Multiple Testing and the FDA

Lucy Gao

January 16, 2024

Multiple Testing

My definition: Whenever we conduct more than one hypothesis test in a single analysis, and look at the outcome of all of them.

- Does this cause problems? If so, what are they?
- Do we need to adjust for multiple testing? If so, how should we adjust?

> **"I don't know, it depends on the scientific problem at hand."** - My annoying partner, whenever I ask him ANY question involving statistics.

Today's science

Multiple Endpoints in Clinical Trials

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

Recall: FDA

- An experimental intervention is approved if applicant can show sufficient evidence of efficacy, with evidence gathered by hypothesis testing
- Approval by the FDA is (essentially) the last barrier before the intervention goes to market
- Mistakes are very hard to correct

When deciding how to map data analysis results to regulatory decisions, the FDA's first priority is controlling the proportion of approvals when reviewing N_0 treatments with no benefit (N_0 big).

Recall: clinical trials

Heavily simplified procedure:

- Recruit n patients from the population
- Randomize 50% of them to control, and randomize 50% of them to treatment
- Measure a single outcome, e.g. blood pressure.
- Apply a two-tailed test of equality of some functional (e.g. mean, proportion, hazard rate) with $\alpha = 0.05$ to the outcome measurements
- Construct the 95% confidence interval as well to go with it

Across N_0 trials for treatments with no benefit, no more than $0.05N_0$ approvals (N_0 big).

Here, "benefit" means treatment has an effect on the single outcome.

FDA Guidelines for > 1 Outcome

- Study can be planned to measure M outcomes for $M > 1$
- M outcomes lead to M tests and M confidence intervals
- But the FDA needs to lay out how those M hypothesis testing results map to approval

Two cases, depending on definition of the word “benefit”:

1. “Benefit” means “has an effect on all M outcomes”
2. “Benefit” means “has an effect on at least one outcome out of M total outcomes”

How should we define benefit?

Scientific considerations

Benefit = has an effect on all outcomes

Example 1: Migraine

- Pain, but also “extra stuff”: nausea, light sensitivity, sound sensitivity
- Past: benefit = changed all extra stuff
- Now: benefit = changed pain and changed the patient’s most hated “extra” symptom

Example 2: Combination vaccine (MMR, DTaP, etc.)

- Is it really sensible to approve if it doesn’t have efficacy against all diseases?

Scientific considerations

Benefit = has an effect on at least one outcome

At the time of trial design:

- We might not know what aspect of the disease will respond to the treatment
- The community might not agree on which aspect(s) must be addressed for the intervention to be clinically meaningful

We might feel comfortable saying that if the treatment has an effect on *any* of the outcomes, then it can be safely declared "clinically effective".

Regulatory mapping

Multiple outcomes version

In the following, we will explore …

- Different mappings between data analysis results for M hypothesis tests to decisions (approval or not)
- Which mappings honour the FDA's goals: across N_0 trials for treatments with no benefit, no more than $0.05N_0$ approvals (N_0 big).

Two cases: "all outcomes" and "$\geq$ 1" outcome

All outcomes, naive approach

Suppose that we have M outcomes, and are interested in testing M null hypotheses of no difference in a pre-specified functional (e.g. the mean) across groups.

A naive idea: Approve the treatment if all M p-values are below α.

Your turn: Do you think this policy controls the proportion of approving treatments with no benefit?

- Remember: benefit means “has an effect on all outcomes”
- Turn it around: what does “no benefit” mean?

All outcomes, naive approach

Statistical setup

Let $X \sim F_X$ represent the data with which we test $H_{0j} : T_j(F_X) = 0$ for all $j = 1, 2, \ldots, M$.

Let $p_j(x)$ be the p-value function for testing H_{0j}. Assume that it's valid and based on a pivot:

$$\mathbb{P}_{F_X}(p_j(X) \leq \alpha) = \alpha, \quad \text{for all } F_X \in \mathcal{F}_{0j} \equiv \{F : T_j(F_X) = 0\}.$$

"Benefit" means F_X satisfies all of $H_{11}, \ldots, H_{1M}$. So the (worst-case) probability of approving a treatment with no benefit is:

$$\sup_{F_X \in \left[\bigcap_{j=1}^{M} \mathcal{F}_{0j}^C\right]^C} \mathbb{P}_{F_X}(p_1(X) \leq \alpha, \ldots, p_M(X) \leq \alpha)$$

All outcomes, naive approach

Statistical setup

Suppose that $F_X \in \left[\bigcap_{j=1}^{M} \mathcal{F}_{0j}^C\right]^C = \bigcup_{j=1}^{M} \mathcal{F}_{0j}$. Then, let j^* be an index for which $F_X \in \mathcal{F}_{0j^*}$.

Follows that:

$$\mathbb{P}_{F_X}(p_1(X) \leq \alpha, \ldots, p_M(X) \leq \alpha) \leq \mathbb{P}_{F_X}(p_{j^*}(X) \leq \alpha) = \alpha.$$

Choice of F_X was arbitrary.

Conclusion: the (worst-case) probability of approving a treatment with no benefit is no more than α!

All outcomes

What is the FDA policy?

Exactly the “naive” approach:

- For each of M outcomes, test for no difference in the mean across treatment and control groups
- Report raw p-values
- Approve if all M raw p-values are $\leq \alpha$.

We have shown that this still controls the proportion of approvals when reviewing N_0 treatments with no benefit (N_0 big).

All outcomes

What about power?

Suppose that under a specific F_X where none of H_{01}, H_{02}, H_{03} are true:

- $p_1(X), p_2(X), p_3(X)$ are mutually independent
- $\mathbb{P}_{F_X}(p_1(X) \leq \alpha) = 0.8, \mathbb{P}_{F_X}(p_2(X) \leq \alpha) = 0.9, \mathbb{P}_{F_X}(p_3(X) \leq \alpha) = 0.95.$

(The idea is that we have chosen the sample size to attain 80% power against H_{01}.)

Then, the probability of approving this (beneficial) treatment is:

$$\mathbb{P}_{F_X}(p_1(X) \leq \alpha, \dots, p_3(X) \leq \alpha) = \prod_{j=1}^{3} \mathbb{P}_{F_X}(p_k(X) \leq \alpha) = 0.8 * 0.9 * 0.95 = 0.684$$

All outcomes

Should we do anything about this?

Recall that when H_{0j^*} is true (so that the treatment has no clinical benefit), we have

$$\mathbb{P}_{F_X}(p_1(X) \leq \alpha, \ldots, p_M(X) \leq \alpha) \leq \mathbb{P}_{F_X}(p_{j^*}(X) \leq \alpha) = \alpha.$$

This bound is only tight when there is one endpoint with no treatment effect, and the rest have infinitely large treatment effects.

Often clinically implausible! Feels like there's room to "safely" reject when all p-values are below α^*, for $\alpha^* \geq \alpha$. This would offset power loss.

All outcomes

FDA's stance

> **There have been suggestions that** α for each co-primary endpoint could be increased from 0.05 to accommodate the loss in statistical power… This is not acceptable because doing so may undermine the ability to interpret a treatment effect on each disease aspect considered critical to show that the drug is effective in support of approval.

My interpretation:

- It's clinically important to not just know that the treatment is effective, but to *also understand the treatment effect*.
- For this we look to the M $100(1-\alpha)$% confidence intervals, and 95% confidence intervals are de rigueur.

≥ 1 outcomes, naive approach

Suppose that we have M outcomes, and are interested in testing M null hypotheses of no difference in a pre-specified functional (e.g. the mean) across groups.

The naive idea: Approve the treatment if **any** of the M p-values are below α.

Your turn: Do you think this policy controls the proportion of approving treatments with no benefit?

- Remember: benefit now means “has an effect on at least one outcomes”
- Turn it around: what does “no benefit” mean?

≥ 1 outcomes, naive approach

Statistical setup

Let $X \sim F_X$ represent the data with which we test $H_{0j} : T_j(F_X) = 0$ for all $j = 1, 2, \ldots, M$.

Let $p_j(x)$ be the p-value function for testing H_{0j}. Assume that it's valid and based on a pivot:

$$\mathbb{P}_{F_X}(p_j(X) \leq \alpha) = \alpha, \quad \text{for all } F_X \in \mathcal{F}_{0j} \equiv \{F : T_j(F_X) = 0\}.$$

"Benefit" means F_X satisfies at least one of $H_{11}, \ldots, H_{1M}$. So the (worst-case) probability of approving a treatment with no benefit is:

$$\sup_{F_X \in \bigcap_{j=1}^{M} \mathcal{F}_{0j}} \mathbb{P}_{F_X}\left(\bigcup_{j=1}^{M} \{p_j(X) \leq \alpha\}\right)$$

≥ 1 outcomes, naive approach

Statistical setup

Let $F_X \in \bigcap_{j=1}^{M} \mathcal{F}_{0j}$. **Assume that** $p_j(X)$ are mutually independent for $X \sim F_X$. Then,

$$\mathbb{P}_{F_X}\left(\bigcup_{j=1}^{M}\{p_j(X) \leq \alpha\}\right) = 1 - \mathbb{P}_{F_X}\left(\bigcap_{j=1}^{M}\{p_j(X) \geq \alpha\}\right)$$

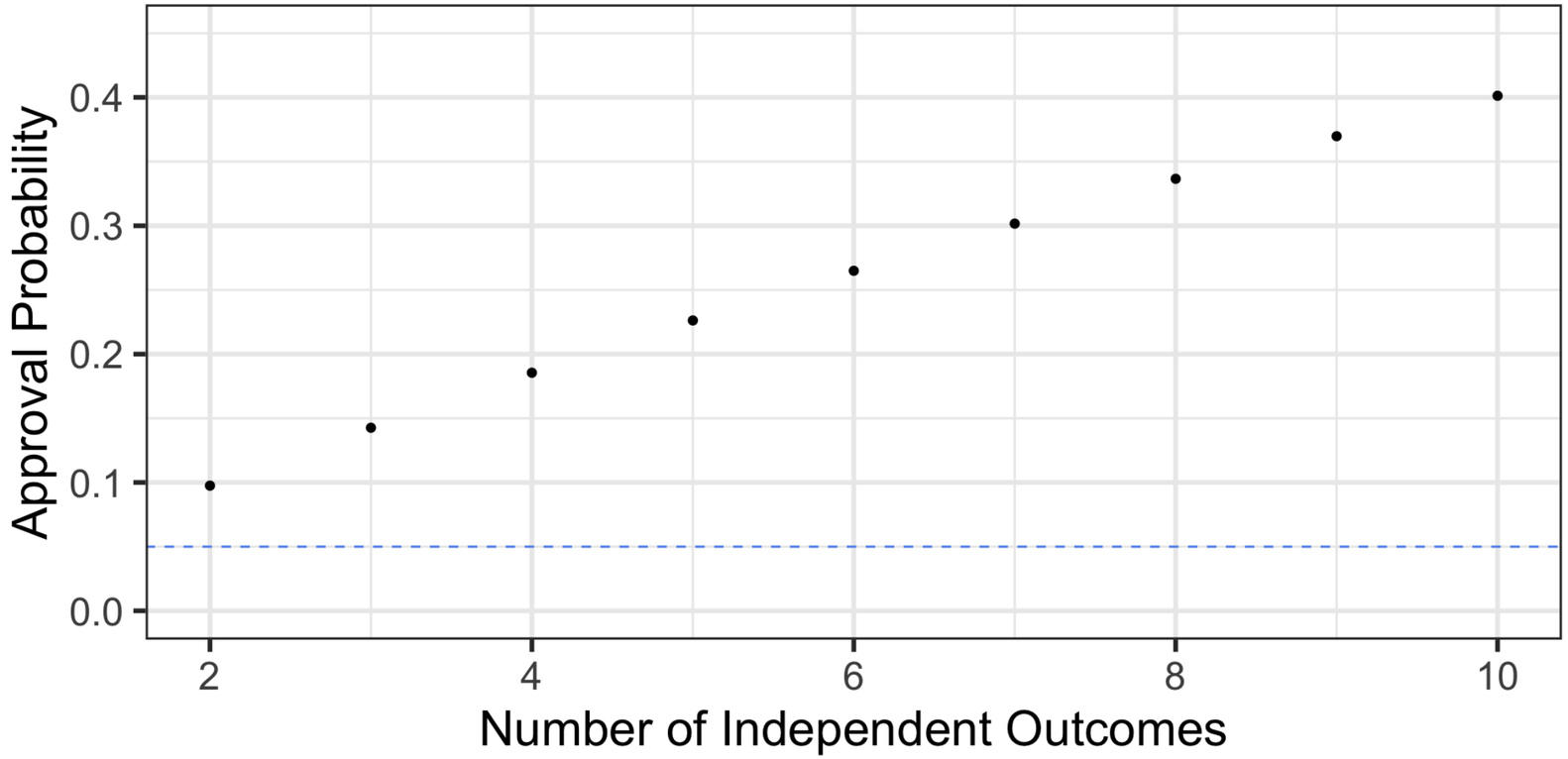
$$\overset{ind}{=} 1 - \prod_{j=1}^{M} \mathbb{P}_{F_X}(p_j(X) \geq \alpha) = 1 - (1-\alpha)^M$$

Under independent outcomes, the (worst-case) probability of approving a treatment with no benefit is BIGGER than α!

≥ 1 outcomes, naive approach

How bad can it get? (independent outcomes)

Let $F_X \in \bigcap_{j=1}^{M} \mathcal{F}_{0j}$. **Assume that** $p_j(X)$ are mutually independent for $X \sim F_X$.



≥ 1 outcomes, naive approach

Suppose that we have M outcomes, and are interested in testing M null hypotheses of no difference in a pre-specified functional (e.g. the mean) across groups.

The naive idea: Approve the treatment if **any** of the M p-values are below α.

If the M outcomes are mutually independent, then over N_0 treatments with no benefit reviewed, naive policy would approve much more than $0.05N_0$ of them!

- Note that if outcomes are positively correlated, then we can expect this effect to be less dramatic … more on this on Thursday!

≥ 1 outcomes, naive approach

Another problem: labelling

It's clinically important to not just know that the treatment is effective, but to *also understand the treatment effect*.

- The treatment is effective against at least one outcome. *Which one?*
- The naive idea: Approve the treatment being labelled as being effective for outcome j if the j th p-value is $\leq \alpha$.

New idea: Even a single false statement about efficacy in a single trial is a catastrophe.

- The naive approach also does not control the proportion of making at least one false statement about efficacy.

≥ 1 outcomes, naive approach

Statistical formulation

We can only make a false statement about efficacy when $F_X \in \bigcup_{j=1}^{M} \mathcal{F}_{0j}$.

So, worst case probability of making ≥ 1 false statement about efficacy under the naive approach:

$$\sup_{F_X \in \bigcup_{j=1}^{M} \mathcal{F}_{0j}} \underbrace{\mathbb{P}_{F_X}\left(\bigcup_{j=1}^{M} \{p_j(X) \leq \alpha\}\right)}_{\text{Family wise error rate}}$$

This is bigger than the worst case probability of approving a treatment with no benefit, which the naive approach doesn't control!

≥ 1 outcomes

So what's the FDA policy?

- FDA requires applicants to report the M p-values and the results of an *adjustment* procedure that takes those M p-values as input, and outputs which H_{0j} to reject.

Procedure must control the family wise error rate at level α.

- This means that the probability of falsely rejecting at least one null hypothesis must be below α, regardless of which and how many outcomes the treatment has no effect on.
- Keeps the proportions of two types of catastrophes low: (1) Approving non-beneficial treatments (2) labelling a treatment as beneficial for an outcome when it isn't

≥ 1 outcomes

Do we lose power by adjusting? (Senn and Bretz, 2007)

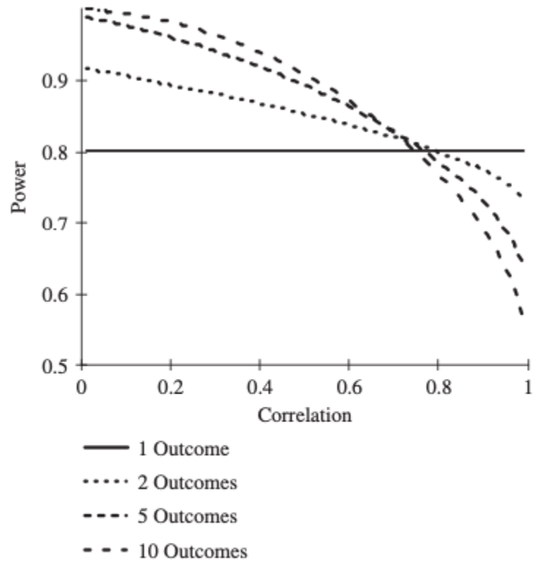


Figure 1. Disjunctive power for trials with multiple equally correlated endpoints where the Bonferroni adjustment has been applied and the power for each individual test at the 2.5% level for a one-sided (unadjusted) test would be 80%.

- The power against each of the **individual** M outcomes will be lower than the power of an unadjusted test
- But the power to detect an effect on **at least one outcome** need not be lower than the power of an unadjusted test
- When your outcomes are independent or moderately correlated, may in fact be higher than than the power of an unadjusted test

Family wise error rate (FWER)

Let $X \sim F_X$ represent the data with which we test $H_{0j} : T_j(F_X) = 0$ for all $j = 1, 2, \ldots, M$.

Let $\text{Reject}(p_1, \ldots, p_M) \subseteq \{1, 2, \ldots, M\}$ be a procedure mapping p-values for H_{0j} to indices to be rejected.

Let $R(X) = |\text{Reject}(p_1(X), \ldots, p_M(X))|$. Then, the family wise error rate is:

$$\mathbb{P}_{F_X}(R(X) \geq 1).$$

We say that the procedure $\text{Reject}(\cdot)$ has (strong) control of the FWER if:

$$\sup_{F_X \in \bigcup_{j=1}^{M} \mathcal{F}_{0j}} \mathbb{P}_{F_X}(R(X) \geq 1) \leq \alpha, \quad \text{for all } 0 \leq \alpha \leq 1.$$

Bonferroni correction

The idea: Reject H_{0j} using X if $p_j(X) \leq \frac{\alpha}{M}$, for all $j = 1, 2, \ldots, M$.

Corresponds to the following rule:

$$\text{Reject}(p_1, \ldots, p_M) = \left\{ j : p_j \leq \frac{\alpha}{M} \right\}.$$

Guarantees FWER control, so long as p-values are valid.

- Doesn't matter what the dependence structure between $p_1(X), \ldots, p_M(X)$ is
- Straightforward proof

FWER bound for Bonferroni

Let $F_X \in \bigcup_{j=1}^{M} \mathcal{F}_{0j}$. Assume for simplicity that p-values are based on a pivot.

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}_{F_X}(R(X) \geq 1) \\
&= \mathbb{P}_{F_X}\left(\bigcup_{j:F_X \in \mathcal{F}_{0j}} \{j \in \text{Reject}(X)\} \right) \\
&\leq \sum_{j:F_X \in \mathcal{F}_{0j}} \mathbb{P}_{F_X}\left(j \in \text{Reject}(X)\right) \\
&= \sum_{j:F_X \in \mathcal{F}_{0j}} \mathbb{P}_{F_X}\left(p_j(X) \leq \alpha/M\right) \\
&= \sum_{j:F_X \in \mathcal{F}_{0j}} \alpha/M = \frac{\alpha|\{j : F_X \in \mathcal{F}_{0j}\}|}{M} \leq \frac{\alpha M}{M} = \alpha.
\end{aligned}
$$

Insights from Bonferroni bound

Let's first look at the first inequality (union bound):

$$\mathbb{P}_{F_X}\left(\bigcup_{j:F_X\in\mathcal{F}_{0j}}\{j\in\text{Reject}(X)\}\right)\leq\sum_{j:F_X\in\mathcal{F}_{0j}}\mathbb{P}_{F_X}\left(j\in\text{Reject}(X)\right)$$

Assumption free; tight when events in $\{j\in\text{Reject}(X)\}_{F_X\in\mathcal{F}_{0j}}$ are disjoint.

- Loose if p-values are independent
- Even looser if p-values are positively dependent
- Tighter if p-values are negatively dependent, tight if perfectly negatively dependent

Insights from Bonferroni bound

Let's look at the second inequality now:

$$\frac{\alpha|\{j : F_X \in \mathcal{F}_{0j}\}|}{M} \leq \frac{\alpha M}{M} = \alpha$$

Tight when $|\{j : F_X \in \mathcal{F}_{0j}\}| = M$, i.e. F_X satisfies *all* nulls. Otherwise loose:

# True Nulls	**1**	**2**	**3**	**4**
LHS	0.0125	0.0250	0.0375	0.0500
RHS	0.05	0.05	0.05	0.05

Note: If we knew the number of true nulls M_0, then we could safely divide by M_0 instead of M.

An idea for improving Bonferroni

Suppose $p_1(x) \leq \alpha/M$. Let $M_0 = |\{j : F_X \in \mathcal{F}_{0j}\}|$. Consider two cases:

1. If H_{01} false, then $M_0 \leq M - 1$; could safely divide by $M - 1$ instead of M for $p_2(x), \ldots, p_m(x)$
2. If H_{01} true, then dividing by $M - 1$ instead of M for $p_2(x), \ldots, p_m$ may not give a valid bound. **But who cares**? $R(x) \geq 1$ no matter what we do with the rest of the p-values …

Key idea:

- FWER control doesn't care how many false rejections we make.
- Once we make a false rejection, we're dead to FWER.
- So let's just make our rejections, and then behave as if all of those rejections are correct!

Holm's method

The idea: Order p-values $p_{(1)} \leq \ldots \leq p_{(M)}$:

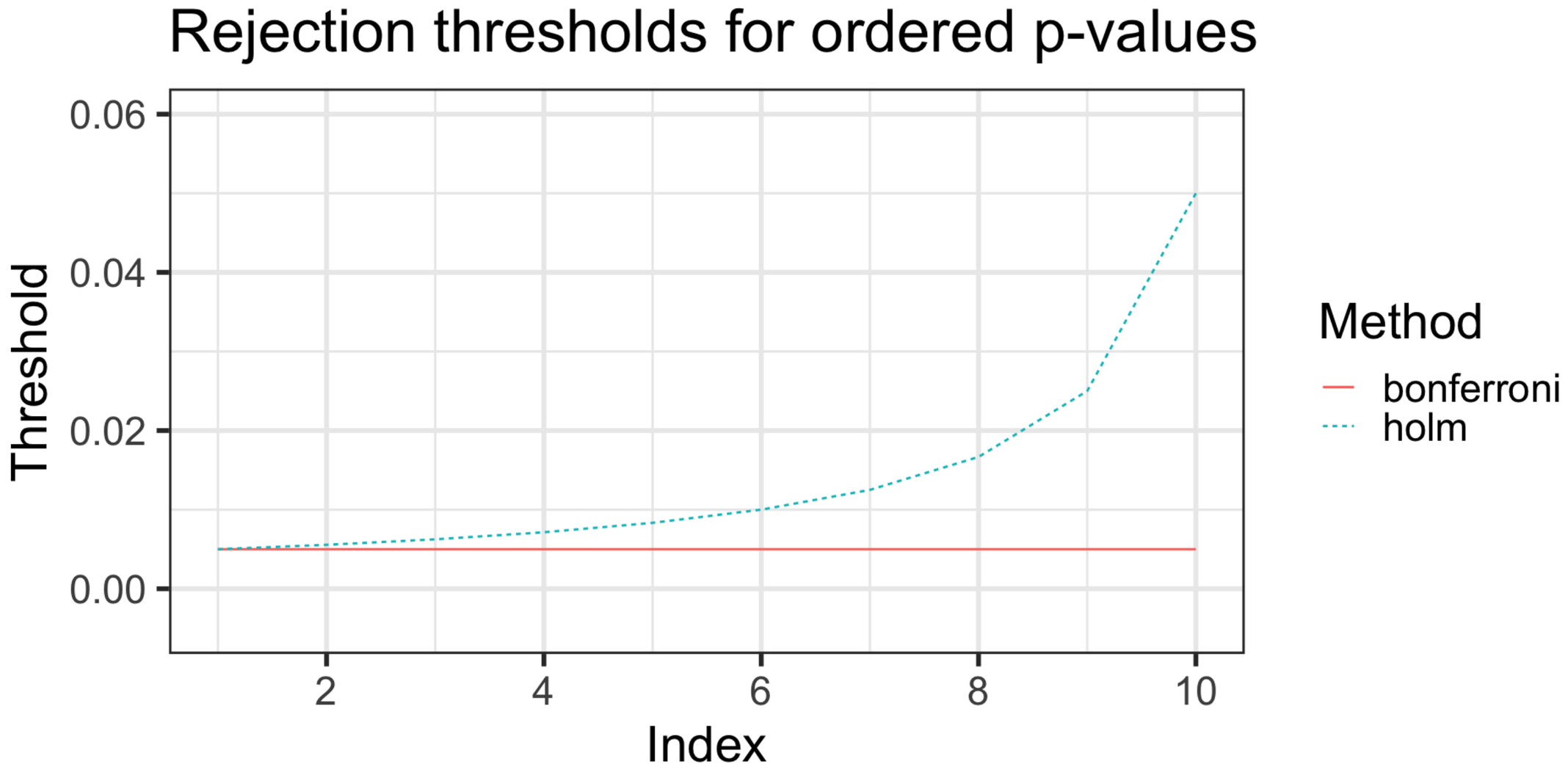
- If $p_{(1)} \leq \alpha/M$, reject $H_{0(1)}$; otherwise stop
- If $p_{(2)} \leq \alpha/(M-1)$, reject $H_{0(2)}$; otherwise stop
- Keep going, one p-value at a time

Corresponds to the following rule:

$$\text{Reject}(p_1, \ldots, p_M) = \left\{ j : p_j \leq \min\{p_{(j)} : p_{(j)} > \alpha/(M+1-j)\} \right\}.$$

Still guarantees FWER control, so long as p-values are valid.

Holm is always more powerful



Adjusted p-values

The idea: Convert $p_1, \ldots, p_M$ to $\tilde{p}_1, \ldots, \tilde{p}_M$ so that

$$\text{Reject}(p_1, \ldots, p_M) = \{j : \tilde{p}_j \leq \alpha\}$$

See e.g. `stats::p.adjust()`.

More on this topic next week!

Back to the beginning …

In the context of clinical trials and the FDA: does multiple testing cause problems? What problems? Do we need to "adjust for multiple testing"? How should we adjust?

- If scientifically reasonable to think that the treatment needs to have an effect on all outcomes, then **no major problems** and **no adjustment needed**.
- If scientifically reasonable to think that the treatment just needs to have an effect on one or more outcomes, then **the catastrophe proportion is no longer guaranteed to be low**, and **we need to adjust in a way that controls FWER**.

> **"I don't know, it depends on the scientific problem at hand."** - My annoying partner, whenever I ask him ANY question involving statistics.

What about other scientific contexts? We will see one more next week …

What about confidence intervals?

> **The emphasis of this guidance is not on the confidence interval, but rather on the test of a hypothesis, where the issue is whether a treatment effect on a particular endpoint exists at all. Although confidence intervals are also critical to the interpretation of an effect when one exists, determining the confidence interval with some of the statistical methods for managing multiplicity described in section IV is complex.** - FDA, Multiple Endpoints in Clinical Trials: Guidance for Industry

- FWER definition applied to confidence intervals: probability of at least one interval not covering
- We will discuss this more when we get to selective inference
- Is this relevant to the science at hand? Probably, but FDA punts, and so does the EMA …