STATS305A - Lecture 9

John Duchi
Scribed by Michael Howes

10/19/21

Contents

1 Recap

Recall our testing framework: we compute a test statistic t_{obs}, under the null H_0 t_{obs} follows some distribution T, reject if the p-value is less than some threshold, i.e. $\mathbb{P}_{H_0}(T \geq t_{obs}) \leq \alpha$.

2 Chosing alternative hypotheses

See "Strong Inference" by John Platt 60s/70s in Science.

When designing a test, it is very important to consider alternatives to the null hypothesis and design a tst to *best* distinguish your null H_0 from possible alternatives.

2.1 Neymann-Pearson

Suppose that we have point null and alternative hypotheses and so

$$H_0 : X \sim \mathbb{P}_0,$$
$$H_1 : X \sim \mathbb{P}_1.$$

Question: For a given level $\alpha \in (0, 1)$ (ie $\mathbb{P}_0(\text{reject } H_0) \leq \alpha$), what is the best test to distinguish $\mathbb{P}_0$ and $\mathbb{P}_1$ i.e. which test maximizes the power

$$\beta := \mathbb{P}_1(\text{reject} H_0).$$

Answer: The Neymann-Pearson lemma (NPL). Say P_0, P_1 have densities $p_0(x)$ and $p_1(x)$, define

$$L(x) = \log\left(\frac{p_1(x)}{p_0(x)}\right).$$

The optimal test is given by choosing a threshold t_α such that we

- Reject H_0 (accept H_1) if $L(X) > t_\alpha$.
- Reject H_1 (accpet H_0 (not reject H_0)) if $L(X) < t_\alpha$.
- Randomly reject/accept H_0 with equal probability if $L(X) = t_\alpha$.

Where we choose t_α to be the smallest t such that

$$\mathbb{P}_0(L(X) > t) \leq \alpha.$$

For the proof, see Wikipedia.

How do we use this in practice? We pick a null H_0, we think about possible alternatives, choose a test which distinguishes them and we can use NP to decide on the choice of test.

Example 1. Suppose we want to test a treatment/intervention in some setting. For example we might want to

(a) Create a drug that treats obesity.

(b) Set a harder final exam.

In (a) we are interested in a change in weight. We wish to see if the mean weight of the treated population decreases. In (b) we want to increase the spread to better assign grades. We wish to see if the variance increases.

Suppose that the null in both cases is

$$X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1).$$

The alternative in (a) is the mean decreases to $\mu < 0$ so $X_i \overset{\text{iid}}{\sim} \mathsf{N}(\mu, 1)$. The alternative in (b) is the variance of X_i increases to $\sigma^2 > 1$ so $X_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

Let $x = (x_i)_{i=1}^n$, then in (a)

$$\begin{aligned} L(x) &= -\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2 + \frac{1}{2}\sum_{i=1}^n x_i^2 \\ &= -\sum_{i=1}^n x_i\mu + \frac{n}{2}\mu^2 \\ &= -\frac{n}{2}\mu^2 + n\mu\bar{x}_n. \end{aligned}$$

By NPL the opitimal test is of the form reject H_0 if $L(x) > t_\alpha$ where t_α satisfies

$$\mathbb{P}_0(L(X) \geq t_\alpha) = \alpha.$$

Note that since $\mu < 0$,

$$\begin{aligned} \mathbb{P}_0(L(X) \geq t_\alpha) &= \mathbb{P}_0\left(-\frac{n}{2}\mu^2 + n\bar{X}_n\mu \geq t_\alpha\right) \\ &= \mathbb{P}_0\left(-n\bar{X}_n\mu \leq -\frac{n}{2}\mu^2 - t_\alpha\right) \\ &= \mathbb{P}_0\left(\sqrt{n}\bar{X}_n \leq \frac{\sqrt{n}}{2} - \frac{t_\alpha}{\mu\sqrt{n}}\right). \end{aligned}$$

We know that $\sqrt{n}\bar{X}_n \sim \mathsf{N}(0, 1)$ under H_0. Thus we set $\frac{\sqrt{n}}{2} - \frac{t_\alpha}{\mu\sqrt{n}} = -z_{1-\alpha}$ where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathsf{N}(0, 1)$. Thus NPL says we should reject when $\sqrt{n}\bar{x}_n \leq -z_{1-\alpha}$. Thus we reject when the sample mean is negative which makes sense.

In (b) note that

$$\begin{aligned} L(x) &= -\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{1}{2}\sum_{i=1}^n x_i^2 + \frac{1}{2}\log(\sigma^2) \\ &= \frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right)\sum_{i=1}^n x_i^2 + \log(\sigma). \end{aligned}$$

We know that $\sigma^2 > 1$ and so $1 - \frac{1}{\sigma^2} > 0$. Thus the NP test rejects when $\sum_{i=1}^n x_i^2$ is greater than some threshold. Under the null, $\sum_{i=1}^n X_i^2$ follows a $\chi^2_{(n)}$ distribution. Thus the optimal level α test rejects when

$$\sum_{i=1}^n x_i^2 \geq \chi^2_{n,\alpha}.$$

What if our H_0 is *not* a point null e.g. $H_0 : X_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with σ^2 unknown. Neymann Pearson does not apply. The "solution" is to use a plug-in estimate. That is estimate the unkown parameters, plug them in and then use Neymann-Pearson and other similiar methods. This is the idea behind the t-test when we compute

$$t_n := \frac{\sqrt{n}\bar{X}_n}{S_n},$$

where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{x}_n)^2$.

2.2 ANOVA

Let's us again consider contrasts in the ANOVA model. We have $Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$ where $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. We are interested in the nulls $H_{0,i} : \alpha_i = 0$ where $i = 1, \ldots, k$. We know the estimable constrasts are $\lambda^T \bar{Y}$ where $\lambda \in R^k$ and $\lambda^T \mathbf{1} = 0$ but $\lambda \neq 0$ where

$$\bar{Y} = \begin{bmatrix} \bar{Y}_{1,\bullet} \\ \vdots \\ \bar{Y}_{k,\bullet} \end{bmatrix} \in \mathbb{R}^k.$$

Under $H_0 : \alpha_i = 0$ for all i and $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ we have $\lambda^T \bar{Y} \sim \mathsf{N}(0, \sum_{i=1}^n \lambda_i^2/\sqrt{n_i})$. We can ask what is the most powerful test of $\alpha \neq 0$? It turns out it dos not exist since $H_1 : \alpha \neq 0$ is not point hypothesis.

Definition 1. The *non-central* χ^2 *distribution* is written $\chi^2_{(n),NC}(\|u\|_2^2)$ is the distribution of $\|Z\|_2^2$ where $Z \sim \mathsf{N}(\mu, I_n)$. Note that this distribution only depends on n and $\|u\|_2^2$ since

$$\begin{aligned} \|z\|_2^2 &= \|u + z - u\|_2^2 \\ &= \|z - u\|_2^2 + 2u^T(z - u) + \|u\|_2^2 \\ &= \|z - u\|_2^2 + 2\|u\|_2 \left(\frac{u}{\|u\|_2}\right)^T (z - u) + \|u\|_2^2. \end{aligned}$$

If $\alpha \neq 0$, then

$$\begin{aligned}\frac{(\lambda^T \bar{Y})^2}{\sum_{i=1}^n \frac{\lambda_i^2}{n_i}} &= \left(\sum_{i=1}^n \frac{\lambda_i \alpha_i}{\sqrt{\sum_{i=1}^n \lambda_i^2/n_i}} - \frac{\lambda^T(\bar{Y}-\alpha)^2}{\sqrt{\sum_{i=1}^n \lambda_i^2/n_i}}\right)^2 \\ &\overset{dist}{=} \left(\frac{\lambda^T \alpha}{\sqrt{\sum_{i=1}^n \lambda_i^2/n_i}} + N(0,1)^2\right) \\ &\sim \chi^2_{(1),NC}\left(\left(\frac{\lambda^T \alpha}{\sqrt{\sum_{i=1}^n \lambda_i^2/n_i}}\right)^2\right).\end{aligned}$$

Thus the distribution of our statistic under H_1 depends on α and thus the test which maximizes power also depends on α. (One way "around" this issue: split the data use half to estimate $\widehat{\alpha}$ and use the other half to test $H_0 : \alpha = 0$ against $H_1 : \alpha = \widehat{\alpha}$. This assumes that we can split the data into two independent halves).

3 Multiple testing

Can we correctly get some of the "discoveries" in a sample without making too many mistakes?

3.1 Setting

We have many nulls $H_{0,j}$, $j = 1, \dots k$, correctly rejecting a null equals making a discovery. A typical example is $(H_{0,j} : Y = X\beta + \varepsilon, \beta_j = 0)$. We want to control *false discoveries* where we reject $H_{0,j}$ even though $H_{0,j}$ is true.

3.2 Family wise error rate (FWER)

When we look at the FWER no mistakes are allowed. In this setting we can use Bonferroni correction and reject $H_{0,j}$ with at the level α/k. In this case

$$\mathbb{P}(\text{one or more false rejections}) \leq \sum_{j=1}^k \mathbb{P}(H_{0,j} \text{ falsely rejected}) \leq \frac{k\alpha}{k} = \alpha.$$

This is quite strict. Maybe we should be willing to make some false discoveries if it allows us to make more true discoveries.

3.3 False discovery proportion (FDP)

Define

$$FDP := \frac{\#\{\text{false rejections}\}}{\#\{\text{rejections}\} \vee 1},$$

where $a \vee b = \max\{a, b\}$. Controlling FDP means we are allowing overselves to make some mistakes if it means we make more true discoveries. Define the FDR (false discovery rate) to be

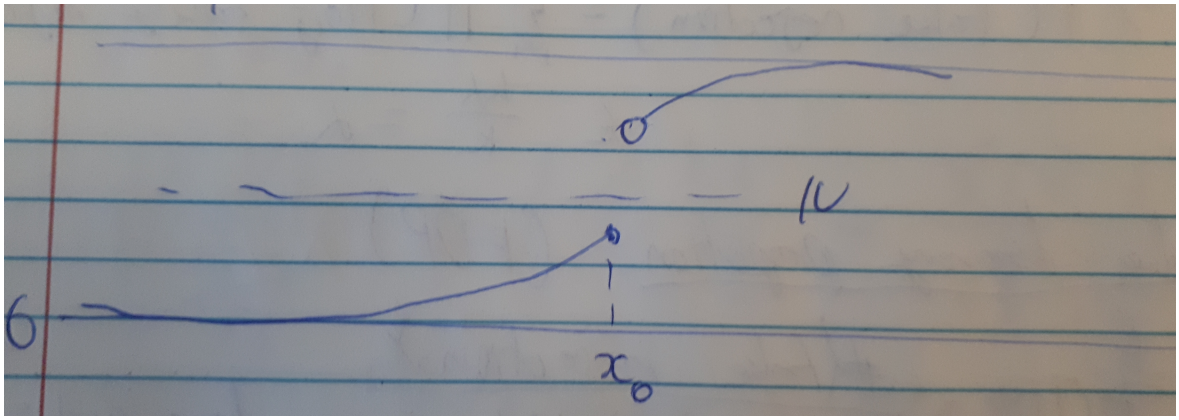
$$FDR = \mathbb{E}[FDP].$$

We will try to control FDR rather than FWER.

Aside 1. What is the distribution of p-value under the null H_0? It should be uniform on $[0,1]$ or "larger". This is because if F is the CDF of a random variable X then

$$\begin{aligned}\mathbb{P}(F(X) \le u) &= \mathbb{P}(X \le F^{-1}(u)) \\ &= F(F^{-1}(u)) \\ &= \begin{cases} u & \text{if } X \text{ has a density,} \\ \ge u & \text{else.} \end{cases}\end{aligned}$$

where $F^{-1}(u) = \inf\{t : F(t) \ge u\}$. See picture for a CDF with a jump at a value x_0.

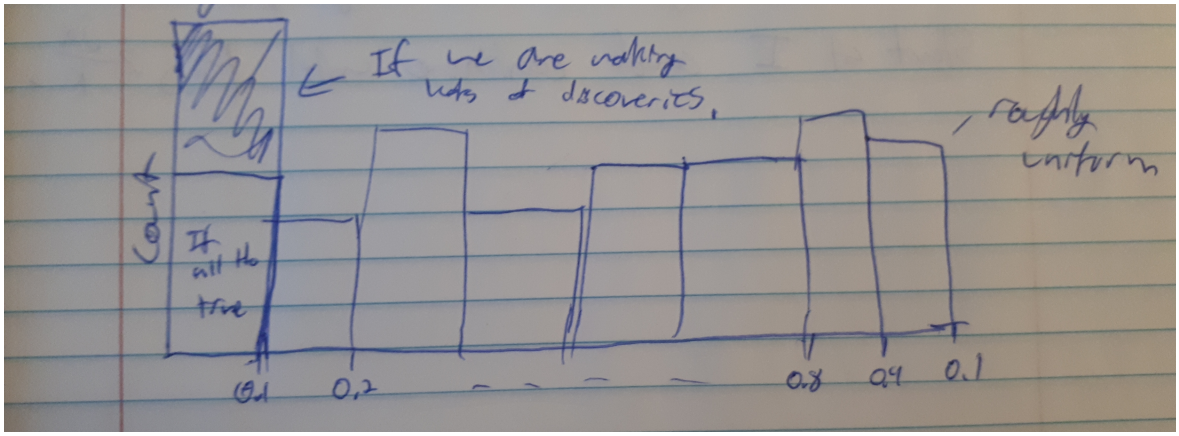


Now if we think about p-values. If F is a CDF of the test-statistic T under H_0 we reject for T large. Thus $p = 1 - F(t_{obs})$ where t_{obs} is our observed test-statistic. Thus

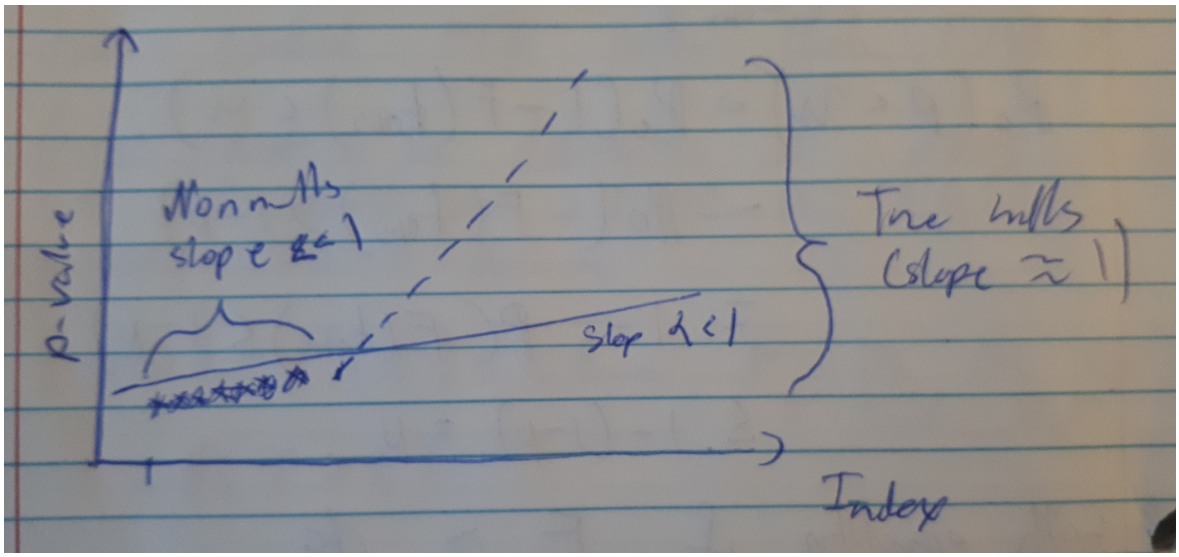
$$\begin{aligned}\mathbb{P}_0(p < u) &= \mathbb{P}_0(1 - F(t_{obs}) < u) \\ &= \mathbb{P}_0(F(t_{obs}) > 1 - u) \\ &= 1 - \mathbb{P}_0(F(t_{obs}) \le 1 - u) \\ &\le 1 - (1 - u) \\ &= u.\end{aligned}$$

And we have equality if F is continuous.

Going back to our multiple testing setting. Suppose we collect a pile of p-values $p_1, \dots, p_k$ and make a histogram. It would look like this



We would expect a roughly uniform distribution for most values of p but for small values there should be a spike corresponding to all the nulls we should reject. If we sort the p-value and plot them we'll get something like this:



For the true nulls, the slope will be approximately 1 but the slope will be much flatter for the p-values corresponding to false nulls. We reject the nulls that fall below the slope α line. This is the *Benjamini Hochberg procedure*. We start at $p_{(1)}$ and reject the nulls until $p_{(j)} \geq \frac{j\alpha}{k}$.