STATS 300C: Theory of Statistics III

Lecturer: Professor Emmanuel Candès

Notes by: Andrew Lin

Spring 2025

Introduction

This is the last course in the STATS 300 series. All of the course information is on Canvas, but we'll do a bit of explaining what the class is about. STATS 300A and 300B are listed prerequisites, but those courses are typically on "classical statistics" (from 40 or 50 years ago), and there's been a tremendous change in the last decade or so. This course basically addresses the changes that have happened.

One big change during Professor Candés' lifetime is that instead of testing for a single hypothesis (separating signal from noise), technology now lets us look at a lot of stuff at once, and we may have many noisy signals to study. So we'll be exposed to modern ideas in statistical theory – the first month will be on large-scale hypothesis testing and inference, and then we'll discuss even more modern topics as we progress. Specifically, we'll consider testing problems in high dimensions and multiple testing problems (leading to the false discovery rate theorem), and we'll see how people address this in difficult scenarios. We'll then consider e-values instead of p-values, do some conditional testing, conformal inference, and so on. The point is "user-friendly theory" that can be useful in practice (which we should know about if we want to function as a professional statistician).

Because we're covering modern topics, we can't really point to a particular textbook (only tangential references), but the course has been offered in the past and there are some past lecture notes available. Slides will be on Canvas as well.

In terms of grading, homework assignments will be distributed on a roughly weekly basis (on Thursdays) – collaboration is encouraged but we should write up solutions on our own and cite sources appropriately. And there will be a final project at the end of the course (which will basically be a take-home exam).

Fact 1

Unless otherwise specified, all images and figures in this document are either taken from the lecture slides or drawn by hand. Also, I was unable to attend Lectures 16 and 18 in person, and so the material there was constructed from course lecture slides and some notes from classmates.

1 April 1, 2025

Today's lecture will focus on **global testing**, particularly building toward Bonferroni's method and Fisher's combination test.

Example 2

One setting in which we might want to do multiple hypothesis testing is in a biological setting: suppose we have n genes (for humans, $n \approx 20000$), and we have data about expression levels for each gene among healthy and sick patients. (This wasn't possible a few decades ago, so technology has really evolved!) Specifically, we have m_0 healthy patients (say a few hundred) and m_1 sick patients (say a hundred), and say that gene i is expressed at level $Y_{ij}^{(0)}$ for the jth healthy patient and $Y_{ij}^{(1)}$ for the jth sick patient. Our goal is to understand which genes are differentially expressed between these two populations (so that we can understand for example how to treat the disease).

Notice that the number of people in the study here is relatively low compared to the number of variables being studied. Formally, we may consider a null hypothesis of the form

$$H_{0,i} : \mathbb{E}\left[Y_{ij}^{(0)}\right] = \mathbb{E}\left[Y_{ij}^{(1)}\right],$$

or perhaps the stronger hypothesis of equality in distribution

$$H_{0,i} : Y_{ij}^{(0)} \stackrel{d}{=} Y_{ij}^{(1)}.$$

In classical statistics, we can construct test statistics to test each of these. For example, we could use a **two-sample t-test**, or if we don't believe in the central limit theorem in the latter case we can do a **permutation test** – this is applicable because we only need to assume **exchangeability** and no other information.

Definition 3

Two random variables X, Y are **exchangeable** if the distributions of (X, Y) and of (Y, X) are identical, or equivalently $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = y, Y = x)$. And we can extend this to finitely many random variables by saying that we have equality in distribution under any permutation, so $\mathbb{P}(\vec{X} = \vec{x}) = \mathbb{P}(\vec{X} = \pi\vec{x})$.

Note that iid random variables are exchangeable, but exchangeable random variables need not be iid – or example, consider $(X, -X)$ for X standard normal, or the top two cards in a shuffled deck of cards.

Fact 4

Here's a "cool math problem": I take a shuffled deck of cards and reveal them from the top one by one, and then you can tell me when to stop. At that point, if the next card is red then you win \$100, and otherwise you lose \$100. The question is whether there is a way to say "stop" so that the expected gain is positive; this turns out to do with exchangeability.

Since we have many variables to consider at once, we may care about the **global null** hypothesis

$$H_0 = \bigcap_{i=1}^{n} H_{0,i}.$$

We'll assume (because we've run our permutation test) that we have p-values for each $H_{0,i}$. Here we'll assume that our p-values have the **super-uniform** property that (here p_i is a random variable coming out of the test)

$$\mathbb{P}(p_i \leq t) \leq t$$

under the null hypothesis. (In fact we'll often assume p_i is uniform on $[0,1]$ for simplicity, but the formal definition is that it's a random variable stochastically dominated by a uniform $[0,1]$.) So our goal is now to combine those p-values together and answer the global question of "is there anything interesting happening."

Definition 5

Let α be some desired significance level (for example 0.05), and suppose we have n different hypotheses. **Bonferroni's global test** rejects the global null if

$$\min_i p_i \le \frac{\alpha}{n}.$$

This test is used thousands of times in leading medical literature.

Proposition 6

If the p-values are super-uniform then Bonferroni's procedure has **size** (that is, chance of Type I error) at most α.

Indeed, the probability of rejecting the null under the null hypothesis is, by a union bound and the super-uniform property

$$\begin{aligned}
\mathbb{P}(\text{reject}) &= \mathbb{P}\left(\bigcup_{i=1}^{n}\left\{p_i \le \frac{\alpha}{n}\right\}\right) \\
&\le \sum_{i=1}^{n}\mathbb{P}\left(p_i \le \frac{\alpha}{n}\right) \\
&\le n \cdot \frac{\alpha}{n} \\
&= \alpha.
\end{aligned}$$

Notice that this does not require any independence assumptions, and in fact if we assume p-values are uniform and independent then as $n \to \infty$ the probability of the type I error is $q(\alpha) = 1 - e^{-\alpha}$, which is very close to α. Here we're doing the calculation

$$\begin{aligned}
\mathbb{P}(\text{no reject}) &= \mathbb{P}\left(\bigcap_{i=1}^{n}\left\{p_i \ge \frac{\alpha}{n}\right\}\right) \\
&= \left(1 - \frac{\alpha}{n}\right)^n \\
&\approx e^{-\alpha} \\
&\approx 1 - \alpha + \frac{\alpha^2}{2} - \cdots.
\end{aligned}$$

So there isn't much room to "change" the Bonferroni method for improvement unless we have correlations, and thus a lot of studies indeed use this threshold.

Fact 7

Graphically, we can visualize what Bonferroni "looks for" by taking our n p-values and sorting them from smallest to largest.

Under the null hypothesis and assuming the p-values are uniform and independent, we expect these p-values to "hug the line" $y = x$ (in fact these order statistics follow beta distributions, so we can explicitly calculate the expected

value and variance). But if we're not under the global null, we will not "hug the line" anymore – sometimes (1) these sorted p-values will have a few that are extremely small and then the rest generally following the line (meaning maybe five hypotheses have very strong signals), and sometimes (2) all of the p-values will be overall deflated (too small). Bonferroni's method is really most useful in case (1), because it will successfully reject due to the very small p-values; it will not reject the null in case (2) and thus is not a very powerful test. Thus we should use it when **we expect strong effects from individual tests** – we call this scenario **sparse alternatives**. In the other case, we can use something different:

Definition 8

Fisher's combination test rejects the global null using a more combined measure. Specifically, we consider the test statistic

$$T = -\sum_{i=1}^{n} 2 \log p_i,$$

and we reject the null if T is large.

This method is actually very frequently used in meta-analysis to combine different studies. The idea is the following:

Proposition 9

Assume that $p_1, \cdots, p_n$ are **independent** and uniform (for example in meta-analysis, suppose we do not have overlapping patients among the studies). Then under the null hypothesis, we have $T \sim \chi^2_{2n}$ (that is, the chi-square distribution with $2n$ degrees of freedom).

Proof. Note that for p_i uniform on $[0, 1]$, $-\log p_i$ is a standard exponential random variable and therefore $-2 \log p_i = 2E$ (which is the chi-square distribution χ^2_2 with 2 degrees of freedom). And the sum of independent chi-square distributions is another chi-square distribution. □

Fisher's combination test is then more powerful in case (2) where we have lots of p-values that are too large rather than a few isolated outliers against the null (which might get absorbed into fluctuations).

Example 10

People often think Bonferroni's method is naive, but it really isn't so naive – there are cases where it is the right thing to do. We want to do power calculations, so we'll specify some distributions. Assume Y_is are distributed independently as $N(\mu_i, 1)$ for $1 \le i \le n$, and we are interested in the n hypotheses $H_{0,i} : \mu_i = 0$. The global null then asserts that (two-sided) all means are zero or (one-sided) that all means are nonpositive, so that under the alternative there is some μ_i which is (two-sided) nonzero or (one-sided) positive.

In the case of sparse alternatives, the alternative hypothesis means that a few of the means are nonzero and the rest are zero, and our goal is to discover this by data analysis. Bonferroni's method would then set the rejection threshold in the one-sided setting if $\max Y_i \ge z\left(1 - \frac{\alpha}{n}\right)$, where $z() = \Phi^{-1}()$ denotes the quantile of the Gaussian. (Indeed, our p-values in this case are $p_i = \Phi(-y_i)$; since Φ is monotone, asking for $p_i \le \frac{\alpha}{n}$ is equivalent to asking $y_i \ge \Phi^{-1}\left(1 - \frac{\alpha}{n}\right)$.) And in the two-sided setting, we replace $\frac{\alpha}{n}$ with $\frac{\alpha}{2n}$.

Given α, n, we can ask a computer to tell us where this threshold actually lies, but mathematically for n large it is roughly $\sqrt{2 \log n}$ (notice that to leading order this doesn't depend on α) – we get this from the expression for the tail of a Gaussian, and more precisely we're saying that if Y_is are iid standard normal we have

$$\frac{\max Y_i}{\sqrt{2 \log n}} \xrightarrow{p} 1.$$

So at this crude rescaled level, the distribution of the maximum becomes deterministic, so it makes sense to set the threshold at roughly this point.

Fact 11

It turns out the distribution of the maximum of iid standard Gaussians is well-approximated as

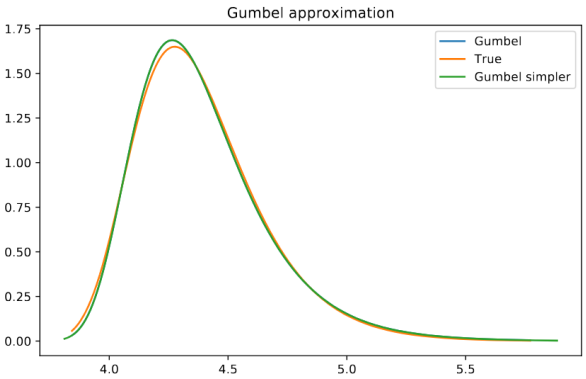
$$\max Y_i \approx \mu_n + \sigma_n G,$$

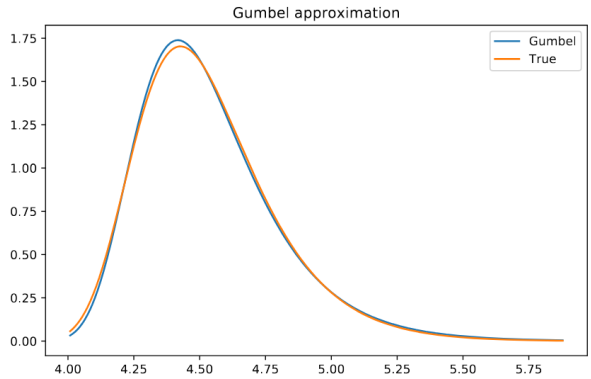
where G is the **Gumbel distribution** (with cdf $\mathbb{P}(G \le x) = \exp(-\exp(-x))$ for $x \in \mathbb{R}$) and

$$\mu_n = z\left(1 - \frac{1}{n}\right), \quad \sigma_n = z\left(1 - \frac{1}{en}\right) - z\left(1 - \frac{1}{n}\right).$$

(The Gumbel distribution often comes up in extreme value statistics like this one, even beyond Gaussians.) And if we replace $\max(Y_i)$ by $\max(|Y_i|)$, we get the same approximation but with $2n$ in place of n.

Note that the mean of the Gumbel distribution is not zero – it's a skew distribution and its mass is roughly concentrated between $[-2, 6]$. And what we're saying is that a shifted, rescaled Gumbel (by some factors related to Gaussian quantiles) gives us the maximum of Gaussians. For size $n = 10^5$, these approximations are quite good (for both the maximum and the maximum of the absolute value):





We can recover this result with the following more precise calculation:

Proposition 12

Let $M_n = \max(Y_1, \cdots, Y_n)$ for Y_i standard Gaussian. Then

$$\mathbb{P}(M_n \le \sqrt{2\log n - \log\log n + z}) = \exp\left(-\frac{\exp(-z/2)}{2\sqrt{\pi}}\right)(1 + o(1)).$$

In particular, this implies the $\sqrt{2\log n}$ convergence in probability from before, and it also gives us some corrective terms.

So we understand now that under the global null, Bonferroni computes the maximum (which concentrates around $\sqrt{2\log n}$ with some fluctuations). And if the Y_is are $N(\mu_i, 1)$ and we don't know which μ_i is positive, we will have power if and only if the mean is above that Bonferroni level – that is, we will be able to spot the "needle in a haystack" if $\mu_i > \sqrt{2\log n}$. If we do a **power plot** where the x-axis plots the nonzero mean $\frac{\mu_i}{\sqrt{2\log n}}$ and the y-axis plots the power, then there is a sharp "phase transition' where the power is very low (α-ish, or more precisely the size $q(\alpha)$ of the test) to the left of 1 and very high (close to 1) to the right of 1. In words, we call this "asymptotic full power above threshold" and "asymptotic powerlessness below threshold" (since we can obtain the same level and power by

just flipping an α-coin); this is strange compared to the usual "smooth transitions" that we might expect in classical statistics.

The question we'll ask next time is then the following: suppose we have a few signals that are "below 1" on this scale, meaning that those nonzero means are (for example) $0.9\sqrt{2\log n}$. Then we want to ask whether there is **any** test which has some power (in other words, whether we can do better than flipping a coin to detect those signals). And the answer turns out to be **no**, so in fact Bonferroni is as good as we can get for this "sparse alternative" model.

2 April 3, 2025

We discussed Bonferroni and Fisher's tests last time – in particular, we considered the "needle in a haystack" problem where one of n means may be nonzero and we want to distinguish whether that indeed occurs. Bonferonni then cares about whether $\max_i Y_i$ (or $\max_i |Y_i|$ if two-sided) is large, and we arrived at the "threshold phenomenon" conclusion that if the "size of the needle" is $h = \sqrt{2r\log n}$, then for $r < 1$ Bonferroni asymptotically has no power, while for $r > 1$ it has full power. We were then wondering whether there is some α-level test with some nontrivial power for $r = 1 - \varepsilon$ (that is, whether we can do better than coin-flipping).

The answer, given by Ibragimov and Hasminski, is **no** – the total variation distance between the hypotheses approaches zero. If we have to do something like this in our research, the crux of the matter is to use the following setup where we know the optimal answer. The point is to avoid having a composite alternative: instead, consider the Bayesian decision problem with null hypothesis

$$H_0 : \mu_i = 0 \text{ for all i}$$

and a "simple" alternative hypothesis

$$H_1 : \{\mu_i\} \sim \pi,$$

where π selects a coordinate uniformly at random and sets its mean to $\mu^{(n)}$, keeping all other means to zero.

The most powerful test in such a setting where H_0, H_1 are both **simple hypotheses** is the **likelihood ratio test**, and if we can show it has no power then everything else will have no power.(This goes under the name **Neyman-Pearson**.) Indeed, we have likelihoods

$$f_0(y) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right), \quad f_1(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i-\mu)^2\right) \prod_{j:j\neq i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right).$$

and we want to reject when $\frac{f_1}{f_0}$ is large. What's nice is that all the terms cancel out except the shifted mean:

$$L = \frac{1}{n}\sum_{i=1}^{n} \exp\left(Y_i\mu - \frac{1}{2}\mu^2\right).$$

Notice that this is different from Bonferroni – it's a softmax instead of looking at the maximum, since if $\mu = \infty$ this would be completely dominated by the maximum Y_i. This is nice because it's just a sum of iid terms of mean 1, and in fact if μ is small enough this likelihood concentrates:

Proposition 13

Under the null hypothesis, if $\mu^{(n)} = (1-\varepsilon)\sqrt{2\log n}$, then $L \to 1$ in probability as $n \to \infty$.

Note that we **can't apply the central limit theorem** to get this result – it's not true that the variance of the likelihood ratio goes to 0, since this sum is dominated by a few terms, namely the large exponentials (so things like

Berry-Esseen don't hold). So there are very rare events where L is very large, but in probability we still converge to 1.

But this fact means that if we set thresholds $T_n(\alpha)$ so that $\mathbb{P}_{H_0}(L \geq T_n(\alpha)) = \alpha$, we see that the probability of a type II error is asymptotically $1 - \alpha$. Indeed, we can do a change of measures

$$\begin{aligned}\mathbb{P}(\text{type II error}) = \mathbb{P}_{H_1}(L \leq T_n(\alpha)) &= \int 1\{L \leq T_n(\alpha)\} dP_{H_1} \\ &= \int L \cdot 1\{L \leq T_n(\alpha)\} dP_{H_0} \\ &= \int 1\{L \leq T_n(\alpha)\} dP_{H_0} + \int (L-1) 1\{L \leq T_n(\alpha)\} dP_{H_0} \\ &\to \alpha + 0,\end{aligned}$$

where we can use the bounded convergence theorem on the latter term because the truncation $(L-1)1\{L \leq T_n(\alpha)\}$ implies that this random variable is bounded, and $L - 1 \to 0$ in probability.

So the power of Neyman-Pearson in this setting is also equal to the power of Bonferroni's test, and this is not that surprising because L is dominated by a few "largest terms." Restated, we see that the optimal test satisfies

$$\mathbb{P}(\text{type I error}) + \mathbb{P}(\text{type II error}) \to 1$$

as $n \to \infty$, and thus we can't do better than coin-tossing – more formally, for **any** test for this needle-in-a-haystack setup, we have

$$\liminf \left(\mathbb{P}_{H_0}(\text{type I error}) + \sup_{H_1} \mathbb{P}(\text{type II error}) \right) \geq 1.$$

(If we set the type I error to any arbitrary α, then the type II error must be $1 - \alpha$ – we cannot ever come up with a better power curve than Bonferroni for this problem.)

Example 14

We'll now turn to the other test, and our goal is now to understand what kind of alternatives it's good for (the "case (2)" in our discussion last lecture). Recall that we want to analyze the sum of the values $-2 \log 2\Phi(|y|)$; for simplicity we're going to instead **replace those with y^2 in this discussion** (which has a very similar curve shape and is used in settings like **analysis of variance** (ANOVA)).

So **instead of Fisher's combination test, we're now looking at the χ^2 test**. Our model is then, as before, that $Y_i = \mu_i + z_i$ for iid standard normals z_i, and the null hypothesis is that all μ_i are zero. The χ^2 statistic is then

$$T = \sum_{i=1}^{n} Y_i^2 = ||Y||^2$$

Under the global null, this is a chi-square with n degrees of freedom, and under the alternative this will be distributed differently (as a non-central chi-square). Specifically, we have the central limit theorem approximation under H_0

$$\frac{T - n}{\sqrt{2n}} \sim N(0,1) \implies \chi^2_n(1-\alpha) = n + \sqrt{2n} z(1-\alpha),$$

where $\chi^2_n()$ and $z()$ denote the corresponding quantiles. Meanwhile, under the alternative hypothesis we have

$$T = \sum_{i=1}^{n} (\mu_i + z_i)^2, \quad \mathbb{E}\left[(\mu_i + z_i)^2\right] = \mu_i^2 + 1, \quad \text{Var}\left[(\mu_i + z_i)^2\right] = 4\mu_i^2 + 2$$

and thus we again have a central limit theorem approximation under H_1

$$\frac{T-(n+||\mu||^2)}{\sqrt{2n+4||\mu||^2}} \sim N(0,1).$$

So the mean is shifted a bit, and the variance is also slightly increased, and all that matters about our alternative means (in this approximation) is the sum of the squares of the means – in fact, $||\mu||^2$ is a sufficient statistic for the non-central χ^2. That means we have our z-score under the null

$$Z = \frac{T-n}{\sqrt{2n}},$$

and defining the parameter $\theta = \frac{||\mu||^2}{\sqrt{2n}}$, we see that our approximation reads

$$H_0 : Z \sim N(0,1), \quad H_1 : Z \sim N\left(\theta, 1+\frac{\theta}{\sqrt{n/8}}\right).$$

So if θ is around 0.1 or 0.2 we won't have a lot of power, but if $\theta = 2$ the power of the test is roughly $\mathbb{P}(N(0,1) > 1.65-2) \approx 66\%$. In other words, the scale at which we can see differences is only where $||\mu||^2 > \sqrt{2n}$ – that's quite large compared to something like Bonferroni! So such sparse alternatives will be drawn into the variance of the chi-square distribution, but if we see a lot of small values we will be able to detect that with χ^2. In slightly different terminology, the main parameter that mattered here was proportional to the **signal-to-noise ratio**

$$\text{SNR} = \frac{\text{signal power}}{\text{expected noise power}} = \frac{||\mu||^2}{\sigma^2 n}$$

(just with an extra factor of $\frac{\sqrt{n}}{2}$).

Remark 15. *We can check that the normal approximation for the shifted chi-square is quite good for a fairly wide range of values of θ (say between 0 and 4) even when $n = 10^4$, so our "simple story" is quite accurate in telling us what's going on.*

Example 16

We now want to ask a similar question as we did for seeing how powerful Bonferroni was: **with the absence of any information about the location of our μ_is, can we find a test with power as $\theta \to 0$?**

Again, the answer is **no**, and we do a very similar thing as before where we set up simple hypotheses. The global null is that $\mu = 0$ identically, and the alternative is that $\mu \sim \pi$ for π distributed uniformly on the sphere of radius $\rho^{(n)}$ (so that the parameter is constrained to be $\theta^{(n)} = \frac{(\rho^{(n)})^2}{\sqrt{2n}}$). We will show that if $\theta^{(n)} \to 0$, then the situation is hopeless.

Again, we do this by writing out the likelihood ratio and showing that it again goes to 1 in probability. Indeed, the conclusion is again that under H_0, if $\theta^{(n)} \to 0$ then $L \to 1$ in probability, meaning that Neyman-Pearson is no better than a coin toss.

Remark 17. *In this case, we do get a power curve in terms of θ: at $\theta = 0$ we have power α, and as θ increases we get a continuously increasing curve approaching 1. And the point is that "we can't do much better than that."*

Example 18

We can now try to compare Bonferroni with χ^2 in some simulations with $n = 10^6$ hypotheses at the level $\alpha = 0.05$. If we have four nonzero means at the Bonferroni level 5.45 (strong sparse effect), then the power of Bonferroni is roughly 94%, while the power of χ^2 is roughly 6%. On the other hand, if we have 2400 nonzero means at 1.1 (mild distributed effect), then the power of Bonferroni is about 6%, while the power of χ^2 is roughly 66%.

In the latter case, the maximum statistic likely actually comes from a null, since the maximum of the $1000000-2400$ nulls is concentrated at a higher value than the 2400 nonnulls. (So it essentially looks like the maximum of 1000000 nulls, and Bonferroni doesn't actually detect the effect.)

So each of χ^2 and Bonferroni is powerful in a different situation, and we would like a method which has the best of both worlds (without double-dipping). Instead of just running both tests separately, we can actually "bridge across sparsity" with a method due to John Tukey called **second-level significance testing** or **higher criticism**.

Example 19

Let $\hat{F}_n(t)$ be the **empirical cdf** of our n p-values $p_1, \cdots, p_n$, meaning that $\hat{F}_n(t) = \frac{1}{n}\#\{i : p_i \leq t\}$. Under the global null and assuming that p-values are uniform, we have $\mathbb{E}\left[\hat{F}_n(t)\right] = t$; furthermore under independence we have $n\hat{F}_n(t)$ binomially distributed with parameters (n, t).

One useful test statistic in this setting is the **Kolmogorov-Smirnov statistic**

$$\sup_t \left|\hat{F}_n(t) - t\right|,$$

which tells us about deviations from the expected cdf. We then reject the null if this quantity is above some critical value – in our case if we want to detect when p-values are unusually small (that is, $\hat{F}_n$ is too big), we will check whether $\sup_t(\hat{F}_n(t) - t)$ is too big.

However, note that the Kolmogorov-Smirnov statistic isn't great because we expect the fluctuations to generally be large near $t = \frac{1}{2}$. So then we generally need to set the critical value based on what happens in the bulk, and that's not where we should be looking if we have a lot of small p-values early on. Instead, we should standardize by the standard deviation and look at the **standarized** value of "how many significant tests we see at level α," considering

$$\frac{n\hat{F}_n(\alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}} = \frac{\hat{F}_n(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/n}},$$

and we use this to define the **higher criticism statistic**

$$HC_n^* = \max_{0<\alpha\leq\alpha_0} \frac{\hat{F}_n(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/n}}.$$

So we scan across significance levels and see whether there's a level at which the number of significant tests is unusually high relative to the binomial distribution. And in a paper by Donoho and Jin, this was actually used for detecting sparse heterogeneous mixtures – it turns out to be a very good statistic for bridging the different kinds of effects we've discussed, since it will be able to detect things at different αs.

3 April 8, 2025

We'll spend most of today's lecture on Tukey's second-level significance test (the higher criticism) and also see some other tests in passing. Then we'll move on to proper multiple testing and understand how to accept some and reject other hypotheses.

We left off last time by considering the test statistic HC_n^* as an attempt to be "always successful" whether we're strong and sparse (where Bonferroni is good) or mild and distributed (where chi-square is good). The idea is that we scan over different significance levels α, and when we see $\hat{F}_n(\alpha)$ particularly large, that means there's unusual behavior compared to the null hypothesis.

To study this new model, we'll move away from the needle-in-a-haystack setting here.

Example 20

Our original model was that we had independent statistics X_i which are $N(0,1)$ under the null hypothesis and $N(\mu_i, 1)$ under the alternative, but now we want to extend to a setting where we have a small fraction of non-null hypotheses. Thus we will now use a simple model where

$$H_0 : X_i \overset{\text{iid}}{\sim} N(0,1), \quad H_1 : X_i \overset{\text{iid}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\mu, 1).$$

For example if $\varepsilon = \frac{1}{n}$, this is essentially like the needle-in-a-haystack, but if ε is constant we have a growing number of non-null hypotheses. We'll end up scaling ε in a very particular way.

Given the values of ε, μ, we can write down the Neyman-Pearson test and compute likelihoods. ε encodes sparsity and μ encodes signal strength, so we can write down a 2D plot of the values

$$L = \prod_{i=1}^{n} \left((1-\varepsilon) + \varepsilon \exp\left(\mu X_i - \frac{\mu^2}{2} \right) \right).$$

It turns out that we again get a sharp transition. To illustrate this, in the literature we historically parameterize

$$\varepsilon_n = n^{-\beta}, \quad \frac{1}{2} < \beta < 1,$$

$$\mu_n = \sqrt{2r \log n}, \quad 0 < r < 1.$$

(So the needle-in-a-haystack problem has $\beta = 1, r = 1$.) Putting r on the y-axis (representing signal strength) and β on the x-axis (representing sparsity), we've already studied the point $(\beta, r) = (1, 1)$ and found that the problem is easy above that point and hard below it. We want to figure out how this generalizes to other β (since the more non-nulls we have, the easier the problem should be).

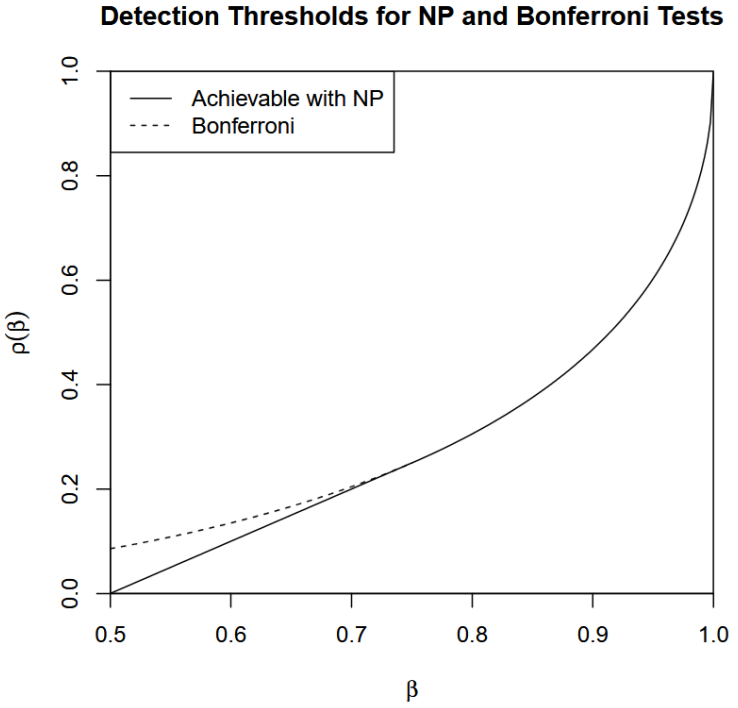
Fact 21 (Ingster '99)

It turns out that we have a threshold curve

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \le \frac{3}{4}, \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} \le \beta \le 1, \end{cases}$$

such that Neyman-Pearson has full power for $r > \rho^*(\beta)$ (that is, we can adjust the test so that the sum of type I and type II error probabilities approaches 0) and no power for $r < \rho^*(\beta)$ (that is, for any test, the limiting sum of type I and type II error is at least 1).

The plot of the curve $\rho^*(\beta)$ is shown below in the solid line – what we're saying is that above that line we have full power with Neyman-Pearson, and below that line we are powerless. And the dashed line shows the level at which Bonferroni's test can effectively distinguish the non-null hypothesis – it instead follows $(1-\sqrt{1-\beta})^2$ all the way until $\beta = \frac{1}{2}$. In other words, Bonferroni is a very good test as long as we have fewer than $n^{1/4}$ non-nulls, but it's suboptimal with less sparsity.



We can do a quick analysis to calculate the Bonferroni threshold with a crude calculation, asking for the probability that the maximum comes from a null hypothesis – if this is overwhelmingly likely, then we cannot have power. Indeed, we get power if

$$\max_{\text{non-null}} X_i \approx \sqrt{2r\log n} + \sqrt{2\log n^{1-\beta}} > \sqrt{2\log n},$$

since the $\sqrt{2r\log n}$ term is the mean of the non-nulls and we have $n^{1-\beta}$ of them. Dividing through by the $\sqrt{2\log n}$ factors, we thus see that

$$\sqrt{r} + \sqrt{1-\beta} > 1 \implies r > (1-\sqrt{1-\beta})^2.$$

What's interesting is that this actually coincides with Neyman-Pearson – Bonferroni doesn't need the values of ε and μ and still achieves the right threshold if we are sparse enough. And indeed in general, the whole point is that in global testing we do not know ε and μ and thus cannot use the NP test in the first place, so what's nice is that **Tukey's higher criticism asymptotically achieves the same threshold** without needing those parameters. Indeed,

$$HC_n^* = \max_{0<\alpha\le\alpha_0} \frac{F_n(\alpha)-\alpha}{\sqrt{\alpha(1-\alpha)/n}}$$

only needs the p-values to compute F_n. (And we should really restrict to scanning over small p-values, e.g. $\alpha_0 = 0.2$.)

Fact 22

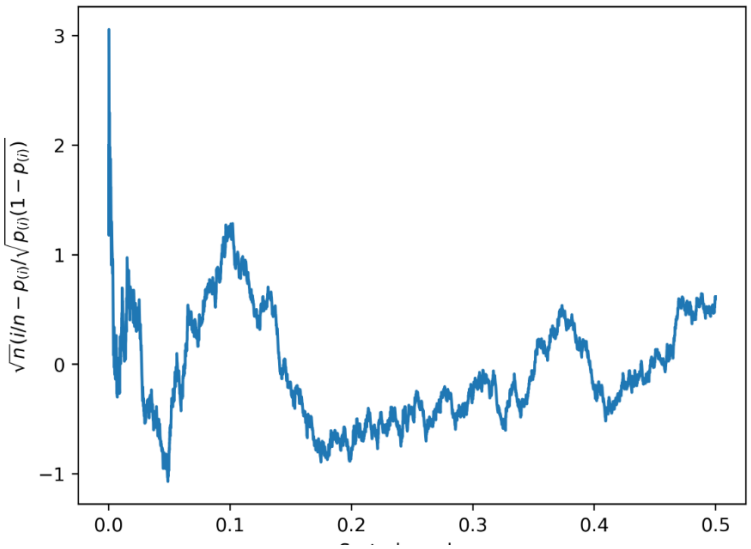
The process $\sqrt{n}(F_n(t) - t)$ will converge in distribution to a Brownian bridge, and the maximum of this rescaled Brownian bridge on $(1/n, \alpha_0)$ converges in probability to $\sqrt{2 \log \log n}$ (this is the law of the iterated logarithm). And a result of Donoho and Jin says that rejecting when $HC_n^* \geq \sqrt{(1+\varepsilon)2 \log \log n}$ gives us $\mathbb{P}_0(\text{type I}) + \mathbb{P}_1(\text{type II}) \to 0$ for any r above the detection threshold.

In more intuitive words, getting something like $HC_n^* = 2$ is not super surprising – for a given α being two standard deviations above the expected number might be somewhat surprising, but when we scan over all possibilities it's still reasonable under the null. But if we get something like 10, then something is definitely going on. (A Poisson(1) random variable can reasonably take on the value 5, but a standard normal will not.)

(And in practice, we won't rely on asymptotics – we will just simulate. For example, we get something like 3.6 for $n = 10^9$.)

Example 23

However, there are issues with the higher criticism. In particular, we can simulate the curve $W_n(t)$ corresponding to the higher criticism statistic (meaning that we sort our p-values, and at each one we get a jump in the empirical CDF and can thus plot a value of $\frac{F_n(t)-t}{\sqrt{t(1-t)/n}}$). (The maximum will be realized at one of those p-value jumps) We can see a sample plot of this below.



The main problem is that near $t = 0$, we are no longer approximately normal – $B(n, p)$ is nicely approximated by a Gaussian if np is not too small, but if np is small we get something that's Poisson, which has much heavier tails. Thus even though we are still mean-zero, variance-one near $t = 0$, we are very likely to achieve the supremum near 0 instead of somewhere else; thus the threshold is still being dominated by the behavior there and often the maximum comes very early. The point is that calibrating a statistic to get the maximum to occur at a uniform location is often very difficult, and this test does not manage to do so.

Example 24

The **Burk-Jones statistic** is an attempt to resolve this problem, but it is not fully effective.

The idea is that for each t (this is the significance level α we're scanning), we can test whether $n\hat{F}_n(t) < t$ using

a likelihood ratio test

$$\log LR_n(t) = \begin{cases} nD(\hat{F}_n(t), t) & 0 \le t \le F_n(t), \\ 0 & \text{otherwise,} \end{cases}$$

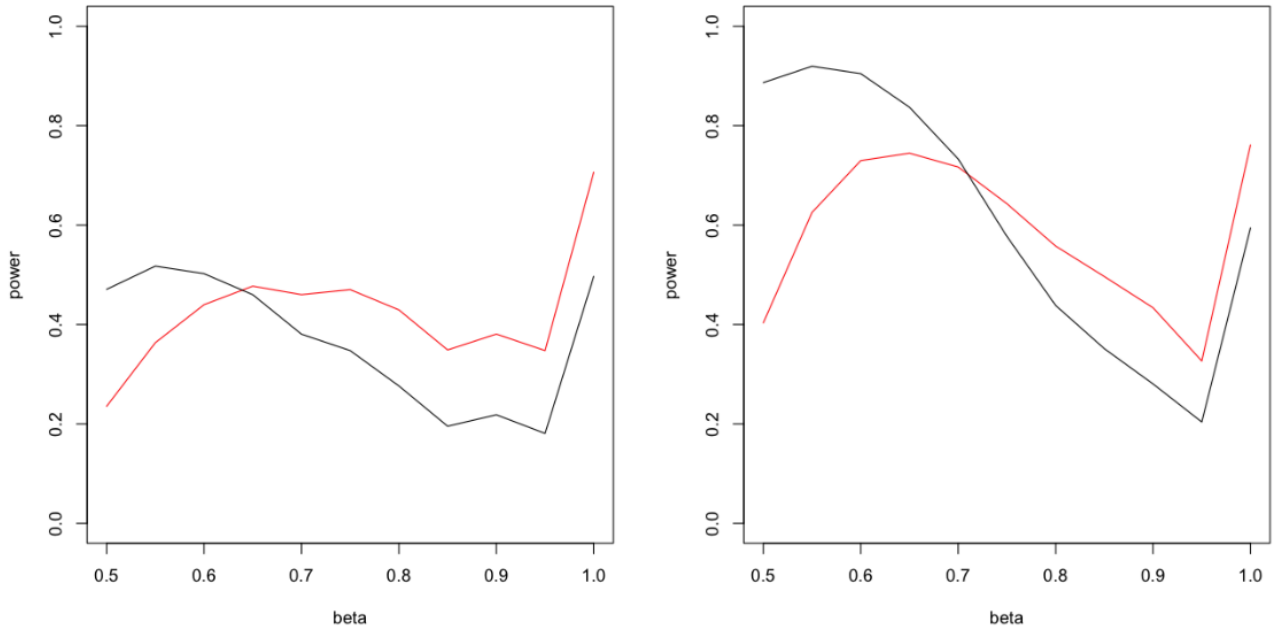
where $D(p_0, p_1) = p_0 \log \frac{p_0}{p_1} + q_0 \log \frac{q_0}{q_1}$), $q_i = 1 - p_i$ is the Kullback-leibler divergence. We then define

$$BJ^+ = \max_{1 \le i \le n/2} nD\left(p_{(i)}, \frac{i}{n}\right)$$

that is, at what significance level we detect a divergence between what we see an what we expect. We can simulate this and see that we also attain the optimal detection boundary, and it has better finite sample properties in the less sparse regime:

Fact 25

If we set $r = 1.2\rho^*(\beta)$ (so 20 percent above the theoretical detection threshold), we don't actually get full power for either HC_n^* or BJ_n^+. The plots below show some finite-size simulations for $n = 10^4$ and $n = 10^6$ (red is higher criticism, while black is Burk-Jones):

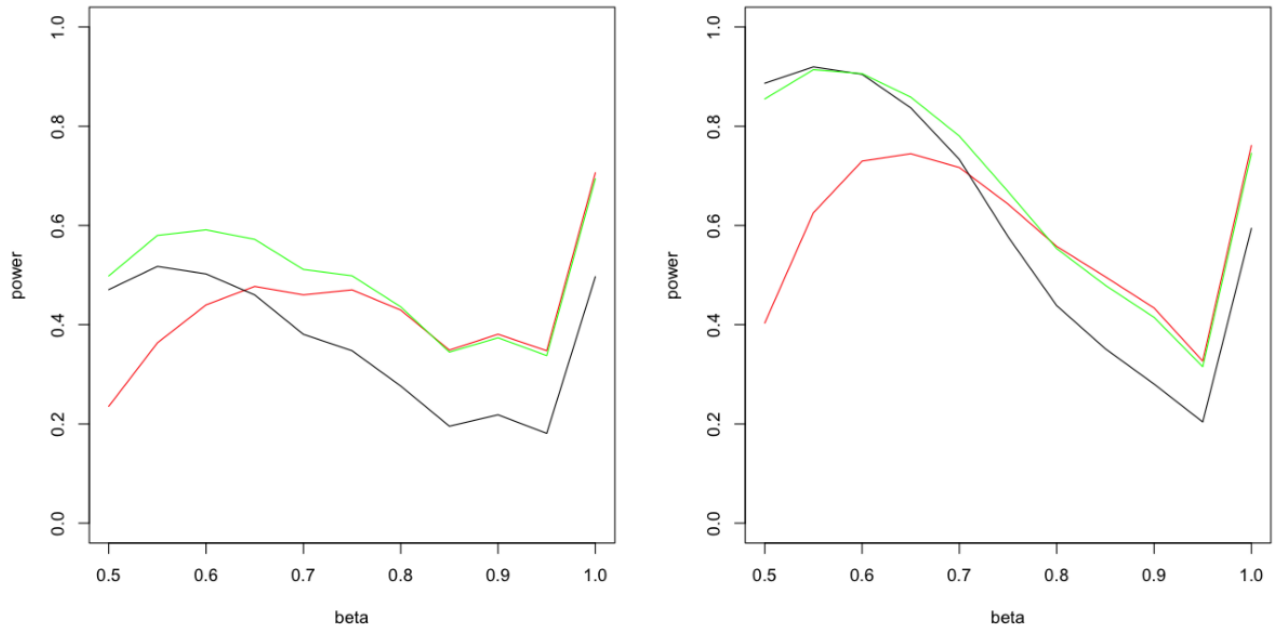


In the latter case, we see that HC_n^* and BJ_n^+ are both nowhere near 100 percent power; from $\beta = 0.5$ to $\beta = 0.7$, HC_n^*'s power roughly goes from 0.4 to 0.7, while BJ_n^+ goes from about 0.9 to 0.7 (meaning the latter is better in the less sparse regime). But then higher criticism does better after that.

Again, we run into a situation where we "want to be like HC on one side and BJ on the other," so that we can get the best of both worlds. One heuristic for this was proposed by Walther in 2011 – the idea is that instead of looking at the maximum of likelihood statistics, we can look at a weighted average, called the **average likelihood ratio**

$$ALR = \frac{1}{2} LR_1 + \frac{1}{2} \sum_{i=2}^{n/2} \frac{1}{i \log(n/3)} LR_i, \quad LR_i = \exp\left(nD\left(p_{(i)}, \frac{i}{n}\right)\right)$$

where LR_1 is the first p-value (so basically Bonferroni) and the other weights are cooked up from some Bayesian argument. (Basically in the remaining case, we look at weights decaying like $\frac{1}{t}$.) This does turn out to do better in finite-size simulations (the green curve shows the power of ALR)



Example 26

There are some other tests that we can also cover (Anderson-Darling, Kolmogorov-Smirnov), but we'll skip over them here. Instead, returning to the higher criticism, we can get a sense of what it is doing and where it is looking.

Define $\alpha(z) = \mathbb{P}(N(0,1) > z)$ and let $z(q) = \sqrt{2q \log n}$. The values of the higher-criticism statistic is then

$$HC_n = \max_{q \geq 0} V_n(q), \quad V_n(q) = \frac{\#\{X_i > z(q)\} - n\alpha(z(q))}{\sqrt{n\alpha(z(q))(1 - \alpha(z(q)))}}.$$

Our question now is what value of q causes $V_n(q)$ to grow fastest **under the alternative hypothesis**. To study this, note that by the tails of a Gaussian $\mathbb{P}(Z > t) \leq \frac{\phi(t)}{t}$ for ϕ the density of the Gaussian (in fact we are bounded from below by a factor $1 - \frac{1}{t^2}$), and thus plugging in $t = \sqrt{2q \log n}$ we find

$$\alpha(z(q)) = L_n n^{-q}$$

for some logarithmic factor L_n. Therefore for a non-null normal, if we have $r < q \leq 1$, then

$$\mathbb{P}(N(\mu_n, 1) > z(q)) = L_n n^{-(\sqrt{q} - \sqrt{r})^2}.$$

Therefore, under this crude approximation we have (plugging in all of those expressions)

$$\begin{aligned} \mathbb{E}_{H_1}[V_n(q)] &= \frac{\mathbb{E}_{H_1}[\#\{X_i > z(q)\}] - nL_n n^{-q}}{\sqrt{nL_n n^{-q}}} \\ &= L_n n^{(1+q)/2 - \beta - (\sqrt{q} - \sqrt{r})^2}. \end{aligned}$$

Thus to find what q this grows fastest for, we can optimize over q and thus find where we expect the higher criticism statistic to peak. Doing the maximization, we find that the optimal for $r < \frac{1}{4}$ is $q^* = 4r$, and we get that

$$\mathbb{E}_{H_1}[V_n(q)] = L_n n^{r - (\beta - 1/2)}.$$

So if $r > \beta - \frac{1}{2}$ we will indeed see the alternative hypothesis using HC. interestingly, this means we're looking at a

signal at $2\mu_n$ rather than μ_n (we're looking at $\sqrt{8r\log n}$ instead of $\sqrt{2r\log n}$).

Meanwhile if $r \geq \frac{1}{4}$, then the optimal q turns out to be the Bonferroni threshold $q^* = 1$. In such a case, we then instead get

$$\mathbb{E}_{H_1}[V_n(q)] = L_n n^{(1-\beta)-(1-\sqrt{r})^2},$$

and that completes the detection threshold curve. So overall, we stop before the Bonferroni level for $\beta < \frac{3}{4}$ because that's where the ratio between non-nulls and nulls is favorable, but for $\beta > \frac{3}{4}$ we essentially look at Bonferroni.

4 April 10, 2025

Today, the topic will shift from global testing to **multiple testing and comparison problems**. We'll learn about the familywise error rate and see basic ways of controlling it, seeing connections with global testing even in this different setting.

Example 27

We'll return to Example 2 from the beginning of the course now where we have $n = 6033$ genes measured on 102 patients (50 control, 52 cases) and want to analyze the expression levels $Y_{ij}^{(0)}$, $Y_{ij}^{(1)}$ among the healthy versus sick patients. Recall that the null hypothesis $H_{0,i}$ is that gene i is null and that the alternative $H_{1,i}$ is that it is not null.

We've seen ways of formulating this more quantitatively; for example we can test whether the mean is nonzero by considering

$$p_i = \mathbb{P}(|t_{100}| > |T_i|), \quad T_i = \frac{\text{avg(sick)} - \text{avg(control)}}{\text{estimated standard error}},$$

since T_i is distributed as t_{100}. But the point is that we have 6000 of these hypotheses and want to know what to do with them all at once (that is, which genes have to do with the sickness). In multiple testing, it's worth thinking about the following two-way table:

	accepted	rejected	total
true	U	V	n_0
false	T	S	$n - n_0$
total	$n - R$	R	n

We do not know the values of U, V, T, S, but they will be populated by numbers (based on how many hypotheses we accept or reject of each type) – T is the number of false negatives and V is the number of false positives, and depending on the sample we collect we get different values. So it would be nice to know the distribution of those (unobserved) random variables, but we don't even know n_0; we only really know the column totals R and $n - R$.

Fact 28

We are particularly interested in V (the number of false discoveries – that is, the papers we write that should not have been written) and how it compares to R (the total number of discoveries).

We can thus consider two error metrics, and we'll look at them in historical order:

Definition 29

The **familywise error rate (FWER)** is defined to be $\mathbb{P}(V \geq 1)$ (that is, the probability of **any** false positive).

The idea is that often in genome-wide studies, the threshold for significance is often set at something like 5×10^{-8}. This is meant to be a way to control the probability of reporting anything false at all, and the question is "given p-values, how can we make rejections in such a way that FWER is controlled at level α?", and we want this to be true for **any** configuration of true and false hypotheses. (There's also variations on this, such as the k-familywise error rate $\mathbb{P}(V \geq k)$, also called k-FWER.)

We've already seen one way we can do this: recalling that Bonferroni's method rejects all hypotheses with p-value below $\frac{\alpha}{n}$, each null leads to a false positive with probability $\frac{\alpha}{n}$, so that we get the following:

Theorem 30

Bonferroni's method controls FWER at level α; more specifically, if there are $n_0 \leq n$ null hypotheses, we get

$$\text{FWER} \leq \mathbb{E}[V] = \frac{n_0}{n}\alpha.$$

Proof. Since V is nonnegative-integer-valued, we have $\mathbb{P}(V \geq 1) \leq \mathbb{E}[V]$. And letting $\mathcal{N}_0$ denote the set of null hypotheses, we have

$$\mathbb{E}[V] = \mathbb{E}\left[\sum_{i \in \mathcal{N}_0} 1\left\{p_i \leq \frac{\alpha}{n}\right\}\right] = \sum_{i \in \mathcal{N}_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{n}\right) = n_0 \cdot \frac{\alpha}{n},$$

□

So we don't need any independence between p-values – as long as our nulls are (super)-uniform, this works. And even if the p-values are independent, we've already seen that the scale at which we can set the threshold α_0 is pretty close to just using the naive Bonferroni (for example instead of $\frac{0.05}{n}$ we would use $\frac{0.0512}{n}$ under independence).

Remark 31. *Note that when people talk about multiple testing, there are actually two types of control. We can consider a two-step procedure for achieving control on FWER proposed by Fisher in 1934: first, do a global test for the global null $H_0 = \bigcap_{i=1}^n H_{0,i}$, and if we say that something interesting is happening, then we test each hypothesis at level α. So then the protection only happens at the first stage, and this controls the familywise error rate* **weakly** *(meaning that* **under the global null** *the chance of a single rejection is α). However, this does not control FWER in general, since we're in serious trouble if we have even one strong signal. For example, suppose we have 1000 hypotheses and 10 of them are nonnull. We will likely pass the global test, and then we test each hypotheses at level $\alpha = 0.05$. Then*

$$\mathbb{P}(V \geq 1) = \textit{minimum null p-value} \leq \alpha,$$

which is extremely high ($1 - 0.95^{990} \approx 10^{-22}$).

We won't be so interested in this notion, but it's something we might see in the literature.

Another procedure (which is useful to know even if it's not such a big deal) is Holm's procedure:

Example 32

Suppose we have n hypotheses $H_{(1)}, \cdots, H_{(n)}$ corresponding to the **ordered** p-values $p_{(1)} \leq \cdots \leq p_{(n)}$ (so we choose $H_{(1)}$ to be the hypotheses with the most surprising p-value, and so on). Now we will compare p-values with an adaptive threshold based on what we've seen so far.

What we do here is called a **step-down procedure**:

- First, compare with Bonferroni's threshold: if $p_{(1)} \leq \frac{\alpha}{n}$, then reject $H_{(1)}$ and move to the next step. Otherwise, reject nothing (accept all null hypotheses).

- Now in general for step i, if $p_{(i)} \leq \frac{\alpha}{n-i+1}$, then reject $H_{(i)}$ and go to the next step $(i+1)$. Otherwise, accept $H_{(i)}, H_{(i+1)}, \cdots, H_{(n)}$ and stop.
- Finally, if $p_{(n)} \leq \alpha$, then we reject $H_{(n)}$; otherwise we accept it.

This adaptive threshold for the p-values is particularly useful if we have something like $n = 12$ total treatments, since the p-value does indeed noticeably change as the procedure proceeds. So the p-value gets larger over time (the comparison gets more favorable over time, meaning it's strictly less conservative than Bonferroni), but it's still called a step-down procedure because the z-score thresholds (for our test statistic) are decreasing. In a sentence, we basically reject all the small p-values until the critical value where $p_{(i)} > \alpha_i = \frac{\alpha}{n-i+1}$.

Theorem 33

Holm's procedure controls the FWER strongly.

Proof. Consider the sorted p-values $p_{(1)}, \cdots, p_{(n)}$. We want to study the event where we make a false rejection; this must happen at some index, and we define i_0 to be the first null hypothesis we encounter (so that $p(i_0) = \min\{p_i : i \in \mathcal{N}_0\}$). We then have

$$\{V \geq 1\} = \{\text{procedure reached } i_0\} \cap \left\{ p(i_0) \leq \frac{\alpha}{n - i_0 + 1} \right\},$$

since if we don't reject this first null we must have stopped before it, and if we do then we get at least one false positive. (Note however that both i_0 and V are random variables.) But now $\frac{\alpha}{n-i_0+1} \leq \frac{\alpha}{n_0}$, since $n - i_0 + 1$ is maximized if we see all the non-nulls first (that is, $i_0 \leq n_1 + 1 = n - n_0 + 1$). Thus the event of false rejection is certainly contained in $\{p(i_0) \leq \frac{\alpha}{n_0}\}$, which has probability bounded by $n_0 \cdot \frac{\alpha}{n_0} = \alpha$ as desired. □

Example 34

We can now see a general method which turns any global test into a multiple test that controls FWER – this is called the **closure principle**. We have a family of hypotheses as usual, and we now want to think about the intersection nulls

$$H_I = \bigcap_{i \in I} H_i \text{ for all nonempty } I \subset \{1, 2, \cdots, n\}.$$

We thus have 2^{n-1} different such joint hypotheses to think about – our global null is that all H_Is are true. The principle is then that for each I, we assume that we can test (by higher criticism or Anderson-Darling or anything else we've seen so far) H_I at level α with some test ϕ_I. In notation (if we say the test rejects the null with $\phi_I = 1$),

$$\mathbb{P}(\phi_I = 1 | H_I) \leq \alpha.$$

Definition 35

The **closure procedure** then rejects H_I if and only if H_J is rejected at level α **for all** $J \supseteq I$. Mathematically, we consider

$$T_I = \min_{J \supseteq I} \phi_J,$$

and if this is 1 (meaning all ϕ_J were 1) then we reject the intersection null I.

Example 36

For example for $n = 3$, we have global tests for $H_1, H_2, H_3, H_{12}, H_{13}, H_{23}$, and H_{123}. And in order to reject H_2, we need global tests to actually reject H_2, H_{12}, H_{23}, and H_{123}.

On the other hand if we have $n = 4$ and we see that global tests reject $H_1, H_2, H_{12}, H_{13}, H_{14}, H_{123}, H_{124}, H_{134}, H_{1234}$ at level $\alpha = 0.05$, then we can reject H_1 (because any hypothesis with a 1 is rejected) but not H_2 (because H_{23} is not rejected), even though H_2 is significant.

Theorem 37

This closure principle described above controls the FWER strongly.

This is nice in that it provides a generic recipe for translating any global test into a procedure for rejecting or accepting various null hypotheses.

Proof. The event $\{V \geq 1\}$ (of a false rejection) is contained in the event $\{\mathcal{H}_0$ was rejected$\}$, but we have a global test at level α which shows that this happens with probability at most α. □

The main problem with this principle is that you have to do an exponential number of tests, so the computational cost is not manageable. But sometimes we can try to find shortcuts, and this is where the research is really interesting: "is it possible to find a procedure which is a bit more conservative than the closure, say in polynomial time?"

Example 38

The suggestion is that we can use Bonferroni as our global test and try to close it – it will turn out that we can use some strategies for speeding it up.

Our global test is then saying that for any index set I,

$$\phi_I = 1 \iff \inf\{p_i : i \in I\} \leq \frac{\alpha}{|I|}.$$

The key claim now is that **closing Bonferroni is exactly Holm's procedure**, so this is a second argument that shows Holm's procedure does control FWER. Indeed, sort the p-values as usual so that $p_{(1)} \leq \cdots \leq p_{(n)}$ correspond to the hypotheses $H_{(1)}, \cdots, H_{(n)}$.

- In order to reject $H_{(1)}$, we must reject $H_{\{1,2,\cdots,n\}}$, meaning that the minimum of the p-values satisfies $\boxed{p_{(1)} \leq \frac{\alpha}{n}}$, and any other comparison is strictly easier than that. And if the intersection is not rejected, then every hypothesis is not, as it is in Holm's procedure.
- Next, in order to reject $H_{(2)}$, there are two kinds of conditions for sets I containing 2: we must have $p_{(1)} \leq \frac{\alpha}{|I|}$ if I also contains 1, and otherwise we must have $p_{(2)} \leq \frac{\alpha}{|I|}$. So in particular we must have to reject $H_{(1)}$ already (so $\boxed{p_{(1)} \leq \frac{\alpha}{n}}$), and then the remaining hardest comparison is for $I = \{2, \cdots, n\}$, meaning that we must also have $\boxed{p_{(2)} \leq \frac{\alpha}{n-1}}$.
- Similarly for the third hypothesis, we need to consier the cases where I contains 1, or doesn't contain 1 but does contain 2, or doesn't contain 1 or 2. These correspond to the conditions on $p_{(1)}, p_{(2)}$, and $p_{(3)}$ respectively.

So we don't need 2^n comparisons in this case! But in general, we might get more unlucky because closing global tests is hard. (And what's useful is that a polynomial-time procedure which rejects strictly less than the closure also controls FWER; in contrast, this is not true for false discovery rate.)

Example 39

Last lecture, we left out a global test, called **Simes' test**. For this, we reject the global null if and only if the ordered p-values satisfy

$$p_{(i)} \leq \frac{\alpha i}{n}.$$

In other words, we want that $\min_i \frac{np_{(i)}}{i} \leq \alpha$. This is less conservative than Bonferroni, and this has level α **under independence**.

Closing this test is a bit of a mess, but it turns out it can be bounded by something called **Hochberg's procedure** – it's Holm's procedure but in the reverse order. We examine the p-values with the same critical thresholds as before, but we start at the largest $p_{(n)}$ and go down until we're below the threshold, rather than starting at the smallest $p_{(1)}$ and going up until we're above the threshold.

This therefore gives us a more liberal procedure (which is very similar to the **Benjamini-Hochberg procedure**), and it will reject more while still controlling FWER because it's still more conservative than the closure principle. So under independence this is a better thing to do!

5 April 15, 2025

We'll introduce another notion of error today, the **false discovery rate (FDR)**, which has become popular in the last 25 years as a replacement for the familywise error rate. We'll understand some procedures for controlling the FDR, especially under independence. The original paper had a lot of trouble being published, but it's since had a large influence (being cited over 115,000 times as of today!).

Fact 40

The familywise error rate makes a lot of sense when testing a small number of hypotheses and where the consequences or cost of a single false rejection is high (for example, comparing treatments and suggesting one for the market which actually hurts people). However, it will be difficult to achieve high power this way if we are testing many hypotheses, and the way we do science has changed. For example in genome-wide association studies today, we test millions of different hypotheses simultaneously (and there is lots of data available everywhere), and it's often "not the end of the world" now if we make a false discovery.

So in today's world, FWER is so stringent that we often return nothing if we require FWER control. There's always a tradeoff – we can't publish things that can't be replicated, but if we require too strong a chance of replication, we won't get anything at all. It would thus often be better to return some false positives and give scientist a chance to follow these primary leads. Maybe "it's okay if there's an irreproducibility chance of 10 percent" or "a lab spends 10 percent of its time on incorrect hypotheses," and this was advanced by Benjamini and Hochberg in 1995:

Remark 41. *Technology "had not caught up" back when this paper first (it wasn't getting citations for the first five years), but over time our data has caught up and this has become quite important.*

Recall the following table from last lecture:

	accepted	rejected	total
true	U	V	n_0
false	T	S	$n - n_0$
total	$n - R$	R	n

In the language of this table, last lecture we studied ways to minimize the probability $\mathbb{P}(V \geq 1)$ (or $\mathbb{P}(V \geq k)$.

Definition 42

The **false discovery proportion (FDP)** is given by

$$\text{FDP} = \frac{V}{\max(R, 1)} = \begin{cases} V/R & R \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

(this last case is just by convention so that we can control this quantity). In general this random variable is unobserved, and the **false discovery rate (FDR)** is the expectation $\text{FDR} = \mathbb{E}[\text{FDP}]$ of this quantity.

The proposal is that if we control this number at 10 percent, we can design methods where on average, at least 90 percent of our discoveries are indeed correct.

One objection that people had with this criterion when it was first proposed was that it is weak: first of all, it's a weaker notion than FWER, but also controlling the expectation of a random variable doesn't actually say much about an individual study. A statement like $\mathbb{P}(\text{FDP} \geq 0.1) \leq 0.05$ would mean that unless we are unlucky, our study has a good chance of doing well, but just applying a procedure with FDR 0.1 doesn't let us say anything – all we can say is that by Markov's inequality we have $\mathbb{P}(\text{FDP} \leq 0.2) \leq 0.5$, for example. But the point is that "science as a whole is correct," and luckily, we'll see that often (with independent p-values) FDP is concentrated around its mean, so we can say something stronger about individual studies as well.

Here are some important properties of the false discovery rate:

- Under the global null, the FDR is equivalent to the FWER. (Indeed, under the global null, every rejection is false, so the variable $1\{V \geq 1\}$ is the same as $\frac{V}{R \vee 1}$ – here $\vee$ means maximum.)
- Thus, a false discovery procedure has to look at Bonferroni a little bit and make comparisons of that sort, since it must control FWER weakly.
- In general we instead have the inequality $1\{V \geq 1\} \geq \frac{V}{R \vee 1}$, so FWER is at least FDR; thus controlling the FWER also controls the FDR.

Example 43 (Benjamini-Hochberg procedure)

We'll now consider a procedure generally more powerful than Holm's procedure – instead of Hochberg's procedure, we consider a step-up procedure with critical values $\alpha_i = \frac{\alpha i}{n}$, which is far less conservative than $\frac{\alpha}{n-i+1}$.

Basically, we draw a line at slope α, and we look at the rightmost (largest) p-value below this line. Then we stop and reject the null for that p-value and anything smaller. (This is somewhat like Simes' test but it's a multiple comparison test instead of a global test.) This threshold is also adaptive, in the sense that a particular fixed p-value is more likely to be rejected if there are many low p-values (that is, if it's ranked less significantly).

Remark 44. *Remember that in a step-up procedure, if our p-values cross the critical threshold multiple times, what matters is the rightmost crossing (which is the more liberal approach). In a step-down procedure, we'd instead care about the leftmost crossing.*

Theorem 45

Suppose our test statistics are independent (so p-values are independent). Then the Benjamini-Hochberg controls the FDR at level α. More precisely, we actually have the expression

$$\text{FDR} = \frac{n_0}{n}\alpha \leq \alpha.$$

The proof in the original paper is a bit contested, but we'll see a proof here which Professor Candés developed with a former student. What's nice is that it's powerful and that it can also be used in a more general setting or to compute moments:

Proof. Let $V_i = \{H_i \text{ rejected}\}$ be the indicator function for hypothesis i being rejected. By definition, we have

$$\text{FDP} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R \vee 1}.$$

Now, it suffices to show that for any null i, we have $\mathbb{E}\left[\frac{V_i}{R \vee 1}\right] = \frac{\alpha}{n}$. (This is somehow "the only answer we can get" because the nulls are uniform and thus the random variables are exchangeable.) To prove this claim, notice that we can do casework over the value of R and write

$$\frac{V_i}{R \vee 1} = \sum_{k=1}^{n} \frac{V_i 1\{R = k\}}{k} = \sum_{k=1}^{n} \frac{1\left\{p_i \leq \frac{\alpha k}{n}\right\} 1\{R = k\}}{k},$$

since assuming $R = k$, we know the threshold for rejection is $\frac{\alpha k}{n}$. Notice that on the event $p_i \leq \frac{\alpha k}{n}$, changing p_i to zero doesn't change the threshold, meaning whenever we reject H_i, the number (and identity) of rejections is the same. So we can write the above expression as

$$\sum_{k=1}^{n} \frac{1\left\{p_i \leq \frac{\alpha k}{n}\right\} 1\{R(p_i \to 0) = k\}}{k}$$

where this notation means that we set this null p-value to zero. Now we can take the expectation of this quantity conditioned on all other p-values – the only randomness is in p_i here, so

$$\begin{aligned}\mathbb{E}\left[\frac{V_i}{R \vee 1}\middle| p_1, \cdots, p_{i-1}, p_{i+1}, \cdots, p_n\right] &= \sum_{k=1}^{n} \frac{\frac{\alpha k}{n} 1\{R(p_i \to 0) = k\}}{k} \\ &= \sum_{k=1}^{n} \frac{\alpha}{n} 1\{R(p_i \to 0) = k\} \\ &= \frac{\alpha}{n}.\end{aligned}$$

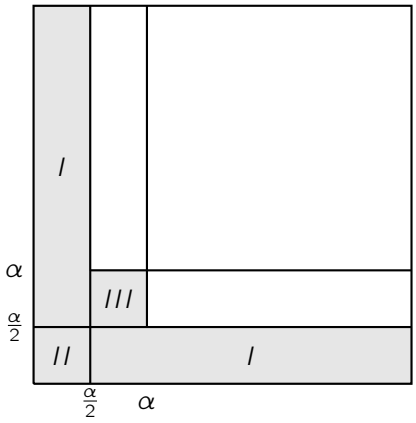
Then finally taking the expectation over the last p-value yields the result. □

Remember that for the tenth smallest p-value, Hochberg's procedure is looking at $\frac{\alpha}{n-9}$, which is essentially still Bonferroni, while Benjamini-Hochberg is looking at $\frac{10\alpha}{n}$, which is ten times larger. Thus this is letting us reject far more liberally.

What we really needed here is that the distribution of any **null** p_i does not depend on the distribution of the other p-values – we could even condition on the non-null p-values as long as they are independent from the nulls. But under dependence, we might get FDR inflation, and we might also find that the distribution of the FDP is less concentrated around its mean, and we'll talk about both of these now.

Example 46

Suppose $n = 2$, meaning we have two p-values which we will represent on the x and y-axis. Assume that we're under the global null. Benjamini-Hochberg commits a false rejection if the smaller p-value is below $\frac{\alpha}{2}$ **or** if the larger p-value is below α. Thus there's a region of situations where we get a rejection:



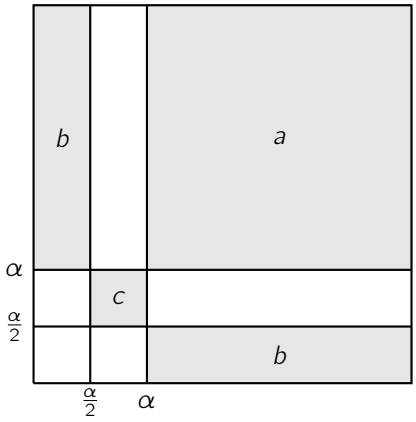
Graphically, we see that if the p-values are uniform and independent, then this chance of a false rejection (the area of these regions) is exactly α. But we can inflate that value and make it larger, because in general we have

$$\text{FDR} = \mathbb{P}(I) + \mathbb{P}(II) + \mathbb{P}(III) = \alpha + \mathbb{P}(III) - \mathbb{P}(II),$$

and now the best bound we can get is

$$\text{FDR} \leq \alpha + \mathbb{P}(III) \leq \alpha + \mathbb{P}\left(\frac{\alpha}{2} < p_1 < \alpha\right) = \frac{3\alpha}{2}.$$

And these inequalities can all be made tight (we can make $\mathbb{P}(III) = \frac{\alpha}{2}$ and $\mathbb{P}(II) = 0$) by considering a joint distribution over these marked regions and assigning the density to be zero in the other regions:



Specifically, if we set the densities to be constant on each region with $b = \frac{1}{1-\alpha}$, $c = \frac{2}{\alpha}$, and $a = b(1 - \frac{b\alpha}{2})$, then the marginal densities (the individual p-values) are each uniform, and indeed that gets us the "inflation" to $\frac{3\alpha}{2}$.

Of course, we're not running this procedure when $n = 2$ in practice – what we're curious about is what the inflation looks like for larger n:

Theorem 47 (Guo-Rao '08)
Let $S(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \log n + 0.577$ be the nth harmonic number. Then there are joint distributions where the FDR of the Benjamini-Hochberg procedure $BH(\alpha)$ is at least $\min(1, \alpha S(n))$.

(This proof essentially comes down to constructing similar examples as the one we did above for $n = 2$.) So if we test 10000 things, we can inflate things by a factor of 10. But conversely, we have tightness (meaning this is the most adversarial situation):

Theorem 48 (Benjamini-Yekutieli '01)
Under dependence of p-values, the $BH(\alpha)$ procedure does control at level $\alpha S(n)$; in fact,

$$\text{FDR} \le \alpha S(n) \cdot \frac{n_0}{n}.$$

The following proof is also by Professor Candés and a former student:

Proof. Let $\alpha_i = \frac{i\alpha}{n}$; much like in the proof before, it suffices to show that $\mathbb{E}\left[\frac{V_i}{R \vee 1}\right] = \frac{\alpha}{n} S(n)$. We again have

$$\frac{V_i}{R \vee 1} = \sum_{k=1}^{n} \frac{1\{p_i \le \alpha_k\} 1\{R = k\}}{k},$$

and we look at where p_i can fall. Summing over the possible ranks it can take on, we have

$$\sum_{k=1}^{n} \sum_{\ell=1}^{k} \frac{1\{\alpha_{\ell-1} \le p_i \le \alpha_\ell\} 1\{R = k\}}{k} = \sum_{\ell=1}^{k} \sum_{k \ge \ell} \frac{1\{\alpha_{\ell-1} \le p_i \le \alpha_\ell\} 1\{R = k\}}{k}$$

just by swapping the order of summation. But now if we do the k-sum first, we're just looking at the probability of getting a particularly high number of rejections, so this simplifies to

$$\sum_{\ell=1}^{n} \frac{1\{R \ge \ell\}}{R} 1\{p_i \in [\alpha_{\ell-1}, \alpha_\ell]\}.$$

Everything so far has been an equality, so "nothing interesting" has happened yet. But now we can simplify the first fraction to be bounded by $\frac{1}{\ell}$,

$$\sum_{\ell=1}^{n} \frac{1}{\ell} 1\{p_i \in [\alpha_{\ell-1}, \alpha_\ell]\} = \sum_{\ell=1}^{n} \frac{1}{\ell} \frac{\alpha}{n} = S(n) \frac{\alpha}{n},$$

and then the rest of the proof proceeds as before. What's surprising is that the result of Guo and Rao shows that there are distributions of p-values for which this inequality is indeed tight! □

So this inflation can indeed get bad, but here's a question to think about: if we are testing the means of a Gaussian using absolute z-values, and the Gaussians are correlated in some arbitrary way (so the test statistics are elliptical), what is the worst possible inflation of FDR when we apply Benjamini-Hochberg? People believe that in cases with unimodal distributions, FDR is still controlled or close to controlled, and numerical simulation shows that this seems to be the case as well. But it's still an open question whether we can indeed prove that FDR inflation is very limited.

6 April 17, 2025

We'll see a different perspective on Benjamini-Hochberg today, thinking about the **empirical process viewpoint**, and then understanding how to improve on BH. While we do have a proof that this gives us control on FDR, we don't really fully understand it yet, and today's lecture will make it more intuitive. This lecture will mostly follow a paper of Storey, Siegmund, and Taylor from 2004.

Example 49

In today's lecture, we'll switch the x- and y-axis in plotting p-values. Previously, we had points at x-coordinates $\frac{i}{n}$ and corresponding y-coordinate the p-values $p_{(i)}$, which we can think of as plotting the cdf on the horizontal axis and the inputs on the vertical axis. But now if we consider the empirical CDF $\hat{F}_n(t) = \frac{\#\{i : p_i \le t\}}{n}$, then plotting $\hat{F}_n(t)$ will have p-values on the x-axis and indices on the y-axis. Then instead of a slope of line α, the critical threshold line will be of slope $\frac{1}{\alpha}$.

Recall that Benjamini-Hochberg sorts the p-values and computes

$$i_0 : \max\left\{ i : p(i) \le \frac{\alpha i}{n} \right\},$$

rejecting all p-values below $p(i_0)$. In other words, we define the critical p-value

$$\begin{aligned} p^* &= \max\left\{ p_{(i)} : p_{(i)} \le \frac{\alpha i}{n} \right\} \\ &= \max\left\{ p_{(i)} : p_{(i)} \le \alpha \hat{F}_n(p_{(i)}) \right\} \\ &= \max\left\{ t \in \{p_1, \cdots, p_n\} : t \le \alpha \hat{F}_n(t) \right\} \end{aligned}$$

(where by convention we can just set $p^* = \frac{\alpha}{n}$ if we don't reject anything). Thus we can rewrite this as saying that we reject all hypotheses H_i for p_i below the adaptive threshold

$$\tau_{\text{BH}} = \max\left\{ t : \frac{t}{\hat{F}_n(t) \vee 1/n} \le \alpha \right\}.$$

Now we will think about how this compares to rejecting hypotheses below a **fixed** threshold t. We then end up with an outcome table of the following form, depending on the threshold t:

	accepted	rejected	total
true	$U(t)$	$V(t)$	n_0
false	$T(t)$	$S(t)$	$n - n_0$
total	$n - R(t)$	$R(t)$	n

The false discovery proportion is then $\text{FDP}(t) = \frac{V(t)}{\max(R(t),1)}$, and we are interested in bounding its expectation FDR. If we could estimate $\text{FDR}(t)$ for all t, then we want to use the most liberal threshold possible, which would be

$$\tau = \sup\left\{ t \le 1 : \widehat{\text{FDR}}(t) \le \alpha \right\};$$

that is, among all procedures, our estimate says that we should be fine, and then we take the best threshold possible at that point. To obtain such an estimate of $\frac{V(t)}{\max(R(t),1)}$, we know the number of rejections (hence the denominator) but not the number of false rejections (the numerator). So we'll estimate using the expectation and do the crude

bound $\mathbb{E}[V(t)] = n_0 t \le nt$. Thus we get the estimate

$$\widehat{\text{FDR}}(t) = \frac{nt}{\max(R(t), 1)} = \frac{t}{\max(\hat{F}_n(t), 1/n)},$$

which is exactly what we have in Benjamini-Hochberg! So the optimization over t for our adaptive threshold makes sense: τ_{BH} is basically estimating the false discovery rate we get for each threshold t, and we pick the highest t so that this statistic is below α.

Theorem 50

If p-values are independent, our estimator is actually biased upward (which is good); that is, $\mathbb{E}[\widehat{\text{FDR}}(t)] \ge \text{FDR}(t)$ for all t.

So even in the case where n_0 is very close to n (so we only have a very small number of alternative hypotheses), we're being a bit conservative, so that in general the false discovery rate is at least as good as what we're complaining. (It would be good to have an unbiased estimator, but it's not clear how to get one or what we can actually do with it.)

We can also take this estimate and rephrase it in terms of control on the FDR (which is what we did last lecture): we showed that $\mathbb{E}[\text{FDR}(\tau_{\text{BH}})] = \frac{\alpha n_0}{n}$, but we will show the proof in a different way now (understanding why it works).

Proof. We want to show that $\mathbb{E}\left[\frac{V(\tau_{\text{BH}})}{R(\tau_{\text{BH}})\wedge 1}\right] = \frac{\alpha n_0}{n}$, and remembering that we're "finding the first t that's above a particular line," we'll try to make a martingale argument where we start at $t = 1$ and go down towards $t = \frac{\alpha}{n}$.

Let $\mathcal{F}_t$ be the sigma-algebra of all information about $R(s)$ (the number of rejections at level s) and $V(s)$ (the number of nulls we would have rejected) **for all** $s \ge t$. At the start, $V(1) = n_0$ (because we would be rejecting all nulls) and $R(n) = n$. Each time we cross a p-value, R goes down by one, and we'll reveal whether it was null or non-null (which tells us V); thus we can keep a running count and have the values of $V(s)$ and $R(s)$ all the way down to t.

We claim that $\frac{V(t)}{t}$ is a martingale with respect to the given information. Indeed, at some $s \le t$ (any real numbers, not necessarily p-values), we have

$$\mathbb{E}\left[V(s)|\mathcal{F}_t\right] = \frac{s}{t}V(t),$$

since conditioned on $\mathcal{F}(t)$ we know how many nulls are left to reveal, and $V(s)$ looks at a fraction $\frac{s}{t}$ of the interval. Thus $\mathbb{E}\left[\frac{V(s)}{s}\middle|\mathcal{F}_t\right] = \frac{V(t)}{t}$, meaning we have a martingale and can stop it using the stopping time τ_{BH}. Indeed,

$$\begin{aligned}\mathbb{E}\left[\frac{V(\tau_{\text{BH}})}{R(\tau_{\text{BH}})\wedge 1}\right] &= \mathbb{E}\left[\frac{\tau_{\text{BH}}}{R(\tau_{\text{BH}})\wedge 1}\cdot\frac{V(\tau_{\text{BH}})}{\tau_{\text{BH}}}\right] \\ &= \frac{\alpha}{n}\mathbb{E}\left[\frac{V(\tau)}{\tau}\right]\end{aligned}$$

(last step by definition of τ_{BH}), and by the optional stopping theorem this is $\frac{\alpha}{n}\mathbb{E}\left[\frac{V(1)}{1}\right] = \frac{\alpha}{n}n_0$, as desired. □

Example 51

We can now try to get closer to getting an unbiased estimator of FDR, but we still need to be able to invert it to actually get control. We were only conservative here because we bounded n_0 by n (since we don't know the value of n_0); it would be nice to save the factor of $\pi_0 = \frac{n_0}{n}$ if we knew what that quantity is.

What's nice is that we **can** estimate it from the distribution of p-values that we observe. Fix a constant λ (say $\frac{1}{2}$)

and consider

$$\hat{\pi}_0^\lambda = \frac{n - R(\lambda)}{(1-\lambda)n}.$$

This is our estimate of "fraction of null hypotheses" – the idea is that if we look at the p-values between $\frac{1}{2}$ and 1, we probably don't have many non-nulls. Since that part is dominated by nulls, if we see that 40 percent of the p-values are in the top half, then we expect about 80 percent of the hypotheses to be null. (So in other words, our estimate of n_0 is the number of p-values in this upper interval, divided by the length of the interval.) This is also upward biased, because we can in fact sometimes see non-nulls and that means $\hat{\pi}_0^\lambda$ will typically be a little larger than π_0. So $\mathbb{E}[\hat{\pi}_0^\lambda] \geq \frac{n_0}{n} = \pi_0$ implies that we can make a new FDR estimate

$$\widehat{\mathrm{FDR}}^\lambda(t) = \frac{\hat{\pi}_0^\lambda nt}{\max(R(t), 1)};$$

we recover the usual BH procedure if we set $\lambda = 0$, but in general it's a little less conservative. We thus want to prove that "putting together these two biased-upward estimates" is good, meaning that setting the threshold to

$$\tau = \sup\left\{t \leq 1 : \frac{\hat{\pi}_0^\lambda nt}{\max(R(t), 1)} \leq \alpha\right\}$$

still yields FDR control at level α. Unfortunately, this is not the case, but luckily, there is a slight variation that works. With $\lambda = 1/2$ we would have $\hat{\pi}_0 = \frac{n - R(1/2)}{n/2}$ (in particular, it's possible for this quantity to be zero), and we just need to inflate this a tiny bit by adding a 1 to the numerator:

Theorem 52 (Storey's procedure)

Consider the estimate

$$\widehat{\mathrm{FDR}}(t) = \frac{n + 1 - R(1/2)}{n/2} \cdot \frac{nt}{1 \wedge R(t)}.$$

This is like "adding one to the number of p-values we see in the interval $[1/2, 1]$ when estimating the proportion of nulls. Then if we reject p_i below the threshold

$$\tau = \sup\left\{t \leq 1/2 : \widehat{\mathrm{FDR}}(t) \leq \alpha\right\}$$

(so basically we only use the points above $1/2$ to estimate π_0; they're never getting rejected), then we control the FDR at level α.

Proof. We will make a similar argument to the martingale one from before, but now starting at $t = \frac{1}{2}$ and coming down. (So we get to condition everything here on $\mathcal{F}_{1/2}$.) We then get

$$\begin{aligned}
\mathbb{E}\left[\mathrm{FDP}(\tau)\middle|\mathcal{F}_{1/2}\right] &= \mathbb{E}\left[\frac{V(\tau)}{R(\tau) \vee 1}\middle|\mathcal{F}_{1/2}\right] \\
&= \mathbb{E}\left[\frac{V(\tau)}{n\tau} \cdot \frac{n\tau}{R(\tau) \vee 1} \cdot \frac{n + 1 - R(1/2)}{n/2} \cdot \frac{n/2}{n + 1 - R(1/2)}\middle|\mathcal{F}_{1/2}\right] \\
&= \alpha\mathbb{E}\left[\frac{V(\tau)}{\tau} \cdot \frac{1/2}{1 + n - R(1/2)}\middle|\mathcal{F}_{1/2}\right]
\end{aligned}$$

where we use that the blue quantity is always exactly α when we stop. Now $\frac{V(\tau)}{\tau}$ is still a martingale, so the optional stopping theorem tells us that this is also equal to $\alpha\frac{V(1/2)}{1/2} \cdot \frac{1/2}{1+n-R(1/2)}$.

So Storey's procedure satisfies (now we take the expectation overall)

$$\mathrm{FDR}(\tau) = \alpha\mathbb{E}\left[\frac{V(1/2)}{n + 1 - R(1/2)}\right] = \alpha\mathbb{E}\left[\frac{V(1/2)}{n + 1 - S(1/2) - V(1/2)}\right].$$

Now $S(1/2) \leq n_1$, so $n - S(1/2) \geq n_0$ and thus

$$\text{FDR}(\tau) \leq \alpha \mathbb{E}\left[\frac{V(1/2)}{1 + n_0 - V(1/2)}\right].$$

Now $V(1/2)$ is binomial with parameters $(n_0, 1/2)$, and we can check that this expected value is indeed at most 1 (in fact it is exactly $1 - 2^{-n_0}$), completing the proof. □

If we didn't add 1 in the numerator, this last step would not work (in particular because $V(1/2)$ might be exactly equal to n_0) – if we replace the 1 with any smaller real number the expectation will be larger than 1. And we do in fact see FDR inflation in real-life situations if we don't have this correction term as well – in regimes where π_0 is substantially away from 1, this is in fact better than Benjamini-Hochberg.

Remark 53. *We can choose other values of λ as well, but we still need this same $+1$ in the numerator in Storey's procedure (the same binomial calculation works out).*

There are two problems with correlations in p-values – first of all, the FDR might be inflated (as we discussed last time), but also the distribution of the false discovery proportion might not be tightly concentrated around its mean, meaning control of the FDR might not actually be informative. (For example if all p-values are exactly equal, then we have to either reject everything or nothing, and that's not useful because it means our FDP is either 0 or 1.) There are tests of independence (via information-theoretic measures), but unfortunately in this situation they don't help very much.

7 April 22, 2025

Today, we'll talk about **controlled variable selection** – we're moving further away from classical stuff now. We'll introduce problems that we care about, for which it's difficult to construct p-values; the point is that there are important problems out there where this is not clear.

Fact 54

There was a paper from 2005 which claimed that most published research (about 80 percent) is false. Regardless of the status of this paper, the fact that science runs into trouble has been noted by lots of people – this is the **replicability crisis**.

There was a Nature paper by Begley and Ellis in 2012 which showed that out of 53 landmark studies in basic cancer science published in top journals, Amgen (a biotech company) could only replicate 6 of them. Similarly, HealthCare (a similar company in Germany) could only replicate about a quarter of 67 seminal studies in their field of research, and generally systematic attempts to replicate widely cited priming experiments in psychology have failed. And other areas of science go through phases of clinical trials – even phase III trials for the FA still end up in failure about half the time, even though they're supposed to be far in the research process alreay.

So there are many different components to consider here – there's of course a publishing culture at fault (since journals want extraordinary results), and there's also pressure to promise a lot to granting agencies and work on "big science" (where there isn't a clear understanding of the pipeline from beginning to end). But there are smaller-scale problems as well, such as computational reproducability (in terms of not publishing data or code) or statistical methodology. And those last issues are what we can address more directly: we can decide when to report a finding and enhance replicability.

Remark 55. *We've hinted at this before, but the scientific method has been turned upside down: in the new scientific paradigm, we collect data first and then ask questions afterward. But usually in hypothesis-driven research, we're supposed to observe the world and make a hypothesis, design an experiment for it, and then only collect data and interpret it once we have the hypothesis in mind. Instead, we now have AI agents that sift through data for us, and that's a big driving factor as well.*

Example 56

In Professor Candés' field of research, the activity went from "testing an experiment for a particular gene" to an explosion of technology where we can now test thousands of genes as well (so that we have a small number of samples and a high number of variables). So it's not fraud or that we're dishonest; instead there are enormous data sets, and most of what we look at is null and the "look-everywhere effect" needs to be addressed accordingly.

The human genome project is the epitome of what we're talking about here: we sequence the genome because we're hoping that by doing so, we can formulate interesting questions and get the data-driven paradigm. So we might have a response variable Y (such as Alzheimer's disease status) which is a phenotype, and we have hundreds of thousands of genotype information variables X. Our goal is then to understand the relationship between mutations and phenotype, and we have no theory – we just want to figure out which variations do affect these traits, or what profiles determine the severity of a tumor.

Thus, our goal is to select variables without creating too many false positives, so that we have a low rate of irreproducibility and increase our credibility. Perhaps we have a sample of individual human beings from a population, and we can collect the values of X and Y from them. (There's in fact lots of datasets like this that are available, such as the UK Biobank.) We then want to understand which variables X are most important. Here we might have $n \approx 5000$ people in a large study but about 500000 variables, and we might want to know which 500 (for example) are actually important for the conditional distribution of $Y|X$.

Definition 57

To make this well-posed, we'll say that a variable is a discovery if

$$p(\text{response}|\text{variable, other variables}) \neq p(\text{response}|\text{other variables});$$

that is, j is null if and only if $Y \perp\!\!\!\perp X_j|X_{-j}$, where X_{-j} is the set of all other variables.

We thus want to do a **conditional test** for whether $Y \perp\!\!\!\perp X_j|X_{-j}$. Notice that this is different from just testing the natural question of $Y \perp\!\!\!\perp X_j$ – we are interested in whether X_j provides information beyond the information from the others. (Of course, we have to address the pathological case where two variables are extremely correlated or actually identical, but we'll do that later on.) Most literature just tests for independence via **marginal tests**, but that doesn't make much sense because of the strong correlation between nearby parts of our chromosome (**linkage disequilibrium**). So if X_1, X_2 are strongly correlated but it's because X_2 is caused by X_1, we might think that both X_2 and X_1 are both discoveries. A conditional test would ideally then only detect X_1 as a discovery but not X_2. (See Example 60 below.)

If we were in an introductory stats course, the equivalent of this would be positing in a linear model that $Y \sim N(\beta^T x, \sigma^2 I)$ and checking whether $\beta_j = 0$. So this is a natural question to do, but we can't do that because the model is way overfit in this case (far too many variables).

Example 58

Instead, we can reframe the question in terms of graphical models. We can consider the dependency graph between our random variables, and then the interesting variables are just the neighbors (the Markov blanket) of Y.

To find interesting variables, we can now try to score them. We live in an era of machine learning and deep learning and so on, and it would be nice to be able to use some of these tools to make scientific discoveries. What we do is feed in our matrix of data (with rows indexed by our samples, and columns indexed by the features X along with the response Y) and then get some black-box scores for feature importance. For example, in a random-forest algorithm, we might ask how many times we used a given feature to split the dataset. But then we would need to ask the question of whether we'd get the same range of importance if we replicate the study and how likely things are to pay off.

The main problem is that (at least in genetics experiments) we don't know how Y actually depends on X, and we don't know how to compute the distribution of these statistical estimates we get, and we also don't know how to get p-values. (For example if we see a lasso coefficient through lasso regression, we don't know the null distribution because we don't know what to compare it to. And bootstrapping doesn't work at all because we're in such high dimensions and even fitting linear models will just randomly interpolate with some subset of the variables without a clear interpretation – inference in high dimensions is hard.)

Definition 59

In the **conditional randomization test (CRT)** (of Professor Candés, Fan, Janson, and Lv in 2016), we can test the hypothesis $X_j \perp\!\!\!\perp Y | X_{-j}$ **if we assume** that we know the joint distribution of X. (This is okay in genetics because we have lots of models which help us detect errors and fill in missing data.) What we do is sample $\tilde{X}_j$ with a "synthetic null," meaning that $\tilde{X}_j \sim X_j | (X_{-j} = x_{-j})$. We then check whether or not

$$(X_1, \cdots, X_j, \cdots, X_p, Y) \stackrel{d}{=} (X_1, \cdots, \tilde{X}_j, \cdots, X_p, Y).$$

If we forget the Y here, then by definition this is always true. But the point is that with the Ys it might not be true if there is dependence – we're detecting an asymmetry that can test our hypothesis. Indeed, the point is that X_j and $\tilde{X}_j$ are equally consistent with X_{-j}s by construction, and asymmetry means the variables are directly connected in our graphical model. We call this an **imputation**.

The method from here is very simple: assuming we can do this resampling because we know the distribution (and if we don't have a good model, we fit one with deep learning), the conditional randomization test does the following:

1. First construct the score (test statistic) $t^* = T(X_j, X_{-j}, Y)$ on our observed data. For example if we do lasso regression, pick λ by cross-validation and let $T = |\hat{\beta}_j(\lambda_{\text{cv}})|$.
2. Now for each of the K patients in the sample, score them on the new imputed value by sampling $\tilde{X}_j$ conditional on X_{-j}; this gives us a score $t_k = T(\tilde{X}_j, X_{-j}, Y)$.
3. We can now get a **finite-sample p-value**

$$p = \frac{1 + \#\{k : t^* \le t_k\}}{K+1}.$$

Indeed, the distribution of this random variable is $\frac{1}{K+1}, \frac{2}{K+1}, \cdots, \frac{K+1}{K+1}$ uniformly under the null hypothesis (since t^* has an equal chance to be ranked anywhere among $\{t^*, t_1, \cdots, t_K\}$, since they're all iid samples), so this is a valid p-value.

Another way to say this is that we want to see whether the test statistic $T(X_j, X_{-j}, Y)$ is extreme compared to the distribution $T(\tilde{X}_j, X_{-j}, Y)$ from resampling. Importantly, we do need to sample from the conditional $X_1|X_{-1}$, and this is **different from doing a permutation test** in which we would just randomly permute the values of X_j around our sample to get our new values of T. That would be incorrect, because we would be essentially resampling from the marginal and then we're not preserving associations of X_j with **other** variables:

Example 60

Suppose we have a linear model with two standardized variables X_1, X_2 (mean 0, variance 1) so that $\text{corr}(X_1, X_2) = 0.5$, but $Y = X_2 + \varepsilon$ is just a noisy observation of X_2. Then we should care about X_2 but not X_1, even though X_1 happens to be correlated with Y.

The marginal correlations would then be $\mathbb{E}[YX_2] = 1$ and $\mathbb{E}[YX_1] = 0.5$, so we get a marginal association even though X_1 is not interesting. So if our statistic just calculates the correlation between the phenotype and a given variable, we'll find 0.5 for something uninteresting, and if we resample X_1 from the marginal we suddenly get a correlation of 0. This means that X_1' would not be a valid control sample – we'd think that we're very significant because 0.5 is so different from the "fake control" of 0.

Fact 61

There are problematic limitations of the CRT that are important to note. First of all, we need to do our tests a large number of times, and the Bonferroni threshold is low so this is very computationally expensive. And furthermore, we're reusing the same data repeatedly, so the p-values we get for different variables are in fact not independent.

Lots of papers have tried to improve on this and take shortcuts, but instead it's better to try to parallelize and do comparisons in one step. There are instead **knockoffs**, which we'll talk about more next lecture. In short, the idea is to take a different point of view and resample all variables at the same time and compute for instance just a single lasso:

- For each variable X_i being considered, make a knockoff version of that variable. (So if we start with a data matrix with rows indexed by patients and columns indexed by SNPs, we append to X a new data matrix with fake SNPs, one per real variable.)
- Instead of running the lasso on just X, we run it on the augmented matrix $(X, \tilde{X})$. In particular, we get a lasso estimate of the importance $|\hat{\beta}_j|$ for each variable X_j, but we also get one for the corresponding knockoff variable $|\hat{\beta}_{j+p}|$.
- If these variables were null, then these statistics would have the same distribution since they are exchangeable. So we can do a comparison to get a sense of whether we're important. Basically for any null variable, the two lasso coefficients should have equal chance to be larger, so the procedure sets a threshold so that there are some real data points above it but very few false data points; the FDR is then well-estimated by the ratio of the points above the threshold, and there is a clever filter (called **SeqStep**) to make this work for finite samples.
- We then select all variables which are above this threshold and larger than their knockoff variant.

We'll see the mechanics of this next time, but we need to understand how to create this fake data matrix which looks like a SNP matrix. For comparison, we could think about taking unrelated people and permuting their genomes to form $\tilde{X}$, but **that is a very bad idea** because the distribution of $\tilde{X}$ is the same as X and then the distribution of

feature importance for the permuted data will look nothing like the distribution of feature importance for the real data. Instead, we point out that the property we need satisfied is that

$$(X, \tilde{X}) \stackrel{d}{=} (X, \tilde{X})_{\text{swap}(j)},$$

where the right-hand side takes the jth column of X and the jth column of $\tilde{X}$ and swaps them.(And of course, we should not be looking at Y here.) If this is the case, then the scores we get out for the importance of our real and fake variables should be **exchangeable for nulls** (that is, if X_j does not influence the responds beyond what we know for the other variables, then our statistics $Z_j, \tilde{Z}_j$ will be exchangeable.) This will turn out to be sufficient for what we want to do, and we'll see how this process works in detail next time! What's nice is that lots of this is model-free and will fall under "non-parametric methods."

8 April 24, 2025

Remark 62. *As a quick remark about high-dimensional inference (continuing on the comments from last lecture), it's generally very hard to get things to work with the number of parameters p is large. For example, it's very hard to get a confidence interval on a particular β for a linear model $y = x\beta + \varepsilon$ with $n = 10000$, $p = 50000$, or to get a bound on variance on β for a logistic model with $n = 1000$, $p = 100$.*

We'll continue our discussion of the conditional randomization test today – we talked last time about how it's computationally expensive and that we need to find some shortcuts. This lecture will discuss **model X knockoffs**.

Recall that the test we want to do is whether our phenotype Y is independent from a particular gene X_j when conditioned on all other variables X_{-j}. (This doesn't give causality, but in some applications it's getting a lot closer to causality than just doing a marginal test – the arrow is much more clear in something like genetics.) We'll discuss the technical details of those model-X knockoffs today and next lecture.

Example 63

Our setting is that we have a matrix of covariates, and our goal is to create "fake genotypes" $\tilde{X}$ in such a way that $(X, \tilde{X})$ has the same distribution as if we swap the jth columns of X and $\tilde{X}$. (That is, we can't tell unless we look at Y whether we're looking at X or $\tilde{X}$.)

So it's good to keep in mind that $\tilde{X}$ is "null" in the sense that

$$\tilde{X}_j \perp\!\!\!\perp Y | X, \tilde{X}_{-j}$$

(since we created $\tilde{X}$ without looking at Y). So the idea is that for any null variable Z_j, **the null scores will be exchangeable** regardless of what our blackbox algorithm $\mathcal{A}$ – taking in as input $[X, \tilde{X}], Y$ and trying to score the importances $Z_1, \cdots, Z_p, \tilde{Z}_1, \cdots, \tilde{Z}_p$ of our variables – does. Formally,

$$Z = \mathcal{A}([X, \tilde{X}], Y) \stackrel{d}{=} \mathcal{A}([X, \tilde{X}]_{\text{swap}(j)}, Y)_{\text{swap}(j)}.$$

(We can think of fitting a lasso and getting coefficients $\hat{\beta}_1, \cdots, \hat{\beta}_p, \tilde{\beta}_1, \cdots, \tilde{\beta}_p$; the point is that if we swapped the real and fake j, we would get the same distribution back except with $\hat{\beta}_j$ and $\tilde{\beta}_j$.) For something like the lasso we have this deterministically, but even with stochastic gradient descent or something it will still hold in distribution.) And that fact is all we're going to need, so we can accordingly choose whatever algorithm $\mathcal{A}$ we want.

So nothing in the score will tell us about X or $\tilde{X}$ if variable j is null (because Y cannot give us any more information): we will have $(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$ for any null j.

Fact 64

It's important to note that we can only use this method if we know the distribution of X, so that we can create $\tilde{X}$ by sampling the rows from it. And in something like genetics, we do have a very good sense of the distribution of genotypes.

Note however that there are some complications because $\tilde{X}$ needs to still depend on X, so we do not just resample independently from the distribution of X:

Example 65

Suppose X_1, X_2 are normal with mean $\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$. (So each of X and $\tilde{X}$ should have two columns. We'll assume that we get iid samples, so each row is independent.) Now to create our knockoff $\tilde{X}$, we need to create a function which samples $\tilde{X}|X$ in such a way that our exchangeability condition holds.

In particular, this means X_1 and $\tilde{X}_2$ still need to have nontrivial correlation, and so do X_2 and $\tilde{X}_1$. We can think of this by creating a 4-dimensional Gaussian; the means necessarily must be $\begin{bmatrix}\mu_1 \\ \mu_2 \\ \mu_1 \\ \mu_2\end{bmatrix}$ (because otherwise we would be able to tell which column is which). For the covariance matrix, by exchangeability it must take the form

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12} & * & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{21} & * \\ * & \Sigma_{12} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & * & \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

and **these are the only conditions we need to have** because the density here is symmetric in x_1 and $\tilde{x}_1$, and also in x_2 and $\tilde{x}_2$ So if we can find a matrix which is positive semidefinite of this form, with any values of $*$ which keep the matrix symmetric, this will be valid and it will be a valid joint distribution for $(X, \tilde{X})$. (Then we have plenty of standard tools – regression formula and Schur complement – to figure out the conditional distribution for $\tilde{X}_1, \tilde{X}_2$ given X_1, X_2).

All that remains is putting something on the diagonal – this is asking us about the covariances $\text{Cov}(\tilde{X}_1, X_1)$ and $\text{Cov}(\tilde{X}_2, X_2)$ between our fake and true SNPs. We wouldn't want to make them equal to Σ_{11} and Σ_{12}, because that would just be duplicating X and that won't create any contrast. Instead, it would make the most sense to set those values to 0, but we don't know whether that would make our matrix positive semidefinite. Thus this ends up coming down to a **convex programming problem**, making those diagonal $*$ entries as small as possible while still ensuring that we're PSD – this is a semidefinite program.

Remark 66. *Notice that this proposal is different from the following alternative way to construct knockoffs, which works regardless of the joint distribution: choose $\tilde{X}_1$ to be sampled from the conditional distribution $X_1|X_2$, and then choose X_2 from $X_2|X_1, \tilde{X}_1$. (we need to do this "Gibbs sampling" where we also include $\tilde{X}_1$ as a conditional, or else we won't get the correct distribution). This bypasses the semidefinite programming, but it's expensive because we have to keep track of more and more variables to get $\tilde{X}_1, \cdots, \tilde{X}_n$.*

We can now think about how we might use this in real life: this construction is nice because it asks questions like "do we know the distribution of X," rather than making up models for randomness and asking for precise facts about "made-up model." (When Professor Candés took Stats 305, he had to model height on length of the church in England, but it's not really clear what the point of that is because we can't make any inferences if we have the full census of all the churches already.) The point is that we've often "assumed away" a lot of stuff, but here we can put the randomness where we know it occurs (in genetics, this is during meiosis, since recombination spots are quite random) and we can use that to make inferences about the actually unknown dependence $Y|X$.

Fact 67

With something like the UK Biobank dataset, we have 500000 individuals (some of which are related) with 500000 SNPs documented and ancestry recorded. We can then build a hidden Markov model and build knockoffs using that. It turns out that if we calculate the first two principal components of our genotype, individuals cluster by ancestry, and this occurs even with our knockoffs as well. And (this is what we called linkage disequilibrium) we see that $\text{corr}(X_j, X_k)$ ranges heavily from 0 to 1 depending on the SNPs, but those values are quite close to $\text{corr}(\tilde{X}_j, \tilde{X}_k)$ and also to $\text{corr}(\tilde{X}_j, X_k)$. So the exchangeability assumption does indeed hold up. The key fact is that even though it's difficult to generate knockoffs in general, it's easy for Markov chains or hidden Markov models because we don't get the same complexity in Gibbs sampling. There are some complications with relatedness (we need to be careful about siblings and parents), but if we do so carefully everything works out.

We'll return to the theory now, developing a test for the conditional hypothesis. We can say something more than that $(Z_j, \tilde{Z}_j)$ are exchangeable:

Proposition 68

Define a test statistic $W_j = w_j(Z_j, \tilde{Z}_j)$ such that $w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$. (For example, we could just let W_j be $Z_j - \tilde{Z}_j$.) Then conditional on $|W|$, the signs of the null W_js are iid coin flips.

The idea is that if we want to know whether variable 1 is important, we can fit a lasso at some point λ and get regression coefficients $\hat{\beta}_1, \tilde{\beta}_1$. We can then let our score be

$$W_1 = |\hat{\beta}_1| - |\tilde{\beta}_1|,$$

and if this quantity is large then we think variable 1 may be important. Thus if we compute all of our test statistics and plot the values of $|W_j|$ for all variables on a number line, then (regardless of the blackbox we use) the sign of each test statistic is an independent coin flip for each null.

That fact will be enough to get FDR control, since we'll apply the program we see before: the candidates for good discoveries are the ones where W_j is large (for example if the lasso thinks X_1 matters but the "control" $\tilde{X}_1$ does not). Thus we need to select the variables j where $W_j \geq t$ for some t, and we need to determine what t is. Much like before, we now have

$$\text{FDP}(t) = \frac{\#\{j \text{ null} : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1}.$$

We don't know what this quantity is exactly, but because the nulls are equally likely to be positive or negative, we also have by symmetry that

$$\text{FDP}(t) \approx \frac{\#\{j \text{ null} : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}.$$

And now we expect most of the very negative values to come from nulls, so we can make our estimate by bounding

this as

$$\text{FDP}(t) \approx \frac{\#\{j \text{ null} : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le \frac{\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} = \widehat{\text{FDP}}(t)$$

This will turn out to be a pretty good estimate, and we won't even be off by that π_0 factor from before! So now we step up in $|W|$ (at the start our estimate is 0) and we stop the first time it jumps below α.

Theorem 69

This procedure described, with a slight modification like in Storey's procedure (see below), controls the false discovery rate. That is, we get FDR $\le \alpha$ for any user input significance level.

Much like the proofs we saw before, this one comes from martingale theory. And we'll be explicitly clear about what the selection is: we choose

$$\tau = \min\left\{t : \frac{1 + |S^-(t)|}{|S^+(t)| \vee 1} \le \alpha\right\}$$

with $S^+(t) = \{j : W_j \ge t\}$ and $S^-(t) = \{j : W_j \le -t\}$, and then our selection of nonnulls is $S^+(\tau)$.

9 April 29, 2025

We'll wrap up conditional testing today and then move on to other topics – at the end, we'll talk a bit about causality and why we're "close to it."

Recall from last time that we're testing whether $Y \perp\!\!\!\perp X_j | X_{-j}$ (that is, X_j helps us in predicting Y, given that we already know about all the other variables). We've seen a framework for this now: we create fake variables in an intelligent way (so that swapping variables X_j and $\tilde{X}_j$ just swaps the black-box scores Z_j, $\tilde{Z}_j$ that we get out), and then we combine the scores into a single constant W_j for each real variable j. (This constant needs to have the "flip-sign property so that conditional on the magnitudes, the nulls' signs are always iid coin flips.)

We will then use this to assess importance: we typically expect the high values of $|W_j|$ to come from non-nulls and be positive, and so we have a selection procedure which selects these large positive scores (the black-box thinks it's important, and it's quite a bit larger than the corresponding control). So we start from 0 and go from left-to-right, and we want to set a threshold t which is the first time the false discovery proportion estimate is below α. (So we compare the number of minuses to the number of pluses, and if the number of pluses is 20 times as large we stop.) Precisely, let $S^+(t) = \{j : W_j \ge t\}$ and $S^-(t) = \{j : W_j \le -t\}$. Then our estimate of FDR is $\frac{1+|S^-(t)|}{|S^+(t)\vee 1|}$, and we let τ be the minimum t where this ratio is at most α.

Proof sketch for FDR control. When the procedure stops, we have some false discovery proportion which is equal to

$$\text{FDP}(\tau) = \frac{\#\{j \text{ null and } j \in S^+(\tau)\}}{\#\{j : j \in S^+(\tau)\} \vee 1}.$$

Using a similar stopping rule as before, we can rewrite this as

$$\text{FDP}(\tau) = \frac{\#\{j \text{ null and } j \in S^+(\tau)\}}{\#\{j : j \in S^-(\tau)\} \vee 1} \cdot \frac{\#\{j : j \in S^-(\tau)\} \vee 1}{\#\{j : j \in S^+(\tau)\} \vee 1}.$$

But now the latter fraction is at most α when we stop, so the FDP is at most α times the former fraction. Then

$$\frac{\#\{j \text{ null and } j \in S^+(\tau)\}}{\#\{j : j \in S^-(\tau)\} \vee 1} \le \frac{V^+(\tau)}{1 + V^-(\tau)},$$

and now because our nulls are symmetrically distributed we expect $V^+(\tau)$ and $V^-(\tau)$ to be approximately equal; indeed

the expected value of this quantity is bounded by 1 by a martingale argument. We can check that with respect to the sigma-algebra $\mathcal{F}_t = \sigma(\{V^{\pm}(u))_{u\le t}\})$, $\frac{V^+(t)}{1+V^-(t)}$ is a **supermartingale** (meaning expectations are nonincreasing) using some hypergeometric random variable calculations. (As an illustration of the kind of calculation we need to do, suppose we have an urn with 17 balls, and we know 10 of them are marked with $+$s and 7 of them are marked with $-$s. Then we remove 3 of them uniformly at random; we need to show that the expected value of $\frac{(+\text{ balls})}{1+(-\text{ balls})}$ is at most $\frac{10}{8}$. In fact, it'll be exactly equal.) Thus

$$\mathbb{E}\left[\frac{V^+(s)}{1+V^-(s)}\middle| V^{\pm}(t), V^+(s)+V^-(S)\right] \le \frac{V^+(t)}{1+V^-(t)},$$

and so this calculation is also true marginally. Therefore

$$\text{FDR} \le \alpha\mathbb{E}\left[\frac{V^+(\tau)}{1+V^-(\tau)}\right] \le \alpha\mathbb{E}\left[\frac{V^+(0)}{1+V^-(0)}\right] \le \alpha,$$

where this last step comes from the symmetry of the nulls, meaning $V^+(0)$ is binomial with parameters $(n_0, \frac{1}{2})$ (so we reduce to a calculation we've seen in a previous lecture). □

What's remarkable about this argument is that throughout most of it, we are very close to equality, so we aren't losing a factor like π_0 in Benjamini-Hochberg.

Remark 70. *To summarize what we have so far, we're introducing a different kind of testing to what we might be used to. Previously, we might write down a linear model and run a t-test, but those kinds of assumptions are faulty and often give us a fixed set of observations X to work with. What's nice about the knockoffs framework is that it works in any dimension (including $p > n$) and with any model for $Y|X$ and any black-box. The main thing is just needing access to the law P_X so that we can construct knockoffs, and then we think of observations of X as random. This is generally more appropriate in "big data applications," and it turns out that even if the model isn't exactly correct, we still get useful inference.*

Fact 71

Overall, the tradeoff is a **shift in burden of knowledge**: we need knowledge of X rather than knowledge of the dependence $Y|X$ to do these tests correctly.

In something like genetics, it's useful because we in fact do not know how phenotypes depend on genotypes, but we do actually have very good understanding of the distribution of genotypes in the population. And then we can use this new framework regardless of what model for $Y|X$ we want to use – there's no inherent risk to power because we can choose whatever machine learning algorithm we think represents the dependence correctly. And we can tell if our representation of P_X is good via data imputation.

Example 72

Some researchers were trying to study the effect of policies on number of COVID cases. They had a tensor with state we live in, the number of COVID cases by date, and dates of certain policies. The questions of interest then look like "what would happen if California implemented a different policy."

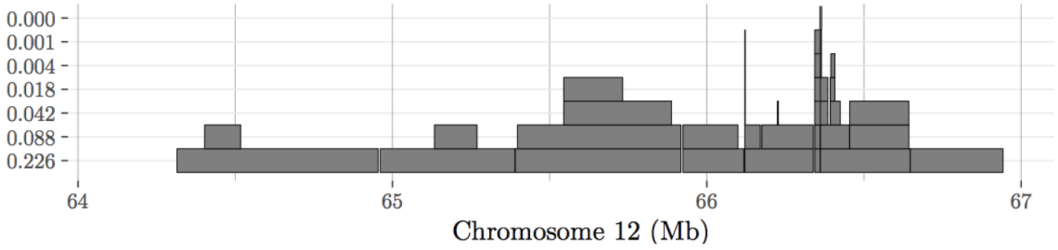
This is a tensor completion problem, and so people tend to make a story about low rank tensor approximations "plus some additional error." But in order to make this make sense, we need to understand "where the randomness ω is:" we've seen all the real data already, so we can't control randomness to make real conclusions. (Professor Candés

is skeptical about being able to answer those kinds of "what if" questions, especially in terms of burying things in layers of math. In particular, sometimes it's useful just to use statistics for description and summary rather than dubious inference.) In contrast, in something like genetics, we know randomness actually comes from inheritance and thus we can model that properly.

We'll now forget a bit about knockoffs and focus more on the conditional testing aspect, specifically **accounting for correlations**. Recall that nearby SNPs are very correlated, and perhaps there is a causal relationship between one of them and the phenotype, while the other one is null. When we collect data, it'll be difficult to distinguish the two SNPs, but we don't want to get the answer wrong.

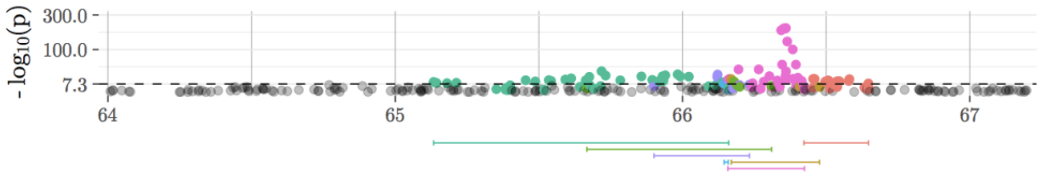
One solution is testing over regions – we can compute a correlation matrix between SNPs and do **hierarchical clustering** (so at the finest level they're all in their own clusters, and then based on correlations they cluster together). Once we have these clusters $C_1, \cdots, C_m$, we can then try to answer the question of $Y \perp\!\!\!\perp X_c | X_{-c}$. We then have a tradeoff, where the question is easier to answer if the cluster is larger but the information we get from it is weaker.

In the image below, there are "seven levels of clustering:" at the very top we have just SNPs, and at the bottom we have large clusters. We can then do knockoff analysis for each one (this is essentially the same as what we discussed before, except that we have more freedom in actually building knockoffs because we don't need exchangeability **among** the cluster). This image shows that we have something interesting in each large region, but as we get finer we localize the regions of interest more and more. (The stronger the effects, the more carefully we can observe their exact locations.)



Remark 73. *Based on everything we've said in this class, note that we cannot take "every highest peak" and mark that down as an independent discovery, because FDR control is only done layer by layer. (Except that we can in fact combine information and still control FDR due to some recent work, and we might hear more about this later on.)*

In contrast, we can do marginal analysis and do univariate regression, as shown below. But it's harder to make sense of that – we still have to do additional processing and understand what the overlapping clumps really mean.



The point is that this strategy (going under the name KGWAS) manages to control the FDR, and it can make precise discoveries and pinpoint the causal things at "low width" (that is, localizing the sources quite well).

Fact 74

Even in genetics, where we know molecular biology and thus can say that editing the genome actually changes the phenotype (and not the other way around), what we're doing is not sufficient for causality yet.

It's rather philosophical to even get an exact definition of causality, but in this setting we can kind of ask it in the context of "if we edit the genome, do we get a different distribution for Y?". Forgetting the knockoffs now, rejecting

the null hypothesis $Y \perp\!\!\!\perp X_j | X_{-j}$ tells us that X_j influences Y, but it could be that there is some other confounding variable U which determines both X_j and Y. But when we say this kind of thing, we need concrete examples:

Example 75

For an easy case, it's possible that we do not end up sampling every spot on the genome (that is, we don't type all of the mutations), and thus we're not getting causality because we're only taking the nearest neighbors that are most correlated. But for a more interesting example, suppose gene A is a "good parent" gene which makes you encourage your children to go outside, and gene B makes your muscles lean. Then if Y is the phenotype for running ability, then clearly B causes Y (and it will be found by this kind of analysis). But this analysis will also find A, because knowing what your parents make you do influences Y (the children of good parents will tend to run faster). However performing an intervention on gene A will not change much; it's an interesting discovery but it's not causal in this interventional sense.

The point is that **randomization** is what gives causal inference, because it corrects for unknown confounders, and genome-wide association data is not randomized. That being said, we can add an additional step to get to causal inference via the **transmission disequilibrium test** – we can condition on the genome of parents, since now we are a randomized experiment coming from thermodynamic fluctuations during meiosis and basically nothing else. In such a test, we let our data be the haplotypes of a subject and their parents, so that we have our genotype $X = (X^m, X^f)$ and our ancestors $A = (M^a, M^b, F^a, F^b)$ (where each of $X^m, X^f, M^a, M^b, F^a, F^b$ are valued in $\{0,1\}^{n \times p}$).

Definition 76

With the notation above, we call a variable Z an **external confounder** if

$$X|(A = a, Z = z) \stackrel{d}{=} X|(A = a, Z = z')$$

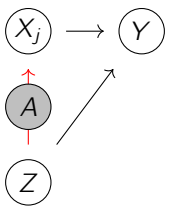
for any value of a, z, z'; that is, we don't change the recombination based on the value of Z.

(It's hard to come up with variables which aren't external confounders, because it's hard to think of anything else which would cause meiosis to behave differently; any environmental condition after conception is external.)

Theorem 77

For any external confounder Z, any valid test of the hypothesis $H_0 : Y \perp\!\!\!\perp X_j | (X_{-j}, A)$ is also a valid test of $H_0' : Y \perp\!\!\!\perp X_j | (X_{-j}, A, Z)$.

Thus, we're showing that our test is valid regardless of confounders Z (even if we don't know what they are!). So conditioning on X_{-j} and A means we end up also conditioning on all other external variables as well, and that's enough for causality (the effect $X_j \to Y$ can't be due to diet or location or anything else, unless it somehow strangely affects meiosis). In the graphical model picture, we're basically saying that conditioning on A blocks the potential effect of Z, so we must have a real causal arrow $X_j \to Y$.



10 May 1, 2025

Last time, we introduced the concept of an **external confounder** (in the context of genetics, this means that the distribution of X conditioned on the ancestors A does not change if we additionally condition on any external confounders Z; that is, $X|A, Z \overset{d}{=} X|A$). What's powerful is that if we have a test $Y \perp\!\!\!\perp X_j|(X_{-j}, A)$, we are then implicitly testing the more general $Y \perp\!\!\!\perp X_j|(X_{-j}, A, Z)$ (meaning we still get an α-level test). So when people say that we cannot do causal inference because of external confounders, we're largely managing to get around that here. It's very well-described statistically how to get the distribution of X given A, so this isn't too difficult to do.

Today we're moving on to a new topic, e-**values**. This is a relatively new topic in the course, and we'll start today by explaining what they are and what we can do with them (controlling the FDR).

Remark 78. *For some references, we can see Grünwald's paper "E is the new P: Tests that are safe under optional stopping, with an application to time-to-event data," as well as (for next lecture) Shafer's "Testing by betting: A strategy for statistical and scientific communication" and Wang's "Game-theoretic statistical inference E-values vs p-values, calibration, combination, and closed testing."*

Example 79

Consider the following situation (which is real): research group A tests a medication and gets a promising but not conclusive result (whatever that means – perhaps it's risky to go to trial). Then research group B tests again on new data, but it's still not clear, so research group C tests next (again on new data). We then want to understand how to combine these test results together even when we aren't following a fixed plan.

The point is that this is different from what we did at the beginning of the course: if research group A got a p-value of 0.6, then maybe research group B wouldn't have done a test at all, so it's not like we're deciding ahead of time to do some fixed number of tests and use some Bonferroni-type threshold. So this is a dynamic thing and continuing tests might depend on various factors, not necessarily under our control.

The current method is typically to sweep all of the data together and recalculate the p-value, but here we have no plan and the ideas of meta-analysis or sequential analysis do not clearly apply mathematically. And what we want to avoid is p-hacking by giving us too many chances to notice something significant.

The e-value is a generic replacement of the p-value which will handle this problem of optional continuation:

Definition 80

Recal that a null hypothesis $\mathcal{H}_0$ is basically a collection of probability measures. An **e-variable E for testing $\mathcal{H}_0$** is a nonnegative random variable such that

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{E}_{P_0}[E] \leq 1.$$

A realization of an e-variable is called an **e-value**. Meanwhile, a **p-variable for testing $\mathcal{H}_0$** is a nonnegative random variable that satisfies

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{P}_{P_0}(P \leq \alpha) \leq \alpha$$

for all $\alpha \in (0, 1)$, and a realized value of a p-variable is a p-value.

The idea is that the constraint for being a p-variable is controlling the cdf, while the constraint for an e-variable is much weaker (just controlling the expectation). And in a simple hypothesis (So there's only a single probability measure), we're just asking for the variable to have mean at most 1.

Proposition 81

For any e-value E, the random variable E^{-1} is a conservative p-value (meaning that it is a p-value with wiggle room).

Indeed, we have

$$\mathbb{P}\left(\frac{1}{E} \leq \alpha\right) = \mathbb{P}\left(E \geq \frac{1}{\alpha}\right) \leq \frac{\mathbb{E}[E]}{1/\alpha} \leq \alpha.$$

(There are other simple functions of E which work as as well.) The main problem is that there can be a fair bit of wiggle room if we test by rejecting the null if $E \geq \frac{1}{\alpha}$ – we have no description of tails and thus knowing whether we're in the tail can be sometimes weak. But this will still turn out to be useful under certain perspectives, specifically if we construct the correct e-values.

Example 82

We can connect this material with likelihood ratios in the following way. Suppose we have a set of null hypotheses $\mathcal{H}_0 = \{p_\theta : \theta \in \Theta_0\}$ and a set of alternatives $\mathcal{H}_1 = \{p_\theta : \theta \in \Theta_1\}$.

We might be interested in the **Bayes factor**

$$\frac{p_{W_1}(X)}{p_{W_0}(X)}, \quad \text{where} \quad p_{W_0}(X) = \int_{\theta \in \Theta_0} p_\theta(X) dW_0(\theta), \quad \int_{\theta \in \Theta_1} p_\theta(X) dW_1(\theta),$$

which essentially gives us a sense of how much more likely X is to occur in $\mathcal{H}_0$ versus $\mathcal{H}_1$ and it's like a generalized likelihood test. In particular, if we have a simple null, then $\mathcal{H}_0 = \{p_0\}$ is a single point and our Bayes factor is of the form

$$M(X) = \frac{p_{W_1}(X)}{p_0(X)}.$$

The point is that **under a simple null**, this quantity is an e-value – regardless of the prior W_1, we have

$$\mathbb{E}_{X \sim p_0}[M(X)] = 1$$

(this is the same argument as how the expected value of the likelihood ratio is 1). And if our alternative is also just a point hypothesis, our e-value looks like

$$E(X) = \frac{p_1(X)}{p_0(X)},$$

and we want to reject this for large values of E (for example if $E \geq 20$). But this "safe test" is not quite the same as Neyman-Pearson testing, which would reject at a much smaller threshold $E \geq \frac{1}{B}$ where $\mathbb{P}_{p_0}(E \geq B) = \alpha$ (in words, whatever the tail of the likelihood ratio actually is under the null).

On the other hand, for any e-variable E, we can also interpret the e-value as a likelihood ratio by defining the alternative in terms of the equation $\frac{p_1(X)}{p_0(X)} = E$ (and then p_1 will integrate to 1). So we can really think of this as "alternative view of the world."

Example 83

Suppose we're doing a simple case of classical testing, where $X = (X_1, \cdots, X_n)$ are iid $N(\mu, 1)$ random variables. Suppose the null is that $\mu = 0$ and the alternative is that $\mu = \mu_1$.

We can then calculate explicitly (as we've already seen)

$$E = \prod_{i=1}^{n} \exp\left(\mu X_i - \frac{\mu^2}{2}\right).$$

The safe test is then rejecting when

$$\sum_{i=1}^{n} \left(\mu X_i - \frac{\mu^2}{2}\right) \geq \ln 20 \approx 3,$$

so if we reparameterize $\mu = \frac{t}{\sqrt{n}}$, the threshold for Neyman-Pearson to reject is $\overline{X} \geq \frac{1.64}{\sqrt{n}}$, while the threshold for the e-value to reject is $\overline{X} \geq \frac{1}{\sqrt{n}}\left(\frac{3}{t} + \frac{t}{2}\right) \geq \frac{\sqrt{6}}{\sqrt{n}}$. Thus we will lose some power.

More generally, suppose our alternative is now $H_1 : \mu \in \Theta_1$ for a more general location family. We could then have some prior $w(\mu) \propto \exp\left(-\frac{\mu^2}{2}\right)$ on the alternatives, and the Bayes factor now looks like

$$E = \frac{\int_\mu p_\mu(X) w(\mu) d\mu}{p_0(X)}.$$

This is an e-value, and we calculate with a straightforward calculation that

$$\log E = -\frac{1}{2}\log(n+1) + \frac{1}{2}(n+1)\breve{\mu}_n^2, \quad \breve{\mu}_n = \frac{n}{n+1}\overline{X},$$

where $\breve{\mu}_n$ is our Bayes MAP (maximum a posteriori) estimator. So we'll indeed reject if $\overline{X}$ is large, which is a reasonable thing to do, but the question is how much it has to deviate. It turns out that our safe test now rejects the null when $|\tilde{\mu}_n| \geq \sqrt{\frac{2\ln 20 + \log(n+1)}{n+1}}$, which isn't quite right – we'd like to see 1.96 in the numerator.

Remark 84. *When we look at the log-likelihood ratio under the alternative, it's a sum of independent terms and thus* $\log E$ *will concentrate around its mean (while* $\breve{\mu}_n$ *converges to* 0*). So more data is still good, but the question we're often asking is "what is the loss in effective sample size to detect a specific effect size compared to Neyman-Pearson," and that's where the* $\log n$ *factor comes in.*

Fact 85

The advantage of e-values is that we don't know how to compute p-values for a lot of problems (and to compute an e-value, we just need to control the mean). So in high-dimensional settings or more irregular models, it's often useful to use them. And as we alluded to at the start of the lecture, it will also help us with more sequential studies – the dependence will not be a problem because expectation is much easier to deal with (using stopping times) than entire probability distributions.

Example 86

We'll be looking at **safety under optional continuation** in this lecture: suppose $(X_1, Z_1), (X_2, Z_2), \cdots$ is our data, where Z_i is some "side information" (for example, whether we have enough money to keep running experiments). Suppose that the data comes in batches of size $n_1, n_2, \cdots$, and $N_t = \sum_{i=1}^{t} n_i$ is the size of the data we have so far.

We'll establish an e-value E_1 on the first batch. From there, we evaluate an e-value E_2 on the next batch, but only if the outcome is in a certain range (promising but not conclusive) and the external factors take on certain values (things that we cannot plan) – otherwise we stop early. Then depending on the outcomes and external factors up until the second batch, we decide whether or not to compute E_3, and so on. But the point is that after τ total data

batches, **the final result we report is the product**

$$V_\tau = \prod_{i=1}^{\tau} E_i.$$

In particular, we're allowed to choose whether to continue on depending on whether each individual E_i is above some threshold of our choice, and the point is that we'll still be able to control the type I error:

Theorem 87

Regardless of the stop-continue rule, as long as τ is a stopping time, V_τ is itself an e-value. More formally, let $\mathcal{F}_t$ be a filtration. Suppose that for all t, the conditional e-variable E_t is a nonnegative random variable which is $\mathcal{F}_t$-measurable, and such that for all $P_0 \in \mathcal{H}_0$ we have $\mathbb{E}_{P_0}[E_t|\mathcal{F}_{t-1}] = 1$. (This is easy to check in practice.) Then $V_t = \prod_{i\le t} E_i$ is a nonnegative supermartingale under the null. Thus by the optional stopping theorem, for any stopping time τ with respect to the filtration, $V_\tau = \prod_{t=1}^{\tau} E_t$ is an e-value. (In particular, V_t is an e-value for each fixed t.)

This even means we can accumulate evidence until $V_t \ge 20$ and let our stopping time τ be the first t where this occurs, and we will still control the type I error! And indeed, if we actually have a non-null hypothesis, $\frac{p_1}{p_0}$ will have expectation greater than 1, and so we will get a compounding effect and with enough data the ratio will eventually be large enough. (So here we can think of the process as "positing a better distribution and betting on it;" furthermore, we can redesign p_1 adaptively as we continue on, since we're only conditioning on the past.) The idea is that any nonnegative supermartingale which starts at 1 can only ever globally achieve 20 with probability $\frac{1}{20}$, and we can be completely adaptive with the procedure as long as we satisfy that specific inequality $\mathbb{E}_{P_0}[E_t|\mathcal{F}_{t-1}] = 1$.

Example 88

A paper by Professor Candes along with Huang, Jin, Li, Li, and Leskovec called "Automated Hypothesis Validation with Agentic Sequential Falsifications" (appearing in ICML this year) designs an automated AI agent which validates experiments of various sub-hypotheses given a main hypothesis and some type I error rate. This AI agent basically collects data (Or finds it online), and automatically implements and computes p-values which are converted to e-values (via $e = \kappa \times p^{\kappa-1}$). It then decides whether to stop based on whether $\prod E_i \ge \frac{1}{\alpha}$. (It turns out this agent got very similar conclusions to expert biologists but in a much shorter amount of time...) And this use of e-values was important for controlling FDR.

11 May 6, 2025

Last time, we showed how to aggregate evidence across trials with an optional continuation concept, where we keep track of an e-value $V_t = \prod_{i=1}^{t} E_i$. Then V_t is a nonnegative supermartingale, and thus if we stop it at $\frac{1}{\alpha}$ we can safely control the FDR. The point is that under the null, the event $\left\{\sup V_t \ge \frac{1}{\alpha}\right\}$ has probability at most α by Markov's inequality as a consequence of the optional stopping theorem; more formally, $\tau = \inf\left\{t : V_t \ge \frac{1}{\alpha}\right\}$ is a stopping time, and a nonnegative martingale V_t converges to some random variable V_∞ almost surely. So

$$\begin{aligned}\mathbb{E}[V_0] &\ge \mathbb{E}\left[V_\tau 1\{\tau < \infty\} + V_\infty 1\{\tau = \infty\}\right] \\ &\ge \mathbb{E}\left[V_\tau 1\{\tau < \infty\}\right] \\ &\ge \frac{1}{\alpha}\mathbb{P}(\tau < \infty),\end{aligned}$$

so rearranging this yields $\mathbb{P}(\tau < \infty) \leq \alpha$.

We'll discuss e-values even more today, first introducing a point of view of "testing by betting' and using the GROW criterion as an analog of power. (The references for this lecture are Shafer's "Testing by betting" paper, as well as work by Wang and Ramdas and also by Ren and Barber.

Example 89

Let's think about conventional hypothesis testing in the way we typically learn about it: we have a hypothesis that P describes a random variable X, and we want to use instances $X = x$ to test the hypothesis P. Conventionally (the Fisherian answer), we pick a significance level α, pick an event A with $\mathbb{P}(A) = \alpha$, and then we reject the null if $x \in A$. But we can also interpret this in terms of betting: we put a dollar on the event A, and we get back \$20 if A occurs (discrediting the hypothesis P) and no money otherwise.

To actually **measure the strength** of that evidence against a hypothesized distribution P, conventionally we construct a p-value which is the smallest α value for which the test (via a test statistic) rejects. So in other words, the smaller the p-value, the more evidence against P we have.

On the other hand, a bet in the e-value language is that we have a function $E(X)$ that can pay many different values depending on our outcome. We choose E so that $\mathbb{E}_P[E(X)] = 1$, and so we pay 1 dollar and get back $E(X)$ dollars. So the larger E is, again the more evidence we accumulate against P. So this $E(x)$ is a betting score, and in fact we can think about it as a likelihood ratio as discussed last lecture:

Lemma 90

A random variable E is a betting score if and only if $E(X) = \frac{dQ}{dP}(X)$ for some other distribution Q.

(Indeed, we're saying that if $E(x) \geq 0$ and $\mathbb{E}_p[E(x)] = 1$, then we can write $q(x) = E(x)p(x)$ and check that q is a probability distribution.) So proposing E is like proposing that data follows Q rather than P:

Theorem 91

Suppose that I think Q describes X. Then $E = \frac{dQ}{dP}$ maximizes the expected log wealth growth $\mathbb{E}[\log E]$, meaning that for any other distribution R (and independently of your prediction P),

$$\mathbb{E}_Q\left[\log \frac{dQ}{dP}(X)\right] \geq \mathbb{E}_Q\left[\log \frac{dR}{dP}(X)\right].$$

This is a form of "Kelly gambling," which says that there's only one way to bet intelligently and maximize the exponential rate of growth of money in expectation. (Maximizing rate of growth makes more sense than maximizing the expected value of the wealth itself – the latter might optimize for situations where you have very high wealth of very low probability, and that introduces too much risk.)

Proof. We can write the difference between the two sides as

$$\mathbb{E}_Q\left[\log \frac{dQ}{dP}(X) - \log \frac{dR}{dP}(X)\right] = \mathbb{E}_Q\left[\log \frac{dQ}{dR}(X)\right],$$

and this is nonnegative because this is the Kullback-Liebler divergence between the random variables Q and R (whose nonnegativity is just Jensen's inequality):

$$\mathbb{E}_Q\left[\log \frac{dQ}{dR}(X)\right] = \mathbb{E}_R\left[\frac{dQ}{dR}(X) \log \frac{dQ}{dR}(X)\right]$$

and now $Y \mapsto Y \log Y$ is convex and thus this is at most $\mathbb{E}_R\left[\frac{dQ}{dR}(X)\right] \log \mathbb{E}_R\left[\frac{dQ}{dR}(X)\right] = 1 \log 1 = 0$. □

So the point is that any test $E = \frac{dQ}{dP}$ expects to grow like $\exp(\mathbb{E}_Q[\log E]) = \exp(\mathbb{E}_P[E \log E])$, and we call this the "implied target." What's nice is that this quantity (if we know what P and Q are) can be evaluated in advance. The interpretaion is that if we expect the wealth to grow at a very slow rate even in an optimal world, then it doesn't make sense to invest in it (and in fact we should be reluctant to publish a paper if this implied target is very low).

Example 92

We talked last class also about Bayes factors $M(X) = \frac{p_{W_1}(X)}{p_{W_0}(X)} = \frac{\int p_\theta(x) d_{w_1}(\theta)}{\int p_\theta(x) dw_0(\theta)}$, where W_0 is some set of nulls and W_1 is some set of alternatives.

This is an e-value if there was just a single null hypothesis $p_0(x)$ contributing to the denominator, but **not in general**: we can only guarantee that coming from the mixture of nulls, $\mathbb{E}_{X \sim P_{W_0}}[M(X)] \le 1$. So this is a problem, since for example in vaccine testing we have contingency tables and we might want to average over various possibilities.

One way to get around this is the following: assume we have some prior W_1 on the alternatives Θ_1. Then for every θ_0 in the null W_0, we want the expected value of the e-value to be 1, so we want to find some Q with $\mathbb{E}_{X \sim P_{\theta_0}}\left[\frac{P_{W_1}(X)}{Q(X)}\right] \le 1$. This turns out to be possible: geometrically, we will find a **mixture of components of Θ_0 that is closest** to P_{W_1} in a KL sense. We have this notion of Kullback-Leibler divergence $D(P||Q) = \mathbb{E}_{X \sim P}\left[\log \frac{p(X)}{q(X)}\right]$, and we'll find a mixture W_0^* such that $D(P_{W_1}||P_{W_0^*})$ is minimized:

$$W_0^* = \text{argmin}_{W_0 \text{ distribution on } \Theta_0} D(P_{W_1}||P_{W_0}).$$

Once we solve this optimization problem, $P_{W_0^*}$ is called the **reverse information projection** of P_{W_1} on the set $\overline{H}_0$, which is the set of P_W over distributions W of Θ_0. In very special cases, this W_0^* will actually be a point mass at some θ.

Theorem 93 (Li '99, Barron–Li '00, Grünwald et al. '19)

Suppose this W_0^* exists (so the optimization problem has a solution). Then $\frac{p_{W_1}(X)}{p_{W_0^*}(X)}$ is an e-variable, and in fact it is the one we should use in the sense that it is the **GROW e-variable relative to W_1**, meaning that it maximizes $\log E$ under P_{W_1}:

$$\frac{p_{W_1}(X)}{p_{W_0^*}(X)} = \max_{E \text{ an } e\text{-variable for } \mathcal{H}_0} \mathbb{E}_{X \sim P_{W_1}}[\log E].$$

We can now think about computations (and we'll have some homework to do these computations as well):

Example 94

Let's make the setting discrete to make our lives easier, and suppose Θ_0 is finite (otherwise we do some discretization). We then want to find w_is (the weights on the null θ_is) to minimize the quantity

$$\sum_x p_{w_1}(x) \log \frac{p_{w_1}(x)}{\sum w_i p_{\theta_i}(x)},$$

subject to the condition that $w_i \ge 0$ and $\sum w_i = 1$.

We can simplify the objective because the $p_{w_1} \log p_{w_1}$ doesn't affect the minimum: our goal is really to minimize the quantity

$$\sum_x -p_{w_1}(x) \log\left(\sum w_i p_{\theta_i}(x)\right),$$

and this expression is actually convex in the w_is, so it's not too bad to do.

So the "GROW e-variable" maximizes the rate of growth of our "money" E – under the null we have $\mathbb{E}[E] \le 1$, and looking for GROWs is the same as looking for power since we're picking the one that "GROWs fastest" under W_1.

Example 95

We can do a lot with e-values, and in fact we can come up with an analog of the Benjamini–Hochberg procedure and control the FDR. Suppose we have n realized e-values $e_1, \cdots, e_n$ associated to $H_1, \cdots, H_n$, and we want to combine them.

As usual, we order them from most promising to least promising, so we have $e_{(1)} \ge \cdots \ge e_{(n)}$.

Definition 96

In the ***e*-BH procedure**, we reject the hypotheses of the largest $\hat{k}$ e-values, where

$$\hat{k} = \max\left\{ i : \frac{i e_{(i)}}{n} \ge \frac{1}{\alpha} \right\}.$$

Numerically, recall that we can create p-values out of e-values by using $\frac{1}{e_i}$, and this is **equivalent to applying Benjamini-Hochberg on those p-values** (since it's the last time $p_{(i)}$ is below $\frac{\alpha i}{n}$).

Fact 97

We should be careful that $\frac{1}{p}$ is not an e-value in general (in fact its expectation need not be finite), even though $\frac{1}{e}$ is always a p-value.

Theorem 98 (Wang–Ramdas '20)

The e-BH procedure has FDR at most $\frac{n_0 \alpha}{n}$, and no independence assumption is required.

In fact, if we work with e-values in our research, it's more generally true that

$$\text{FDR} \le \frac{\alpha}{n} \sum_{i \in \mathcal{H}_0} \mathbb{E}_{H_i}[E_i]]$$

even if we don't have "valid" e-values (we just need the sum of the expectations to be at most n, not that each individual one has expectation bounded by 1).

Proof. We can calculate the false discovery proportion; as usual, let R be the number of rejections, which is a random variable. We have

$$\begin{aligned}
\text{FDP} &= \frac{\sum_{i \in \mathcal{H}_0} 1\left\{ e_i \ge \frac{n}{\alpha(R \vee 1)} \right\}}{R \vee 1} \\
&= 1\{R \ge 1\} \frac{\sum_{i \in \mathcal{H}_0} 1\left\{ e_i \ge \frac{n}{\alpha R} \right\}}{R} \\
&\le 1\{R \ge 1\} \frac{\sum_{i \in \mathcal{H}_0} e_i / (\frac{n}{\alpha R})}{R} \\
&= \frac{\alpha}{n} \left(\sum_{i \in \mathcal{H}_0} e_i \right),
\end{aligned}$$

and taking expectations yields the result. □

The main problem is that this requires us to construct e-values that are quite large, say $\frac{n}{\alpha}$ – that's the main problem with the field, since it's difficult to cross even 20 (so 20000 will be very difficult). And even though this result looks fantastic and works even for dependence, it's quite conservative – we won't get much power out of this. In general, e-values are less likely to product rejections and thus underperforms for "0-1 decisions" – they only focus on expectation and not the tail behavior.

Example 99

It turns out that we can reinterpret what we've already seen and do a lot more with these e-values – for example, we can connect this idea with knockoffs. Specifically, letting T_α be the knockoff threshold and supposing we have p hypotheses, we can define the e-values

$$e_j = \frac{p \cdot 1\{W_j \geq T_\alpha\}}{1 + \sum_{k \in [p]} 1\{W_k \leq -T_\alpha\}}.$$

This random variable is always nonnegative, and if we sum over null js we get

$$\sum_{\text{null } j} \mathbb{E}[e_j] = p\mathbb{E}\left[\frac{1\{W_j \geq T_\alpha\}}{1 + \sum_{k \in [p]} 1\{W_k \leq -T_\alpha\}}\right],$$

and this expectation on the right-hand side is at most 1 (the same binomial calculation we've done often before). So we do get FDR control at level α if we use these as e-values, by the theorem of Wang and Ramdas above. (And in fact the two αs here – for FDR control and for T_α – need not be the same.)

Do note however that this isn't so useful for us – the e-values are all one of two values (positive or zero), but the knockoffs procedure already told us those are the ones to select. So we end up getting the exact same procedure as before, just studied differently – knockoffs are essentially doing e-BH on a particular selection of e-values $\{e_1, \cdots, e_p\}$.

Fact 100

Averages of e-values (even dependent ones) are e-values, but averages of p-values are not p-values. Notice that when we do the knockoff procedure and create random knockoffs $\tilde{X}^{(i)}$s, we might get different results depending on the randomness. But now with e-values, we can derandomize knockoffs by repeating this procedure repeatedly M times and averaging.

The idea is that for each $\tilde{X}^{(i)}$ we get some e-values $e_j^{(i)}$ (for $1 \leq j \leq p$). For each fixed i doing e-BH is just selecting our knockoffs in the usual way, but now if we average the (very dependent) $e_j^{(i)}$s over all runs i (say 1000 trials), we get some values $\overline{e}_1, \cdots, \overline{e}_p$ and we can do e-BH on that last result. We'll still control the FDR, and no other assumptions are needed. (Lots of other stuff can be derandomized with this principle as well!)

12 May 8, 2025

Today's lecture will mostly focus on limitations of e-values, but we'll see one last good thing first.

Example 101

We've been discussing how we can do analysis at different levels of granularity to get more precise rejections when signal-to-noise is strong. Naively we can't pick the finest discoveries at all levels and get FDR control overall, but it turns out that we can do e-value multi-resolution analysis and that will still work.

Last time, we showed e-BH controls the FDR, and in general it turns out the idea is that if we reject $R_{\mathcal{F}}$ out of the $|\mathcal{H}|$ hypotheses, all we need is the **self-consistency** (at level α) property

$$e_k \geq \frac{|\mathcal{H}|}{\alpha R_{\mathcal{F}}}$$

for all rejected hypotheses H_k. (We tuned this in one particular way for e-BH, but there are various other ones too, and the point is that self-consistency is a purely algorithmic property.) And we can enforce this by doing an optimization problem – let x_H be indicators for rejecting hypotheses H, and let $w(H)$ be some corresponding fixed weights for how valuable rejections are. (This is then a vector which tells us something about quality of rejections – w will be proportional to the resolution.) We can then solve the maximization problem

$$\max_{\{x_H : H \in \mathcal{H}\}} w(H) x_H \quad \text{given} \quad |\mathcal{H}| - \alpha e_H \sum_{H \in \mathcal{H}} x_H \leq |\mathcal{H}|(1 - x_H) \text{ for all } H \in \mathcal{H}.$$

In other words, we don't need to know any statistics at this point – we're getting FDR control just by imposing this constraint, and we can maximize same weighted version of power and "go as high as possible" in resolution. So e-values allow us to do this control much more easily than p-values, and we can even optimize for things like quality or discovery.

Remark 102. *One problem is that solving this is a Boolean optimization problem, which is very hard computationally in general. If we could relax each x_H to the full interval $[0, 1]$, we would just have a linear program and could solve the problem almost instantly. What we can do is find a relaxation problem and round it in some way; the canonical thing to do is sample ξ_i Bernoulli with probability x_i and check whether it satisfies our inequality (and then from there do some more fiddling around).*

The point is that we're still asking very hard questions like "which parts of the genome affect this particular phenotype" and solve some problems at different resolutions all at once, so this is real progress! Of course, we're focused on a genetics example where we have an ordering coming from contiguous parts of our chromosomes, but in general we can always cluster our variables and do this process on clusters instead.

Example 103

As we've said many times, e-values "shouldn't be that hard to find" since we only need to control the mean instead of the distribution. Suppose we're observing a logistic regression (y_i, X_i) to the model

$$\mathbb{P}(y_i = 1 | X_i) = \sigma(X_i^T \theta^*) = \frac{\exp(X_i^T \theta^*)}{1 + \exp(X_i^T \theta^*)},$$

and we want to test whether a coefficient $\theta_1^* = 0$.

There are several ways that we typically learn to do this: one is to look at the likelihood function

$$L(\theta) = \prod_{i=1}^{n} \sigma(X_i^T \theta)^{y_i} (1 - \sigma(X_i^T \theta))^{1 - y_i}$$

and we calculate the log-likelihood ratio

$$\text{LLR} = 2 \log \left(\frac{\sup_{\theta \in \mathbb{R}^p} L(\theta)}{\sup_{\theta \in \mathbb{R}^p : \theta_1 = 0} L(\theta)} \right);$$

since we're just testing one coefficient it will converge to a chi-square by Wilks' theorem. So if the sample size goes to infinity, this log-likelihood should converge to χ_1^2.

But the question is whether this is actually good in the regime where $p \sim n$ – the answer turns out to be no. If we apply the cdf of the chi-square, we should get our p-values and get something close to uniform under the null; instead it turns out things are spiked significantly near $p = 0$. That's a problem, because this means we'll be falsely rejecting much more than we should (and we can't even do Bonferroni correction properly!).

Very similarly, we can use the **Wald statistic**, which says that for p fixed and $n \to \infty$, we have $\sqrt{n}(\hat{\theta}_{\text{MLE}} \to \theta^*)$ converges to a normal with variance $I^{-1}(\theta^*)$. (Statistical packages that compute p-values for logistic regression will basically always use one of these.) We then get standard errors by plugging in $I^{-1}(\hat{\theta}_{\text{MLE}})$ in place of $I^{-1}(\theta^*)$, and we're again curious if this still holds up when p scales with n. Once again, the answer is no – the standard errors are off by a significant amount.

So the point is that "what we get in R is wrong," and in fact we don't really have any way to calculate p-values for high dimensions. The bootstrap method doesn't help us out either, whether it's parametric or non-parametric: the real distribution of the MLE does not match up with the parametric and pairs bootstrap (which look similar to each other but are both wrong!). It turns out that even if we sample with something like $p = 8, n = 80$, we're still running into problems where these limit statements aren't working out.

Thus we're still stuck with the problem of constructing a finite-sample test for the null hypothesis $\Theta_0 = \{\theta \in \mathbb{R}^p : \theta_1 = 0\}$. There is an idea of Wasserman et all. from 2020 which goes as follows:

- Do split sampling, so write $[n] = \mathcal{D}_0 \cup \mathcal{D}_1$. With $\mathcal{D}_1$, compute any estimator (MLE, logistical lasso, etc.) $\hat{\theta}_1$ of θ. We don't know anything about the sampling distribution of $\hat{\theta}$, so we can't form accurate p-values.
- Now look at $\mathcal{D}_0$ and consider
$$T_{\text{split-LRT}} = \frac{\prod_{i \in \mathcal{D}_0} p_{\hat{\theta}_1}(y_i|X_i)}{\sup_{\theta \in \Theta_0} \prod_{i \in \mathcal{D}_0} p_\theta(y_i|X_i)}.$$
If this quantity is large, that means that under the likely value of θ_1, the likelihood ratio is large and we have evidence against the null. So we reject if this is larger than $\frac{1}{\alpha}$.

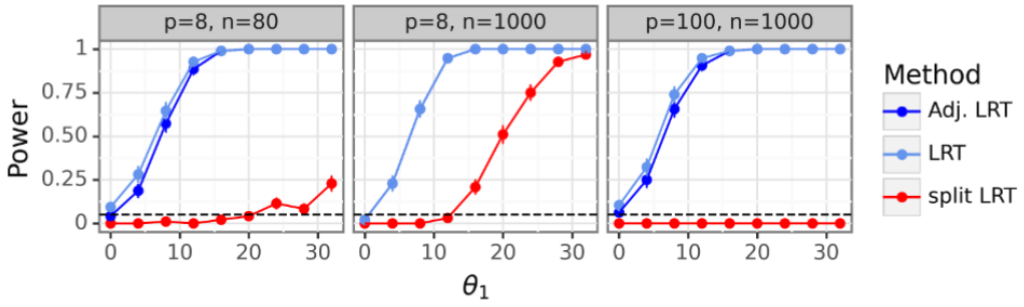
The point is that **what we have here is an e-value**, so we actually end up with an α-level test! Indeed, conditioned on any value of $\mathcal{D}_1$ and under the null, the expectation of $T_{\text{split-LRT}}$ is at most 1, since letting θ^* be the true parameter and $\hat{\theta}$ be our estimate, we have (replacing the blue quantity above)

$$\mathbb{E}_{\theta^*}\left[T_{\text{split-LRT}} | \vec{y}_{\mathcal{D}_1}\right] \le \mathbb{E}_{\theta^*}\left[\frac{\prod_{i \in \mathcal{D}_0} p_{\hat{\theta}}(y_i|X_i)}{\prod_{i \in \mathcal{D}_0} p_{\theta^*}(y_i|X_i)} \middle| \vec{y}_{\mathcal{D}_1}\right] = 1,$$

This then also lets us get a confidence interval by inverting our test, and we can do anything we want by duality.

Remark 104. *This strategy is advantageous because it works for a wide range of problems – we could have done so for any parametric family, we're allowed to use regularized estimators of θ, and it's exactly valid for finite samples. But the main disadvantage is that we have low power – indeed, we lose data because of sample splitting, and Markov's inequality is conservative (the usual e-value concern), but more concerningly using a supremum over all of Θ_0 is in general extremely conservative compared to the exact likelihood ratio – people are reporting that this is exponentially conservative as the dimension of Θ_0 grows.*

If we do power plots, we in fact see that the split LRT has incredibly low power – methods like the ordinary and adjusted likelihood ratio test are detecting much more than the split LRT.



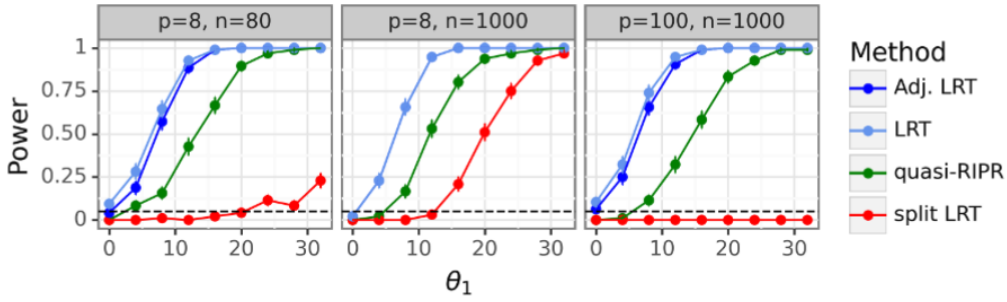
So an open question is whether we can do e-values and construct logistic regressions which have power closer to the blue curves! And if not, then it's "not so clear why we keep talking about them."

Remark 105. *We could be less conservative by using the reverse information projection instead of a supremum, so that we consider*

$$T_{RIPR} = \frac{p_{\hat{\theta},\mathcal{D}_0}(\vec{y}_{\mathcal{D}_0}|X)}{p_{ripr,\mathcal{D}_0}(\vec{y}_{\mathcal{D}_0}|X)}$$

This is an e-value, so we can validly reject when this is greater than $\frac{1}{\alpha}$. Unfortunately this RIPR is difficult to compute in general – we want to optimize over mixture distributions on $p_{\theta,\mathcal{D}_0}$, and there aren't known ways to do that even for something like logistic regression.

We can cheat a bit and use a quasi-RIPR, where instead of finding the best mixture in KL divergence, we find **the** distribution in Θ_0 (that is, a point mass) which is closest. This is **not** an e-value, but it still controls the size of the test because there's so much slack everywhere else. This becomes a maximium likelihood problem and we can do it, and the result is shown below:



It's somewhat better but still not as good as other tests, and we've gotten to the heart of the matter – we need to get good e-values so that e-BH or something similar will actually get rejections." The examples where we do get good e-values already leverage other frameworks, and doing things from scratch is not simple.

Fact 106

There are methods of getting concentration out of this framework too. Suppose we have bounded random variables X_i with mean μ, and we want to produce a confidence interval for μ with $\mathbb{P}(\mu \in C_n) \geq 1-\alpha$. (It turns out that if our random variables are not bounded, we cannot do this unless C_n is the whole real line.) We can do this with Azuma-Hoeffding, but that turns out to be a bit silly – the interval is way too wide especially for small variance, for example if X_i is Beta$(10, 30)$ instead of Bernoulli. (Azuma-Hoeffding doesn't even take the variance of the random variables into account!) A paper by Waudby-Smith and Ramdas goes into a way to adapt to the distribution in more detail, and it gives a narrower interval than other methods that have been proposed in the past.

The key idea is that Hoeffding is a large-deviations inequality which allows us an interval of the form

$$C_n^H = \left\{ \{m \in [0,1] : \prod \exp\left(\lambda(X_i - m) - \frac{\lambda^2}{8}\right) < \frac{1}{\alpha} \right\},$$

optimized over λ. Waudby-Smith and Ramdas basically **chooses a different functional**

$$C_m = \left\{ m \in [0,1] : \prod_{i=1}^{n} (1 + \lambda_i(X_i - m)) < \frac{1}{\alpha} \right\},$$

and by cleverly choosing λ_i with a "gambling strategy" we can get something still valid but much narrower. But we won't go into this in more detail this year.

13 May 13, 2025

Today's lecture will be a blackboard talk (just to change things up a bit); it will cover **permutation tests** and have some interesting questions for us. We'll be looking ahead to conformal prediction, hoping to understand the key ideas from the point of view of permutation tests, before we do it in more detail in the coming lectures.

Problem 107

The following fundamental problem is an old idea going back to Fisher. Suppose $X_1, \cdots, X_n$ are iid samples from some probability distribution P, and $Y_1, \cdots, Y_m$ are iid samples from some other probability distribution Q. Our goal is to test the hypothesis that $P = Q$ even if we don't know either of the distributions.

The strategy we learn in early statistics is that we should choose a test statistic T (for example we can let $T = |\overline{X} - \overline{Y}|$) and reject the null based on whether T is unusually large. The question we need to ask is "how large does T need to be," and typically this is wher permutation tests are introduced: we compare T to how the statistic would look **if we were to permute the data**.

Formally, we introduce the following (randomized) **permutation distribution**: choose M uniform permutations $\sigma_1, \cdots, \sigma_M$ from S_{n+m} (the set of permutations acting on a list of $n+m$ objects), and we will have these permutations act on the vector

$$Z = (X_1, \cdots, X_n, Y_1, \cdots, Y_m).$$

Specifically, the σ_is will shuffle the entries of Z, and then we evaluate the test statistic on the permuted entry. So we always take the average of the first n entries and the average of the last m entries after permuting, and we find their absolute difference; in other words, we compare

$$T(Z) = \left| \overline{Z_{1:n}} - \overline{Z_{n+1:n+m}} \right|$$

to the corresponding values of $T(Z_\sigma)$, and the p-value will essentially be the relative rank of $T(Z)$ compared to $T(Z_{\sigma_1})$ through $T(Z_{\sigma_M})$:

$$p = \frac{1 + \sum_{i=1}^{M} 1\{T(Z_{\sigma_i}) \geq T(Z)\}}{1 + M}.$$

Under the null, this is either uniformly distributed on $\left\{\frac{1}{M+1}, \frac{2}{M+1}, \cdots, \frac{M+1}{M+1}\right\}$ or biased upward due to ties (because under the null we have exchangeability of the vector Z). Thus it is indeed a p-value, and notice that this doesn't rely on us needing to know the distribution of T at all.

Example 108

The discussion above has been a motivating example, and from now on, we'll assume that $Z = (Z_1, \cdots, Z_n)$ is an exchangeable vector of random variables. That is, the distribution is symmetric in its arguments – for example for two variables, we might want the pdf or density to satisfy $p(z_1, z_2) = p(z_2, z_1)$.

There's another way to characterize exchangeability as well: suppose we have a "bag" $\{Z\}$ of numbers, where some of the numbers may appear multiple times. (So it's an unordered multiset of the values that appear.) Then our random variables are exchangeable if

$$\mathbb{P}(Z = z_\sigma | \{z\}) \text{ is constant over all } \sigma;$$

that is, we cannot distinguish ordering, and every permutation is equally likely. We're going to go through some key facts now:

Example 109

Given a set of permutations S, consider the quantity

$$p = \frac{1 + \sum_{\sigma \in S} 1\{T(Z_\sigma) \geq T(Z)\}}{1 + |S|}.$$

If the variables are exchangeable, then Z_σ has the same distribution as Z and thus $T(Z_\sigma)$ has the same distribution as $T(Z)$. We are interested in what sets of permutations S yield a valid p-value.

Here are the answers:

S	p-value?
all permutations (S_n)	Yes
iid samples from S_n	Yes
an arbitrary fixed subset of S_n	No
iid samples from an arbitrary fixed subset	No
a subgroup of S_n	Yes
iid samples from a subgroup of S_n	Yes

Fact 110

Here "subgroup" means that we have a subset S of elements in S_n which is closed under composition and inverses (and in particular therefore also contains the identity permutation).

In other words, we can only depart a little bit from taking all permutations or uniform samples from S_n. And first, we'll show why we can't let S just be some arbitrary fixed subset of permutations with a counterexample:

Example 111

Consider $Z_1, \cdots, Z_n$ iid standard normal, and let S be the set of permutations

$$S = \{\sigma \in S_n : \{\sigma(n-1), \sigma(n)\} = \{1, 2\}\}.$$

Let our test statistic be $T(Z) = Z_1 + Z_2$.

We will show that this does not yield a valid p-value. Indeed, we now have

$$p = \frac{1 + \sum_{\sigma \in S} \{Z_{n-1} + Z_n \geq Z_1 + Z_2\}}{1 + |S|}.$$

But now this inner sum is equal for all σ, which means that our p-value either takes the value $\frac{1}{1+|S|}$ or 1, each with probability $\frac{1}{2}$. And this does not stochastically dominate a uniform random variable, since it takes on a very small value with probability $\frac{1}{2}$; the idea is that "all of the Z_σs are conspiring."

We'll now show that if we use all permutations instead, then we're okay, and in fact we don't even need the $+1$ in the numerator and denominator:

Proposition 112

The quantity

$$p = \frac{\sum_{\sigma \in S_n} 1\{T(Z_\sigma) \geq T(Z)\}}{n!}$$

is a p-value.

Proof. We'll make use of the following "probability integral transform" lemma:

Lemma 113

Let the test statistic T have cdf F. Then $F(T)$ stochastically dominates U (in fact it is equal in distribution if T is continuous, but larger if the variable takes on discrete values). Equivalently, if $G(t) = \mathbb{P}(T \geq t)$, then $G(T)$ stochastically dominates U.

With this, let σ_0 be an arbitrary permutation in S_n, and let $Z' = Z_{\sigma_0}$. Then

$$\frac{1}{n!} \sum_{\sigma \in S_n} 1\{T(Z_\sigma) \geq T(Z_{\sigma_0})\} = \frac{1}{n!} \sum_{\sigma \in S_n} 1\left\{T(Z'_{\sigma \circ \sigma_0^{-1}}) \geq T(Z')\right\},$$

and now because we are summing over all permutations, we can do a change of variables and sum over $\tau = \sigma \circ \sigma_0^{-1}$ instead. Then τ also ranges over S_n again, so this is also equal to

$$\frac{1}{n!} \sum_{\tau \in S_n} 1\{T(Z'_\tau) \geq T(Z')\}.$$

But $Z' \overset{d}{=} Z$ by exchangeability, so this quantity is also equal in distribution to $\frac{1}{n!} \sum_{\tau \in S_n} 1\{T(Z_\tau) \geq T(Z)\}$. That means that if we let σ_0 be a random draw from S_n, then $T(Z_{\sigma_0})$ is a random draw from $T(Z_\sigma)$ and thus we can apply our probability integral transform lemma. More precisely,

$$p \overset{d}{=} \frac{1}{n!} \sum_{\sigma \in S_n} 1\left\{T(Z_\sigma) \geq T_{Z_{\sigma_0}}\right\} \overset{\text{sto}}{\geq} U,$$

as desired. □

Corollary 114

We can also adapt this to random permutations drawn from S_n, and we will now put the 1 back. Letting σ_i be M iid permutations drawn from S_n,

$$p = \frac{1 + \sum_{i=1}^{M} 1\{T(Z_{\sigma_i}) \geq T(Z)\}}{1 + M}$$

is a p-value.

Proof. Use the same argument as before: let σ_0 be uniform over S_n and define

$$q = \frac{\sum_{i=0}^{M} 1\{T(Z_{\sigma_i}) \geq T(Z_{\sigma_0})\}}{1 + M}$$

But now q stochastically dominates a uniform random variable (by the same logic as before and the fact that σ_0 is a random draw from $\{\sigma_0, \cdots, \sigma_M\}$), and we claim that q and p have the same distribution. Indeed, by the same change-of-variable as before,

$$\sum_{i=1}^{M} 1\{T(Z_{\sigma_i}) \geq T(Z_{\sigma_0})\} = \sum_{i=1}^{M} 1\{T(Z'_{\sigma_i \circ \sigma_0^{-1}} \geq T(Z')\} \stackrel{d}{=} \sum_{i=1}^{M} 1\left\{T(Z_{\sigma_i \circ \sigma_0^{-1}}) \geq T(Z)\right\}$$

and because the distribution of $\sigma_i \circ \sigma_0^{-1}$ is the same as the distribution of σ_i itself, this also has the same distribution as $\sum_{i=1}^{M} 1\{T(Z_{\sigma_i}) \geq T(Z)\}$. Plugging this into the expressions for p and q yields the result. □

Fact 115

The two proofs above also apply if we replace S_n with any subgroup of S_n. Indeed, instead of summing over S_n in the proofs we sum over the subgroup H, and we choose σ_0 to be first an arbitrary, then a random draw from H. The subgroup property is necessary here so that summing over $\sigma \circ \sigma_0^{-1} \in H$ is equivalent to summing over $\sigma \in H$.

We'll close with a brief note on conformal inference:

Example 116

In predictive inference, we take some set of training samples $(X_1, Y_1), \cdots, (X_n, Y_n)$ which are iid from some distribution P. We then get a test sample X_{n+1}, and our goal is to construct, from the training samples, a prediction interval for Y_{n+1} with prescribed coverage. That is, we want some C so that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha.$$

Note that this is not a confidence interval – this is an observation we have not seen yet, and it's interesting that we can solve this problem at all. In fact, the way we can do so is through permutation tests! We hypothesize a value of Y_{n+1} and test for exchangeability by computing

$$p^y = \frac{1}{(n+1)!} \sum_{\sigma \in S_{n+1}} 1\{T(Z^y_\sigma) \geq T(Z^y)\},$$

where $Z^y = \{(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, y)\}$. Then setting $y \in C(X_{n+1})$ if and only if $p^y \geq \alpha$, we claim this is a valid prediction interval regardless of our choice of T. Indeed, Y_{n+1} itself will be in the confidence interval if and only if $p^{Y_{n+1}} \geq \alpha$, but we've already shown that $p^{Y_{n+1}} \stackrel{\text{sto}}{\geq} U$ and thus we are done.

Remark 117. *Conformal inference requires some additional assumptions, and we'll see that in the coming lectures. The point is just that we can turn permutation tests into prediction intervals, but the caveat is that we require a sum over $(n+1)!$ possibilities. So by introducing symmetry in the function T (specifically in its first n arguments) we can collapse the calculation and make it practical (we only need to sum over $n+1$ things instead of $(n+1)!$):*

$$p^y = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\left\{T(Z^y_{-i}, Z^y_i) \geq T(Z^y_{-(n+1)}, Z^y_{n+1})\right\}.$$

So we can use non-symmetric test statistics at the cost of additional computation.

14 May 20, 2025

Our topic for the next two lectures will be **conformal prediction**: we'll discuss prediction intervals and different flavors for how to get them, as well as the new ideas of jackknife+ and CV+.

Example 118

The motivation for conformal prediction is that we want some uncertainty in our prediction and some way of quantifying accuracy. For example, suppose we have partial data of percentage vote change between 2016 and 2020 in some counties. Then for example at the Washington Post, there will be issued statements of predicted ranges for unreported counties, and the cost of being wrong there is rather high.

There is some uncertainty quantification being used there – it's not exactly the conformal prediction concept we'll talk about here, but there are some similar ideas. We use machine learning more and more in a lot of sensitive applications (for example treatment being received or punishments for crimes), and we outsource a lot to decision algorithms. Our goal is then to get some kind of confidence prediction even if we don't understand the algorithm fully.

Fact 119

As professionals, we use data to make predictions – we have some input features and we predict some output feature. Questions we need to be able to answer include "how certain are we of a prediction," "how does that uncertainty affect the eventual decision," and "can the model be safely deployed." It's not so useful to have a blackbox otherwise.

Summarizing the idea from last time, we have a training data of n points (X_i, Y_i) as well as a test point $(X_{n+1}, ?)$. If we assume these points are all exchangeable (for example iid from the distribution P_{XY}), we have some hope of predicting Y_{n+1}, and our goal is to construct a marginal prediction interval $C(X_{n+1})$ so that $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$ for any distribution P_{XY} and any sample size n. (This is different from a confidence interval, which is just measuring population parameters and thus is sometimes harder to interpret if we're for example specifying a linear model.) One solution comes via permutation tests: letting $Z_i = (X_i, Y_i)$ and $Z^y_{n+1} = (X_{n+1}, y)$, we basically consider the vector $Z^y = (Z_1, \cdots, Z_n, Z^y_{n+1})$ with an imputed value y and test for exchangeability via some arbitrary statistic T:

$$p^y = \frac{1}{(n+1)!}\sum_{\sigma \in S_{n+1}} 1\left\{T(Z^y_\sigma) \geq T(Z^y)\right\}$$

(This is a test for whether we can detect that we are not exchangeable and thus should be surprised.) We then include y in the confidence interval if and only if $p^y \geq \alpha$; the point is that we have a valid p-value, so $p^{Y_{n+1}}$ stochastically dominates a uniform $[0,1]$, and thus Y_{n+1} is in the prediction interval with probability $1-\alpha$.

The calculations here turn out to simplify if we use a test statistic with some symmetry (and that's where conformal prediction comes in) – if we let $T(Z_1, \cdots, Z_n, Z_{n+1})$ be assumed to be symmetric in its first n arguments, then all that matters is the value of $\sigma(n+1)$. So the permutation p-value simplifies in this case to

$$p^y = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\left\{T(Z^y_{-i}, Z^y_i) \geq T(Z^y_{-(n+1)}, Z^y_{n+1})\right\}$$

and again we include y in the interval if and only if $p^y \geq \alpha$. (We can check that all of this works even if T is randomized.) We will call this **full conformal invariance**.

Example 120

The way this is typically done as follows: we fit a model $\hat{\mu}(\cdot) = \mathcal{A}(Z^y)$ to $Z^y = ((X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, y)$ which satisfies the symmetry assumption (for example in random forests, it doesn't matter what order we pass in the data, and this is true of most algorithms), and we define $T(Z^y) = |y - \hat{\mu}(X_{n+1})|$ to be the residual. Replacing T with what we have, we thus get a p-value

$$p^y = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\left\{|Y_i - \hat{\mu}(X_i)| \geq |y - \hat{\mu}(X_{n+1})|\right\} = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\{R^y_i \geq R^y_{n+1}\}.$$

Note that it's okay (and correct) here for us to **fit the prediction model to the imputed value** as well, since what we're ultimately doing is comparing the residual we get there to the residual of the other data points and we won't get a valid comparison if we don't include everything. (For example, if our model interpolated between the valid data points and we only trained it on Z_1 through Z_n, our first n "training residuals" would all be zero and we wouldn't be able to say anything about the "test residual" R^y_{n+1}.)

And now if R_{n+1} is very large, this quantity is quite small (it doesn't conform), so instead we want to include a point y in the prediction interval if the test residual is quite small. More precisely, the condition for including y is that

$$R^y_{n+1} \leq \text{Quantile}\left(1-\alpha; R^y_1, \cdots, R^y_{n+1}\right).$$

Remark 121. *Interpolation is not actually a good thing to do here, though – in this framework, a neural network will just fit a model through all of the $n+1$ points, and so it will include any y that we choose and have a very large prediction interval. We'll talk about this more soon.*

The point is that if all conformity scores are almost surely distinct, then in fact we get

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \in \left[1-\alpha, 1-\alpha+\frac{1}{n+1}\right]$$

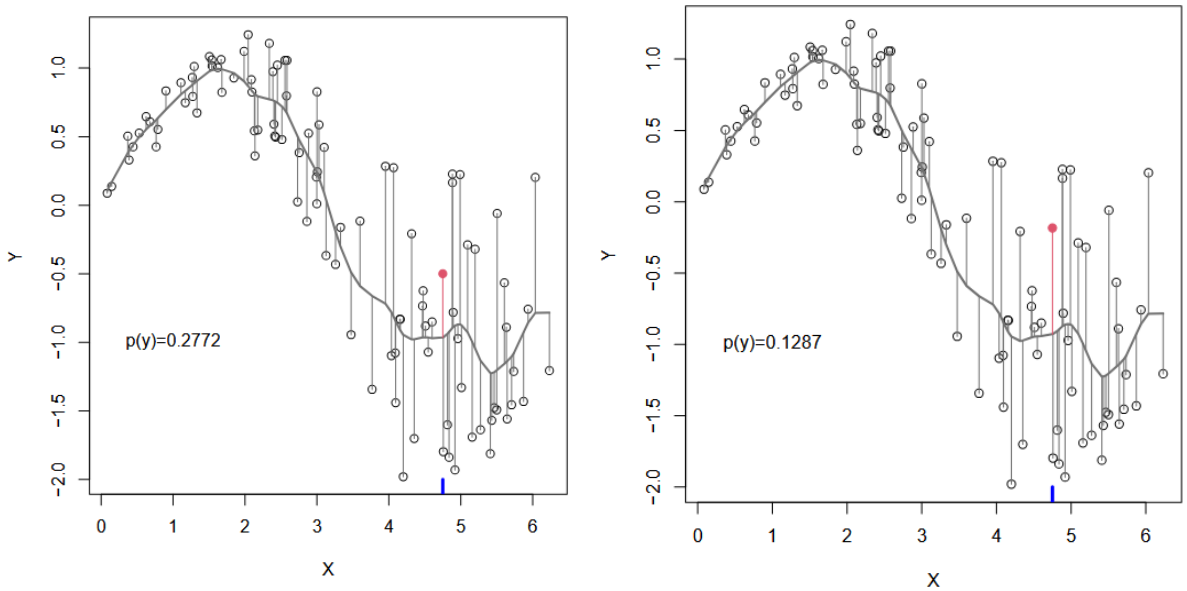
(the p-value we get out of this is almost exactly uniform). And this is a statement of expectation over the training set and the test set – one criticism against conformal inference is that it's **not conditional on** X_{n+1} – we can say "we'll be 90 percent accurate about the next county we randomly pull," but we can't say anything if it's a specific county.

Example 122

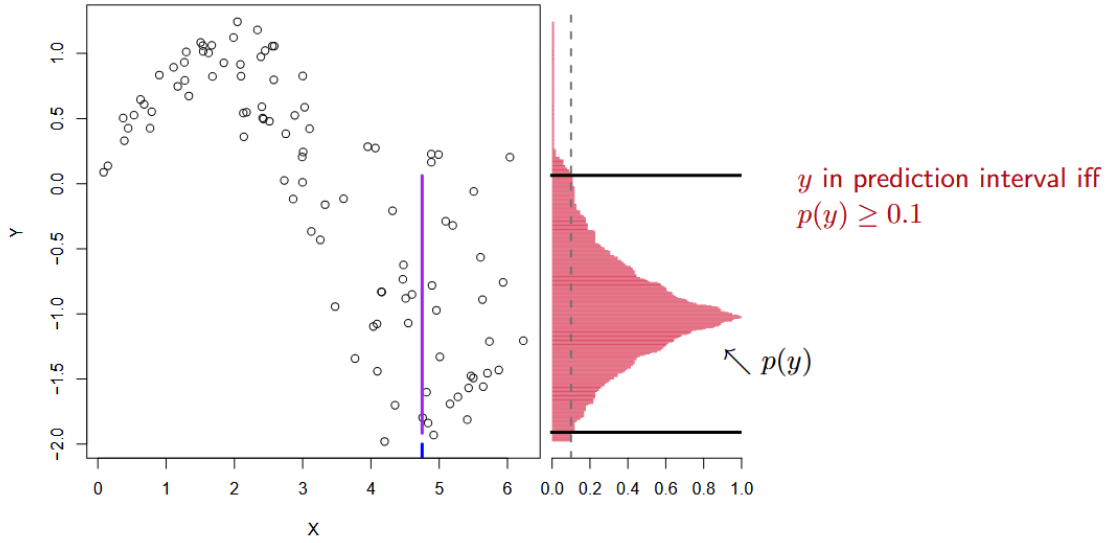
Suppose we have a training dataset as shown below, and we want a prediction interval at the blue tick shown below, $X_{n+1} = 4.7$.

What we do is ask "can Y_{n+1} be equal to y?" by fitting a model with that value and calculating $p(y)$, the fraction of residuals with larger magnitude than for our imputed data point. And once $p(y)$ goes below some value, say 0.1,

it "stops conforming appropriately' and leaves the prediction interval. (And note that as our point varies, the fitted curve also slightly shifts.)



The prediction interval we end up with is shown below:



This is all computationally intensive – every time we fit the model we might need to do a lot of computation, and in fact the prediction interval can be a disjoint set of intervals instead. But there are computationally tractable ways of doing this. We can check how this works when we take, say, 1200 of the counties in the election and try to predict the other 1900, and when we want 90 percent coverage we do in fact get very close to 90 percent coverage! (This is a bit too good to be true since we're not really testing the theorem – we're getting 1900 test points to compute residual quantiles and we're using the "same training set" for all of them, and we're really supposed to rerandomize for each county. But it gets the idea across.)

Example 123

We'll now understand the various "flavors" of conformal inference, starting with **split conformal invariance**.

This is a special case where we have n data points and we do sample splitting: we learn a model $\hat{\mu}$ with the first split (also called a "fold"), and on the second split we calculate out-of-sample residuals (that is, learn the distribution

of the residuals $R_i = |Y_i - \hat{\mu}(X_i)|$). Then the test residual relates to this second split by keeping track of quantiles, and the point is that we separately do training and calibration and form our interval from points that have all not been used for training.

Formally, we compute a score function $\mathcal{S}(x, y) = |y - \hat{\mu}(x)|$ by fitting a model on an independent training set. Once we have this, we use a distinct calibration set of size n to find typical size of residuals

$$S_i = \mathcal{S}(X_i, Y_i), \quad S_{n+1}^y = \mathcal{S}(X_{n+1}, y).$$

The point is that if $y = Y_{n+1}$ these points should all be indistinguishable (they're from the same distribution), and now we include y if

$$p^y = \frac{1}{n+1} \sum_{i=1}^{n} 1\left\{S_i \geq S_{n+1}^y\right\} \geq \alpha.$$

And this result now holds conditionally on the training set, since for all purposes $\hat{\mu}$ is fixed.

This is better computationally, but it isn't quite "as good as full conformal inference" in that it's not cross-validation and thus we need larger samples to get both a good model and good calibration points – it has the drawbacks of sample splitting.

Example 124

So if data is scarce, we might want something that lets us both fit and calibrate while still being efficient, and that's going under the name jackknife+/CV+. This will be a bit more subtle than the previous methods.

We want to construct a prediction interval, and **naively** we might "fit a model $\hat{\mu}$, measure $\hat{\mu}(X_{n+1})$, and then do a $\pm$ interval corresponding to the 90th percentile of residuals"

$$\left[\text{10th percentile of } \left\{\hat{\mu}(X_{n+1}) - R_i\right\}, \quad \text{90th percentile of } \left\{\hat{\mu}(X_{n+1}) + R_i\right\}\right].$$

But we know training residuals are too small compared to test residuals, and an alternative method (called the **jackknife**) is to take the regression value $\hat{\mu}(X_{n+1})$ and then do an interval corresponding to 90th percentile of the **leave-one-out residuals**

$$R_i^{\text{LOO}} = |Y_i - \hat{\mu}_{-i}(X_i)|.$$

That is, the interval we end up with is

$$\left[\text{10th percentile of } \left\{\hat{\mu}(X_{n+1}) - R_i^{\text{LOO}}\right\}, \quad \text{90th percentile of } \left\{\hat{\mu}(X_{n+1}) + R_i^{\text{LOO}}\right\}\right].$$

But this still doesn't quite work – it's forgetting something important, and instead what we should do is the **jackknife+** where we recenter each time:

$$\left[\text{10th percentile of } \left\{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\right\}, \quad \text{90th percentile of } \left\{\hat{\mu}_{-i}(X_{n+1}) + R_i^{\text{LOO}}\right\}\right].$$

If $\hat{\mu}$ and $\hat{\mu}_{-i}$ are very similar (the fit is stable), then it'll be essentially the same. But if the fit depends crucially on a few data points, this could be significant. More intuitively, jackknife always has the same centering point while jackknife+ has shifted centers.

Fact 125

Note that now we don't have to fit the value for every hypothesized value y – we still need to fit the model n times for each removed data point in either jackknife or jackknife+, but it's still computationally feasible.

Theorem 126 (Barber, Candés, Ramdas, Tibshirani 2019)

With exchangeable data points, we have

$$\mathbb{P}(Y_{n+1} \in C^{\text{jackknife+}}) \geq 1 - 2\alpha$$

(in practice this looks like $1-\alpha$ for any real dataset, but there are counterexamples where the prediction interval is smaller).

Interestingly, it's possible to construct examples where the jackknife actually has zero percent coverage for the prediction interval, so this strategy is much better!

Remark 127. *Professor Candés did a simple example where $Y|X$ is linear, we do regularized least-squares, and we have* 100 *samples and* 100 *features, the jackknife method yields a coverage of* 0.475 *and jackknife+ yields a coverage of* 0.913. *And the idea is that when the number of features and samples are close, fitting a linear model means $\hat{\mu}$ is not very stable.*

Example 128

To further make this computationally tractable, we can do "leave-one-out folds" instead of "leave-one-out residuals" (say with $K = 10$ equal-sized folds). We use 9 folds to train a model, and we use the last one to calculate $R_i^{\text{CV}} = |Y_i - \hat{\mu}_{-S_{k(i)}}(X_i)|$ and find a prediction interval. This is called the **CV+ method**, and now we only need to train $K = 10$ models (and if we can't even do, say, 5, we should abandon doing uncertainty quantification).

What we haven't done in all of this is discuss discrete labels – everything has been continuous, but everything extends to discrete labels as well.

Fact 129

We can take the fashion-MNIST and MNIST datasets – these have 10 classes of images, 50 features obtained by PCA, and 1000 training examples. Below are the values of coverage and set size for different methods and classifiers (all of them have a type I error guarantee, but we want to figure out which one to use).

Classifier	Method	Fashion-MNIST		MNIST	
		Coverage	Set Size	Coverage	Set Size
Logistic	Full conformal	0.897	1.623	0.900	1.248
	Split conformal	0.893	1.552	0.898	1.387
	Jackknife+	0.897	**1.407**	0.897	**1.177**
Random Forests	Full conformal	0.895	1.497	0.901	1.324
	Split conformal	0.895	1.651	0.901	1.590
	Jackknife+	0.903	**1.473**	0.909	**1.288**
Kernel SVM	Full conformal	0.898	1.901	0.904	1.098
	Split conformal	0.894	1.382	0.899	1.092
	Jackknife+	0.897	**1.266**	0.898	**0.966**
Neural Net	Full conformal	0.899	3.942	0.898	2.733
	Split conformal	0.893	1.818	0.897	1.270
	Jackknife+	0.915	**1.499**	0.913	**1.041**

The coverage numbers are all roughly 0.9, which is exactly what we want. So now we want to return small prediction sets (that's more power), and the bold numbers show that jackknife+ does the best.

To understand why full conformal does not do too well here, these models are usually overparameterized and interpolate between data points. The idea is that test residuals tend to be rather low, so no matter how we pick y, we'll see a small outcome due to overfitting. Thus it makes sense to instead use jackknife+ for something like a neural network.

More generally, let's phrase the discrete problem that we're trying to solve: suppose we want to categorize into a discrete, unordered set of labels $\mathcal{Y}$, and we want to construct a prediction set using a conformity score. One thing we can do is take an algorithm $\hat{\pi}(y|x)$, estimating the probability of $Y = y$ given $X = x$ (for example the output of a softmax layer of a neural net); our prediction set is then

$$\hat{S}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^{n} 1\left\{ \hat{\pi}_{-i}(y|X_{n+1}) \geq \hat{\pi}_{-i}(Y_i|X_i) \right\} \geq \alpha(n+1) \right\}.$$

In English, we're saying that we're in the predictive set if the predicted probability is in the top 90 percent of the leave-one-out probabilities – $-\hat{\pi}(y|x)$ is taking the place of $|y - \hat{\mu}(x)|$, we train classifiers instead of regressions, and then everything else is the same. And with the exact same proof we show that the probability of being in the predictive interval is again at most $1 - 2\alpha$.

15 May 22, 2025

Fact 130

Notice that in split conformal prediction, the width of the interval does not depend on the new value of X_{n+1} (since our prediction function is $\hat{\mu}(X_{n+1})$ plus or minus some quantity). Letting $S_i = |Y_i - \hat{\mu}(X_i)|$ be our conformity scores in split conformal mode, the width then depends on the quantile of those S_is, but not on the new test point. And this is the case for jackknife and jackknife+ as well, and this will come up later on in the lecture.

We'll continue with conformal prediction today. First, we summarize the idea of discrete prediction intervals, i.e. classification: we now want to output a prediction set which includes the true classification with probability 90 percent, and everything stays the same except with a newly specified conformity score $T(Z^y) = \frac{1}{\hat{\pi}(y|X_{n+1})}$ (or alternatively $-\hat{\pi}(y|X_{n+1})$); again p-values depend on what proportion of these conformity scores we're above, and we include points as long as $\hat{\pi}$ is larger than some specified quantile. (There's split conformal and jackknife versions of this, which play out in the exact same time. The main advantage of jackknife and jackknife+ is that we don't need to train models for every y as long as we look at the leave-one-out probabilities, and we can run things in cross-validation mode as well.)

Example 131

What we'll turn to now is **enhanced conformity scores** – conformal prediction lets us choose conformity scores in any way we want, and our goal should be to fit the high and low quantiles of $Y|X$.

If we had an oracle which revealed those quantities to us, our prediction interval would then exactly be (for example) the 5th to the 95th percentile. So it's not clear why we should start by estimating the mean in the first place, especially when the distribution is often not expected to have normal fluctuations and when the length of the interval can really vary greatly depending on X.

Thus, if we want to estimate these true quantiles, we want a way to form adaptive intervals rather than doing

constant-width across all values of x. (Right now we're marginally correct over the full distribution of X, but we should aim for something a bit stronger rather than over-covering in some regimes and under-covering in others.)

The **first idea** comes from Lei, and it is to change the conformity scores to depend on **relative residuals** instead

$$\tilde{R}_i = \frac{|Y_i - \hat{\mu}(X_i)|}{|\hat{\sigma}(X_i)} = \frac{|R_i|}{\hat{\sigma}(X_i)}.$$

(So this requires us to run two machine learning algorithms for fitting instead of one.) We can then get a valid prediction interval by "standardizing"

$$C(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm Q \cdot \hat{\sigma}(X_{n+1}),$$

where Q is the $(1-\alpha)$-quantile of $\tilde{R}_i$ – everything goes through the same way as before, and now the width can change.

However, in practice this doesn't really work very well. One problem is that when we estimate $\hat{\sigma}$, we often underestimate it (for example if we have a tendency to interpolate) and thus end up with huge terms because of fairly unstable quantities. So that means the intervals aren't actually as adaptive as we hope, and we don't end up with great results.

So the **other idea** to go with, which is what people do at the moment, goes under the name conformalized quantile regression. Basically we estimate quantiles, then conformalize them – we change the squared loss to a different loss function

$$f(\cdot) = \text{argmin}_{f \in \mathcal{F}} \sum_i \rho_\alpha(Y_i - f(X_i)) + \mathcal{R}(f)$$

where $\mathcal{R}(f)$ is some regularizer and p_α is the **pinball loss** which is the absolute value function for $\alpha = 0.5$ (hence finding the median) and more generally it is piecewise with slopes α and $-(1-\alpha)$ above and below 0. We can then use these quantiles to find conformity scores, and then we don't need to prove anything new because all we're doing is changing the score function: we now use a score function such as

$$S(x,y) = \max\left\{\hat{q}_{\alpha/2} - y, y - \hat{q}_{1-\alpha/2}(x)\right\},$$

or we can do some kind of standardization (normalizing by width, though in practice it doesn't really improve very much)

$$S(x,y) = \max\left\{\frac{\hat{q}_{\alpha/2} - y}{\hat{q}_{0.5}(x) - \hat{q}_{\alpha/2}(y)}, \frac{y - \hat{q}_{1-\alpha/2}(x)}{\hat{q}_{1-\alpha/2}(x) - \hat{q}_{0.5}(x)}\right\}.$$

We then include y in our prediction interval for X_{n+1} if and only if $S(X_{n+1}, y) \leq Q(1-\alpha, S_i)$ (where depending on the method we use, maybe we are refitting the model, but the idea is still the same).

Remark 132. *To give some intuition, suppose we run this in split conformal mode. So then we calculate quantiles from our first set, and then we measure the distance of a point to the nearest quantile, receiving a $+$ if we're beyond the quantiles and $-$ if we're between them. Ideally we would have 90 percent of our points with a $-$ and 10 percent of our points with a $+$, so the quantiles of our scores would be about 0. But if we were overconfident (as typical machine learning algorithms are), we will have too many positive points and thus the quantile will be positive. Similarly if we are too conservative our quantile will be negative.*

Doing a bit of algebra, this means the values of y that will be in our prediction interval at a fixed x are given by

$$[\mathit{lower}(x) - Q_{1-\alpha}, \mathit{upper}(x) + Q_{1-\alpha}]$$

and so this indeed lets us get an adaptive-width interval.

In practice, both the "ordinary" split conformal method and CQR will give the intended coverage of 90%, but the width of CQR is adaptive and has a smaller average length. So it's better because it has shorter intervals, but it also looks better conditionally. (Of course, we can't actually get any theoretical guarantee like $\mathbb{P}(Y \in C(X_{n+1})|X_{n+1}) = 1-\alpha$, though. To guarantee that in general, we need infinite length intervals or some modeling assumptions.

Example 133

There's a medical expenditure panel survey from 2015, which has 16000 subjects and 140 features (including age, marital status, race, health status, etc.) predicting things like healthcare system utilization and number of visits to the doctor. And regardless of the conformity scores we pick, we'll get 90 percent coverage in our prediction intervals – it turns out CQR gets better conditional coverage (much closer to 90 percent) and shorter length of intervals as well.

It is difficult to measure what conditional coverage actually means – the idea is to measure over slabs of data points in this 140-dimensional space and find the worst-case adversarial coverage on that slab. The point is that it turns out to be very hard to make CQR look bad, but it's pretty easy for other notions to do significantly worse than 90 percent. And there's been more comprehensive studies to show that this works in a variety of contexts, too.

Example 134

There are other things we can do too which are CQR-related, such as **calibration via adaptive coverage**. One thing we can do is start with a data set, and try to essentially predict $Y|X$ via quantile regression (so we do 100 quantile regression problems).

This basically means we've fitted a model and have an uncalibrated guess (for example with $\tau = 0.1$)

$$C^{\text{naive}}(x, 1-\tau) = \left[\hat{F}^{-1}_{Y|X}(\tau/2), \hat{F}^{-1}_{Y|X}(1-\tau/2)\right].$$

This interval won't be good yet (our fitted model is probably overconfident on its training data), so now we conformalize: run things in full conformal or jackknife+ mode, using this as a conformity score. Specifically, in split conformal mode we get quantiles based on some predicted distribution, and then we'll pick a prediction interval **of this kind** but choosing $\hat{\tau}$ so that we achieve 90 percent coverage on the calibration set! So for example we might pick τ at 95 percent so that on test data we get 90 percent coverage. In practice, this ends up being more work but it works fairly well, basically at the same level as CQR.

We can apply this all to discrete labels as well: we first have some conditional probabilities $\hat{\pi}(y|x)$ that we estimate (for example as the output of a softmax layer), for example $50\%, 30\%, 10\%, 5\%, 2\%$ for a, b, c, d, e. It would then be tempting to say that $\{a, b, c\}$ is a good prediction set with 90 percent coverage, but the neural net might be too sure of itself. So instead we can "conformalize" and see how well $\{a, b, c, d\}$ does on the test data, seeing whether we now get 90 percent coverage. (Or maybe $\{a, b\}$ turns out to be enough, too.) This is also useful because it actually reveals uncertainty in a more honest way (we can output the probabilities along with the predictions), and it lets the threshold be adaptive to x. This does mean we return larger sizes in general, but that's mostly because we return fewer empty sets compared to the "original" discrete classification strategy.

Example 135

We'll close this lecture with **weighted conformal inference**: some ideas are that (1) the guarantees of conformal inference are too weak, and we want something closer to conditional guarantees, and (2) there may be shifts in covariates.

For example, in the presidential election results by county, it's a bit too much to expect that counties are pulled randomly from an urn. It's reasonable to assume test (revealed counties) and training data (remaining counties) are exchangeable if all orders are equally likely, but small populations or East Coast counties are typically finished earlier, and in fact there is some distribution on the orders.

So what we need to do often is reweight observed data: suppose we have

$$(X_i, Y_i) \overset{\text{iid}}{\sim} P_X \times P_{Y|X}$$

for known data i, and

$$(X_j, Y_j) \overset{\text{iid}}{\sim} Q_X \times P_{Y|X}$$

for unseen data j, where Q_X has undergone a covariate shift (this is really what's happening in all of the examples we typically think of for why counties report in different orders). Another setting in which this happens is where we assign either treatment or control based on a coin toss X whose probability **may depend on the person**, and so we can observe what happens to people in the treatment group but "not what would have happened if they were in the control," and vice versa.

So the question we might ask is: if we are in the control group, how can we predict how well we would have done if we had been given treatment? This is different from the usual causal inference, and it turns out that we can in fact construct a prediction interval for this because we have lots of data in the other group. We're **not from the same population** (we're not exchangeable with the other group), so we have to be careful to account for the covariate shift, but there is something we can in fact do.

This is called the **counterfactual inference problem**, and the idea is to make use of the **covariate shift**

$$w(x) = \frac{dQ_X}{dP_X}(x).$$

In the context of CQR, we have some histogram of the conformity scores on the calibration side, and now we reweight that distribution of conformity scores according to w by upweighting training points that could have come from the calibration distribution: instead of the empirical distribution we now use

$$\sum_{i=1}^{n} p_i(x)\delta_{S_i} + p_\infty(x)\delta_\infty, \quad p_i(x) = \frac{w(X_i)}{\sum_{i=1}^{n} w(X_i)},$$

and otherwise the method is exactly the same: we return a prediction interval which takes the quantiles in the weighted distribution

$$[\hat{q}_{0.05}(x) - Q(x), \hat{q}_{0.95}(x) + Q(x)].$$

We'll see more of this next time, and the point is that we can again guarantee $(1-\alpha)$ coverage as long as we can find this $\frac{dQ}{dP}$.

16 May 27, 2025

Fact 136

As mentioned at the beginning of the document, this lecture was transcribed via course lecture slides and some notes by Marc Soong and Yifan Zhu.

We'll start by continuing our discussion of the counterfactual inference problem and weighted conformal prediction.

Recall that the idea is that sometimes we have different populations (for example if we've sampled two different groups for treatment and control via some non-iid method) and thus need to account for covariate shift because exchangeability is no longer true. We do so by "upsampling the stuff more likely to be from the test sampling."

Using the notation in the previous lecture, we have some weights $p_i(x) \propto w(X_i) = \frac{dQ_X}{dP_X}(X_i)$.

Proposition 137

Let $P^w = \sum_{i=1}^n p_i(x)\delta_{S_i} + p_{n+1}(x)\delta_\infty$ be the weighted distribution of scores. If we condition on an (unordered) bag of $n+1$ scores, then $S_{n+1}^{Y_{n+1}}$ is a random draw from P^w.

The point is that this means the p-value $p^{Y_{n+1}}$ will stochastically dominate a uniform random variable by the probability integral transform lemma when conditioned on the bag of scores, so in particular it also does so marginally. That means that we can construct the following prediction interval:

Theorem 138 (Barber-Candes-Ramdas-Tibshirani 2019)

For any score function $\mathcal{S}$ and any $\alpha \in (0,1)$, we can construct the prediction interval

$$\hat{C}_n(x) = \left\{ y \in \mathbb{R} : S_{n+1}^{(x,y)} \le Q_{1-\alpha}\left(\sum_{i=1}^n p_i(x)\delta_{S_i^{(x,y)}} + p_{n+1}(x)\delta_\infty \right) \right\}.$$

Then $\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \ge 1-\alpha$.

Proof of Proposition 137. We'll assume for simplicity that all scores are distinct, but the argument can be adapted to collisions easily. We thus have a one-to-one mapping between $S_i = S(X_i, X_i)$ and $Z_i = (X_i, Y_i)$, and we **wish to show** that

$$\mathbb{P}(Z_{n+1} = z_i | \text{bag}(z)) = \mathbb{P}(S_{n+1} = s_i | \text{bag}(s)) = p_i = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}.$$

(In the case where all ws are 1, there is no covariate shift.) For this, let f be the pdf (likelihood) of the data; we have

$$f(z_1, \cdots, z_{n+1}) = p(z_1)\cdots p(z_n)q(z_{n+1})$$

and we want the total density over permutations where the last one is z_i. Thus

$$\begin{aligned}\mathbb{P}(Z_{n+1} = z_i | \text{bag}(z)) &= \frac{\sum_{\sigma:\sigma(n+1)=i} p(z_{\sigma(1)})\cdots p(z_{\sigma(n)})q(z_i)}{\sum_j \sum_{\sigma:\sigma(n+1)=j} p(z_{\sigma(1)})\cdots p(z_{\sigma(n)})q(z_j)} \\ &= \frac{\sum_{\sigma:\sigma(n+1)=i} p(z_{\sigma(1)})\cdots p(z_{\sigma(n)})p(z_i)w(x_i)}{\sum_j \sum_{\sigma:\sigma(n+1)=j} p(z_{\sigma(1)})\cdots p(z_{\sigma(n)})p(z_j)w(x_j)},\end{aligned}$$

but now all terms have the same product of ps and thus this all just simplifies to $\frac{w(x_i)}{\sum_j w(x_j)} = p_i$, as desired. □

Example 139

Sometimes we have a setting where each subject is either treated or control with probability determined by some propensity score $e(x)$, and we may be curious about how to perform ITE (individualized treatment effect) inference. This is a special case of counterfactual inference, and the goal is to predict the values of $Y(1)$ for specific subjects even if those values are not observed. Skipping the details, the point is that the performance of methods like causal forest, X-learner, or BART become much better when conformalized.

The idea is that we reveal something about the bag and then want to see whether the test point is also from the bag as well, so it's the same core idea of conformal inference! We just need to upweight or downweight by the appropriate p_i factor, and all that's necessary is for us to know what the covariate shift actually is.

With this, we'll move on to more on conditional conformal prediction. First, we recap the ideas of split conformal prediction. We construct a conformity score $S(X_i, Y_i)$ (for example $|\hat{Y}_i - Y_i|$), and we compute a (slightly modified) quantile $\hat{c}_{1-\alpha}$ of the values $S_{1:n}$ or of the values $S^y_{1:n+1}$ (where we fill in the $(n+1)$th data point by (X_{n+1}, y)). We then output an interval $\hat{C}(X_{n+1})$ to be those values where $S(X_{n+1}, y) \le \hat{c}_{1-\alpha}$. The point is that **constructing a prediction set is equivalent to quantile estimation**, and the main question we ask is how to make the tradeoff between better score estimates (fitting the model) and better quantile estimates (gathering more test data points) with limited data.

However, everything we've discussed so far has only guaranteed marginal coverage (that is, if X_{n+1} is drawn from the same distribution, then we get 90 percent prediction coverage). This does not account for things like variation over groups, and it doesn't help us deal with the case where test data and training data may be different as well. For example, if we have 100 percent coverage for old people and 0 percent coverage for young people, but 90 percent of all people are old, then we do get 90 percent prediction coverage but possibly not in the most useful way. Thus, we may want to ask for conditional coverage, and here would be the ideal conditions that we would want to hold:

- **Distribution-free**: we should not need to make any assumptions about the distribution of the data.
- **Capable of utilizing a black-box model**: we should be able to feed in models like neural networks which give us scores $S(x, y)$ that may not be fully transparent.
- **Conditional validity**: we have that $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) | X_{n+1}) \ge 1 - \alpha$ for almost all X_{n+1}.

Unfortunately, this is basically too much to ask for:

Theorem 140 (Vovk 2012)
If $\hat{C}$ is distribution-free and conditionally valid, then we must have $\mathbb{E}[|\hat{C}(X_{n+1})|] = \infty$.

Remark 141. *There are some additional details mentioned in the original paper – for example, X_{n+1} must not be an atom of the distribution because otherwise we could just see that value enough times in the training set to get a valid prediction interval. But if it isn't an atom, then in fact it has positive probability of actually having infinite length.*

So what we'll hope for is instead "conditional-ish validity." There's basically a spectrum of guarantees we can aim for – on the one hand we have **marginal validity** (which is what we get from split conformal), which can be equivalently formulated in either of the forms

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) = 1 - \alpha, \quad \mathbb{P}(S_{n+1} \le \hat{c}_{1-\alpha}) = 1 - \alpha.$$

On the other end, we have the (unfortunately impossible) **conditional validity**

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) | X_{n+1}) = 1 - \alpha, \quad \mathbb{P}(S_{n+1} \le \hat{c}_{1-\alpha}(X_{n+1}) | X_{n+1}) = 1 - \alpha.$$

Yet another way to formulate these statements is that marginal validity guarantees

$$\mathbb{E}\left[(1\{S_{n+1} \le \hat{c}_{1-\alpha}\} - (1-\alpha)) \cdot c\right] = 0$$

for any constant, while conditional validity guarantees

$$\mathbb{E}\left[(1\{S_{n+1} \le \hat{c}_{1-\alpha}(X_{n+1})\} - (1-\alpha)) \cdot f(X_{n+1})\right] = 0$$

for all integrable functions f. We'll instead try to guarantee a statement of this kind for all f within some σ-algebra $\mathcal{F}$, which we call a **function class**.

Example 142

As a simple but useful case, suppose G is some pre-specified group of values (for example specifying some characteristics of our population, like being young versus old or white versus nonwhite). We can then let $\mathcal{F}$ be the set of linear combinations of indicator functions of G, which will allow us to get **group-conditional validity**

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) | X_{n+1} \in G) = 1 - \alpha.$$

This framework also lets us get guarantees under covariate shifts as well: we will see with the strategy below that defining a tilt of the form $dP_f = \left(\frac{f}{\mathbb{E}_P[f]} dP_X\right) \times dP_{Y|X}$, we get validity under $\mathbb{P}_f$ if f is a nonnegative function in our function class $\mathcal{F}$. (This idea is due to Gibbs, Cherian, and Candes.)

The key question to ask is then how we change our strategy when estimating $\hat{c}_{1-\alpha}$; the key tool will be to use quantile loss.

Example 143

Suppose we have d covariates $\Phi(X_i)$ for a linear model, which in the "group setting" above may be the indicators for existing in each of the d groups (so $\Phi(X_i) = (1, 1, 0, \cdots)$ would indicate being in the first two groups but not the second). Letting ℓ_α be the quantile loss function, we can then solve the quantile regression problem

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell_\alpha(S_i, \beta^T \Phi(X_i)).$$

With the true value of the linear fit β^*, $\beta_*^T \Phi$ is in fact $\mathcal{F}$-conditionally valid, where $\mathcal{F}$ is the set of linear combinations of the Φ functions:

$$\mathbb{E}\left[\left(1\{S \leq \beta_*^T \Phi(\vec{X})\} - (1-\alpha)\right) \cdot \Phi(\vec{X})\right] = \vec{0}.$$

This yields asymptotic guarantees if we basically plug in $\Phi(X_{n+1})$:

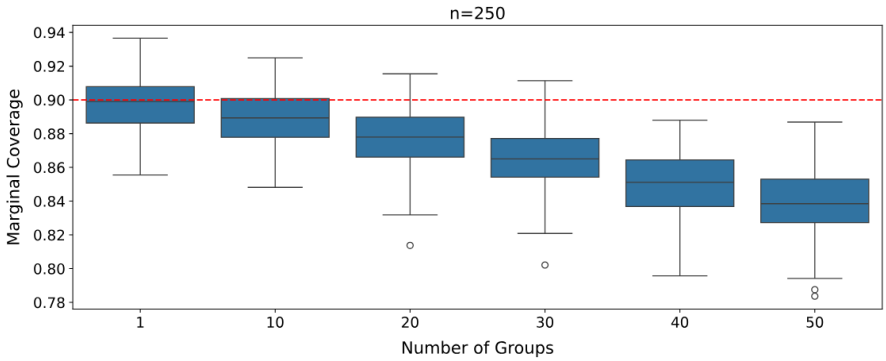
Proposition 144 (Jung et al. 2023)

Let $\Phi(\cdot)$ be the family of indicators coming from some set of groups $\mathcal{G}$, and let $\hat{c}_{1-\alpha}^{\text{QR}}(X_{n+1}) = \hat{\beta}^T \Phi(X_{n+1})$. Then we get asymptotic group-conditional coverage

$$\mathbb{P}\left(S_{n+1} \leq \hat{c}_{1-\alpha}^{\text{QR}}(X_{n+1}) | X_{n+1} \in G\right) \to 1 - \alpha$$

for $G \in \mathcal{G}$.

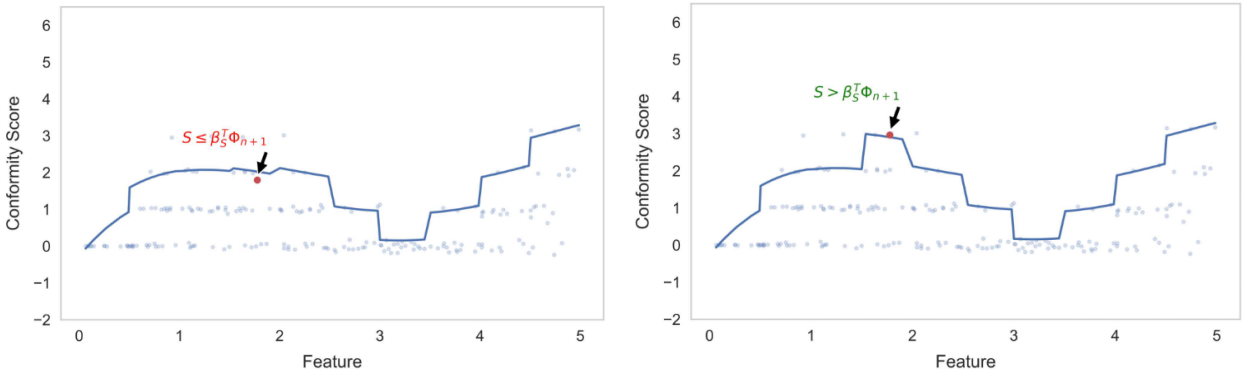
But the point of conformal prediction isn't really to get asymptotic guarantees – indeed, with finite sample sizes and many groups, we start losing coverage (as shown below). The idea is that quantile regression is a linear program and thus an interpolation of a lot of lines, so the 90th quantile will be biased downward and the 10th percentile will be biased upward.



The way to fix this is to include one more term in the loss, yielding the augmented quantile regression estimate

$$\hat{\beta}_S = \text{argmin}_{\beta\in\mathbb{R}^d} \sum_{i=1}^{n} \ell_\alpha(S_i, \beta^T\Phi(X_i)) + \ell_\alpha(S, \beta^T\phi(X_{n+1})).$$

We can interpret this in the following way. We propose a value of S and ask whether it is above the quantile fit $\beta_S^T\Phi(X_{n+1})$; if no, then we raise it a bit. This will correspondingly pull the quantile regression up, and we get our estimate $\hat{c}^{CC}_{1-\alpha}(X_{n+1})$ (CC stands for "conditional calibration") right when S is above $\beta_S^T\Phi(X_{n+1})$ for the first time.



(Naive quantile regression would have just set the threshold when our point crossed the fit on the left.)

Theorem 145

With the notation above, $\hat{c}^{CC}_{1-\alpha}(X_{n+1}; U)$ (where U is some randomization procedure) is $\mathcal{F}$-conditionally valid for $\mathcal{F}$ the class of linear combinations of Φ functions. In particular, this yields group-conditional coverage of the sort we're looking for if Φ are our indicator functions for various G.

We can then test whether this approach indeed works to get group coverage, and indeed conditional calibration gets us the correct amount of coverage on each group (with some randomization to avoid accidentally causing overcoverage).

Remark 146. *With the notation we've set up, we can now see split conformal prediction as quantile regression. Indeed, the* $(1-\alpha)$*th quantile of our scores* $S_1, \cdots, S_n$ *is then* $q^* = \text{argmin}_{q\in\mathbb{R}} \sum_{i=1}^{n}(1-\alpha)(S_i - q)_+ + \alpha(q - S_i)_+$*, and so we can construct the prediction interval by picking all* y *such that* $S(X_{n+1}, y)$ *is less than* $\text{argmin}_q \left[\sum_{i=1}^{n} \ell_\alpha(S_i; q) + \ell_\alpha(S_{n+1}; q)\right]$.

We'll close by briefly mentioning the proof of marginal coverage, which is related to the idea in Jung et al. This

relies on the KKT condition for optimization, which reads

$$0 \in \left\{ \sum_{S_i \neq q^*} (\alpha - 1\{S_i > q^*\}) + \sum_{S_i = q^*} \lambda_i | \lambda_i \in [\alpha - 1, \alpha] \right\}.$$

If this condition holds, then we can rearrange and solve for λ_i to achieve equality, which yields

$$\frac{1}{n+1} \sum_{i=1}^{n+1} (\alpha - 1\{S_i > q^*\}) = \frac{1}{n+1} \sum_{S_i = q^*} (\alpha - \lambda_i) \geq 0.$$

But now by exchangeability of the scores, the left-hand side being nonnegative is exactly equivalent to $\alpha - \mathbb{P}(S_{n+1} > q^*) \geq 0$, which is what we want.

17 May 29, 2025

We'll begin with a few more details about conditional calibration. Last time, we mentioned issues with "using the training points to estimate the quantiles" (which is not useful because empirical quantiles are very biased) and how it only gives us asymptotic guarantees. Instead, we learn quantiles using the features $\Phi(X_i)$ and also include a test point S, which we move so that it's just below the estimated quantile function. This gives us conditional coverage over (possibly overlapping) groups.

Fact 147

In practice, we need to be able to actually compute this value $\hat{c}^{\text{CC}}_{1-\alpha}$, and we do so with the help of LP duality because this problem is LP-representable (the function $\ell_\alpha(t)$ is $(1-\alpha)u_+ + \alpha u_-$, subject to $u_+ - u_- = t$ and $u_+, u_- \geq 0$).

We thus have a loss function of the form

$$\sum_{i=1}^{n+1} (1-\alpha) p_i + \alpha q_i$$

subject to constraints $S_i - \Phi_i^T \beta - p_i + q_i = 0$, $S - \Phi_{n+1}^T \beta - p_{n+1} + q_{n+1} = 0$, and $p_i, q_i \geq 0$. Our goal is to minimize this over p, q, β.

Every such problem has an **LP dual**, and the dual problem that we end up wanting to solve is of the form

$$\max_{\eta} \sum_{i=1}^{n} \eta_i S_i + \eta_{n+1} S, \quad \text{with constraints} \quad -\alpha \leq \eta_i \leq 1-\alpha, \quad \sum_{i=1}^{n+1} \eta_i \Phi_i = 0.$$

But this dual perspective unifies all three quantile approaches if we track the dual variable. Ordinarily for quantile regression we want the largest S such that the dual variable for the imputed point satisfies $\hat{\eta}^S_{n+1} < 0$, but what we actually want under conditional calibration is that $\eta^S_{n+1} < 1 - \alpha$. (All we need to do is check via a "path-following homotopy method" whether η_{n+1} reaches its upper limit constraint.) So it's not computationally hard, and in general under randomization $\hat{c}^{\text{CC}}_{1-\alpha}(X_{n+1}; U)$ then only asks us to have $\hat{\eta}^S_{n+1} < U - \alpha$; this will indeed be what's necessary to get us exactly 90 percent coverage.

As mentioned before, all of this theory can apply to covariate shifts (if our shifts are in the appropriate function class), and it can be used in real-world problems like predicting crime rate from various demographics in ways that don't bias against certain groups. There are various extensions too, such as derandomization, infinite classes, or more

precise estimates of coverage, but we won't get into them here.

Today's main topic is **Stein's phenomenon** and estimation of a multivariate normal mean. This might seem like a silly topic, but we'll see something surprising about estimating the mean and study the correct unbiased risk estimate. (The person who a lot of this is based on, Charles Stein, was a math and statistics professor at Stanford. He didn't publish a lot of papers, but they were extremely groundbreaking.)

Example 148

Consider a Gaussian vector $X \sim N(\mu, \sigma^2 I_p)$ in p dimensions, which we can write as $X_i = \mu_i + \sigma z_i$ for z_i iid standard normal. We will assume that μ is unknown but that σ^2 is known, and our goal is to estimate μ.

We then want to look for estimators that are good at estimating the mean using the quadratic loss function

$$\ell(\hat{\mu}, \mu) = ||\hat{\mu} - \mu||_2^2 = \sum_{i=1}^{p} (\hat{\mu}_i - \mu_i)^2 .$$

This is a random quantity depending on the samples we receive, and our risk function will be the mean-squared error

$$R(\hat{\mu}, \mu) = \mathbb{E}_\mu \left[||\hat{\mu} - \mu||^2 \right].$$

Notice that X_i and X_j are independent here, so perhaps X_i is an estimate of the age of the universe and X_j is the quality of students at Stanford – they don't need to have anything to do with each other. The MLE (maximum likelihood estimator) for this is trivial, since it's just the sample mean X. We can calculate the risk of this estimator – it's unbiased so we'll just see the variance, which is σ^2 per coordinate. More explicitly,

$$R(\hat{\mu}_{\text{MLE}}, \mu) = \mathbb{E}_\mu \left[||X - \mu||_2^2 \right] = \sigma^2 \mathbb{E}[||z||^2] = p\sigma^2.$$

The question is whether this is the smallest possible risk, and that's a mathematical question: "is it possible to find an estimate $\hat{\mu}$ which outperforms this risk, no matter what μ is?". Of course, we can do better for some values of μ (because we can just say $\hat{\mu} = 0$ and then if μ is actually zero, we have no risk), but our goal is to always do better than this constant risk $p\sigma^2$ no matter what.

Fact 149

It turns out that for $p = 1, 2$, this MLE is indeed the best estimator in the sense that we cannot uniformly improve the risk. We say that the sample mean is the best estimator and that it is **admissible**.

Stein proved this for $p = 2$ but couldn't do so for $p = 3$, and it turns out it's because the sample mean is not admissible!

Definition 150

The **James-Stein estimator** is defined by

$$\hat{\mu}_{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right) X.$$

Notice that this is a nonlinear, biased estimator – it shrinks the MLE towards 0 by multiplying it by some specific scalar, and in fact in principle it can actually become negative and go in the opposite direction as our sample mean.

Theorem 151 (James, Stein 1961)

For all $\mu \in \mathbb{R}^p$, we have

$$\mathbb{E}_\mu \left[||\hat{\mu}_{\text{JS}} - \mu||^2 \right] < \mathbb{E}_\mu \left[||\hat{\mu}_{\text{MLE}} - \mu||^2 \right].$$

This may make sense intuitively as a "regularization" term near $\mu = 0$, but it's perhaps surprising that it works for large μ as well. And what this is saying is that if we want to estimate the quality of students X_i at Stanford and also two other quantities X_j, X_k, we should take our estimate for X_i and then use some other information about those other quantities X_j, X_k, even though they are independent.

Remark 152. *Professor Candes met Jim Simons at some award a few years ago, and he had to explain why people care about the theory of statistics. He explained the James-Stein phenomenon and wrote out the proof, and that was what convinced Jim Simons to care more!*

This James-Stein estimator is also not admissible – for example putting a "positive part" on the scalar actually outperforms this quantity as well. But we won't focus so much on that.

Stein's original argument (in 1956) for why the mean should not be the best estimator is that a good estimate should roughly obey $\hat{\mu}_i \approx \mu_i$ and thus $\hat{\mu}_i^2 \approx \mu_i^2$ for all i. That means $\sum_i \hat{\mu}_i^2 \approx \sum_i \mu_i^2$, but that's not what is happening for the MLE: we instead have $\mathbb{E}[\sum X_i^2] = ||\mu||^2 + \sigma^2 p$. Thus this suggests that wherever the MLE is, it will be rather big compared to the true μ – in high dimensions we'll have a problem where we get concentration around the mean for $\mathbb{E}\sum X_i^2$ (because chi-square converges around its mean quite quickly). So shrinking makes sense: in high dimension, the true mean will be on some sphere with squared radius $||\mu||^2$, while the MLE will be extremely close to a sphere of squared radius $||\mu||^2 + \sigma^2 p$. So if we shrink slightly, we'll bring ourselves closer, and indeed the result is as follows:

Theorem 153

If we use the estimator

$$\hat{\mu}_c = \left(1 - c\frac{\sigma^2}{||X||^2}\right) X,$$

then this outperforms the MLE for all $c \in (0, 2(p-2))$.

We'll prove this, and we'll just do the case $c = p - 2$ for simplicity (that's where we get the most savings). For this, we make use of **Stein's unbiased risk estimate (SURE)**. Assuming that X is Gaussian with mean μ and variance $\sigma^2 I$, let's write our estimator as $\hat{\mu} = X + g(X)$ for some **almost differentiable** g (we'll explain this terminology below) which is integrable in the sense that

$$\mathbb{E}\left[\sum_{i=1}^p |\partial_i g_i(X)|\right] < \infty.$$

(Otherwise, the formulas we will write will make no sense – we need to be careful!) We then have the following fact:

Theorem 154 (Stein's identity, 1981)

With the notation above,

$$\mathbb{E}\left[||\hat{\mu} - \mu||^2\right] = p\sigma^2 + \mathbb{E}\left[||g(X)||^2 + 2\sigma^2 \sum_i \partial_i g_i(X)\right].$$

This is actually incredibly useful, because the expectation on the right-hand side does not depend on μ and thus we get an unbiased estimator for the risk! And thus we just need to calculate that quantity (provided that we can differentiate g) and try to find some function g such that the expectation is negative.

Here the assumption we have on g of **almost differentiability** is saying that there exists some function h_i satisfying

$$g_i(x+z) - g_i(x) + \int_0^1 \langle h_i(x+tz), z\rangle dt,$$

which we often write as $h_i = \nabla g_i$. So something like $g(x) = 1\{x > 0\}$ is "differentiable almost everywhere," but it is not "almost differentiable" because we would need h to be a delta function. But something like $g(x) = x\{x > 0\}$ is fine, since we can let h be $1\{x > 0\}$.

The point of this identity is that

$$\text{SURE}(\hat{\mu}) = p\sigma^2 + ||g(X)||^2 + 2\sigma^2 \text{div } g(X)$$

is always an unbiased statistic for the risk.

Proof of Theorem 154. This is essentially "just calculus" and comes down to integration by parts. Expanding out the square,

$$\mathbb{E}\left[||X + g(X) - \mu||^2\right] = \mathbb{E}\left[||X-\mu||^2\right] + 2\mathbb{E}\left[(X-\mu)^T g(X)\right] + \mathbb{E}\left[||g(X)||^2\right],$$

and so we just need to show that the cross-term matches up with the divergence term; that is,

$$\mathbb{E}\left[(X-\mu)^T g(X)\right] \stackrel{?}{=} \sigma^2 \mathbb{E}\left[\text{div } g(X)\right].$$

We'll scale $\sigma = 1$ just to keep notation simpler (nothing in the argument changes otherwise); the left-hand side is a sum of terms

$$\mathbb{E}\left[(X_i - \mu_i) g_i(X)\right] = \int (x_i - \mu_i) g_i(x) \phi(X - \mu) d\vec{x}$$

where $\phi(X-\mu)$ is a p-dimensional product of densities. Separating out the term we care about, we can write this as

$$\int (x_i - \mu_i) g_i(x) \phi(x_i - \mu_i) \prod_{j \neq i} \phi(x_j - \mu_j) d\vec{x},$$

and now we do the integral over x_i by parts. For a Gaussian, **the derivative of a standard Gaussian density is $-x$ times the density itself**, so in particular $\phi'(x_i - \mu_i) = (x_i - \mu_i)\phi(x_i - \mu_i)$. Therefore, we can use integration by parts on the blue terms (the boundary term goes away because of our integrability condition)

$$\int (x_i - \mu_i) g_i(x) \phi(x_i - \mu_i) \prod_{j \neq i} \phi(x_j - \mu_j) d\vec{x} = \int \partial_i g_i(x) \phi(x_i - \mu) \prod_{j \neq i} \phi(x_j - \mu_j) d\vec{x} = \mathbb{E}\left[\partial_i g_i(x)\right],$$

and so the identity holds term-by-term and summing over all i yields the result. □

But now we can use this formula to prove that $\hat{\mu}_{\text{JS}}$ is a better estimate. Again take $\sigma = 1$ for simplicity and write the estimator in the form we've been setting up

$$\hat{\mu}_{\text{JS}} = X - \frac{p-2}{||X||^2} X = X + g(X), \quad \text{where } g(x) = -\frac{(p-2)x}{||x||^2}.$$

The risk of this estimate is then the MLE risk p plus this additional expectation, and that's just calculus: we find that (here each component g_i is $-\frac{(p-2)x_i}{||x||^2}$)

$$||g(x)||^2 = \frac{(p-2)^2}{||x||^2},$$

$$\partial_i g_i(x) = -\frac{p-2}{||x||^2} + \frac{2(p-2)x_i^2}{||x||^4} \implies \text{div } g(x) = \frac{-p(p-2)+2(p-2)}{||x||^2} = \frac{-(p-2)^2}{||x||^2}.$$

So plugging back into Theorem 154 shows that the latter term of the expectation outweighs the former term, and we get

$$\mathbb{E}\left[||\hat{\mu}_{\text{JS}} - \mu||^2\right] = p - \mathbb{E}\left[\frac{(p-2)^2}{||X||^2}\right],$$

which is strictly smaller than p no matter what μ is. Indeed, this correction term is larger when μ is small, and so we're getting more benefits in the regime where regularization makes sense.

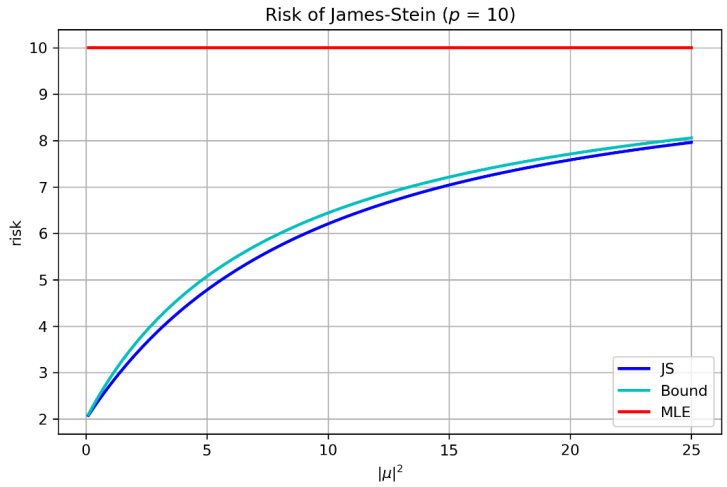
Fact 155

We can in fact calculate some sharper estimates using the inequality

$$\mathbb{E}\left[\frac{1}{||X||^2}\right] \geq \frac{1}{(p-2)+||\mu||^2},$$

with equality when $\mu = 0$. So that tells us exactly how much savings we get compared to the MLE; the risk is at most $p - \frac{p-2}{1+\frac{||\mu||^2}{p-2}}$. For example, if $\mu = 0$, the risk of the MLE is p but the risk of James-Stein is 2.

Below is a plot of risk for James-Stein versus MLE in dimension $p = 10$; it's quite close to the derived bound we just found.



As mentioned before, we can in fact go even lower than the JS line shown above by taking only the positive part of $1 - \frac{p-2}{||X||^2}$, but it's still not admissible. If we do in fact want an admissible estimator, any Bayes estimator will do.

Remark 156. *Intuitively, if we have a bunch of values of $\mu_i + Z_i$, the largest ones are expected to overestimate and the smallest ones are expected to underestimate, so shrinking the values should be likely to give us something closer to the true answer. And so we can think of this as just being regression to the mean in another perspective!*

18 June 3, 2025

Fact 157

As mentioned at the beginning of the document, this lecture was transcribed via course lecture slides and some notes by Salil Goyal and Victor Kolev.

The topic of this last lecture will be the **empirical Bayes** interpretation of James-Stein. The material here partially follows ideas from Efron and Morris's paper "Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach."

Example 158

Consider a model where we sample iid from $\mu_i \sim N(0, \tau^2)$ and then sample our multivariate vector $X \sim N(\mu, \sigma^2 I)$. Here, we should interpret X as our observed variables and μ as our underlying mean, and here everything is a random variable rather than an unknown parameter.

We can calculate the posterior distribution; it turns out that

$$\mu|X \sim N\left(\frac{v}{\sigma^2}X, vI\right), \quad \frac{1}{v} = \frac{1}{\tau^2} + \frac{1}{\sigma^2},$$

or more compactly, that $\mu|X \sim N\left(\frac{\tau^2}{\tau^2+\sigma^2}X, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}I\right)$. So in fact the coordinates of μ are marginally independent, and they are even conditionally independent given X with posterior mean a "shrunken estimate" $\hat{\mu}_B = \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)X$. When the "signal-to-noise ratio" is roughly 1 (that is, when $\tau = \sigma$), this shrinks our mean halfway to zero.

Proposition 159

The Bayes risk associated with this "shrunk mean" posterior is

$$\mathbb{E}\left[||\mu_B - \hat{\mu}||^2\right] = R_{\text{MLE}}\frac{\tau^2}{\tau^2+\sigma^2}.$$

So at $\tau = \sigma$ we are removing half the risk by doing this shrinkage. This is a similar story to James-Stein, though do note here that the quantity we multiply by is always positive.

Proof. Letting $\rho = \frac{\sigma^2}{\sigma^2+\tau^2}$, we have that

$$\hat{\mu}_B - \mu = (1-\rho)(X-\mu) - \rho\mu.$$

We can thus compute the conditional expectation

$$\mathbb{E}\left[||\hat{\mu}_B - \mu||^2|\mu\right] = (1-\rho)^2\mathbb{E}\left[||X-\mu||^2|\mu\right] + \rho^2||\mu||^2,$$

and now because $X - \mu$ is $N(0, \sigma^2 I)$ by definition this simplifies to $(1-\rho)^2 p\sigma^2 + \rho^2||\mu||^2$. If we now marginalize over μ, the overall expectation is thus $(1-\rho)^2 p\sigma^2 + \rho^2 p\tau^2$, which simplifies to $p\sigma^2 \cdot \frac{\tau^2}{\tau^2+\sigma^2}$. Since $p\sigma^2$ is the risk of the MLE, this yields the result. □

We can now use this to interpret James-Stein in this Bayesian way: suppose now that σ is known and that the Bayes model is correct, but τ is not known and thus we cannot compute the shrinkage factor. What we'll do then is to **estimate the factor from data** in the following way. If we know that each $X_i = \mu_i + z_i$ is iid from $N(0, \tau^2+\sigma^2)$, then the distribution of $||X||^2$ follows $(\tau^2+\sigma^2)\chi^2_p$, and the chi-square distribution has the nice property that $\mathbb{E}\left[\frac{p-2}{\chi^2_p}\right] = 1$. Therefore $\frac{(p-2)\sigma^2}{||X||^2} = \frac{\sigma^2}{\tau^2+\sigma^2}\frac{p-2}{\chi^2_p}$ is actually an unbiased estimator for the shrinkage factor, and in fact plugging this in yields exactly the James-Stein expression from last time!

The risk of this estimate (assuming our Bayes model is indeed correct) can also be calculated: it works out to

$$\mathbb{E}\left[||\hat{\mu}_{\text{JS}} - \mu||\right]^2 = p\frac{\sigma^2\tau^2}{\tau^2+\sigma^2} + \frac{2\sigma^4}{\tau^2+\sigma^2} = R_{\hat{\mu}_B}\left(1 + \frac{2\sigma^2}{p\tau^2}\right).$$

So even with τ unknown, we still get something fairly close to the Bayes risk – for example at a signal-to-noise ratio of 1 and with $p = 20$, this is only a 10 percent increase.

Fact 160

In the "regression to the mean" perspective, it's not so surprising that shrinking all of our estimates towards zero outperforms the MLE. But it's interesting that we do better as a frequentist statement – that is, regardless of the distribution of μ (even when the mean is very far away from zero!).

Related to this is the idea that we can even shrink towards any arbitrary value μ_0 and still dominate the MLE; that is, the risk of

$$\mu_0 + \left(1 - \frac{(p-2)\sigma^2}{||X - \mu_0||^2}\right)(X - \mu_0)$$

is lower than for just the sample mean. (Indeed, equivariance means we can intuitively interpret this as just being a translation of sorts.)

Remark 161. *Stein's officemate once asked Stein whether he would actually use this in practice, to which the response was "You've seen the proof. Why wouldn't you?"*

Example 162

In practice, it would make sense to shrink towards the true center and set $\mu_0 = \overline{X}$, and we can now see how that works out. Our model here is similar to before, but now μ_i are iid from some $N(\mu_0, \tau^2)$ and $X|\mu \sim N(\mu, \sigma^2 I)$.

The μ_i here are iid, and the posterior distribution takes the form

$$\mu_i | X_i \sim N(\mu_0 + (1-\rho)(X_i - \mu_0), \sigma^2(1-\rho)), \quad \rho = \frac{\sigma^2}{\tau^2 + \sigma^2}.$$

This time both μ_0 and τ^2 are unknown, so to construct an estimate we need to estimate both quantities: we'll estimate μ_0 by the sample mean and $\tau^2 + \sigma^2$ by $\frac{1}{(p-3)}\sum_{i=1}^p (X_i - \overline{X})^2$, since this time the **sample** variance is distributed as $(\tau^2 + \sigma^2)\chi_{p-1}$ and we have $\mathbb{E}\left[\frac{p-3}{\chi^2_{p-1}}\right] = 1$.

Theorem 163

This new James-Stein-type estimator

$$\hat{\mu}_i = \overline{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum_{i=1}^p (X_i - \overline{X})^2}\right)(X_i - \overline{X})$$

dominates the MLE for all $p > 3$ (and is equal for $p = 3$).

It's again interesting that such a result is purely frequentist even though it arises from Bayesian thinking! And while we've only considered the case where our random variables are independent Gaussians, similar phenomena occur for Poisson or binomial data or correlated Gaussians as well.

Example 164

We'll close the course with a very typical problem in statistics, in which we want to estimate players' true batting ability (via their batting average across a season) via observations made during the first week of play. To be concrete, we can consider the set of all players who had exactly 45 at-bats by a certain day – in April 26, 1970, this was a set of 18 players.

If θ_i denotes the true batting ability of each player, then X_i should be distributed as $\frac{1}{45}\text{Bin}(45, \theta_i)$, which is roughly $N(\theta_i, \frac{1}{45}\theta_i(1-\theta_i))$. This is not the setting of our example where the variance is fixed, but there is a certain **variance stabilization trick** we can use: if instead of considering these X_is, we consider

$$Y_i = \sqrt{45}\arcsin(2X_i - 1),$$

then the result will be approximately normal with variance 1: more precisely, $Y_i \approx N(\sqrt{45}\arcsin(2\theta_i - 1), 1)$. Thus if we have our 18 players, we can apply our estimator and predict that

$$\hat{\mu}_{\text{JS}} = \hat{Y} + \left(1 - \frac{15}{\sum_{i=1}^{18}(Y_i - \overline{Y})^2}\right)(Y - \overline{Y}).$$

Applying this to real historical data yields the following result:

player	$x_i = \hat{\theta}_i^{\text{ML}}$	$\hat{\theta}_i^{\text{JS}}$	θ_i	$y_i = \hat{\psi}_i^{\text{ML}}$	$\hat{\psi}_i^{\text{JS}}$	ψ_i
Roberto Clemente (Pitt, NL)	.400	.290	.346	-1.35	-2.91	-2.10
Frank Robinson (Balt, AL)	.378	.286	.298	-1.66	-2.97	-2.79
Frank Howard (Wash, AL)	.356	.281	.276	-1.97	-3.04	-3.11
Jay Johnstone (Cal, AL)	.333	.277	.222	-2.28	-3.10	-3.96
Ken Berry (Chi, AL)	.311	.273	.273	-2.60	-3.16	-3.17
Jim Spencer (Cal, AL)	.311	.273	.270	-2.60	-3.16	-3.20
Don Kessinger (Chi, NL)	.289	.268	.263	-2.92	-3.24	-3.32
Luis Alvarado (Bos, AL)	.267	.264	.210	-3.26	-3.30	-4.15
Ron Santo (Chi, NL)	.244	.259	.269	-3.60	-3.37	-3.23
Ron Swoboda (NY, NL)	.244	.259	.230	-3.60	-3.37	-3.83
Del Unser (Wash, AL)	.222	.254	.264	-3.95	-3.45	-3.30
Billy Williams (Chi, NL)	.222	.254	.256	-3.95	-3.45	-3.43
George Scott (Bos, AL)	.222	.254	.303	-3.95	-3.45	-2.71
Rico Petrocelli (Bos, AL)	.222	.254	.264	-3.95	-3.45	-3.30
Ellie Rodriguez (KC, AL)	.222	.254	.226	-3.95	-3.45	-3.89
Bert Campaneris (Oak, AL)	.200	.249	.285	-4.32	-3.53	-2.98
Thurman Munson (NY, AL)	.178	.244	.316	-4.70	-3.61	-2.53
Max Alvis (Mil, NL)	.156	.239	.200	-5.10	-3.68	-4.32

Indeed, we find a very large improvement in the mean squared error: for the values of μ we get 5.01 via James-Stein versus 17.56 via the MLE, and when we convert back to batting averages we get a mean squared error of 0.022 versus 0.077 (so improvement by a factor of 3.5).

Fact 165

However, the real question at this point is whether this is actually what we want to do. Indeed, we have increased the squared error for three of the players, and in particular we have significantly shrunk Clemente's (exceptional) average significantly towards the rest of the group. So there's a bit of a fairness dilemma here especially if we care only about individual results.

So overall, the big picture is that James-Stein is effectively fitting a regression of μ on X rather than X on μ. Jensen's inequality tells us that the expected value of the largest X_i will be biased upward from the true value, and so shrinkage corrects that bias.