye is no good? Not necessarily.

You might not [buy into the assumptions](http://bit.ly/quaesita_saddest) your friend made about the [population](http://bit.ly/quaesita_popwrong) of dye (maybe they assumed [zero variance](http://bit.ly/quaesita_gistlist) for all bottles of dye and only tested one, which might have been a bad batch) and/or you might have different tolerance for statistical risks and/or your wouldn’t frame your own decision about dye suppliers the same way, so their findings might have nothing to do with you.

Good questions come from possibilities that make you curious about probabilities.

However, you are welcome to consume your friend’s findings from an [analytics](http://bit.ly/quaesita_battle) perspective… as long as you don’t take them too seriously. There’s plenty of inspiration you could take from other people's work to help you ask better questions and structure your own approach to decision-making. After all, good questions don’t arrive out of nowhere. They come from possibilities that make you curious about probabilities. You need to have some exposure to what might be ***possible*** in order to start asking those good questions. If your friend’s findings make you curious and blue dye is a big deal in your life, you might perhaps be inspired to do your own testing. If not, well, isn't that what noise cancelling headphones are for?

**Thanks for reading! How about a YouTube course?**

If you had fun here and you’re looking for an entire applied AI course designed to be fun for beginners and experts alike, here’s the one I made for your amusement:

**Looking for hands-on ML/AI tutorials?**

Here are some of my favorite 10 minute walkthroughs:

* [AutoML](https://console.cloud.google.com/?walkthrough_id=automl_quickstart)
* [Vertex AI](https://bit.ly/kozvertex)
* [AI notebooks](https://bit.ly/kozvertexnotebooks)
* [ML for tabular data](https://bit.ly/kozvertextables)
* [Text classification](https://bit.ly/kozvertextext)
* [Image classification](https://bit.ly/kozverteximage)
* [Video classification](https://bit.ly/kozvertexvideo)

**\*Interpretation reminder: “Someone was surprised by something”**

While *“statistically significant”* sounds like a cousin of “*important*” or “*meaningful*” …it isn’t. Unfortunately, the term is often abused in precisely this way. This is a trap.

* **Statistically significant** = someone was surprised by something.
* **Significant** = sufficiently great or important to be worthy of attention; noteworthy.

Please don’t confuse a piece of dry statistical jargon with the poetry of a word that means something entirely different.

Welcome to statistics, where The Answer is [p](http://bit.ly/quaesita_puppies) = 0.042 but you don’t know what the question was.

More on this here: