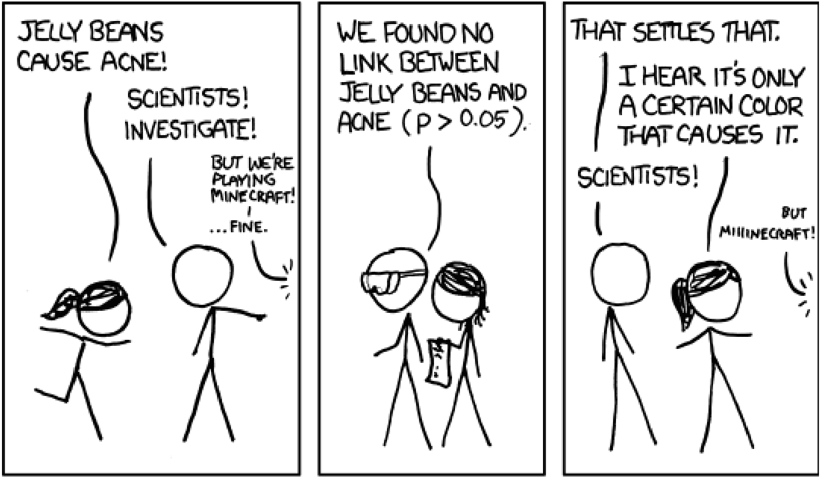
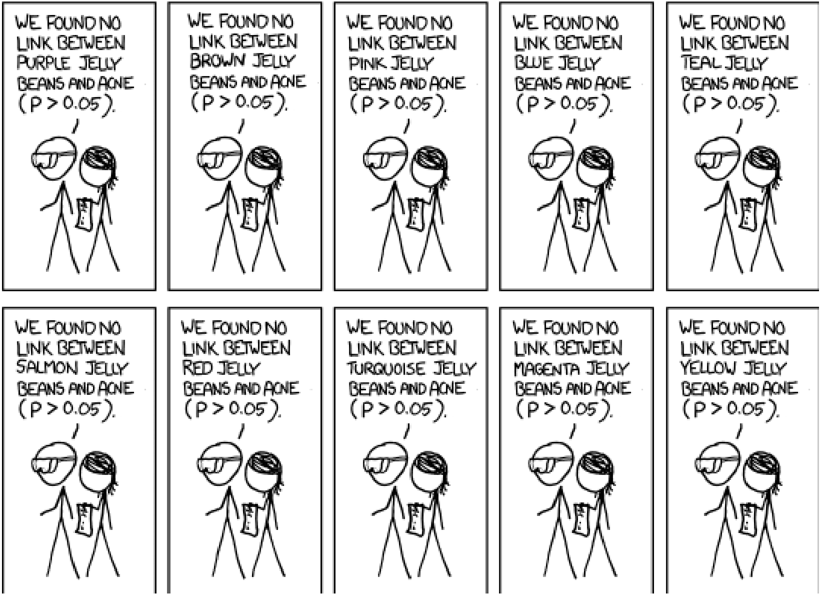
STATS 200: Introduction to Statistical Inference

Lecture 11: Testing multiple hypotheses

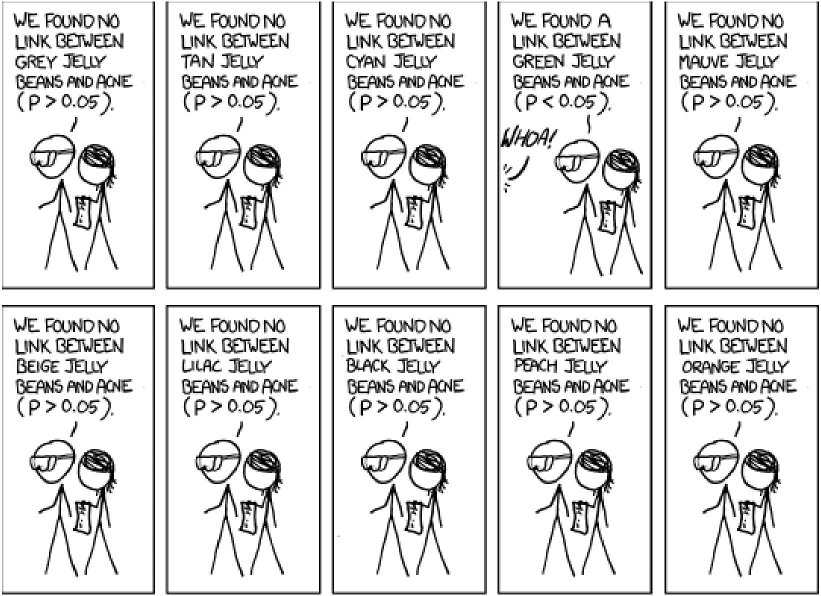
The multiple testing problem


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!

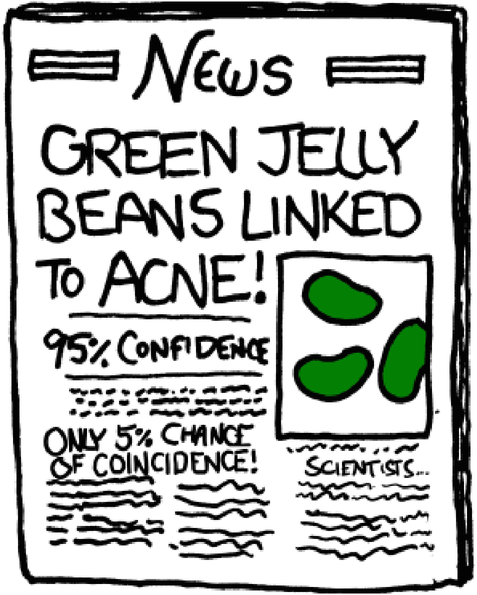
The multiple testing problem


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).

The multiple testing problem

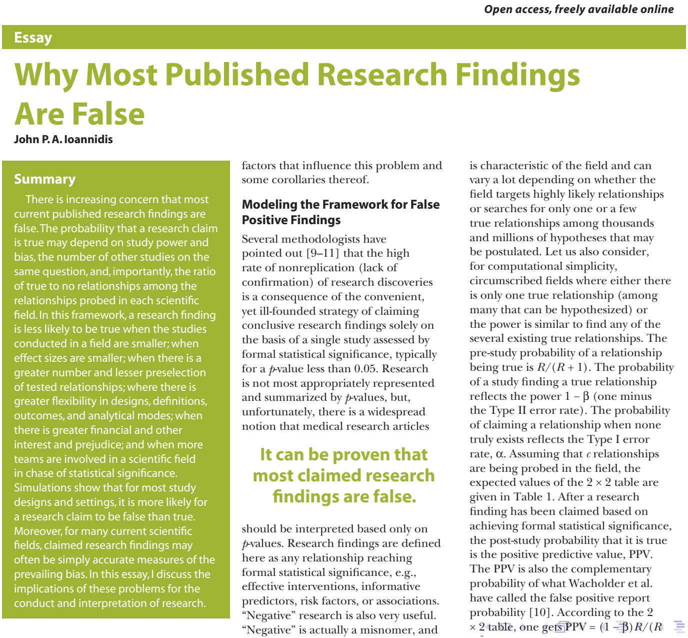

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P < 0.05$).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY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

The multiple testing problem



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

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

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

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 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

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

Examples:

- Test the safety of a drug in terms of a dozen different side effects
- Test whether a disease is related to 10,000 different gene expressions

The multiple testing problem

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

Examples:

- Test the safety of a drug in terms of a dozen different side effects
- Test whether a disease is related to 10,000 different gene expressions

What are some ways we can think about acceptance/rejection errors across multiple hypothesis tests/experiments?

What statistical procedures can control these measures of errors?

The Bonferroni correction

Consider testing n different null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$, all of which are, in fact, true. One goal we might set is to ensure

$$\mathbb{P}[\text{ reject } \textit{any} \text{ null hypothesis }] \leq \alpha.$$

A simple and commonly-used method of achieving this is called the **Bonferroni** method: Perform each test at significance level α/n, instead of level α.

The Bonferroni correction

Consider testing n different null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$, all of which are, in fact, true. One goal we might set is to ensure

$$\mathbb{P}[\text{ reject } \textit{any} \text{ null hypothesis }] \leq \alpha.$$

A simple and commonly-used method of achieving this is called the **Bonferroni** method: Perform each test at significance level α/n, instead of level α. Verification:

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

The Bonferroni correction

Consider testing n different null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$, all of which are, in fact, true. One goal we might set is to ensure

$$\mathbb{P}[\text{ reject } \textit{any} \text{ null hypothesis }] \leq \alpha.$$

A simple and commonly-used method of achieving this is called the **Bonferroni** method: Perform each test at significance level α/n, instead of level α. Verification:

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

Family-wise error rate

More generally, suppose we test n null hypotheses, n_0 of which are true and $n - n_0$ of which are false. Results of the tests might be tabulated as follows:

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n - R$
Total	n_0	$n - n_0$	n

$R = \#$ rejected null hypotheses
$V = \#$ type I errors, $T = \#$ type II errors

Remark: We consider n_0 and $n - n_0$ to be fixed quantities. The number of hypotheses we reject, R, as well as the cell counts U, V, S, T, are random, as they depend on the data observed in each experiment.

Family-wise error rate

The **family-wise error rate (FWER)** is the probability of falsely rejecting at least one true null hypothesis,

$$\mathbb{P}[V \geq 1].$$

A procedure controls FWER at level α if $\mathbb{P}[V \geq 1] \leq \alpha$, regardless of the (possibly unknown) number of true null hypotheses n_0.

Family-wise error rate

The **family-wise error rate (FWER)** is the probability of falsely rejecting at least one true null hypothesis,

$$\mathbb{P}[V \geq 1].$$

A procedure controls FWER at level α if $\mathbb{P}[V \geq 1] \leq \alpha$, regardless of the (possibly unknown) number of true null hypotheses n_0.

Bonferroni controls FWER: Without loss of generality, let $H_0^{(1)}, \ldots, H_0^{(n_0)}$ be the true null hypotheses.

$$\begin{aligned}\mathbb{P}[V \geq 1] &= \mathbb{P}\left[\{\text{reject } H_0^{(1)}\} \cup \ldots \cup \{\text{reject } H_0^{(n_0)}\}\right] \\ &\leq \mathbb{P}\left[\text{reject } H_0^{(1)}\right] + \ldots + \mathbb{P}\left[\text{reject } H_0^{(n_0)}\right] \\ &= \frac{\alpha}{n} + \ldots + \frac{\alpha}{n} = \frac{\alpha n_0}{n} \leq \alpha.\end{aligned}$$

Thinking in terms of p-values

Many multiple-testing procedures are formulated as operating on the p-values returned by individual tests, rather than on the original data or the test statistics that were used.

For example, Bonferroni may be described as follows: Reject those null hypotheses whose corresponding p-values are at most α/n.

Thinking in terms of p-values

Many multiple-testing procedures are formulated as operating on the p-values returned by individual tests, rather than on the original data or the test statistics that were used.

For example, Bonferroni may be described as follows: Reject those null hypotheses whose corresponding p-values are at most α/n.

Advantages:

- Abstracts away details about how individual tests were performed
- Applicable regardless of which tests/test statistics were used for each experiment
- Allows for meta-analyses of previous experiments without access to the original data

The null distribution of a *p*-value

Suppose a null hypothesis H_0 is true, and we perform a statistical test of H_0 and obtain a *p*-value P. What is the distribution of P?

The null distribution of a *p*-value

Suppose a null hypothesis H_0 is true, and we perform a statistical test of H_0 and obtain a *p*-value P. What is the distribution of P?

If our test statistic T has a continuous distribution under H_0 with CDF F, and we reject for small values of T, then the *p*-value is just the lower tail probability

$$P = F(T).$$

The null distribution of a *p*-value

Suppose a null hypothesis H_0 is true, and we perform a statistical test of H_0 and obtain a *p*-value P. What is the distribution of P?

If our test statistic T has a continuous distribution under H_0 with CDF F, and we reject for small values of T, then the *p*-value is just the lower tail probability

$$P = F(T).$$

For any $t \in (0,1)$,

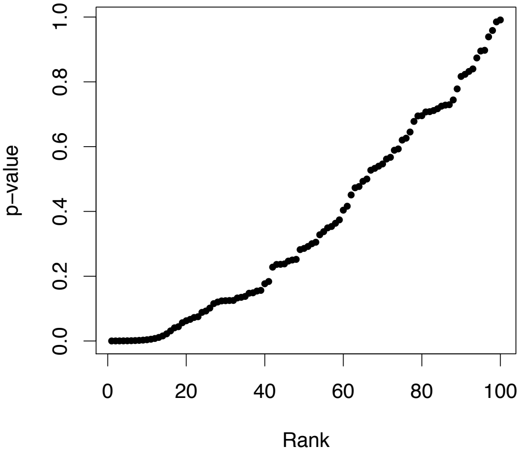
$$\mathbb{P}[P \leq t] = \mathbb{P}[F(T) \leq t] = \mathbb{P}[T \leq F^{-1}(t)] = F(F^{-1}(t)) = t.$$

So $P \sim \text{Uniform}(0,1)$. Similarly $P \sim \text{Uniform}(0,1)$ if we reject for large T, or both large and small T.*

*If T has a discrete distribution under H_0, then so does P, so the null distribution of P wouldn't be exactly $\text{Uniform}(0,1)$.

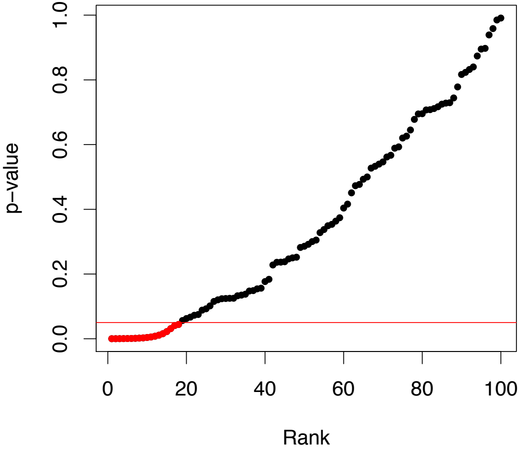
Ordered *p*-value plots

We can understand multiple testing procedures visually in terms of the plot of the ordered *p*-values (sorted from smallest to largest):



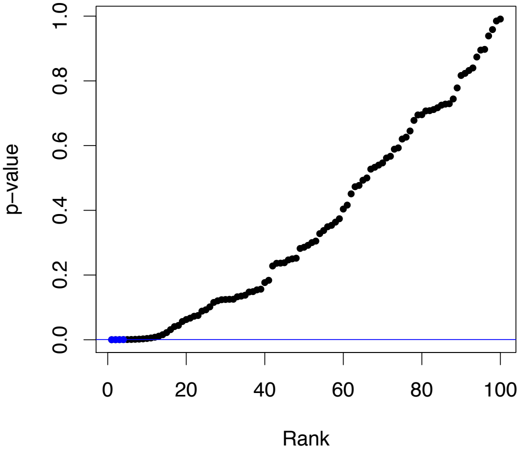
Ordered *p*-value plots

Applying each test at level 0.05, we reject the null hypotheses corresponding to the below 18 red points.



Ordered *p*-value plots

Applying the Bonferroni correction, we reject null hypotheses with *p*-value less than 0.0005, corresponding to the below 4 blue points.



False discovery rate

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n-R$
Total	n_0	$n-n_0$	n

Controlling the FWER $\mathbb{P}[V \geq 1]$ may be too conservative and greatly reduce our power to detect real effects, especially when n (the total number of tested hypotheses) is large.

False discovery rate

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n-R$
Total	n_0	$n-n_0$	n

Controlling the FWER $\mathbb{P}[V \geq 1]$ may be too conservative and greatly reduce our power to detect real effects, especially when n (the total number of tested hypotheses) is large.

In many modern "large-scale testing" applications, focus has shifted to the **false-discovery proportion (FDP)**

$$\mathrm{FDP} = \begin{cases} \frac{V}{R} & R \geq 1 \\ 0 & R = 0, \end{cases}$$

and on procedures that control its expected value $\mathbb{E}[\mathrm{FDP}]$, called the **false-discovery rate (FDR)**.

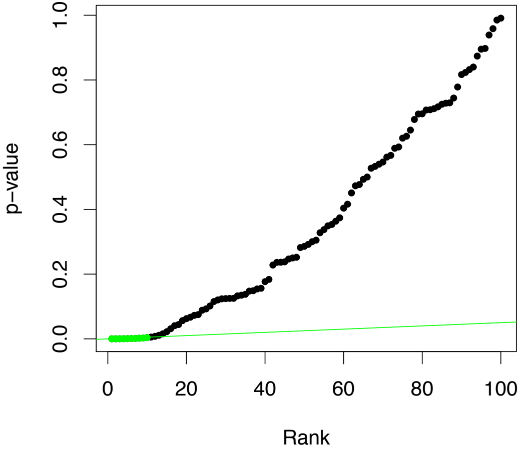
FWER vs. FDR

Controlling FDR is a shift in paradigm—we are willing to tolerate some type I errors (false discoveries), as long as most of the discoveries we make are still true.

It has been argued that in applications where the statistical test is thought of as providing a "definitive answer" for whether an effect is real, FWER control is still the correct objective. In contrast, for applications where the statistical test identifies candidate effects that are likely to be real and which merit further study, it may be better to target FDR control.

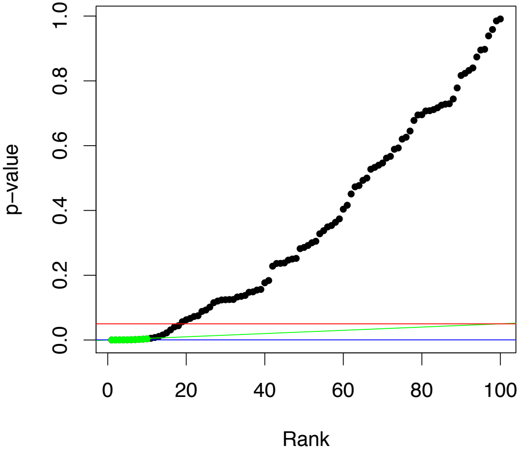
The Benjamini-Hochberg procedure

The **Benjamini-Hochberg (BH)** procedure compares the sorted *p*-values to a diagonal cutoff line, finds the largest *p*-value that still falls below this line, and rejects the null hypotheses for the *p*-values up to and including this one.



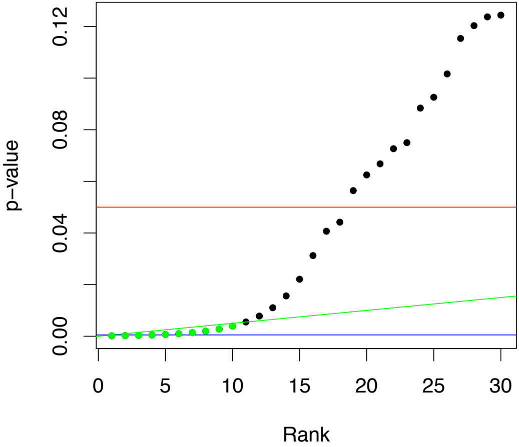
The Benjamini-Hochberg procedure

To control FDR at level q, the diagonal cutoff line is set to equal the Bonferroni level q/n at the smallest p-value and to equal the uncorrected level q at the largest p-value.



The Benjamini-Hochberg procedure

Here's the same picture, zoomed in to the 30 smallest *p*-values. In this example, the BH procedure rejects the 10 null hypotheses corresponding to the points in green.



The Benjamini-Hochberg procedure

Formally, the BH procedure at level q is defined as follows:

1. Sort the p-values. Call them $P_{(1)} \leq \ldots \leq P_{(n)}$.
2. Find the largest r such that $P_{(r)} \leq \frac{qr}{n}$.
3. Reject the null hypotheses $H_{(1)}, \ldots, H_{(r)}$.

The Benjamini-Hochberg procedure

Formally, the BH procedure at level q is defined as follows:

1. Sort the p-values. Call them $P_{(1)} \leq \ldots \leq P_{(n)}$.
2. Find the largest r such that $P_{(r)} \leq \frac{qr}{n}$.
3. Reject the null hypotheses $H_{(1)}, \ldots, H_{(r)}$.

Theorem (Benjamini and Hochberg (1995))

Consider tests of n null hypotheses, n_0 of which are true. If the test statistics (or equivalently, p-values) of these tests are independent, then the FDR of the above procedure satisfies

$$\mathrm{FDR} \leq \frac{n_0 q}{n} \leq q.$$

Note: FDR control is not guaranteed if the test statistics are dependent.

Motivating the BH procedure

For each $\alpha \in (0,1)$, let $R(\alpha)$ be the number of p-values $\leq \alpha$. If we reject hypotheses with p-value $\leq \alpha$, then we expect (on average) to falsely reject αn_0 null hypotheses, since the null p-values are distributed as $\text{Uniform}(0,1)$. So we might estimate the false discovery proportion by

$$\alpha n_0 / R(\alpha).$$

Motivating the BH procedure

For each $\alpha \in (0,1)$, let $R(\alpha)$ be the number of p-values $\leq \alpha$. If we reject hypotheses with p-value $\leq \alpha$, then we expect (on average) to falsely reject αn_0 null hypotheses, since the null p-values are distributed as $\text{Uniform}(0,1)$. So we might estimate the false discovery proportion by

$$\alpha n_0 / R(\alpha).$$

As we don't know n_0, let's take the conservative upper-bound

$$\alpha n / R(\alpha).$$

If we set $\alpha = P_{(r)}$, the rth largest p-value, then $\alpha n / R(\alpha) \leq q$ exactly when $P_{(r)} \leq qr/n$. So the BH procedure chooses α (in a data-dependent way) so as to reject as many hypotheses as possible, subject to the constraint $\alpha n / R(\alpha) \leq q$.

Proof of FDR control

Let's prove (more formally) the previous theorem, that BH controls the FDR. For any event $\mathcal{E}$, we use the indicator notation

$$\mathbb{1}\{\mathcal{E}\} = \begin{cases} 1 & \mathcal{E} \text{ holds} \\ 0 & \mathcal{E} \text{ does not hold.} \end{cases}$$

Without loss of generality, order the n null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$ so that the first n_0 of them are true nulls. Then

$$\begin{aligned} \mathrm{FDR} &= \mathbb{E}[\mathrm{FDP}] \\ &= \mathbb{E}\left[\sum_{r=1}^{n} \frac{V}{r} \mathbb{1}\{R = r\}\right] \\ &= \mathbb{E}\left[\sum_{r=1}^{n} \sum_{j=1}^{n_0} \mathbb{1}\{\text{reject } H_0^{(j)}\} \frac{1}{r} \mathbb{1}\{R = r\}\right], \end{aligned}$$

(where we have noted $V = \sum_{j=1}^{n_0} \mathbb{1}\{\text{reject } H_0^{(j)}\}$).

Proof of FDR control

Applying linearity of expectation,

$$\begin{aligned}\mathrm{FDR} &= \sum_{r=1}^{n}\sum_{j=1}^{n_0}\frac{1}{r}\mathbb{E}\left[\mathbb{1}\{\text{reject } H_0^{(j)}\}\mathbb{1}\{R=r\}\right]\\ &= \sum_{r=1}^{n}\sum_{j=1}^{n_0}\frac{1}{r}\mathbb{P}\left[\text{reject } H_0^{(j)} \text{ and } R=r\right].\end{aligned}$$

For fixed j, let $P^*_{(1)} \leq \ldots \leq P^*_{(n-1)}$ be the sorted $n-1$ p-values other than P_j. Then the BH procedure rejects r total hypotheses including $H_0^{(j)}$ if and only if $P_j \leq \frac{qr}{n}$ and the following event holds:

$$\begin{aligned}\mathcal{E}^{(r)} := \Bigg\{ & P^*_{(1)},\ldots,P^*_{(r-1)} \leq \frac{qr}{n},\\ & P^*_{(r)} > \frac{q(r+1)}{n}, P^*_{(r+1)} > \frac{q(r+2)}{n},\ldots,P^*_{(n-1)} > q\Bigg\}.\end{aligned}$$

Proof of FDR control

As the p-values are independent, P_j is independent of $P^*_{(1)}, \dots, P^*_{(n-1)}$. Furthermore, $P_j \sim \text{Uniform}(0,1)$. So

$$\begin{aligned}
\mathrm{FDR} &= \sum_{r=1}^{n} \sum_{j=1}^{n_0} \frac{1}{r} \mathbb{P}\left[P_j \le \frac{qr}{n} \text{ and } \mathcal{E}^{(r)} \text{ holds}\right] \\
&= \sum_{j=1}^{n_0} \sum_{r=1}^{n} \frac{1}{r} \mathbb{P}\left[P_j \le \frac{qr}{n}\right] \mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] \\
&= \sum_{j=1}^{n_0} \sum_{r=1}^{n} \frac{1}{r} \frac{qr}{n} \mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] \\
&= \frac{q}{n} \sum_{j=1}^{n_0} \sum_{r=1}^{n} \mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right].
\end{aligned}$$

Proof of FDR control

Finally, note that (for any fixed j) the events $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(n)}$ are mutually exclusive—$\mathcal{E}^{(r)}$ holds if and only if the largest index k such that $P^*_{(k)} \leq \frac{q(k+1)}{n}$ is exactly $k = r - 1$ (with $\mathcal{E}^{(1)}$ holding if $P^*_{(k)} > \frac{q(k+1)}{n}$ for all k), and this is true for exactly one value of $r \in \{1, \ldots, n\}$. So

$$\sum_{r=1}^{n} \mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] = 1.$$

Hence

$$\mathrm{FDR} \leq \frac{q}{n} \sum_{j=1}^{n_0} 1 = \frac{q n_0}{n} \leq q.$$