18.650: Fundamentals of Statistics November 1, 2023

Lecture 19 — Multiple hypothesis tests and T tests

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Multiple hypothesis testing

A neuroscientist at Dartmouth put a dead salmon under an fMRI scanner, and showed it images of humans in social situations. He then conducted a hypothesis test at level $\alpha = 0.1\%$ for each of 8000 voxels in its brain, to see whether the voxel was active while the fish was being shown the images. Represent the ith test by $H_{0,i} : \theta_i = 0$ vs $H_{1,i} : \theta_i \neq 0$, $i = 1, \dots, m$ where $m = 8000$, and let p_i be the corresponding p-value. The neuroscientist found that 8 voxels were active, i.e. rejected 8 of the null hypotheses. How can this be? The salmon is dead!

Let p_i be the ith p-value, so we reject the ith null hypothesis if $p_i \leq \alpha$. Since the salmon is dead we definitely know $H_{0,i}$ is true for all i. Under this null distribution, we expect that approximately a proportion $\alpha = 0.001$ of the p_i's are less than α. This is because we can think of the p_i's as 8000 random draws from a probability distribution (since the test statistics they are computed from are random). In fact, one can show this distribution of p values is uniform. Since $8000 \times 0.001 = 8$, it is no surprise we rejected 8 of the null hypotheses.

The message here is that when you conduct multiple hypothesis tests, it is inevitable that you will falsely reject some of them.

How do we fix this?

1.1 Bonferroni correction:

Instead of making sure the probability of type I error of each test is less than α, we can set up our tests so that the *overall type I error* is less than α, i.e. we could try to ensure

$$\mathbb{P}_{H_0}(\exists i \ : \ H_{0,i} \text{ is rejected}) \leq \alpha,$$

where $\mathbb{P}_{H_0}$ denotes the probability when all $H_{0,i}$, $i = 1, \dots, m$ are true.

Definition 1.1: Bonferroni correction

Using the Bonferroni correction means rejecting $H_{0,i}$ if $p_i \leq \alpha/m$, which ensures the overall type I error $\leq \alpha$.

Let's check the overall Type I error is really less than α using the Bonferroni correction:

$$\begin{aligned}\mathbb{P}_{H_0}(\exists i \ : \ H_{0,i} \text{ is rejected}) &= \mathbb{P}_{H_0}(\exists i \ : \ p_i \leq \alpha/m) \\ &= \mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m}\{p_i \leq \alpha/m\}\right) \\ &\leq \sum_{i=1}^{m}\mathbb{P}_{H_{0,i}}(p_i \leq \alpha/m) = \sum_{i=1}^{m}\alpha/m = \alpha.\end{aligned} \tag{1}$$

Here we have used that each of the p_i's is uniform under $H_{0,i}$. To get the inequality in the third line, we used the union bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. This could be a conservative upper bound (it's an equality only if the events $\{p_i \leq \alpha/m\}$ are disjoint, but there is no reason that should be the case.)

To see just how conservative the Bonferroni correction is, note that in the dead salmon study we would reject the ith hypothesis if $p_i \leq 0.001/8000 = 1.25 \times 10^{-5}$, which is below the resolution of statistical software.

1.2 Benjamini-Hochberg (BH) method.

To explain the BH method, we first introduce some terminology. Assume $H_{0,i}$ is rejected. Then we call the ith test a"positive" test. Now suppose $H_{0,i}$ was actually true. Then we call the outcome a "false positive". On the other hand if $H_{0,i}$ was correctly rejected, call this outcome a "true" positive.

Using this terminology, we can revisit the bound we computed with Bonferroni:

$$\begin{aligned}\text{overall type I error } &= \mathbb{P}(\exists \text{ false positive}) \\ &= \mathbb{E}\left[\max_i \mathbb{1}(i \text{ is a false positive})\right] \\ &\leq \mathbb{E}\left[\sum_i \mathbb{1}(i \text{ is a false positive})\right] = \mathbb{E}\left[\#\text{ false positives}\right]\end{aligned}$$

We are just rewriting equation (1) using expectations of indicators rather than probabilities of events. In the last line, we recognize that the sum of the indicators is precisely just the number of false positives out of the m tests.

Thus Bonferroni guarantees that the expected number of false positives is less than α. In contrast, BH guaranteed that the expected *proportion* of false positives out of all positives is less than α. Specifically, define

Definition 1.2: False discovery proportion & rate (FDP & FDR)

$$\text{FDP} = \frac{\#\text{ false positives}}{\#\text{positives}},$$

and

$$\text{FDR} = \mathbb{E}_{H_0}[\text{FDP}],$$

Remark.

We can compute the number of positives but of course not the number of false positives. Therefore, FDP cannot be computed in practice. However, we can still bound FDR.

Remark.

FDR is different from $\mathbb{P}(\exists$ false positive)!

Definition 1.3: BH test

Step 1. Order your p-values, so that $p_{(1)} \leq \cdots \leq p_{(m)}$. Step 2. Let $i_{\max} = \max\{i \in \{1, \ldots, m\} : p_{(i)} \leq \frac{\alpha}{m} i\}$. Reject all tests such that $p_{(i)} \leq p_{(i_{\max})}$.

Theorem 1.4

If the test statistics are independent across all the i's, then FDR$\leq \alpha$ for the BH test.

See Figure 1 for a visualization of the BH test. Note that the BH test is less stringent than Bonferroni, in which we only reject the hypotheses with p values below the horizontal line at height α/m. However, the BH test is more stringent than running the m tests each at level α, which would correspond to rejecting all hypotheses with p values below α.

Remark.

There are many variations of the BH test. For example, some of these variations allow for the possibility that the test statistics used in the individual hypothesis tests are correlated with each other.

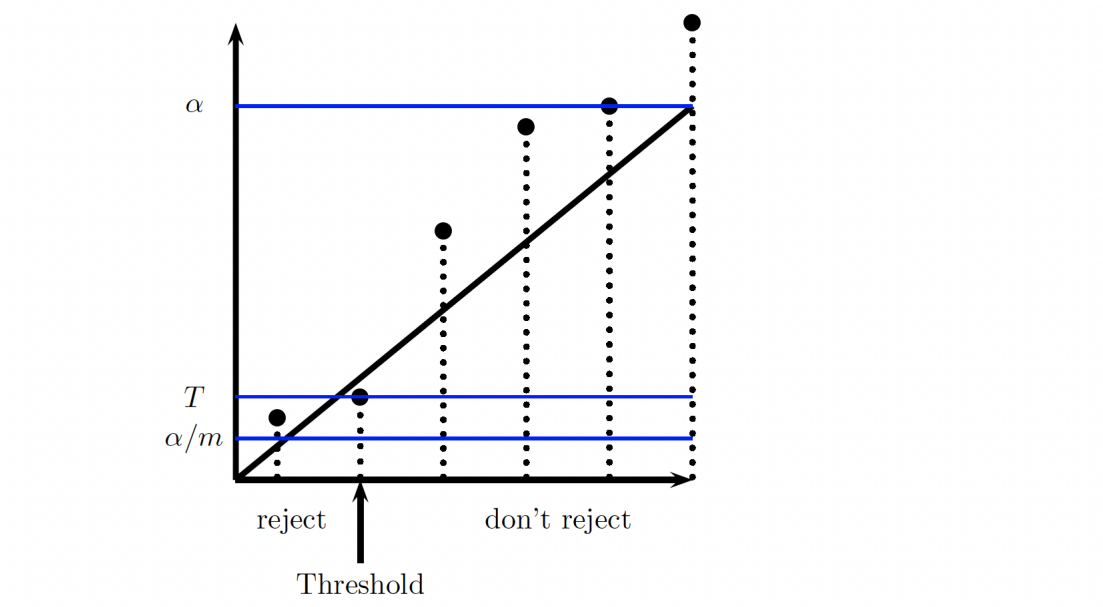

Figure 1: The BH test. We mark the indices (1), (2), ... (m) on the x-axis, and the corresponding ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ on the y-axis (represented by dots in the plot). The upward sloping line has slope α/m. We find $i_{\max}$, the largest i such that $p_{(i)} \leq i\alpha/m$, i.e. is below the upward sloping line. We then reject all hypothesis for which $p_{(i)} \leq p_{(i_{\max})}$.

2 t-test

The t-test is essentially the Wald test for small sample sizes. Recall that in the Wald's test, we used the assumption of large n to make two approximations:

1. $\sqrt{n}(\bar{X}_n - \mu)/\sigma \approx \mathcal{N}(0,1)$ by the CLT,
2. $\sigma \approx \hat{\sigma}$ by the LLN, and therefore $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma} \approx \mathcal{N}(0,1)$ by CLT + Slutsky.

In the t-test, we work with data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, so that the distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is *exactly* $\mathcal{N}(0,1)$, regardless of whether n is small or large. On the other hand, $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$ is *not* exactly $\mathcal{N}(0,1)$. But it turns out that for all n, the distribution of $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$ is precisely given by the Student's t distribution with $n-1$ degrees of freedom.

Definition 2.1: Student's t-distribution

The Student's t-distribution t_ν with ν degrees of freedom is the distribution with pdf

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

See Figure 2 for a visualization of the pdf. The Student's t has heavier tails than the Gaussian distribution, but as ν increases, the pdf of Student's t approaches the standard Gaussian pdf.

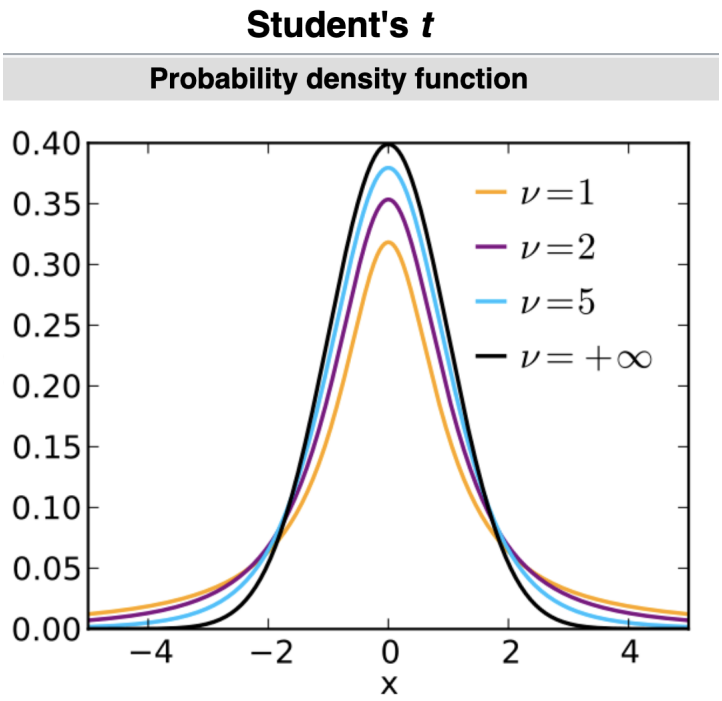

Figure 2: The pdf of the student's t distribution with ν degrees of freedom.

Theorem 2.2

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and let $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Then

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\hat{\sigma}} \sim t_{n-1}.$$

In words, $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$ is exactly distributed according to t_{n-1}, the t distribution with $n-1$ degrees of freedom.

Definition 2.3: Student's t-test

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ are unknown to us. We want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Then the Student's t-test at level α rejects the null if

$$\sqrt{n}\frac{|\bar{X}_n - \mu|}{\hat{\sigma}} \geq t_{n-1,\alpha/2},$$

where $t_{n-1,\alpha/2}$ is the $\alpha/2$ quantile of the t_{n-1} distribution.

Suppose $H_0 : \mu = 2$ and $H_1 : \mu \neq 2$, and $n = 8$, and suppose we measured that

$$\sqrt{8}\frac{\bar{X}_n - 2}{\hat{\sigma}} = 3.2.$$

Then

$$\text{p-val} = \mathbb{P}(|t_7| \geq 3.2) \approx 1.5\%$$

Note that this is larger than what we would get using the assumption of normality, in which case the p value would be $\mathbb{P}(|Z| \geq 3.2) = 0.14\%$.

Remark.

> The price we had to pay to get this exact test which works for small n is the extra assumption that the X_i are Gaussian. If this is not known, we can use the Kolmogorov-Lilliefors test to verify normality (approximately). If normality is confirmed, we can then run the t-test.