The positive predictive value

John Ioannides (2005) writes about how most published research findings are false. At the same time, we have learned that if you set your alpha at 5%, the Type 1 error rate (or the probability of a false positive given that the null hypothesis is true) will not be higher than 5% (in the long run). How are these two statements related? Why aren't 95% of published research findings true?

The trick to understanding the discrepancy between the possibility that more than 50% of published findings are false (positives) and the 5% Type 1 error rate is that two different probabilities are calculated. The Type 1 error rate is the probability of saying there is an effect, when there is no effect. Ioannides calculates the *positive predictive value* (PPV), which is the conditional probability that if a study turns out to show a statistically significant result, there is actually a true effect.

Some definitions

It helps to clearly define the different probabilities that can be calculated when talking about the probability that a published result is a true positive of a false positive.

False Positive (FP): Concluding there is a true effect, when there is a no true effect (H0 is true). This is also referred to as a **Type 1 error**, and indicated by **α**.

False Negative (FN): Concluding there is a no true effect, when there is a true effect (H1 is true). This is also referred to as a **Type 2 error**, and indicated by **β**.

True Negative (TN): Concluding there is no true effect, when there is a no true effect (H0 is true). This is the complement of False Positives, and is thus indicated by**1-α**.

True Positive (TP): Concluding there is a true effect, when there is a true effect (H1 is true). This is the complement of False Negatives, and is thus indicated by **1-β**.

The probability of observing a true positive when there is a true effect is, in the long run, equal to the statistical power of your study. The probability of observing a false positive when the null hypothesis is true is, in the long run, equal to the alpha level you have set, or the Type 1 error rate.

You might also be interested in knowing what the probability is that if you have observed a significant result in an experiment, the result is actually a true positive. In other words, in the long run, how many true positives can we expect, among all true positives and false positives. This is known as the Positive Predictive Value (PPV). We can also calculate how many false positives we can expect, among all true positives and false positives. This is known as the False Positive Report Probability (Wacholder, Chanock, Garcia-Closas, El ghormli, & Rothman, 2004, sometimes also referred to as the False Discovery Rate, Colquhoun, 2017).

$$\textbf{PPV}: \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

$$\textbf{FPRP}: \frac{False\ Positives}{(True\ Positives + False\ Positives)}$$

The PPV and FPRP combine classic Frequentist concepts of statistical power and alpha levels with prior probabilities that H0 and H1 are true. They depend on the proportion of studies you do where there is an effect (H1 is true), and where there is no effect (H0 is true), in addition to the statistical power, and the alpha level. After all, you can only observe a false positive if the null hypothesis is true, and you can only observe a true positive if the alternative hypothesis is true. Whenever you perform a study, you are either operating in a reality where there is a true effect, or you are operating in a reality where there is no effect – but you don't know in which reality you are.

When you perform studies, you will be aware of all outcomes of your studies (both the significant and the non-significant findings). When you read the literature, there is publication bias, and you often only have access to significant results. This is when thinking about the PPV (and the FPRP) becomes important. If we set the alpha level to 5%, in the long run 5% of studies where H0 is true (FP + TN) will be significant. But in a literature with only significant results, we do not have access to all true negatives, and it is possible that the proportion of false positives in the literature is much larger than 5%.

Which outcome can you expect if you perform a study?

Let's assume you perform 200 experiments in your lifetime. Let's also assume that you plan, and miraculously achieve, exactly 80% power in all studies where there is a true effect to be found. Furthermore, you consistently use an alpha level of 5%, and therefore, when there is no true effect to be found, you will observe a false positive 5% of the time. Finally, let's assume that in 50% of the experiments (i.e., 100 out of 200 experiments) the

null hypothesis is true, and in the remaining 50% of the experiments H1 is true. What, on average, can you expect to happen in these 200 studies?

We can calculate the proportion for all experiments that will yield a false positive, a true positive, a true negative, and a false negative, as shown in the table below:

	H0 True (50%)	**H1 True (50%)**
Significant Finding (Positive result) α = 5%, 1-β=80%	**False Positive 5%*50%=2.5% (5 studies)**	**True Positive 80%*50%=40% (80 studies)**
Non-Significant Finding (Negative result) 1-α = 95%, β=20%	**True Negative 95%*50%=47.5% (95 studies)**	**False Negative 20%*50%=10% (20 studies)**

Thus, 2.5% of all studies will be a false positive (a 5% Type 1 error rate, multiplied by a 50% probability that H0 is true). 40% of all studies will be a true positive (80% power multiplied by a 50% probability that H1 is true).

For the 85 positive results (80 + 5), the false positive report probability is 5/85 = 0.0588. At the same time, the alpha of 5% guarantees that (in the long run) 5% of the 100 studies where the null hypothesis is true are Type 1 errors: 5%*100 = 0.05. This is also true. When we do 200 studies, at most 0.05*200 = 10 could possibly be false positives (if H0 was true in all experiments). In the 200 studies we performed (and where H0 was true in only 50% of the studies), the **proportion of false positives for all experiments** is only 2.5%. Thus, for all experiments you do, the proportion of false positives will, in the long run, never be higher than the Type I error rate set by the researcher (e.g., 5% when H0 is true in all experiments), but it can be lower (when H0 is true in less than 100% of the experiments). You can redo these calculations by hand (try them for a scenario where the null hypothesis is true in 40% of the studies, and the alternative hypothesis is true in 60% of the studies).

Q1: We see that we control the Type 1 error rate at 5% by using an alpha of 0.05. Still, the proportion of false positives for all experiments we have performed turns out to be much lower, namely 2.5%, or 0.025. Why?

A) The proportion of false positives for all experiments we have performed is a variable with a distribution around the true error rate – sometimes it's higher, sometimes it's lower, due to random variation.

B) The proportion of false positives for all experiments we have performed is only 5% when H0 is true for all 200 studies.

C) The proportion of false positives for all experiments we have performed is only 5% when you have 50% power – if power increases above 50%, the proportion of false positives for all experiments we have performed becomes smaller.

D) The proportion of false positives for all experiments we have performed is only 5% when you have 100% power, and it becomes smaller if power is lower than 100%.

We can do these calculations by hand, but there is also a great app, made by Felix Schönbrodt, that calculates these probabilities for us. Go to http://shinyapps.org/apps/PPV/. A second version of this online app can be reached at http://shiny.ieis.tue.nl/PPV/. At any moment, either of these app should be online.

Let's recreate the example we discussed above. On the left, you see some sliders. Set the "% of a priori true hypotheses" slider to 50%. Leave the 'α level' slider at 5%. Set the 'Power' slider to 0.8 (or 80%). Leave the '% of p-hacked studies' slider at 0.

We get the following results summary:

true positives: 40%; false negatives: 10%; true negatives: 47.5%; false positives: 2.5%

Positive predictive value (PPV): 94.1% of claimed findings are true

False discovery rate (FDR): 5.9% of claimed findings are false

(Note: The FDR and FPRP are different abbreviations for the same thing)

Q2: First, let's just look at the probability that you will find a true positive (which is often a goal in research). What will make the biggest difference in improving the probability that you will find a true positive? Check your ideas by shifting the sliders

A) Increase the % of a-priori true hypotheses

B) Decrease the % of a-priori true hypotheses

C) Increase the alpha level

D) Decrease the alpha level

E) Increase the power

F) Decrease the power

Increasing the power requires bigger sample sizes, or studying larger effects. Increasing the % of a-priori true hypotheses can be done by making better predictions – for example building on reliable findings, and relying on strong theories. These are useful recommendations if you want to increase the probability of performing studies where you find a statistically significant result.

Q3: Set the "% of a priori true hypotheses" slider to 50%. Leave the 'α level' slider at 5%. Leave the '% of p-hacked studies' slider at 0. The title of Ioannidis' paper is 'why most published research findings are false'. One reason might be that studies often have low power. At which value for power is the PPV 50%. In other words, at which level of power is a significant result just as likely to be true, as that it is false?

A) 80%

B) 50%

C) 20%

D) 5%

It seems low power alone is not the best explanation for why most published findings might be false. Ioannidis (2005) discusses some scenarios under which it becomes likely that most published research findings are false. Some of these assume that 'p-hacked studies', or studies that show a significant result due to bias, enter the literature. There are good reasons to believe this happens, as we discussed in the section on flexibility in the data analysis in this course. In the 'presets by Ioannidis' dropdown menu, you can select some of these situations. Explore all of them, and pay close attention to the ones where the PPV is smaller than 50%.

Q4: In general, when are most published findings false? Interpret 'low' and 'high' in the answer options below in relation to the values in the first example in this assignment of 50% probability H1 is true, 5% alpha, 80% power, and 0% bias.

A) When the probability of examining a true hypothesis is low, combined with either low power or substantial bias (e.g., p-hacking).

B) When the probability of examining a true hypothesis is high, combined with either low power or substantial bias (e.g., p-hacking).

C) When the alpha level is high, combined with either low power or substantial bias (e.g., p-hacking).

D) When power is low and p-hacking is high (regardless of the % of true hypotheses one examines).

Q5: Set the "% of a priori true hypotheses" slider to 0%. Set the "% of p-hacked studies" slider to 0%. Set the "α level" slider to 5%. Play around with the power slider. Which statement is true? Without *p*-hacking, when the alpha level is 5%, and when 0% of the hypotheses are true, ____

A) the proportion of false positives for all experiments we have performed is 100%.

B) the PPV depends on the power of the studies.

C) regardless of the power, the PPV equals the proportion of false positives for all experiments we have performed.

D) regardless of the power, the proportion of false positives for all experiments we have performed is 5%, and the PPV is 0% (all significant results are false positives).

Conclusion

People often say something like: "*Well, we all know 1 in 20 results in the published literature are Type 1 errors*". After this assignment, you should be able to understand this is not true in practice. Even when we use a 5% alpha level, it is quite reasonable to assume much more than 5% of significant findings in the published literature are false positives. When in 100% of the studies you perform, the null hypothesis is true, and all studies are published, only *then* 1 in 20 studies, in the long run, are false positives (and the rest correctly reveal no statistically significant difference). In the scientific literature, the positive predictive value (the probability that given that a statistically significant result is observed, the effect is true) can be quite high, and under specific circumstances, it might even be so high that most published research findings are false. This will happen when researchers examine mostly studies where the null-hypothesis is true, with low power, or when the Type 1 error rate is inflated due to *p*-hacking or other types of bias. Publication bias, power, and Type 1 error rates together determine the probability that significant results in the literature reflect true effects.