**It’s the Data That Should Surprise You, Not the P-Values**

**Your conclusions should not be based on statistical significance tests, nor should they be based on p<0.05**



Photo by [Daniel Tuttle](https://unsplash.com/@danieltuttle?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

Statistics is the field that changes our minds under uncertainty [1]. Our mind (i.e., the preconceived opinion about something) is usually formulated as a null hypothesis that summarizes what…

* The data has to contradict what we believe is true
* It should be unlikely that this contradiction has occurred by chance alone when the null hypothesis is true.

For example, we generally assume that a specific coin is fair, i.e., that there is an equal probability of getting heads and tails. If we now observe, for example, two heads when flipping the coin two times, this contradicts our initial expectation. On the other hand, such an outcome is very likely with a fair coin (25% of the time). Consequently, the data is not surprising enough to reject the null hypothesis. On the contrary, if we flip the coin ten times and observe ten heads, this result might be sufficiently surprising since the chance of such an outcome with a fair coin is only about 0.1%. This leads directly to the inevitable question:

**When exactly is data surprising enough to change our minds, i.e. when should we reject the null hypothesis?**

Statistical hypothesis testing has a simple answer: if the null hypothesis is actually true, then the probability of the resulting outcome (p) should be smaller than some *a priori-defined* significance level (ɑ). Depending on the field, an often-used convention is ɑ=5% — consequently, any findings where p<5% are called statistically significant.

This definition becomes much more apparent when applied to the coin-flip example. We would reject our belief in a fair coin if we observed data that contradicts this belief **and** the commencement that this data would occur in less than 5% of the cases if the coin actually were fair. As we see from the table below, we would reject our belief in a fair coin if observing more than four heads in a row. In other words, we would say that this finding (i.e., the coin is not fair) is statistically significant since p<5%.



Image by author

But what does it actually mean for something to be statistically significant? Let’s assume you would bet considerable money on the next coin flip. Would your decision be different when the four previous coin flips were heads vs. the five previous flips were heads? Probably not, but you would be a bit more confident that the next coin flip will also be heads after five heads, right?

A conclusion does not immediately become true (or more important) on one side of the divide and false (or unimportant) on the other

This is actually one problem with statistical significance testing: for many people, statistically significant results are more convincing than non-significant results. For example, it is easier to publish scientific results when they are “statistically significant”. Consequently, many researchers have tried to turn non-significant findings into something *more convincing* by choosing creative descriptions. The following word cloud shows some often-used expressions from peer-reviewed journal articles where the authors set a threshold of p<5% and failed to achieve it [2].

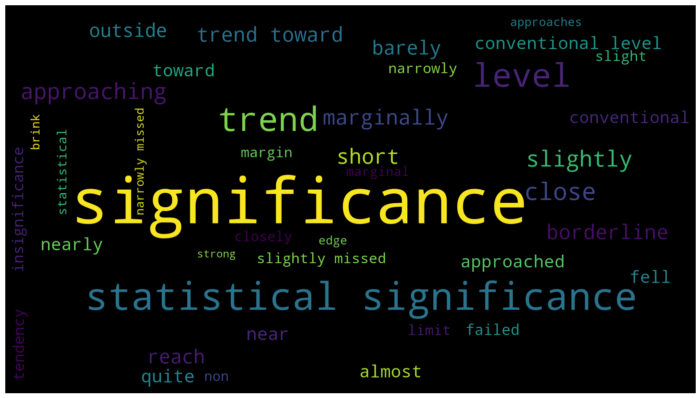


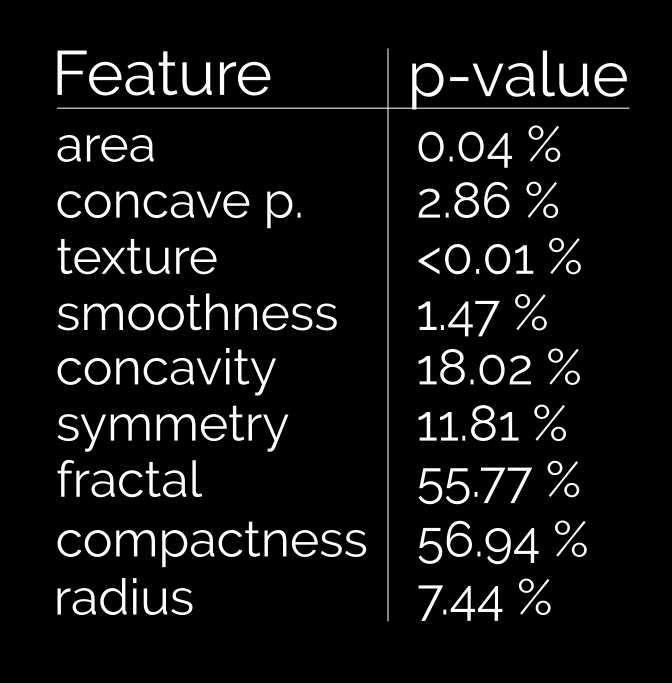
Image by the author, data originate from [1]

The wording is fascinating since it often describes an aspect of the result that does not exist. For example, towards, approaching, or trend towards significance is not true since the p-value is a number without a trend in any direction.

**Significance does not imply importance**

This brings us to another important aspect: the p-value does not imply importance, nor does it reflect the size of an effect.

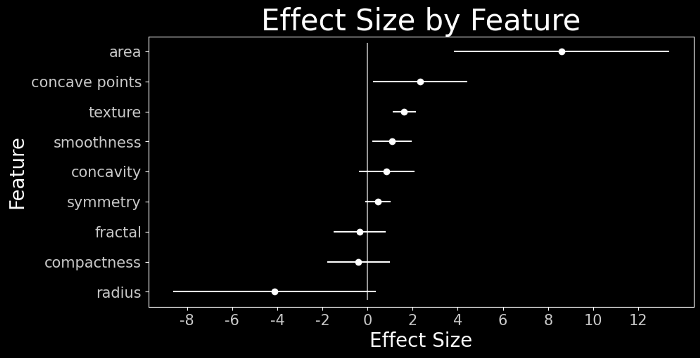
For example, let us investigate the predictors for a malignant or benign tumor in the [Breast Cancer Wisconsin (Diagnosis) Data Set](https://archive-beta.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic) [4] using logistic regression. The following table summarizes the p-values per predictor.



p-values per predictor of the [Breast Cancer Wisconsin (Diagnosis) Data Set](https://archive-beta.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). Image by the author.

A predictor with a low p-value is unlikely not to be related to the response variable (i.e., malignant or benign tumor). From these p-values alone, one might be tempted to judge that, for example, the radius is less important than smoothness since its p-value is larger. However, the p-value does not allow any conclusion on how much these predictors decrease or increase the chance of a malignant tumor.

To answer this question, we need to analyze the effect size: A large absolute effect size means that a research finding has practical meaning. In contrast, a small absolute effect size indicates limited practical applications.



Effect sizes per predictor of the [Breast Cancer Wisconsin (Diagnosis) Data Set](https://archive-beta.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). Image by the author.

In the breast cancer dataset, the effect size with their uncertainties conveys much more information than the p-values alone. For example, we can now conclude that we found strong evidence for a moderate association between smoothness and malignant tumors. For the radius, however, we can conclude that we found moderate evidence for a potentially strong association between the radius and a malignant tumor.   
Consequently, although we are, based on the available data, convinced that the smoothness is strongly associated with the tumor being malignant, its effect is relatively small. Thus, it might be clinically more relevant to focus on the radius, collecting more data to decrease the uncertainty.

In conclusion, small effects with “significant” p-values may have limited practical importance.

**Conclusion**

This leads us back to the first sentence of this blog post (which I copied from one of Cassie Kozyrkov’s excellent blog posts BTW.):

[Statistics is the field that changes our minds under uncertainty](https://towardsdatascience.com/whats-the-point-of-statistics-8163635da56c)

There is, and will always remain uncertainty. This uncertainty cannot be removed by relying on only one single value. Rather, we have to embrace the uncertainty and include them in our conclusions. As the American Statistical Association Statement on p-Value concludes:

[No single index should substitute for scientific reasoning](https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.Y4Dpjn3MKUm)

Thanks for reading!