Data 102: Lecture 4

Moritz Hardt
UC Berkeley, Spring 2020

Part of the slide deck courtesy of Michael Jordan

Announcements

We're moving. Seriously! Check Piazza before coming to class

Last time

Statistical decision-making framework

Data X

Parameter θ

Decision rule $\delta(X)$

Bayesian setting: Prior $P(\theta)$ over parameters, joint distribution $P(\theta, X)$

Frequentist setting: Likelihood $P(X \mid \theta)$

Neyman-Pearson formulation (1932)

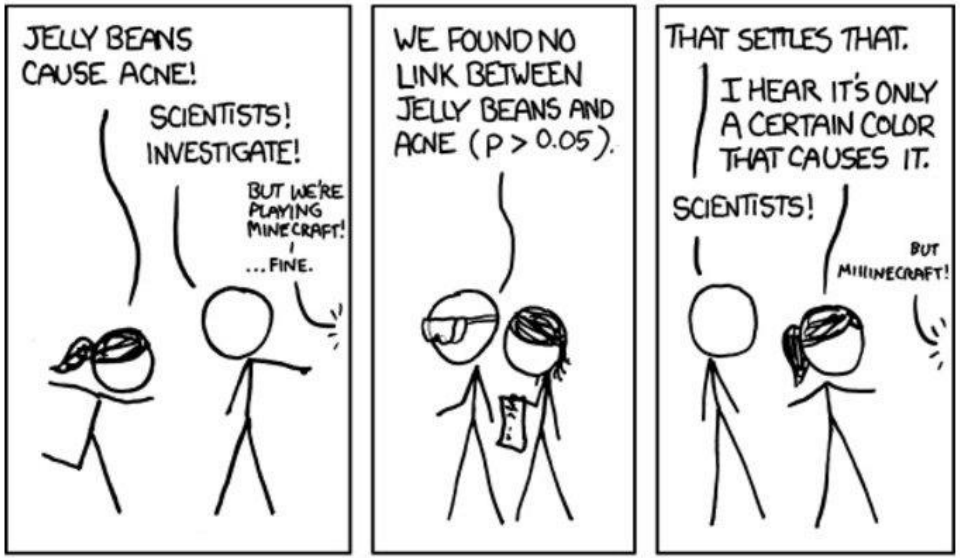
Constrained optimization:

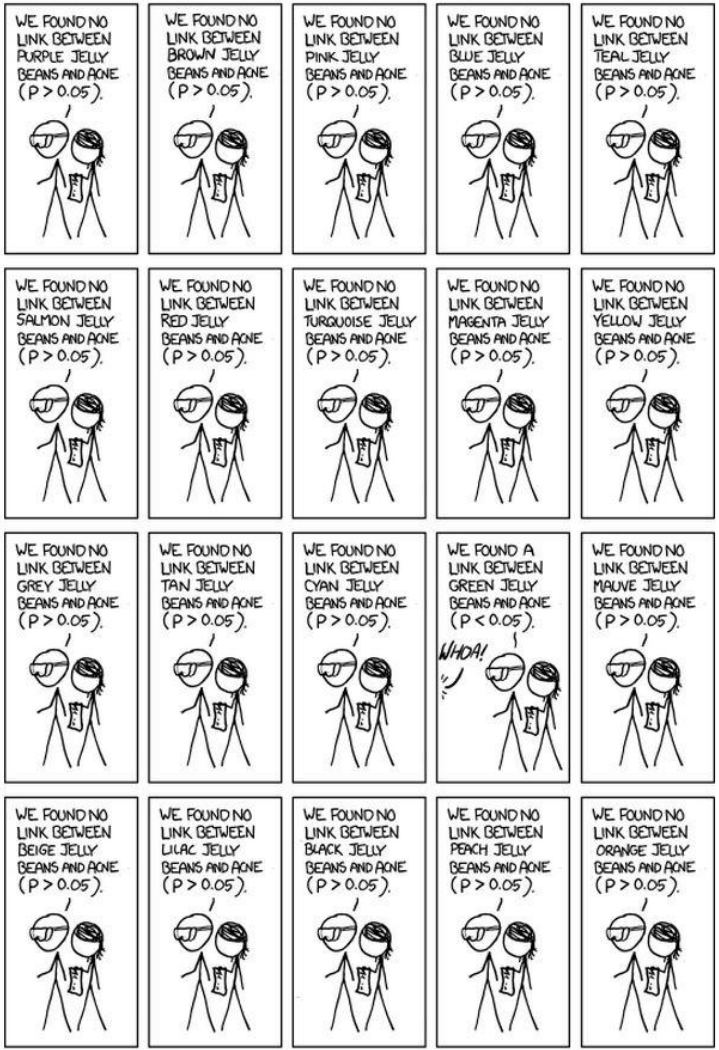
Maximize true positive rate of δ

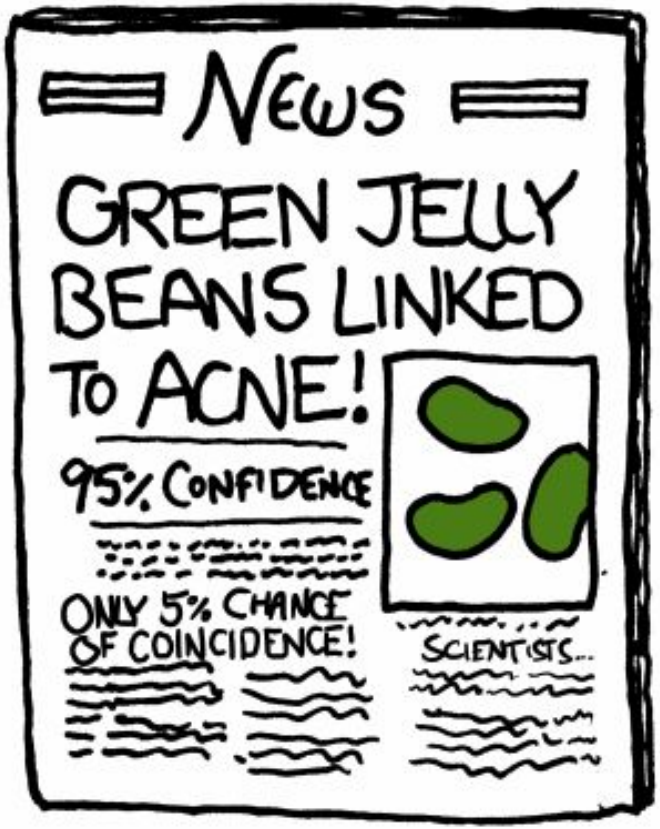
s.t. false positive rate ≤ α (e.g. 0.05)

Tuesday: Neyman-Pearson lemma. Optimal solution is **Likelihood Ratio Test**

Today: Multiple hypothesis testing


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!


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).
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).


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

The reproducibility crisis

The Economist. October 2013.

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

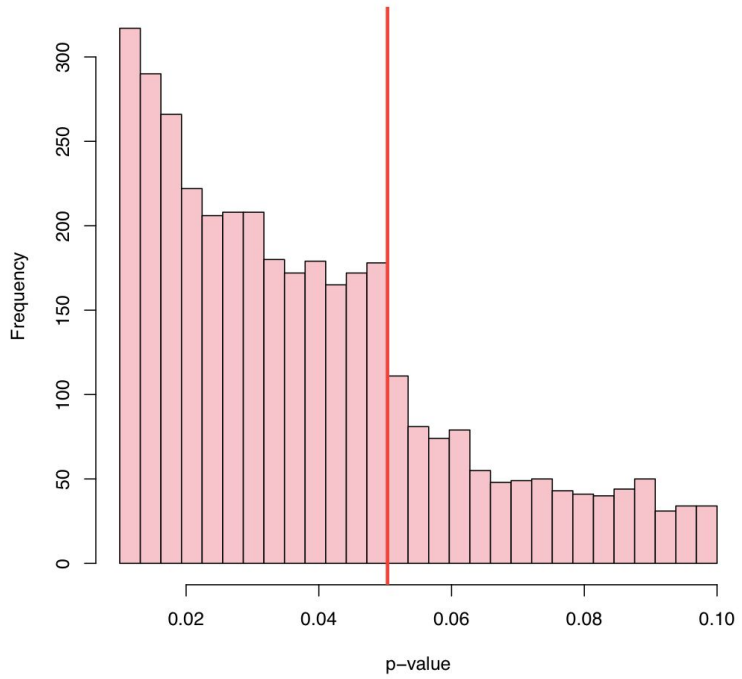
WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY TRANSPORTATION

CHRISTIE ASCHWANDEN IDEAS 11.26.2019 09:00 AM

We're All 'P-Hacking' Now

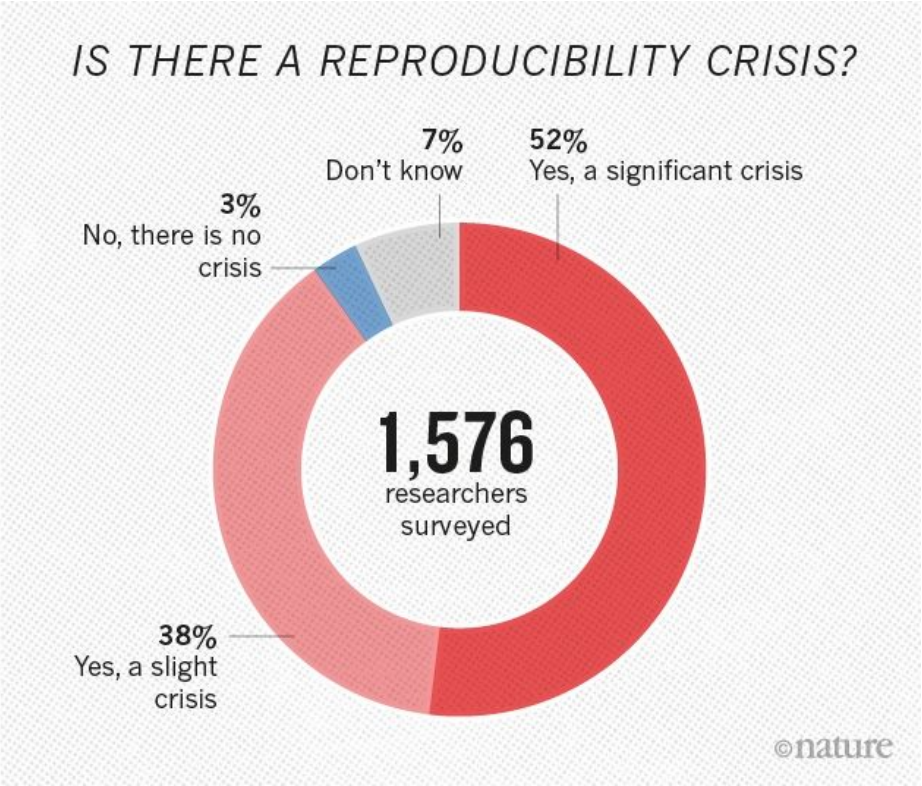
An insiders' term for scientific malpractice has worked its way into pop culture. Is that a good thing?

Distribution of published p-values



Source:
A peculiar prevalence of p values just below .05
Masicampo, Lalande
https://journals.sagepub.com/doi/10.1080/17470218.2012.711335

The reproducibility crisis



NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

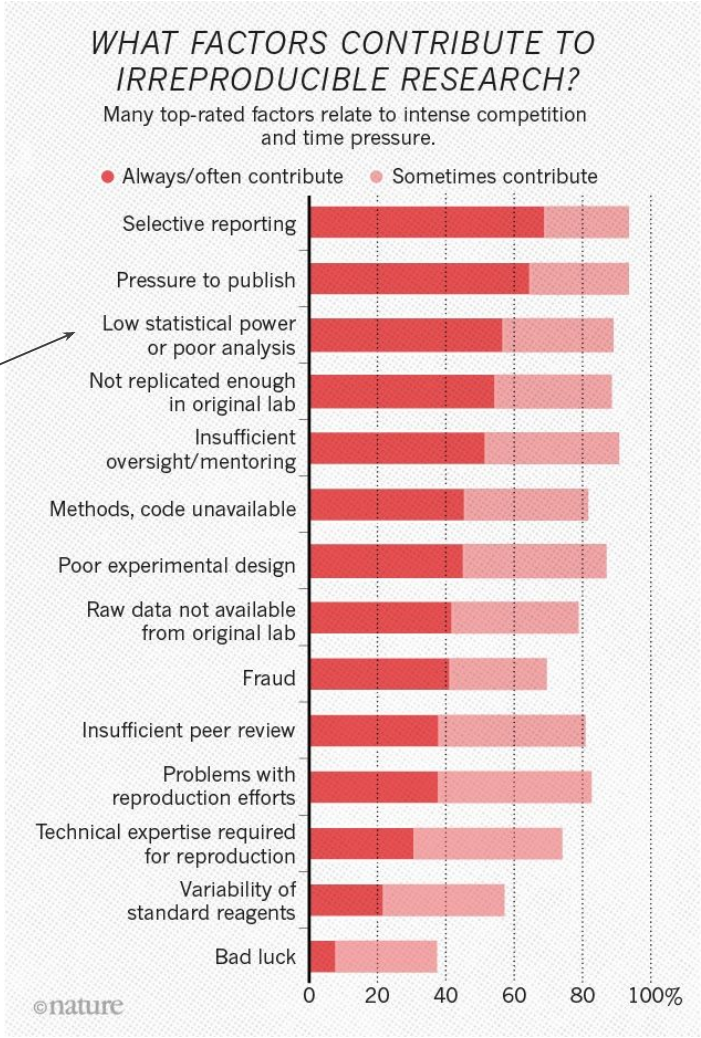
25 May 2016 | Corrected: 28 July 2016

Source: Nature News, 2016. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

The reproducibility crisis

There are *many causes*. We won't touch on all of them in this class.

Primarily our focus.



Source: Nature News, 2016.
https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Should we do hypothesis testing at all?

As a result of the replication crisis, hypothesis testing has often been the scapegoat.

But we saw there are many problems that co-occurred with hypothesis testing

We'll argue that hypothesis testing can still be a useful tool up your sleeve, if you understand it well and use it carefully.

Recap: Hypothesis tests as decision making

Hypothesis H

Reality: Null hypothesis is true ($\theta = 0$), null hypothesis is false ($\theta = 1$)

Decision: Accept null hypothesis ($\delta(X) = 0$), Reject null hypothesis ($\delta(X) = 1$)

Interpret "$\delta(X) = 1$" as declaring a "discovery"

Hence, false positive = *false discovery*.

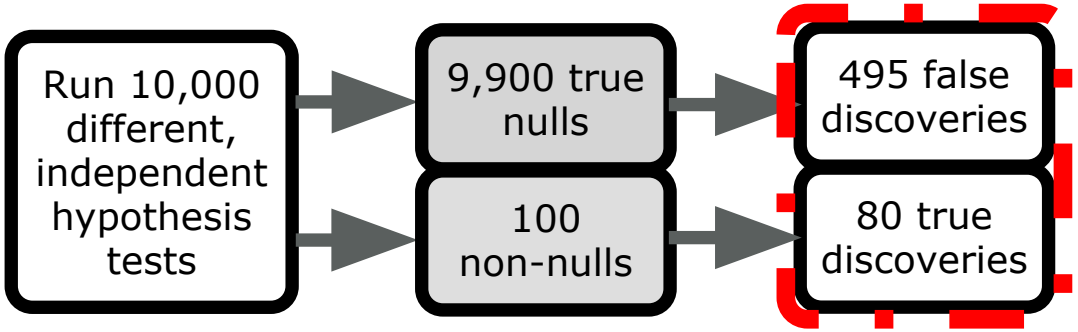
What we'll focus on today

Reality \ Decision	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

false discovery proportion (FDP) $\dfrac{n_{01}}{n_{01}+n_{11}}$

False discovery proportion in hypothesis testing

FPR = Pr(reject | null) = 0.05



false discovery proportion = 495/575

TPR = Pr(reject | non-null) = 0.80

Recap P-values

Consider a null hypothesis ($\theta = 0$) with distribution $P_0(X)$ under the null hypothesis. (This is a shorthand for the likelihood of X under the null.)

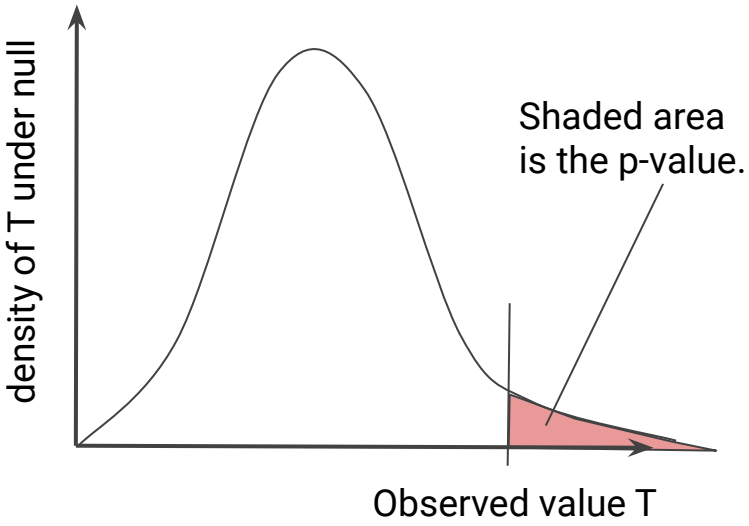
Test statistic $T(X)$ with *tail* cdf $F(t) = P_0(T > t)$

P-value is defined as the random variable $F(T)$

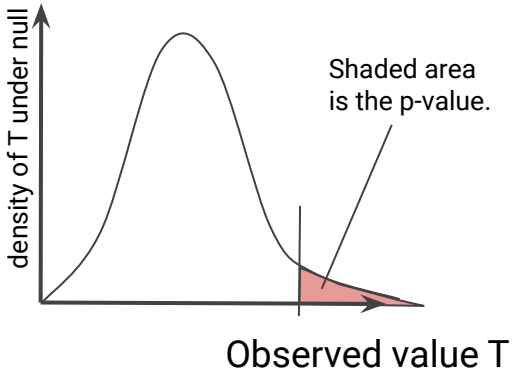
Generic test: $\delta(X)$ = REJECT if $F(T) < \alpha$ and ACCEPT otherwise.

P-values

Data distribution $P_0(X)$ under the null hypothesis.
Test statistic $T(X)$ with cdf $F(t) = P_0(T > t)$, p-value is $P = F(T)$



P-values

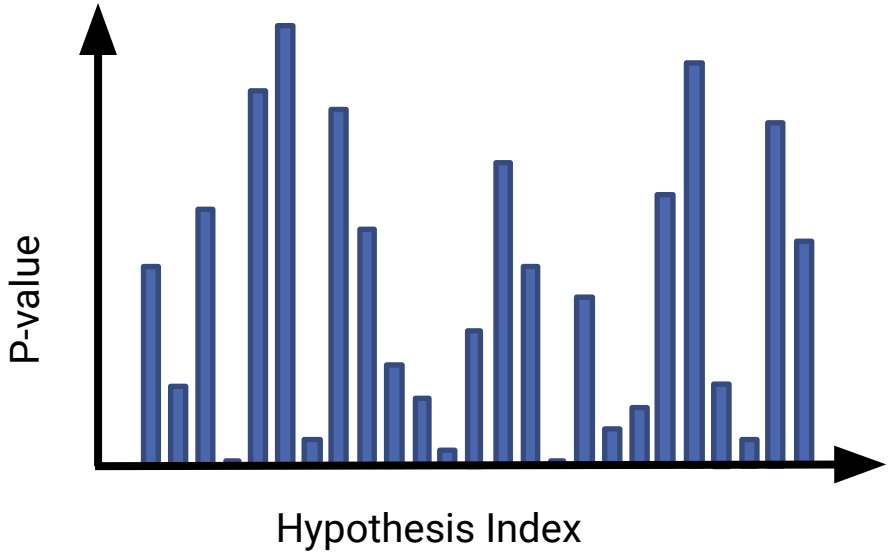


Data distribution $P_0(X)$ under the null hypothesis.
Test statistic $T(X)$ with cdf $F(t) = P_0(T > t)$, p-value is $P = F(T)$

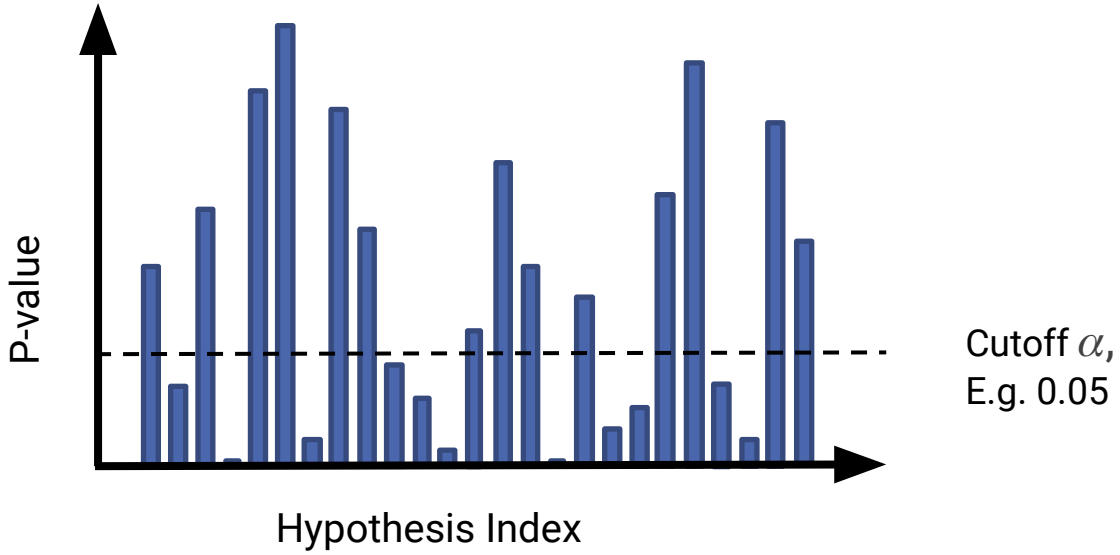
Fact: p-value is uniformly distributed under the null.

Proof:

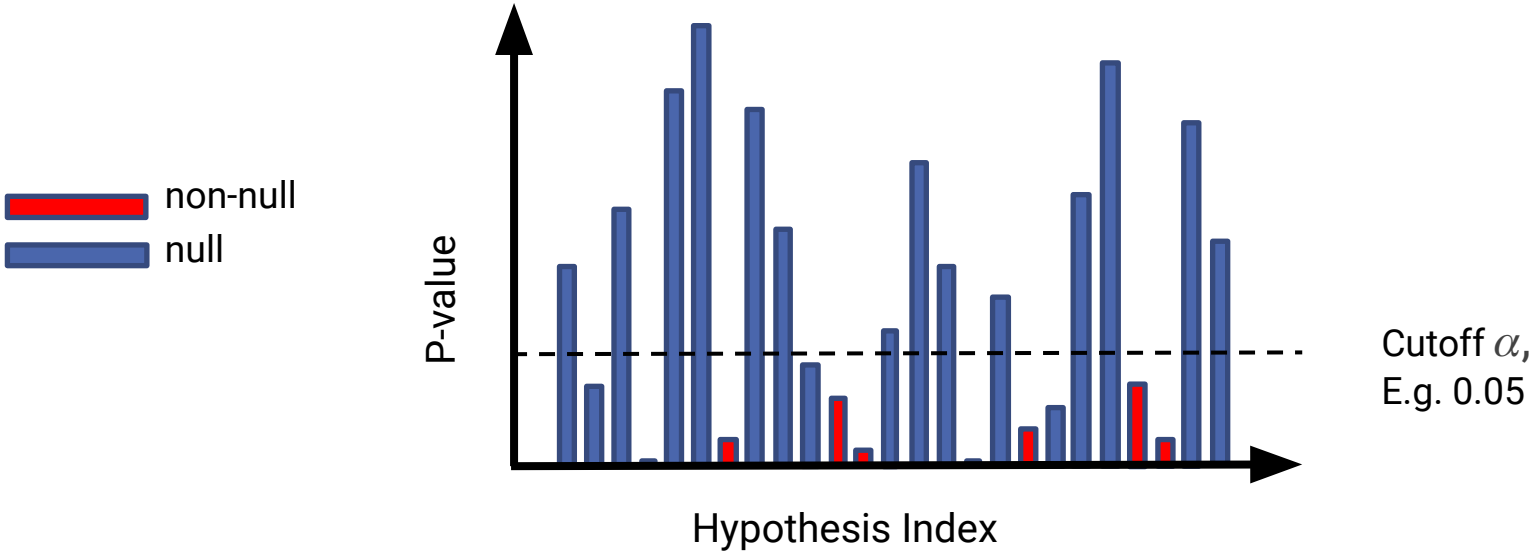
$$P_0(P < p) = P_0(F(T) < p) = P_0(T > F^{-1}(p)) = F(F^{-1}(p)) = p$$



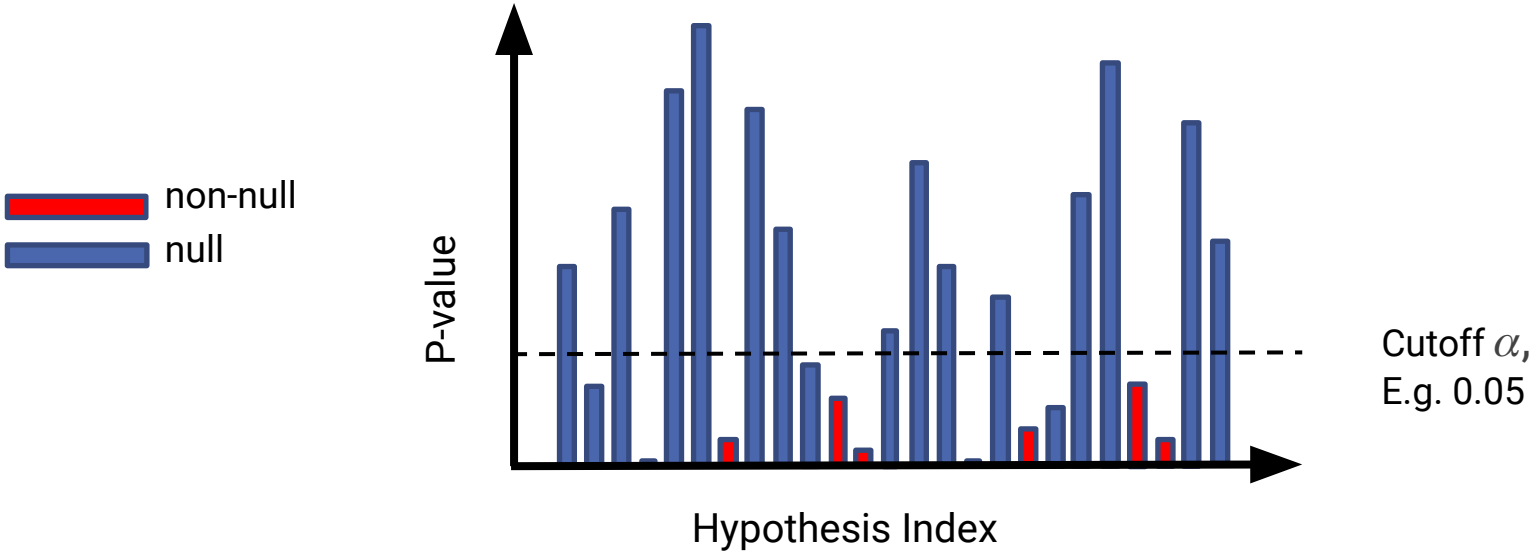
Suppose we run 25 *independent* experiments and record their p-values.



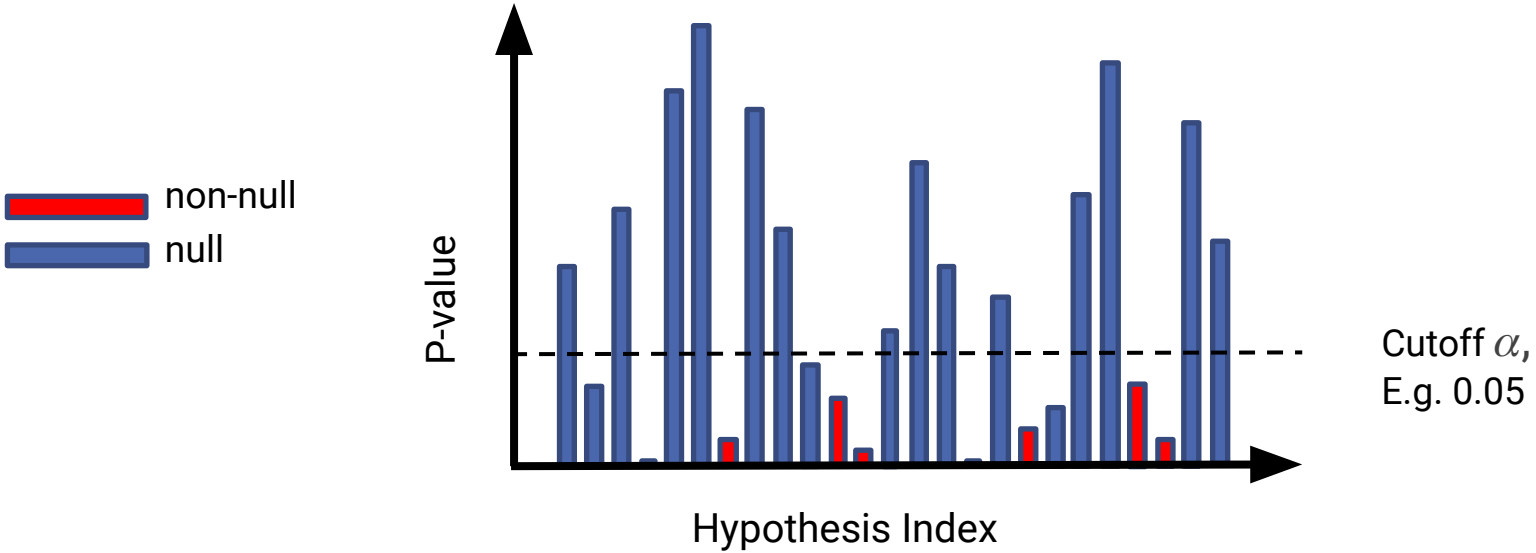
Say we reject all null hypotheses below cutoff.



Suppose highlighted hypotheses are non-nulls (Reality = 1), and blue ones are the true nulls (Reality = 0)



Our fixed cutoff rejects all **6** non-nulls, but it also rejects **5** nulls.



Our false discovery proportion is 5/11. Not so great!

Can we avoid false positives?

Old idea: Bonferroni correction, a.k.a. union bound.

Suppose we make *m* tests. Let *V* be the number of false positives across all tests. Let E_i denote the event of a false positive in the i-th test. These are random variables.

So, we can apply the union bound to P(V > 0)

$P(V > 0) \leq \sum_i P(E_i)$

If each test has FPR $\leq \alpha'$

$P(V > 0) \leq m\,\alpha'$

To get $P(V > 0) \leq \alpha$,
we need $\alpha' \leq \alpha/m$.

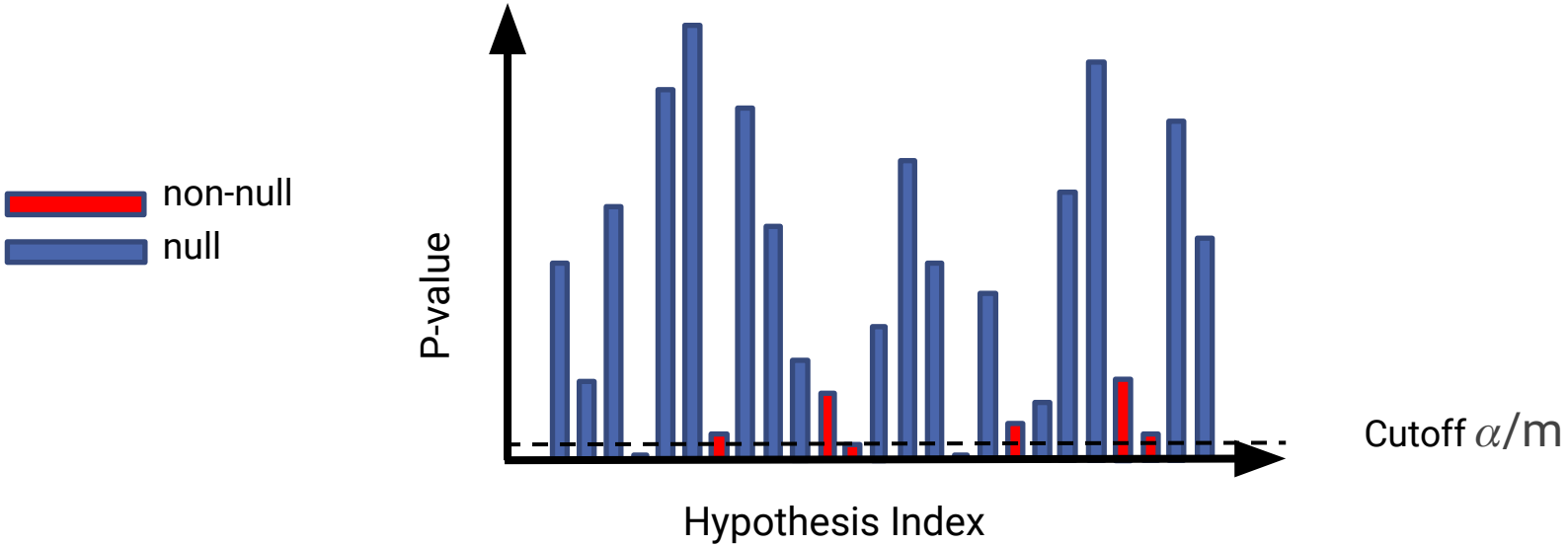
Bonferroni correction

If you make m hypothesis tests, reject each hypothesis if p-value < α/m

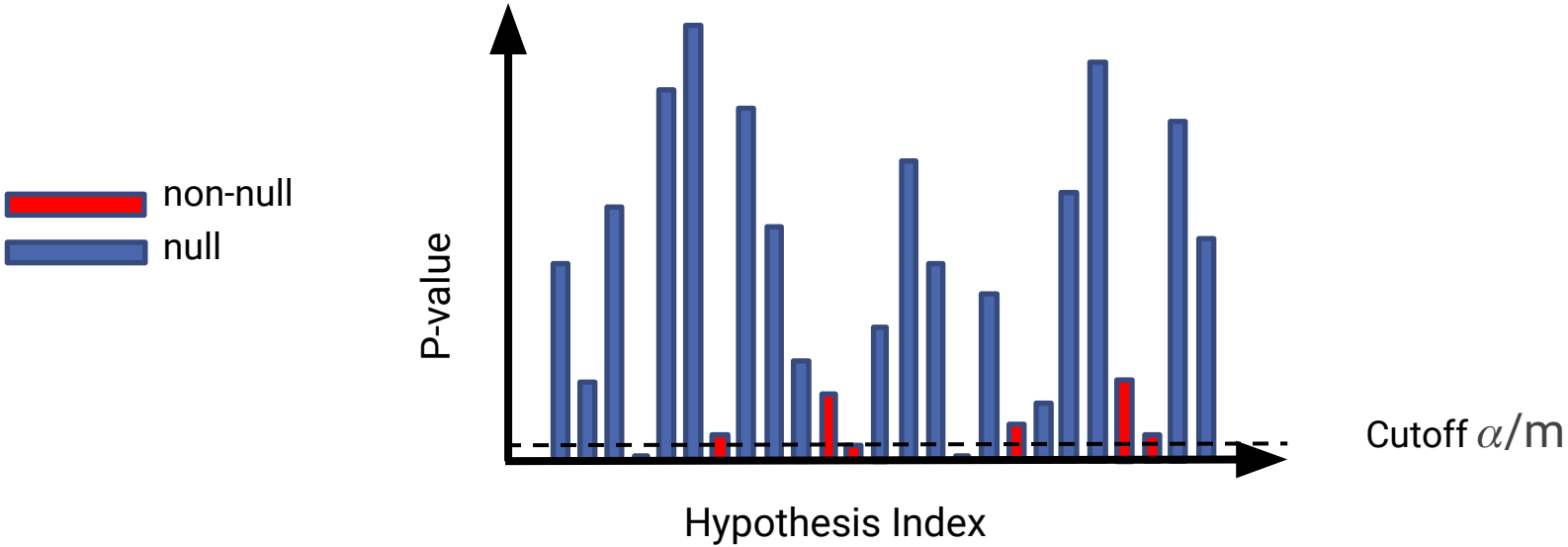
This bounds probability of a single false positive across all tests by α.

Statisticians call this **controlling the family-wise error**.

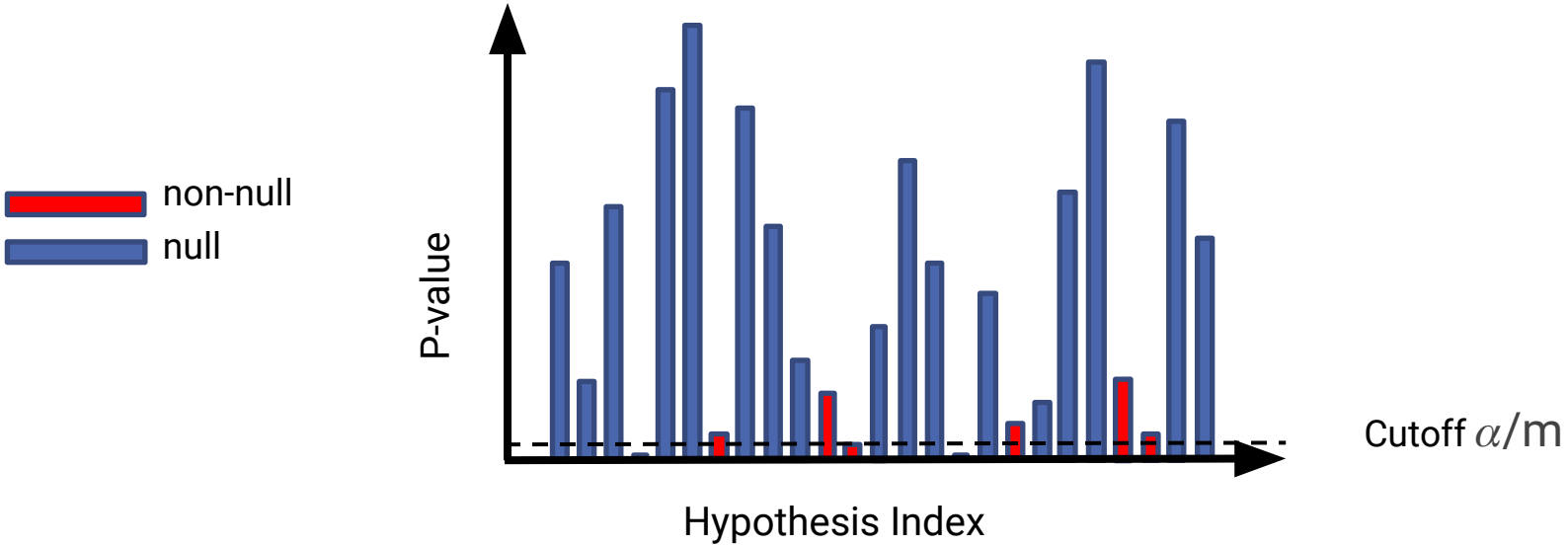
“Controlling” means you have an *a priori* guarantee.



Bonferroni: Divide cutoff by 25 (number of hypotheses).



Now we reject **1 non-null, reject 2 nulls.**
False discovery proportion is now ⅔. Even worse!



Bonferroni avoid false positives at the expense of more false negatives!

Observation

If we want to make any discoveries at all, we cannot guarantee that false discovery proportion is always less than any fixed value strictly less than 1.

Why?

P-values are uniform under the null. There’s some tiny probability that all nulls will have tiny p-value.

False discovery *rate* control

Let V be the number of falsely rejected nulls (“false discoveries”).

Let R be the number of all rejected hypotheses (“discoveries”).

Note that FDP = V/R. Let’s put FDP = 0, if R=0.

Statisticians focus on tests that guarantee **E**[FDP] = **E**[V/R] ≤ α.

This expectation **E**[V/R] is called *false discovery rate* in the research community.

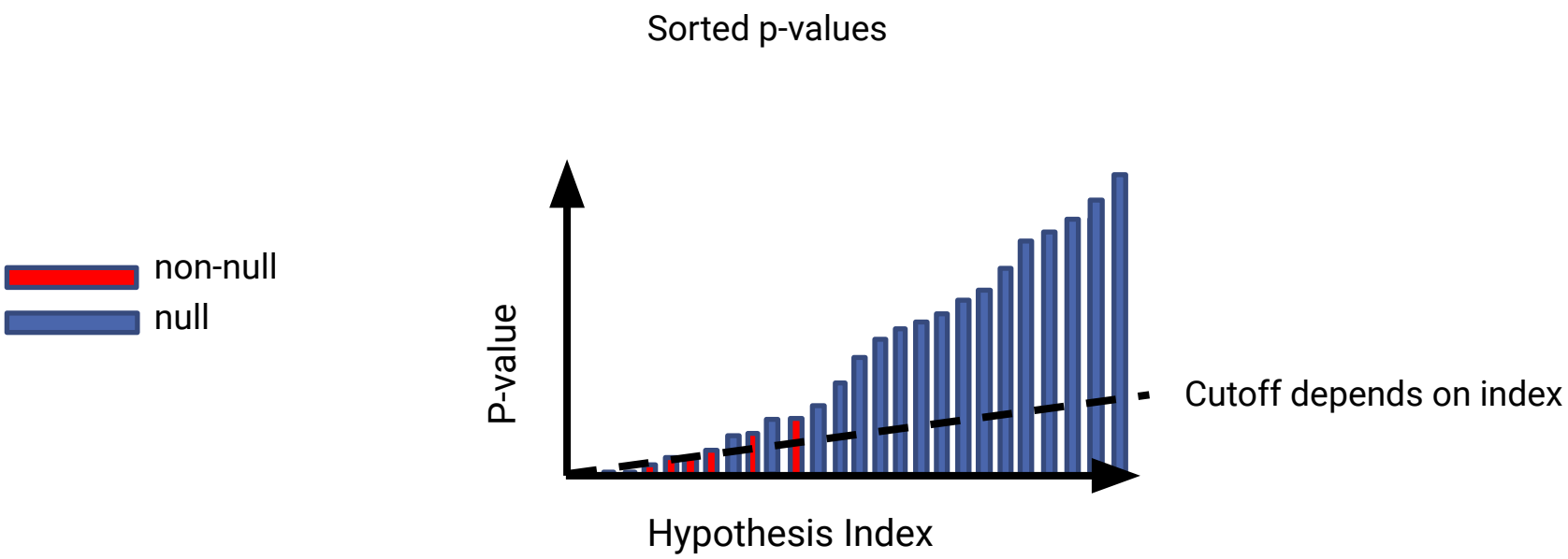
False discovery *rate* control

Let V be the number of falsely rejected nulls (“false discoveries”).
Let R be the number of all rejected hypotheses (“discoveries”).
Note that FDP = V/R. Convention: FDP = 0, if $R = 0$.
FDR = $\mathbf{E}[\text{FDP}] = \mathbf{E}[V/R]$

There are two ways to make FDR small:

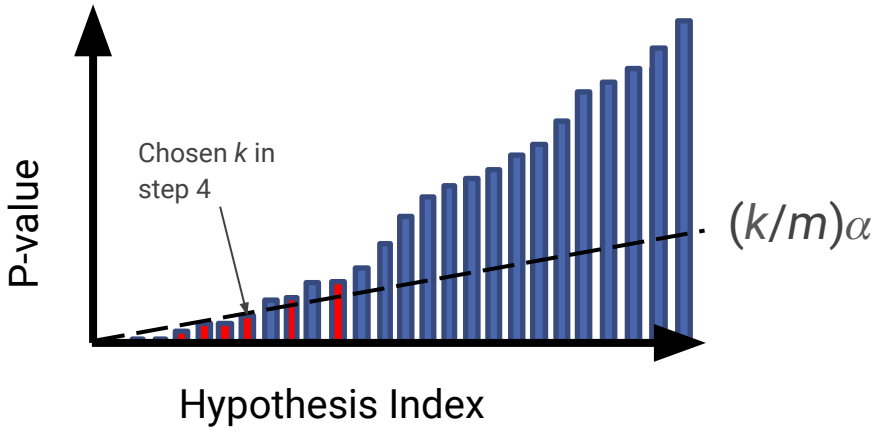
Make V small, or make R large

Safe discoveries should make us more *risk tolerant*!



Later cutoffs more relaxed banking on earlier discoveries!

Benjamini Hochberg



1. Given m tests, obtain m p-values
2. Sort them as $P_1 \leq P_2 \leq \ldots \leq P_m$
3. Find the largest k s.t. $P_k \leq (k/m)\alpha$
4. Reject null hypothesis for all $i \leq k$

Theorem: This procedure controls FDR at level α, i.e., $\mathbf{E}[V/R] \leq \alpha$.

A Bayesian derivation of the BH procedure

Suppose we're in the Bayesian setting: We have a joint distribution P over θ (state of reality, i.e., null vs non-null) and data X.

We think of FDP as estimating the probability P(null | reject) = $P(\theta = 0 \mid \delta(X) = 1)$

Suppose now our goal is to ensure $P(\theta = 0 \mid \delta(X) = 1) \leq \alpha$

This is *not* equivalent to FDR control. This is a Bayesian perspective.

We will see that this perspective naturally recovers the BH procedure.

A Bayesian derivation of the BH procedure

Suppose now our goal is to ensure $P(\theta = 0 \mid \delta(X) = 1) \leq \alpha$

Apply Bayes rule:

$P(\theta = 0 \mid \delta(X) = 1) = P(\delta(X) = 1 \mid \theta = 0)\ (P(\theta=0) / P(\delta(X) = 1))$

Note:

$P(\delta(X) = 1 \mid \theta = 0) = \text{FPR}$ (false positive rate)

$P(\theta=0) \leq 1$ (and in fact, not a bad bound if non-nulls are rare)

$P(\delta(X) = 1) \leq k/m$ (by design of BH procedure)

A Bayesian derivation of the BH procedure

Suppose now our goal is to ensure $P(\theta = 0 \mid \delta(X) = 1) \leq \alpha$

Bayes rule: $P(\theta = 0 \mid \delta(X) = 1) = P(\delta(X) = 1 \mid \theta = 0)\,(P(\theta=0) / P(\delta(X) = 1))$

Note: $P(\delta(X) = 1 \mid \theta = 0) = \text{FPR}$ (false positive rate)
$P(\theta=0) \leq 1$ (and in fact, not a bad bound if non-nulls are rare)
$P(\delta(X) = 1) \leq k/m$ (by design of BH procedure)

So, $P(\theta = 0 \mid \delta(X) = 1) \leq \text{FPR} / (k/m)$

But what is FPR?

A Bayesian derivation of the BH procedure

We have: $P(\theta = 0 \mid \delta(X) = 1) \leq FPR/(k/m)$

But what is FPR?

By design, FPR = P_k *i.e. the cutoff we choose in BH*

Hence, $P(\theta = 0 \mid \delta(X) = 1) \leq P_k / (k/m)$

We can ensure $P(\theta = 0 \mid \delta(X) = 1) \leq \alpha$ by making sure $P_k / (k/m) \leq \alpha$

Equivalently, $P_k \leq (k/m)\alpha$

To be least conservative, pick the largest such k. This is exactly what BH does.

The online problem

The online problem

- Classical statistics, and also the Benjamini & Hochberg algorithm focused on a batch setting in which all data has already been collected
- E.g., for Benjamini & Hochberg, you need all of the p-values before you can get started
- Is is possible to consider methods that make sequences of decisions, and provide FDR control at any moment in time?

A common industry problem: Repeated A/B testing

Decision Rule:

$P_1 \leq \alpha?$

$P_2 \leq \alpha?$

$P_3 \leq \alpha?$

$P_4 \leq \alpha?$

$P_5 \leq \alpha?$

Time

Problem!

Color

Size

Orientation

Style

Logo

What you can do instead

Decision Rule:

Time

Guarantee FDR control at each time step

$P_1 \leq \alpha_1?$  Color

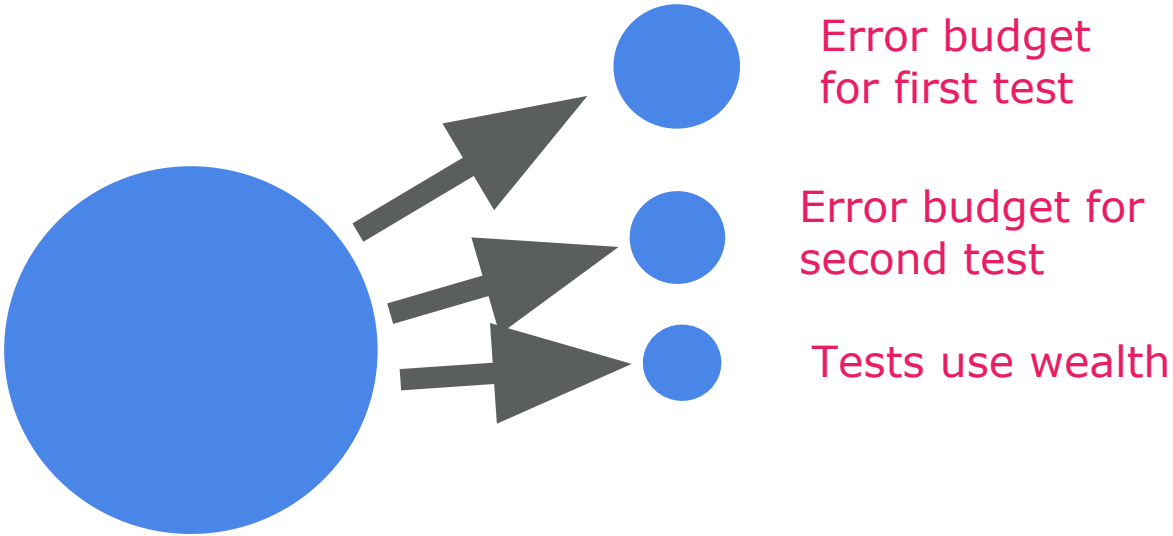
$P_2 \leq \alpha_2?$  Size

$P_3 \leq \alpha_3?$  Orientation

$P_4 \leq \alpha_4?$  Style

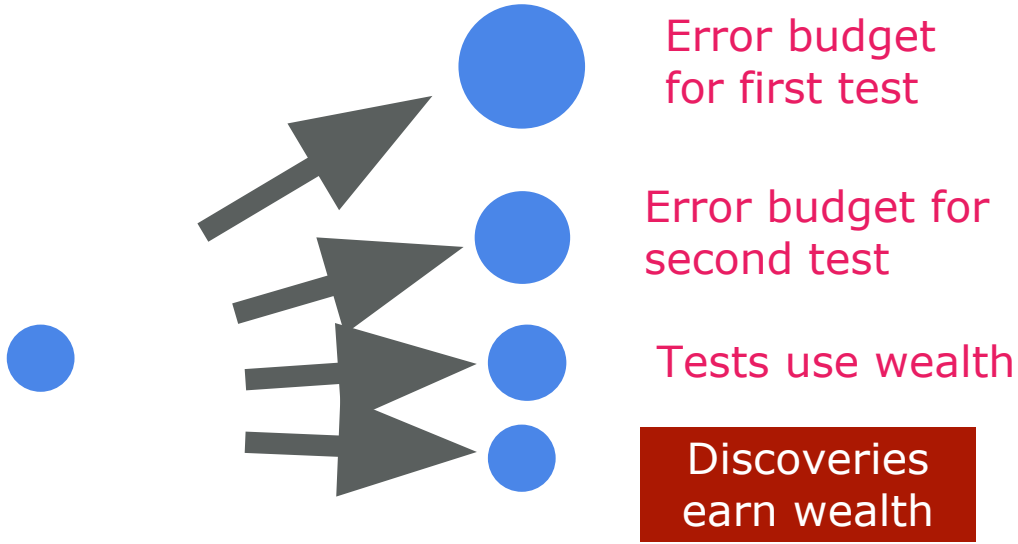
$P_5 \leq \alpha_5?$  Logo

Online FDR control : high-level picture



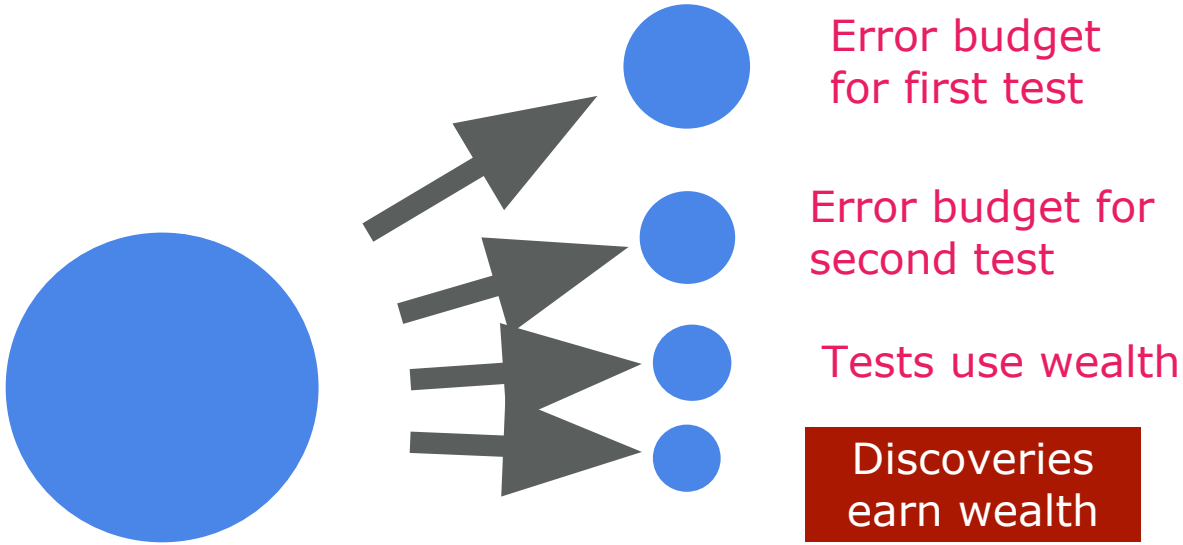
Remaining error budget
or “alpha-wealth”

Online FDR control : high-level picture



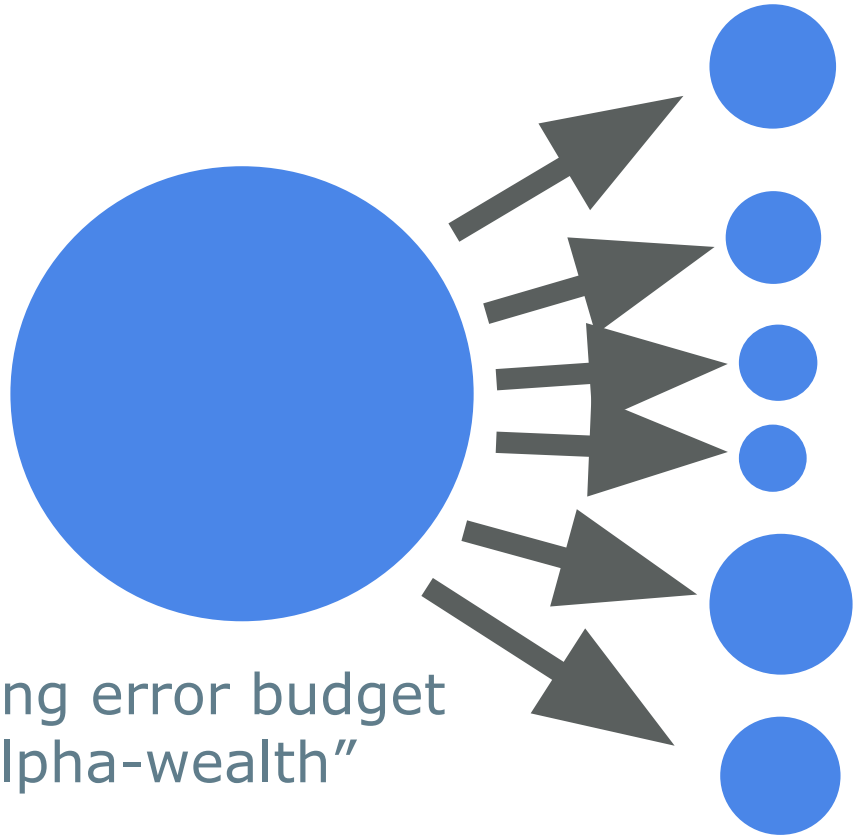
Remaining error budget
or “alpha-wealth”

Online FDR control : high-level picture



Remaining error budget
or “alpha-wealth”

Online FDR control : high-level picture



Error budget
for first test

Error budget for
second test

Tests use wealth

Discoveries
earn wealth

Error budget
is data-dependent

Infinite process

Remaining error budget
or “alpha-wealth”

Online FDR control

- classical FDR literature assumes that the data for all hypotheses is collected at once, and only after all the p-values are available, one can decide which of the hypotheses should be proclaimed discoveries
- in modern testing we often do not know how many hypotheses we want to test in advance
- instead, a possibly infinite sequence of tests (i.e. p-values) arrives *sequentially*
- we have to make decisions *online*, with no knowledge of future tests, in a way that guarantees FDR control under a pre-specified level *at any given time*
- motivating examples: A/B testing, large-scale clinical trials…

Online FDR control is possible

The first online FDR algorithm was due to Foster and Stine (2008)

A more recent (and simpler) online FDR algorithm is due to Javanmard and Montanari, and is called LORD.

We might to a homework problem on this.

Some issues and limitations

Major caveat in everything we saw

All hypotheses are independent

More formally precise statement: All p-values always have to be uniform under the null, regardless of other hypotheses.

This can be relaxed slightly (negative dependence etc.).

Thought experiment

Suppose you get your data.

You start playing around with it, clean it a bit, select some reasonable variables, throw out some others.

Now you do a *single* hypothesis test.

You get a p-value of 0.001. Is it *legit*? Do you need a correction? If so, what?

Inference after selection and adaptivity

What we saw can be a major problem.

Computing p-values after **data-dependent choices** generally breaks the assumptions of your p-value (distribution ***not*** uniform under null).

This was recognized by David Freedman (UC Berkeley) and is known as Freedman's paradox

Now widely recognized and studied as *inference after selection* (in statistics), *adaptive data analysis* (in computer science).

How do we cope?

The easiest way is to collect new data from the same distribution and run hypothesis test on fresh data.

This is safe, but wasteful in terms of sample splitting.

Better approaches are often very sophisticated and not yet very practical.

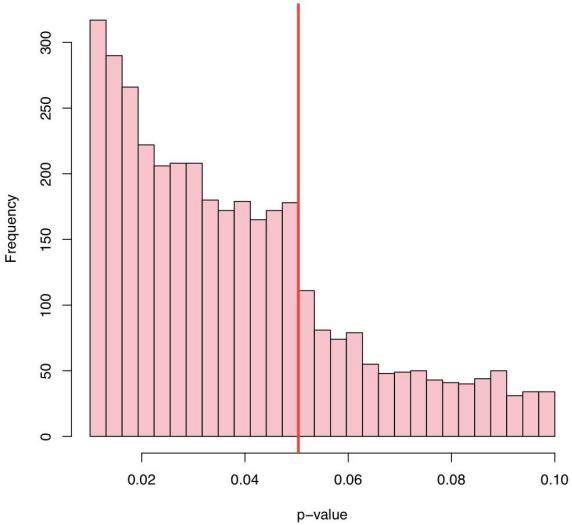
Beware of “implicit comparisons”

A researcher has lots of degrees of freedoms that lead to implicit comparisons favoring one analysis over the other.

These implicit comparisons often happen without being recorded or recognized.

Increasingly, researchers turn to **pre-registration**: Specify your entire experimental setup ahead of time and commit to it before data collection. Run the setup as specified once you have the data. Report outcome no matter what.

Recall from earlier:



nature

Subscribe

Search

NEWS · 24 OCTOBER 2018

First analysis of 'pre-registered' studies shows sharp rise in null findings

Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results.

Matthew Warren

That’s it for today.