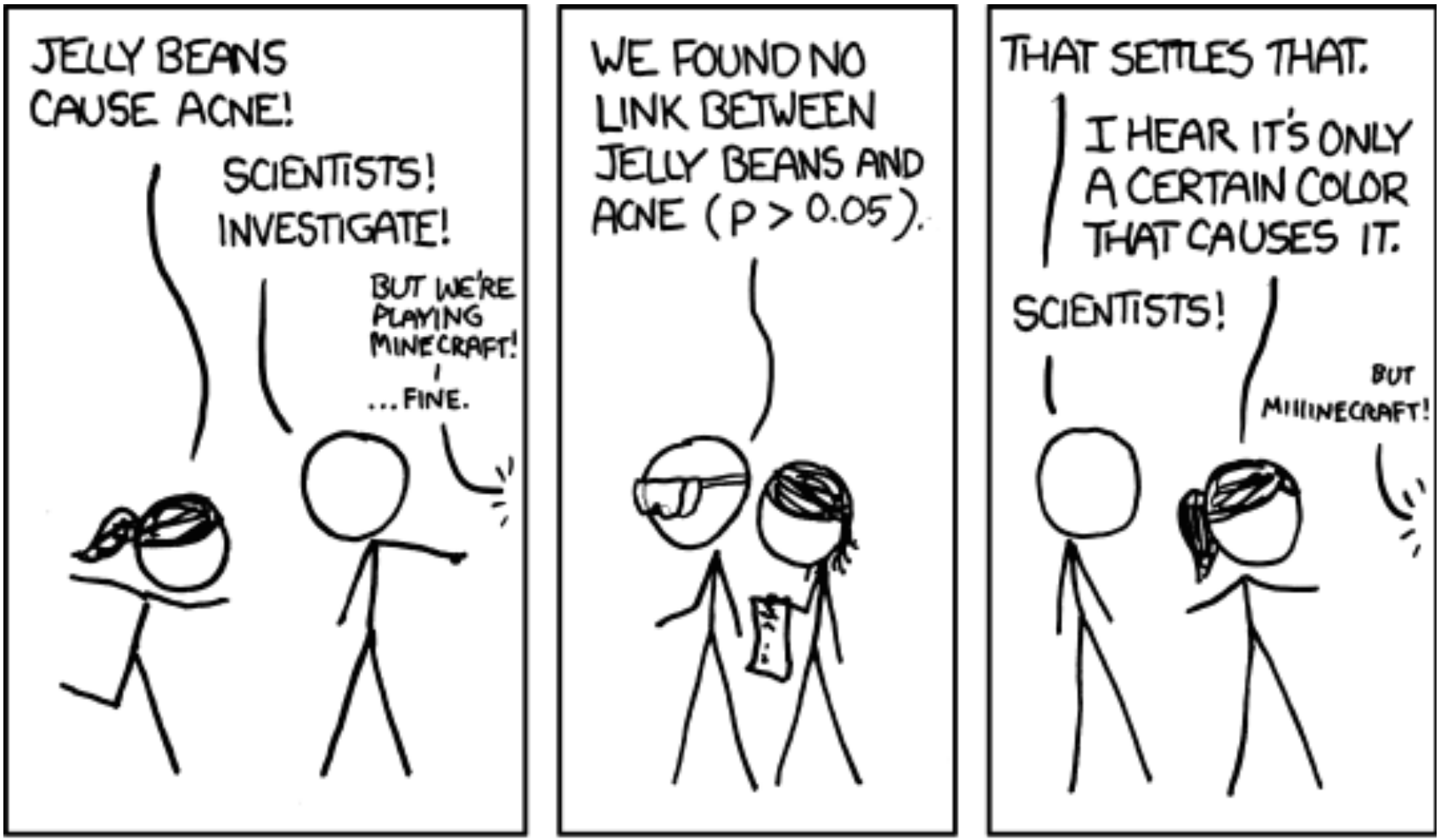
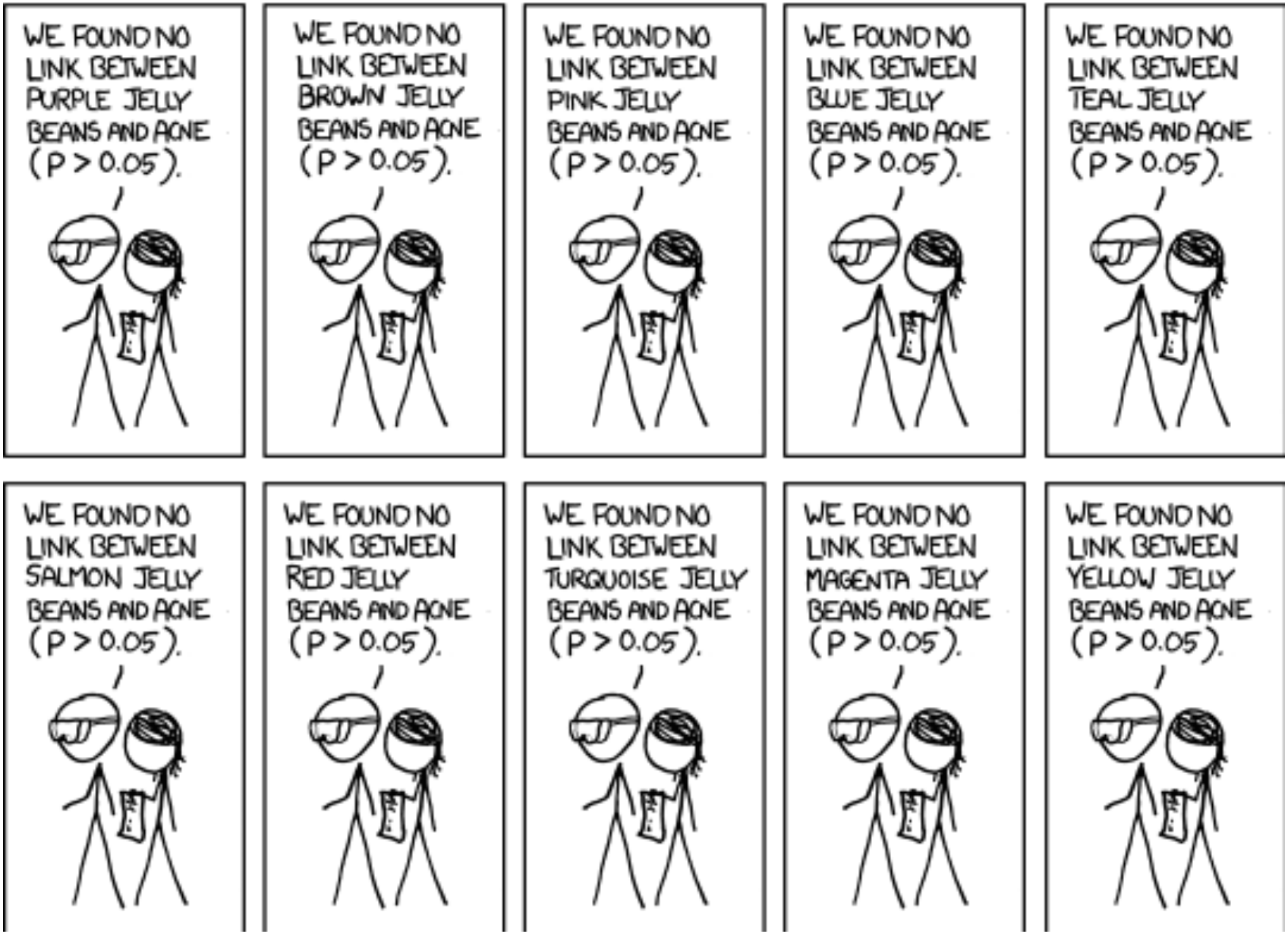
S&DS 242/542: Theory of Statistics

Lecture 10: Testing multiple hypotheses

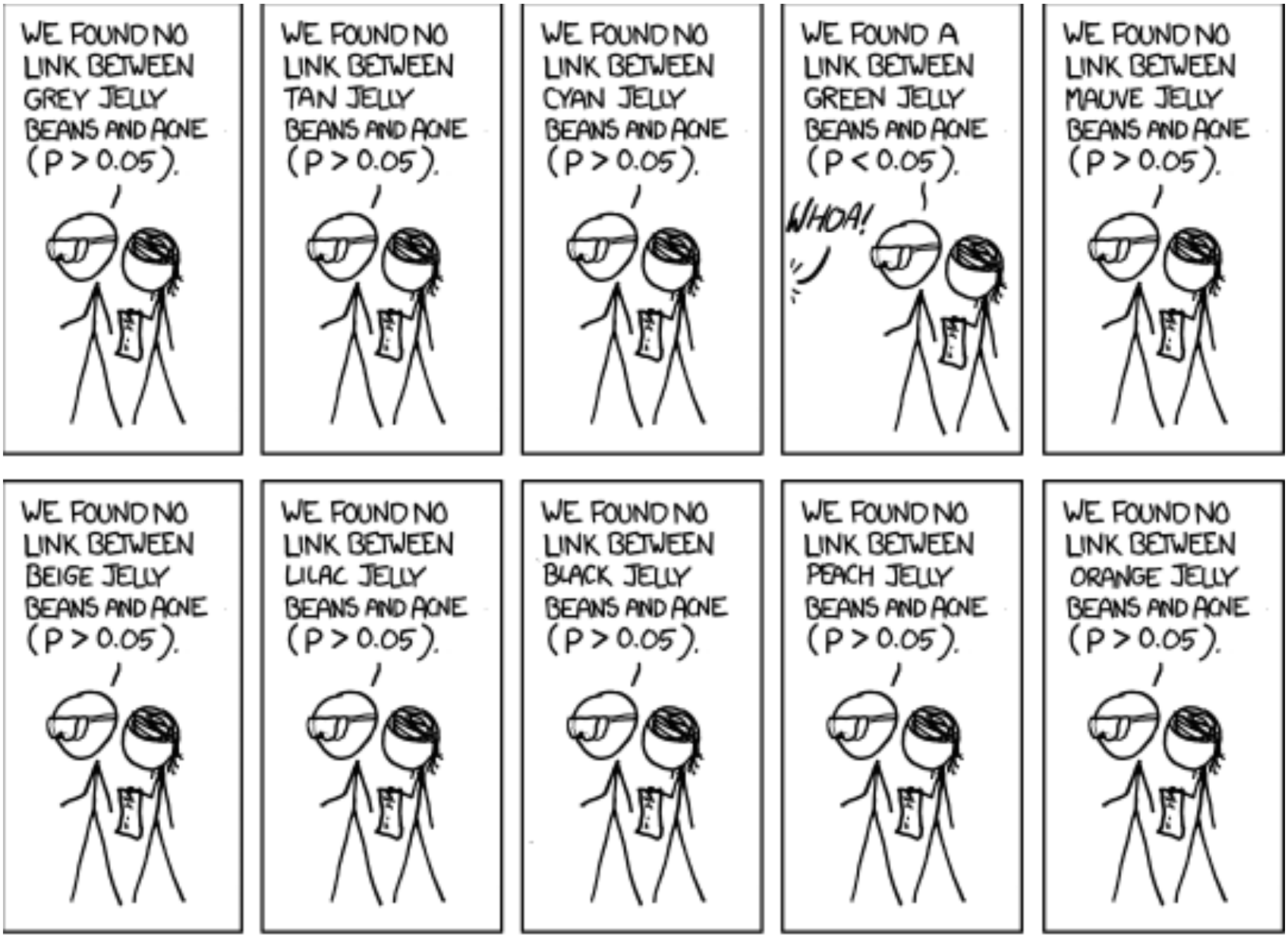
The multiple testing problem



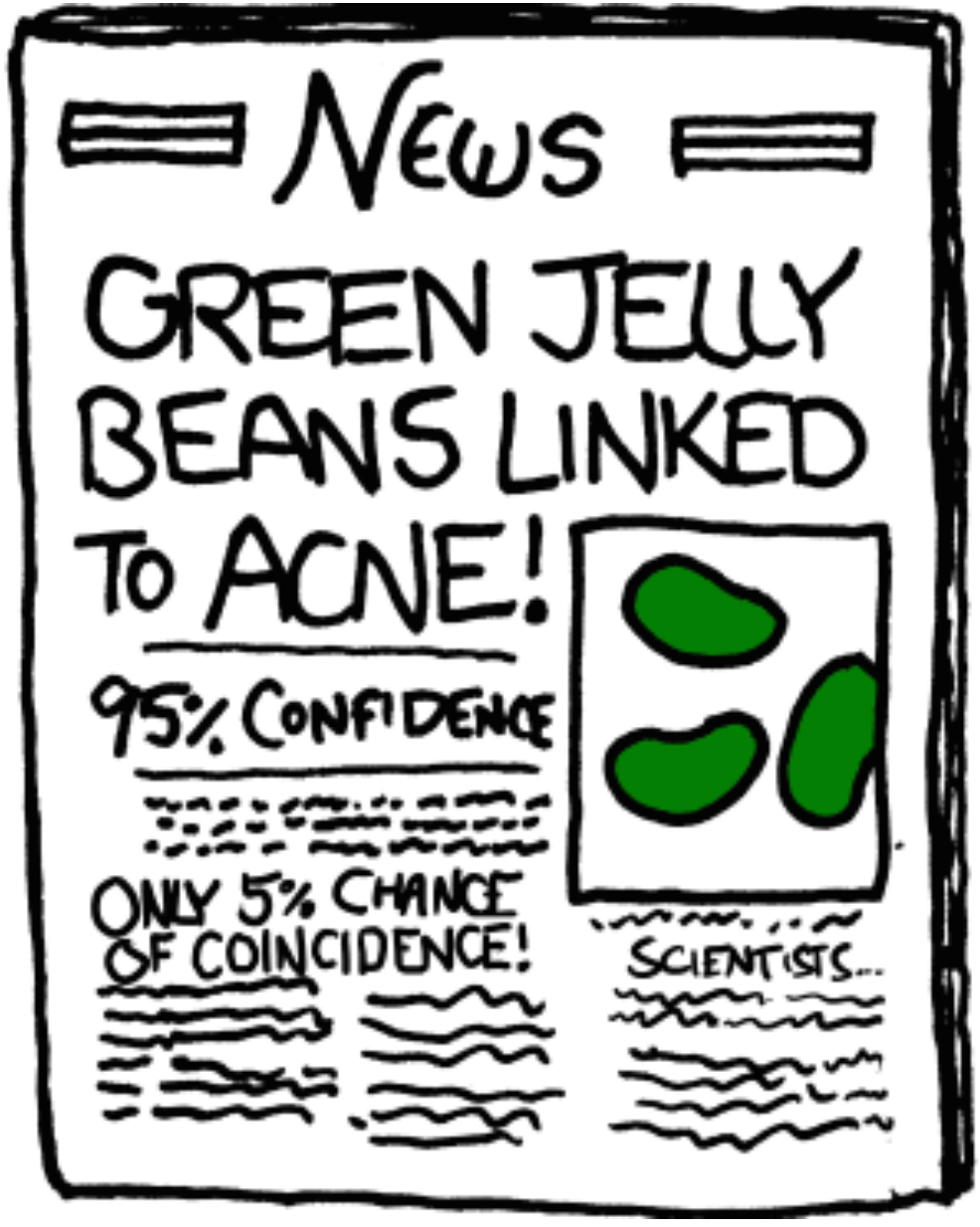
The multiple testing problem



The multiple testing problem



The multiple testing problem



The multiple testing problem

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1 - \beta)R/(R$

The multiple testing problem

"There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships..."

—John P.A. Ioannidis

The multiple testing problem

Multiple testing problem: If I test n null hypotheses at level α, all of which are true, then on average I'll (wrongly) reject αn of them.

Examples:

- Test the safety of a drug in terms of many different side effects
- Test whether a disease is associated to 1,000,000 different genetic markers

What are some aggregate notions of statistical signficance and Type I error across multiple hypothesis tests and experiments?

What statistical procedures can we use to control these aggregate measures of error?

Thinking in terms of p-values

For today, we will think of each individual hypothesis test as returning a p-value, and compare/combine statistical significance using these p-values.

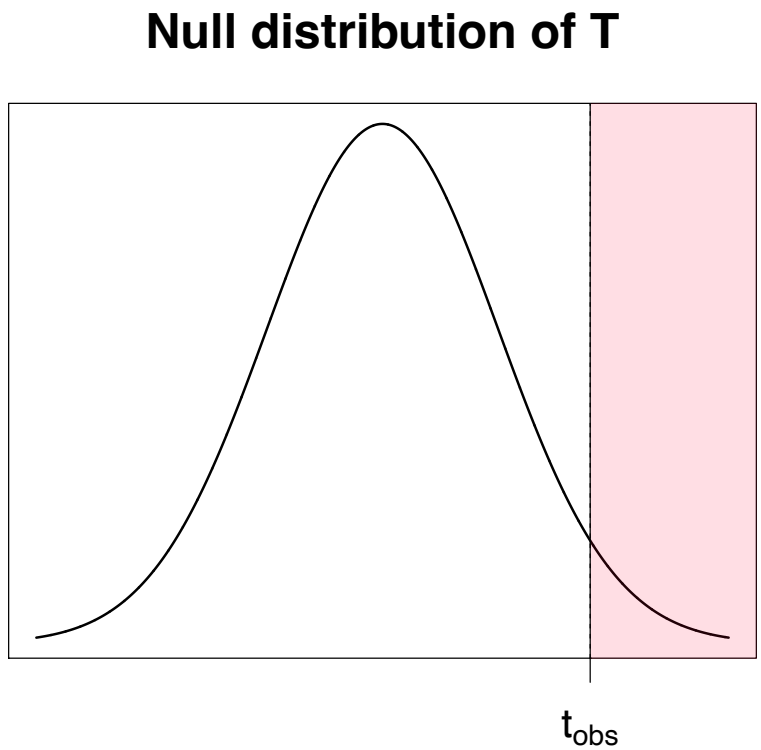
Most statistical multiple-testing procedures take as input these p-values, rather than the original data or the original test statistic used in each experiment.

Advantages:

- Abstracts away details about individual tests
- Allows different experiments to use different test statistics
- Allows for meta-analysis of previous results without needing access to the original data

P-values for one-sided tests

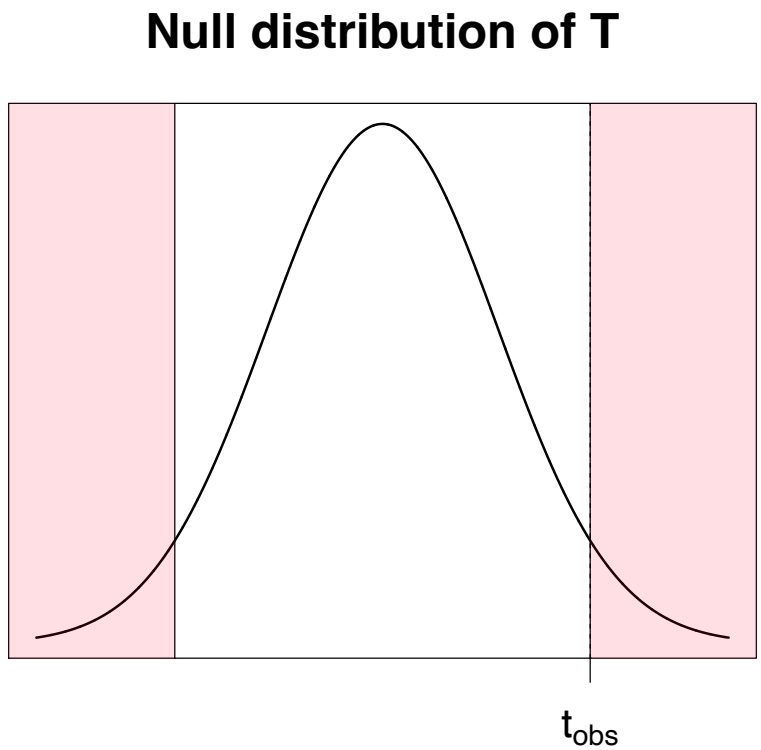
Recall: For a statistical test, the **p-value** is the smallest significance level at which the test rejects the null hypothesis.



For a one-sided test rejecting H_0 for large T, the p-value is the right tail probability at the observed value of T.

P-values for two-sided tests

Recall: For a statistical test, the **p-value** is the smallest significance level at which the test rejects the null hypothesis.



For a two-sided test, the p-value is the sum of left and right tail probabilities when the boundary of the rejection region is at T.

P-values as transformed test statistics

Let $F(t)$ be the CDF of the null distribution of T. For a one-sided test rejecting H_0 for large T, the p-value is

$$P = 1 - F(T)$$

If the null distribution of T is symmetric around 0, for a two-sided test rejecting H_0 for large $|T|$, the p-value is

$$P = [1 - F(|T|)] + F(-|T|)$$

The p-value is a transformation of different test statistics to a common $[0, 1]$ scale, summarizing the amount of statistical evidence against H_0 in a way that admits a common interpretation.

By definition, the test rejects H_0 at significance levels $\alpha \geq P$ and accepts H_0 at significance levels $\alpha < P$. So the rejection event of the test at any fixed significance level $\alpha \in (0, 1)$ is $P \leq \alpha$.

The null distribution of the p-value

Suppose, for each $\alpha \in (0, 1)$, our test at significance level α has Type I error probability exactly equal to α:

$$\alpha = \mathbb{P}_{H_0}[\text{reject } H_0]$$

(This is usually the case for continuous test statistics T.) Since the test rejects H_0 when $P \leq \alpha$, this means

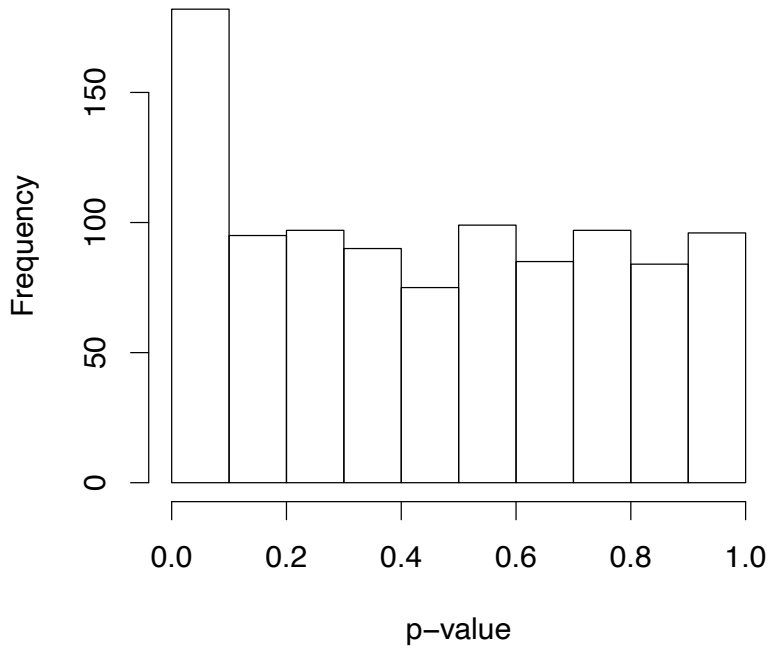
$$\alpha = \mathbb{P}_{H_0}[P \leq \alpha]$$

This holds for every $\alpha \in (0, 1)$, so $P \sim \text{Uniform}(0, 1)$ under H_0.

[One may also verify this by computing the CDF of P from the expression $P = 1 - F(T)$ or $P = [1 - F(|T|)] + F(-|T|)$ in the preceding one-sided and two-sided testing examples.]

P-values across multiple tests

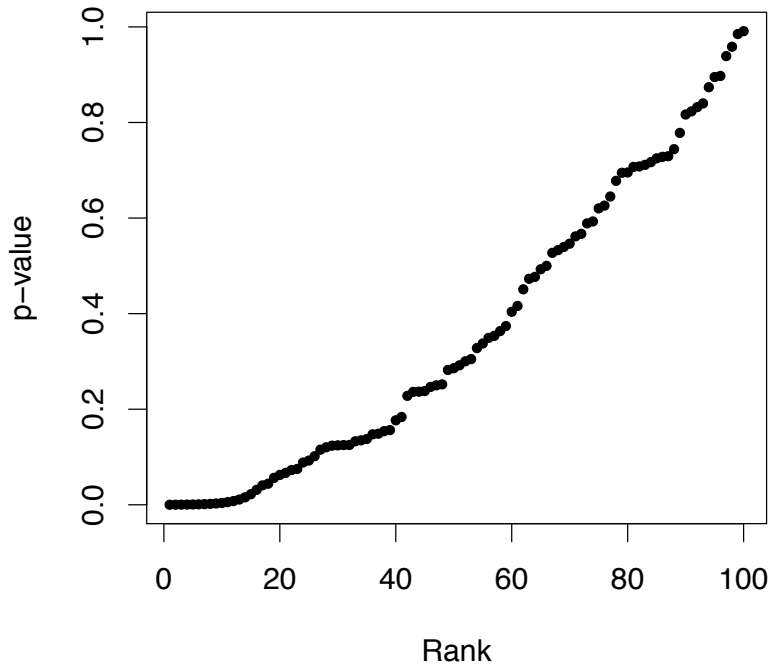
A typical histogram of p-values obtained across many different hypothesis tests may look like the following.



Most tested null hypotheses may be true nulls, and their p-values are uniformly distributed on $[0, 1]$. A small subset of tested null hypotheses may be false nulls, and their p-values are closer to 0.

Ordered p-value plots

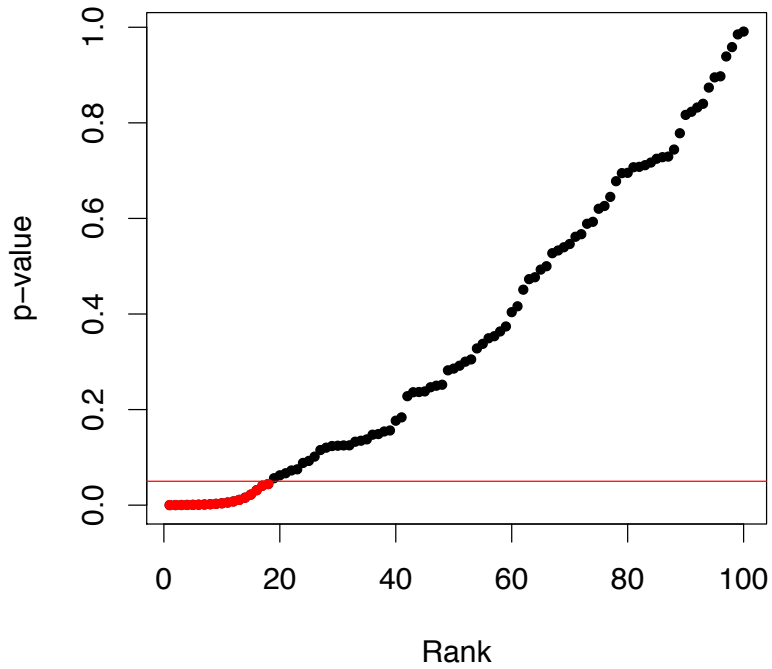
We may also visualize the p-values across many tests by sorting them and plotting them in rank order:



If all tested null hypotheses were true nulls, we should see a diagonal line. Here instead, there seem to be many p-values close to 0, suggesting the presence of some false null hypotheses.

Rejected and accepted null hypotheses

Suppose we apply each hypothesis test at the significance level $\alpha = 0.05$. Then we would reject H_0 in the tests that yielded the 18 red p-values below, and accept H_0 in the remaining 82 tests.



We might suspect that the rejected null hypotheses H_0 with extremely small p-values are correctly rejected, but some of those with p-value closer to the $\alpha = 0.05$ cutoff are incorrectly rejected.

The Bonferroni correction

The simplest multiple-testing correction is the **Bonferroni method**: When testing n different null hypotheses, perform each test at the significance level α/n instead of α.

Justification: Suppose that all n null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$ are in fact true nulls. The Bonferroni method ensures that

$$\begin{aligned}
&\mathbb{P}[\text{ reject } \textit{any} \text{ null hypothesis }] \\
&= \mathbb{P}\left[\{\text{reject } H_0^{(1)}\} \cup \ldots \cup \{\text{reject } H_0^{(n)}\}\right] \\
&\leq \mathbb{P}\left[\text{reject } H_0^{(1)}\right] + \ldots + \mathbb{P}\left[\text{reject } H_0^{(n)}\right] \\
&\leq \frac{\alpha}{n} + \ldots + \frac{\alpha}{n} = \alpha
\end{aligned}$$

The last line holds because each hypothesis is rejected with probability at most α/n, under a test with significance level α/n.

Family-wise error rate

More generally, suppose we test n null hypotheses, n_0 of which are true nulls and $n - n_0$ of which are false nulls. (Each null hypothesis is either true or false — this is unknown to us, but not random.)

The **family-wise error rate (FWER)** is the probability that we reject at least one of the n_0 true null hypotheses:

$$\text{FWER} = \mathbb{P}[\text{ reject any } \textit{true} \text{ null hypothesis }]$$

This is not affected by our decision for the false nulls. If all n tested null hypotheses are false (so $n_0 = 0$), then FWER is trivially 0.

A procedure controls FWER at level α if FWER $\leq \alpha$, regardless of how many (and which) null hypotheses are true and false.

FWER of the Bonferroni method

The Bonferroni method controls FWER at level α.

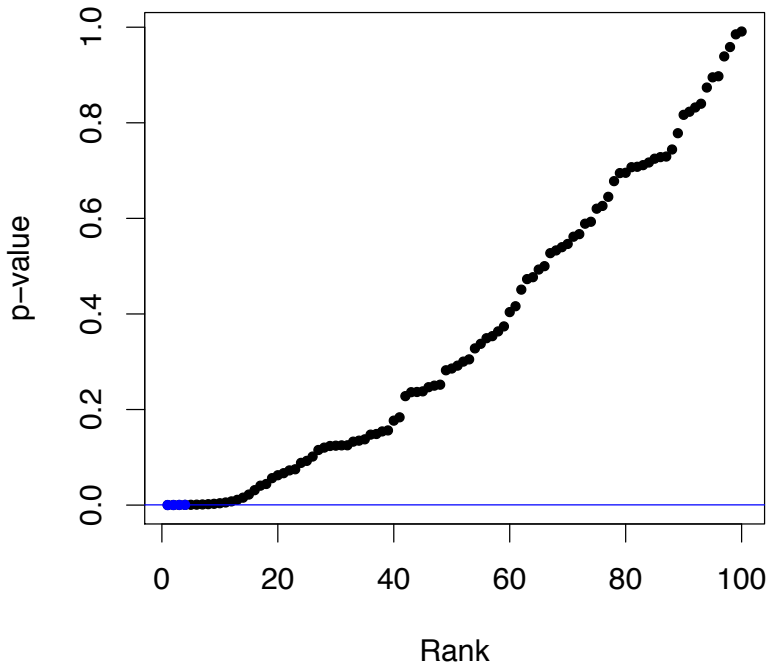
Justification: Suppose now that $H_0^{(1)}, \ldots, H_0^{(n_0)}$ are true, and $H_0^{(n_0+1)}, \ldots, H_0^{(n)}$ are false. Then

$$\begin{aligned}
\text{FWER} &= \mathbb{P}[\text{ reject any } \textit{true} \text{ null hypothesis }] \\
&= \mathbb{P}\left[\{\text{reject } H_0^{(1)}\} \cup \ldots \cup \{\text{reject } H_0^{(n_0)}\}\right] \\
&\leq \mathbb{P}\left[\text{reject } H_0^{(1)}\right] + \ldots + \mathbb{P}\left[\text{reject } H_0^{(n_0)}\right] \\
&= \underbrace{\frac{\alpha}{n} + \ldots + \frac{\alpha}{n}}_{n_0 \text{ times}} = \frac{\alpha n_0}{n} \leq \alpha.
\end{aligned}$$

If we knew the number of true null hypotheses n_0, we could do each individual test at level α/n_0. But we usually don't know n_0, and often n_0 is close to n, so we use the conservative level α/n.

Rejected and accepted null hypotheses

Applying the Bonferroni method to control FWER ≤ 0.05 across 100 tests, we reject the 4 null hypotheses below with p-value less than 0.0005, instead of the previous 18.



FWER is controlled, but we have sacrificed testing power and may be accepting many null hypotheses H_0 which are actually false.

False discovery proportion

In certain applications, we may be tolerant of making a few Type I errors, provided that the proportion of Type I errors among all rejected null hypotheses — the *false discovery proportion (FDP)* — is not too high.

Example: We test 1,000,000 genetic markers, and identify 1,000 of them as associated to a disease. (The null hypothesis H_0 for each marker is that there is no association.) Of these, 950 are truly associated to the disease, and 50 are not. Then our false discovery proportion is

$$\text{FDP} = \frac{50}{1000} = 5\%$$

False discovery rate

Let

V = number of true null hypotheses rejected ("false discoveries")

R = number of total null hypotheses rejected ("total discoveries")

so FDP $= V/R$.

Here, V and R are random quantities depending on the data of each individual hypothesis test. The **false discovery rate** is

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right] = \mathbb{E}\left[\frac{V}{R}\right]$$

with the convention that FDP $= V/R = 0$ if $V = R = 0$.

A procedure controls FDR at level α if FDR $\leq \alpha$, regardless of how many (and which) null hypotheses are true and false.

FWER vs. FDR

Controlling FWER may be appropriate if

- There is a more severe consequence for committing even a single Type I error
- The result of the statistical test is going to be interpreted as a definitive answer for whether the discovery is true

In contrast, controlling FDR may be appropriate if

- The statistical test identifies candidate discoveries out of a large pool, which are then going to be subject to further study
- There is some cost associated to false discoveries, but this is acceptable as long as most of our discoveries are correct

The Benjamini-Hochberg procedure

Suppose, for n hypothesis tests, we observe the final outcome of the tests: Null hypotheses with p-values $\leq t$ were rejected, and those with p-values $> t$ were accepted. Can we estimate the FDP?

Recall FDP $= V/R$. We observe R, the total number of rejections. We don't know which are true and false, so we don't know V.

However, we can estimate V: Recall that p-values corresponding to true null hypotheses have distribution Uniform$(0, 1)$. So for n_0 true nulls, we expect roughly tn_0 of these to have p-value $\leq t$. That is, $V \approx tn_0$, and FDP $\approx tn_0/R$.

We usually don't know n_0. A slightly conservative estimate of FDP (erring on the side of being too large) is

$$\widehat{\text{FDP}} = tn/R$$

The Benjamini-Hochberg procedure

Idea: To control FDR at level α, pick the largest cutoff t such that

$$\widehat{\text{FDP}} = \frac{tn}{R(t)} \leq \alpha$$

Here $R(t)$ is the number of rejected hypotheses using this cutoff t, i.e. the total number of p-values $\leq t$.

Equivalently: Suppose we reject r null hypotheses. Then the cutoff is $t = P_{(r)}$, the r^{th} smallest p-value. Pick the largest r such that

$$\frac{P_{(r)} \cdot n}{r} \leq \alpha \qquad \Longleftrightarrow \qquad P_{(r)} \leq \frac{\alpha r}{n}$$

This is the **Benjamini-Hochberg (BH) procedure**.

The Benjamini-Hochberg procedure

More precisely, the BH procedure at level α is performed as follows:

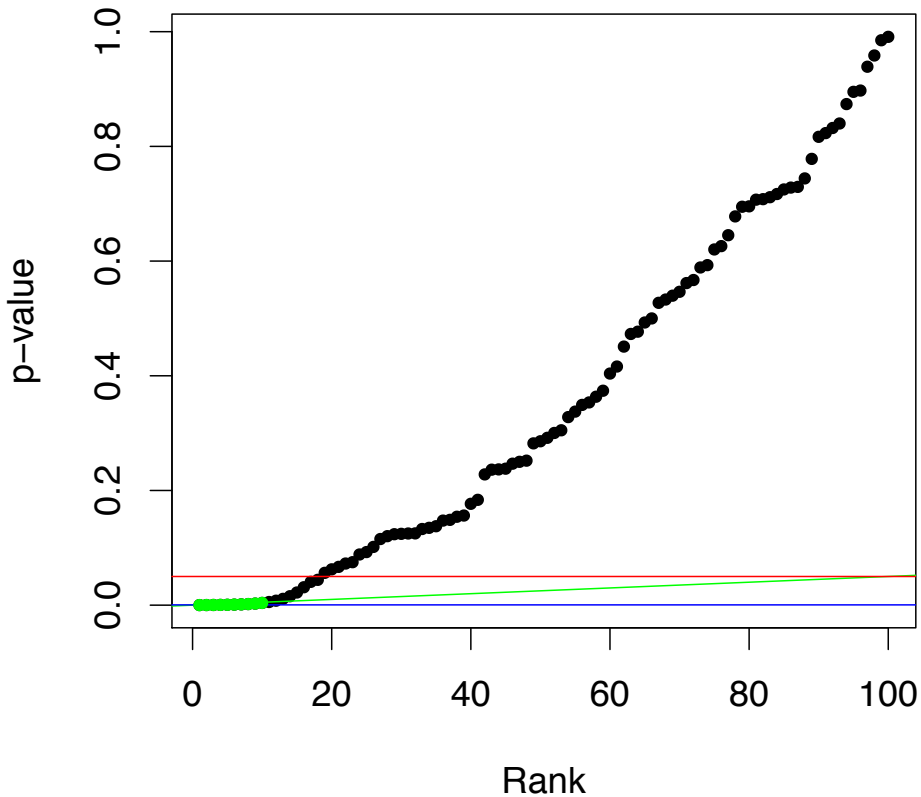
1. Sort the n total p-values from smallest to largest. Denote these by $P_{(1)} \leq \ldots \leq P_{(n)}$.
2. Find the largest r such that $P_{(r)} \leq \frac{\alpha r}{n}$.
3. Reject the null hypotheses corresponding to $P_{(1)}, \ldots, P_{(r)}$.

The smallest p-value $P_{(1)}$ is compared to the Bonferroni level, α/n. However, the next smallest p-value $P_{(2)}$ is compared to $2\alpha/n$, then $P_{(3)}$ to $3\alpha/n$, etc.

If some p-values are extremely small, then there is strong evidence that these null hypotheses are false. It is then allowable to reject a few true nulls and still control the FDR, so the BH procedure uses a more lenient threshold for the remaining p-values.

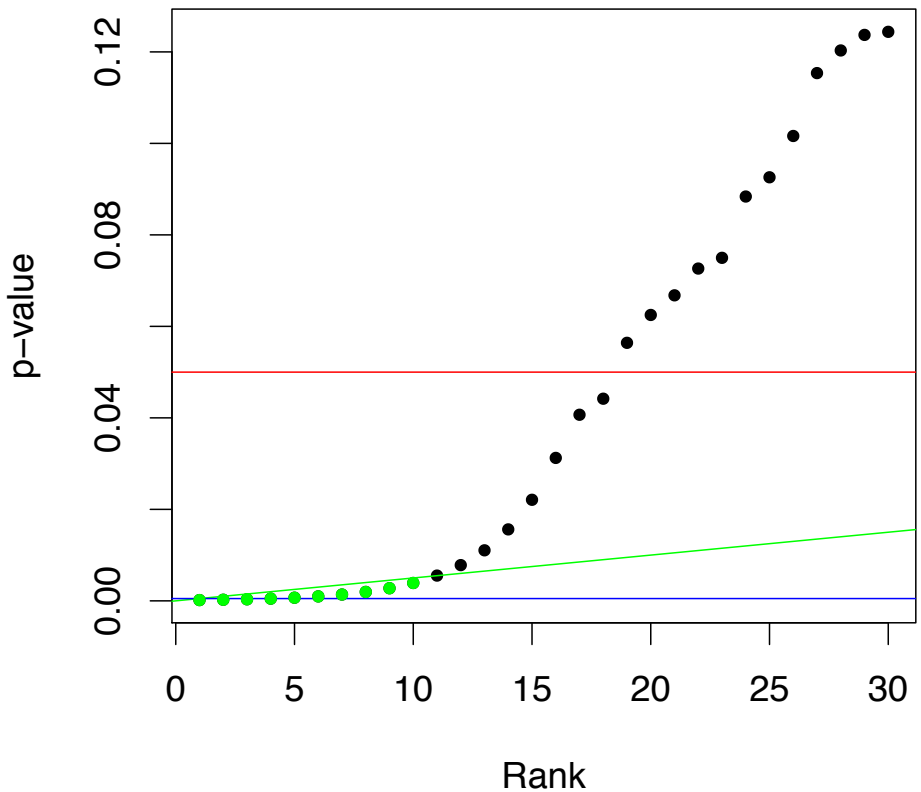
The Benjamini-Hochberg procedure

The BH procedure compares the sorted p-values to a *diagonal* cutoff line $P_{(r)} = \alpha r/n$. This line is equal to the Bonferroni level α/n at $r = 1$ and to the uncorrected level α at $r = n$.



Rejected and accepted null hypotheses

In this example, the BH procedure applied at level $\alpha = 0.05$ rejects 10 null hypotheses, in green. Recall that Bonferroni rejected 4, while naively testing each hypothesis at level $\alpha = 0.05$ rejected 18.



Guarantee for FDR control

Theorem (Benjamini and Hochberg (1995))

Consider tests of n null hypotheses, n_0 of which are true. If the n p-values are independent, then the false discovery rate of the BH procedure applied at level α satisfies

$$\mathrm{FDR} \leq \frac{n_0 \alpha}{n} \leq \alpha$$

- The p-values are independent if the data from the n experiments are independent.
- There are some conditions of *positive dependence* where the BH procedure still controls FDR at level α.
- There are also counterexamples where p-values are dependent and FDR is not controlled at level α. In the worst case, for n hypotheses, the FDR is controlled at level $\alpha(1 + \frac{1}{2} + \ldots + \frac{1}{n})$.