Null Hypothesis Significance Testing

Null Hypothesis Significance Testing

Modern textbooks with titles such as "Statistics for Engineers" and similar do not explicitly teach either Fisher's Test procedure nor the Neyman-Pearson decision-making process, but rather a mixture of both. This is now often called ***Null Hypothesis Significance Testing (NHST)*** and works as follows:

- Two hypotheses, H_0 and H_1 are set up, but H_1 is always the logical negation of H_0
- Then either a "hypothesis test" is performed, whereby a critical region for given α is defined, the test statistic is evaluated and H_0 is either rejected or accepted.
- Alternatively (and more commonly), the test statistic is evaluated immediately, a P-value is found, and H_0 is either rejected or accepted based on that value.
- In either case, there is no meaningful discussion of β, since H_1 is exactly the negation of H_0.

Criticism of NHST

- A small P-value does not guarantee that a large probability that H_0 is false. Fisher did not intend for a small P-value to lead to a clear rejection of H_0, but only to serve as evidence against H_0 if little else is known.
- Rejecting H_0 based on $\alpha = 0.05$ or 0.01 or any other value is arbitrary.
- NHST is actually biased ***against failing to reject*** H_0. From a Bayesian point of view, it is far too easy to reject H_0 because $P[H_0]$ does not enter into NHST.
- A two-sided test such as $H_0 \colon \theta = \theta_0$, $H_1 \colon \theta \neq \theta_0$ is meaningless.
- The power (and β) of the test is not properly defined, since H_1 is just the alternative "not H_0" rather than referring to a distinct value θ_1. Occasionally, this θ_1 is then mentioned indirectly for purposes of power calculations.

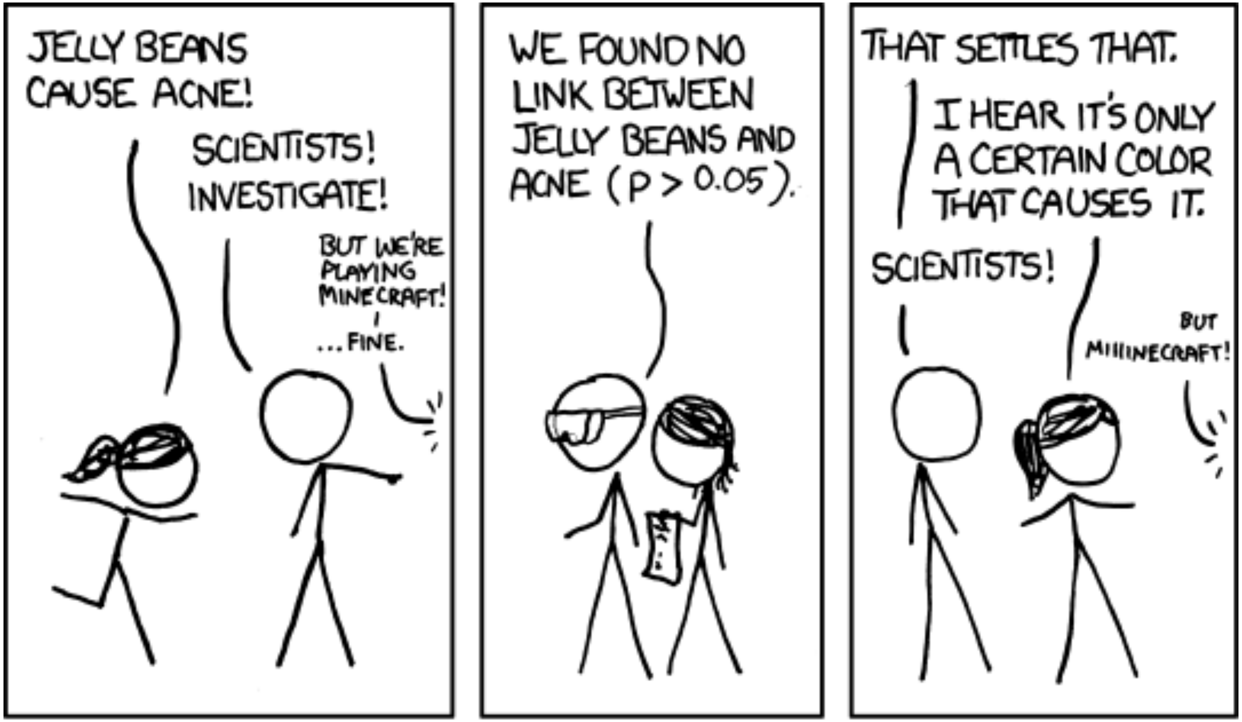
Publication Bias and NHST

NHST is currently the preferred technique for verifying statistical results. In the current academic environment, research papers are only publishable if the results are statistically significant. Editors of scientific journals will usually not publish results where $P = 0.23$, for example.

This means that many interesting studies are not made available to the scientific community because they are considered to be “failed experiments”. However, that does not mean that they are not useful (even if H_0 is true) or that H_0 actually is false (since the experiment may simply not have had enough power).

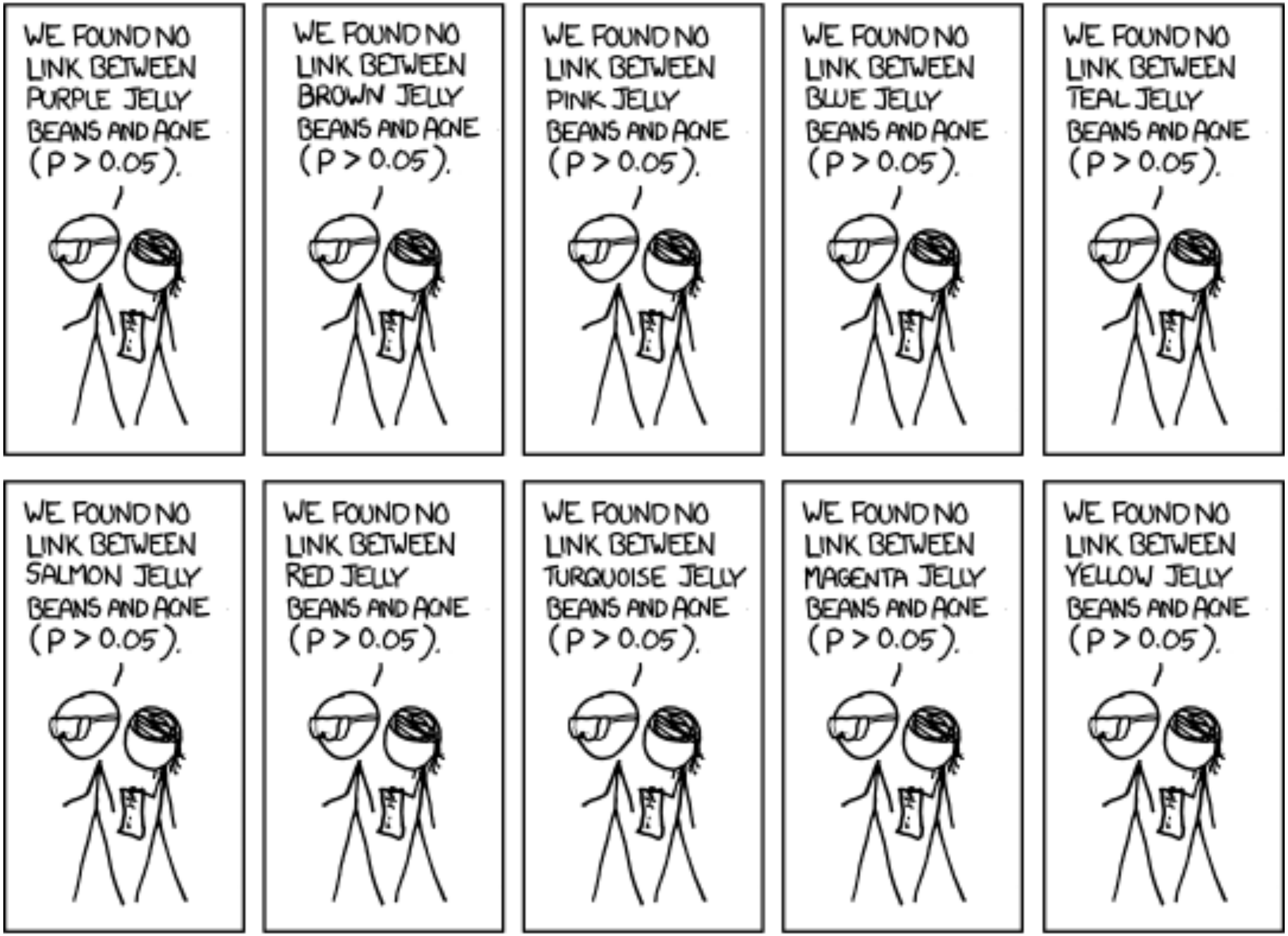
At the same time, this tempts researchers to continue increasing the sample sizes of a study or to do repeated studies until they get a result that is statistically significant and can be published.

Publication Bias and NHST


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

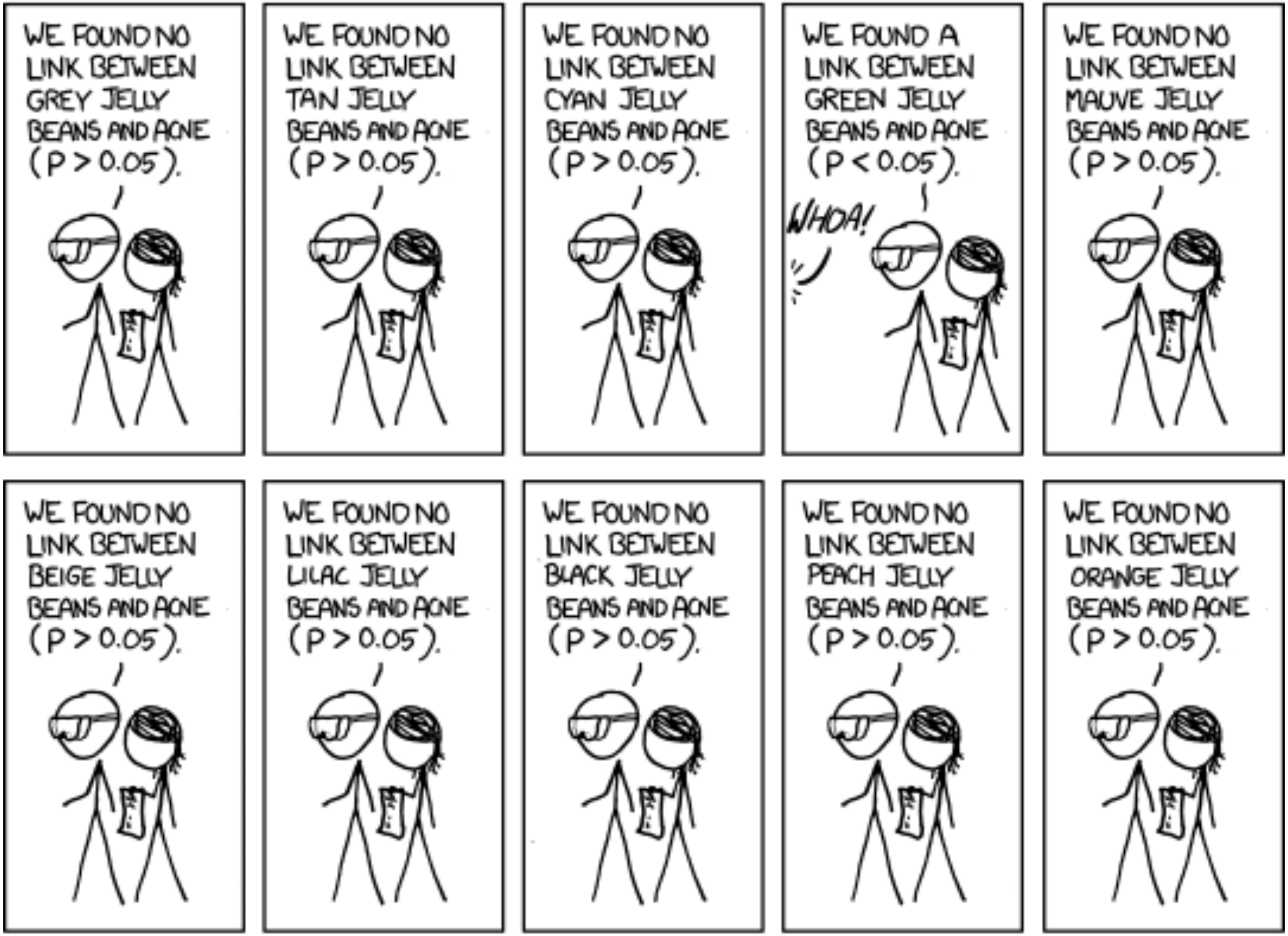
xkcd: significant, Randall Munroe, published on xkcd.com, April 6, 2011

Publication Bias and NHST


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

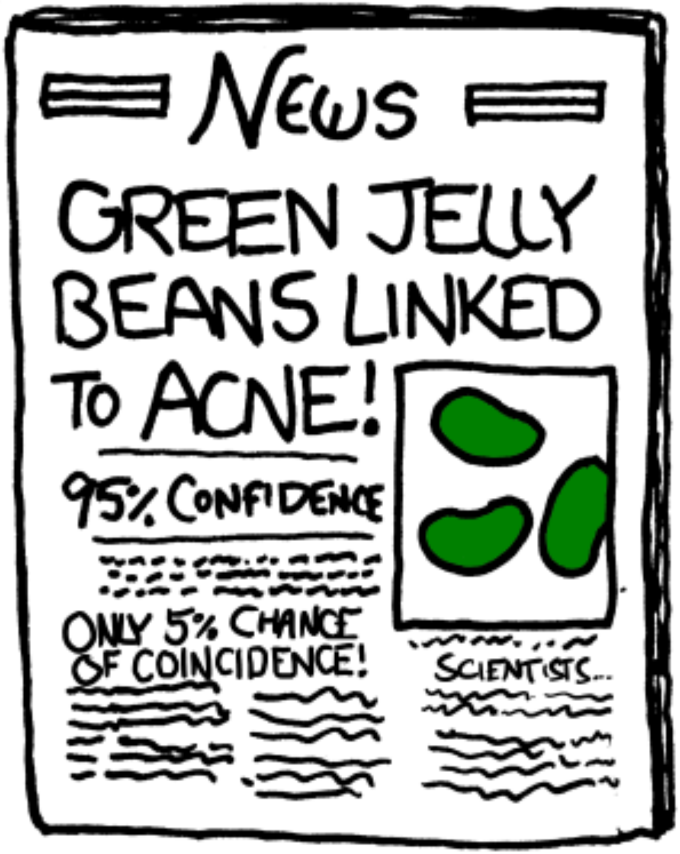
xkcd: significant, Randall Munroe, published on xkcd.com, April 6, 2011

Publication Bias and NHST


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

xkcd: significant, Randall Munroe, published on xkcd.com, April 6, 2011

Publication Bias and NHST


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

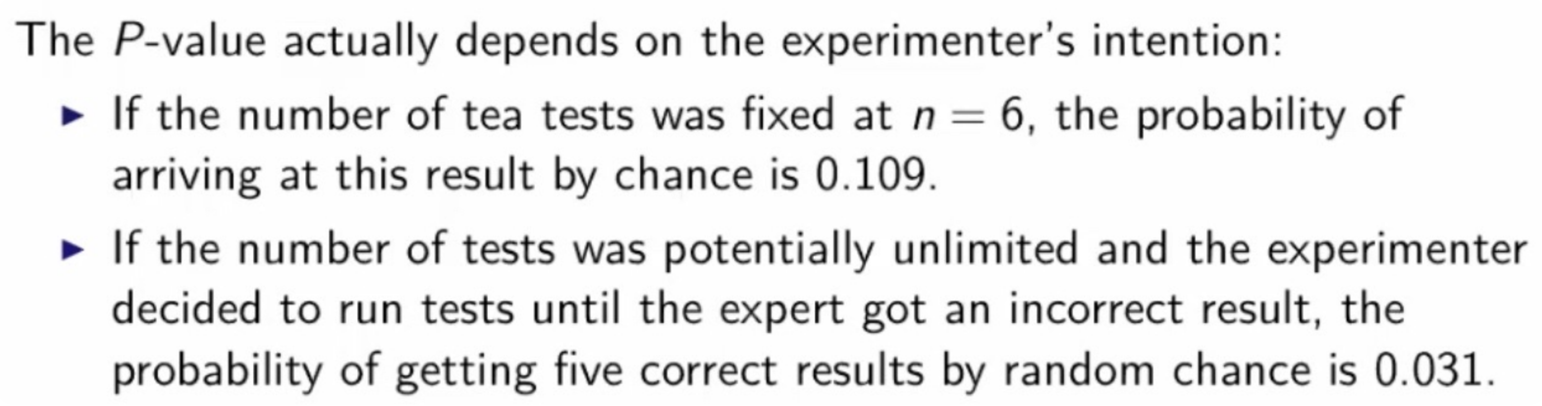
xkcd: significant, Randall Munroe, published on xkcd.com, April 6, 2011

A “Tea-Test”

Consider the following example, given by Fisher in 1935:

> *In England, tea is often drunk together with milk. Suppose a tea expert claims to be able to tell whether the milk or the tea has been poured into a cup first. He is put to the test and is to state whether or not a given cup was produced by pouring milk first. His results are*
>
> *correct, correct, correct, correct, correct, incorrect.*

Question. What is the P-value of this test?



The P-value actually depends on the experimenter’s intention:

- If the number of tea tests was fixed at $n = 6$, the probability of arriving at this result by chance is 0.109.
- If the number of tests was potentially unlimited and the experimenter decided to run tests until the expert got an incorrect result, the probability of getting five correct results by random chance is 0.031.

A “Tea-Test”

There are (at least) two interpretations of this example:

- The intention of the researcher, not just the raw experimental data, may determine the P-value of a test.
- In a hypothesis test, the outcomes that do ***not*** occur are just as important as the outcomes that do occur.

In particular, it is not considered to be good statistical practice to repeat an experiment to reject a null hypothesis until it is successful. To be probabilistically pure in the NHST sense, an experiment should be run once, and if the null hypothesis is not rejected, it should not be repeated.

Implications for Science

This causes problems of another nature - should an experiment that fails due to insufficient power really never be repeated ever again? That appears to be quite contrary to the nature of scientific inquiry.

Of course this is nonsense! In Fisher's approach, data may be obtained as often as desired, a test repeated as often as necessary, since the proof only serves as indirect evidence, not as a definitive rejection of the null hypothesis. In Neyman-Pearson, there are two alternatives and in a given situation, a decision is necessary. Therefore, data is gathered only once and a decision is made in the concrete circumstances. The fact that the alternative hypothesis is usually not just the negation of H_1 ensures that the result is meaningful.

However, since most researchers use the NHST approach, there is a large proportion of Type II errors in unpublished papers and many studies that would have led to good results that could not be obtained due to insufficient power (e.g., small sample size) are abandoned forever.