Math 150 - Methods in Biostatistics - Homework 10

Ethan Ashby

Friday, April 23, 2021

Assignment Summary (Goals)

- understanding PPV
- type I and type II errors
- adjusting p-values to control for multiple comparisons

Q1. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Annika's musical is this weekend! Good luck! Annie's in sour mood. Annie got a research thing! Lian's dad is taking tiny Kermit ornaments on his long backpacking trip to take photos of. And Lian knit little Kermit a mini-backpack!

Q2. Positive Predictive Value Using the Ioannidis paper, explain the details of PPV for the model with multiple researchers. That is, derive the entire PPV equation (you may need to derive most of Table 3).

Let $R = \frac{\#\text{true relationships}}{\#\text{null relationships}}$. Then $\mathbb{P}(\text{study is true}) = \frac{\#\text{true}}{\#\text{true}+\#\text{null}} = \frac{R}{R+1}$. Let $\alpha = \mathbb{P}(\text{T1 error}) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ True})$. Let $\beta = \mathbb{P}(\text{T2 error}) = \mathbb{P}(\text{fail to reject } H_0 | H_0 \text{ False})$. Let C be the number of tests.

Since the probability of a study containing a true result (H_A is true) is $\frac{R}{R+1}$, and there are C experiments performed, the number of studies containing a true result is $\frac{C \cdot R}{R+1}$. Since β is the probability of making a T2 error (failing to accept a true alternative hypothesis), the number of tests that commit a T2 error is $\beta \cdot \frac{C \cdot R}{R+1}$. The number of tests that do not commit a T2 error (therefore identifying true relationships) is the toatl number of true tests minus the number of tests that commit T2 errors: $\frac{C \cdot R}{R+1} - \beta \cdot \frac{C \cdot R}{R+1} = (1-\beta) \cdot \frac{C \cdot R}{R+1}$.

We also know that the probability of containing a negative result (H_0 is true) is $1 - \frac{R}{R+1} = \frac{1}{R+1}$. Since C total tests are performed, the total number of studies with negative results is $\frac{C}{R+1}$. Since α is the probability of a T1 error (falsely rejecting a true H_0), the number of studies that make a T1 error is $\alpha \cdot \frac{C}{R+1}$.

To obtain PPV, we need the number of studies that identify true relationships over the total number of studies that reach significance. The number of studies that successfuly identify true relationships (i.e. achieve signifiance and do not commit T2 errors) is $(1-\beta) \cdot \frac{C \cdot R}{R+1}$. The total number of studies that reach significance include the number of studies that successfuly identify true relationships AND the studies that commit T1 errors; summing these two gives the desired quantity, $(1-\beta) \cdot \frac{C \cdot R}{R+1} + \alpha \cdot \frac{C}{R+1} = \frac{C}{R+1}(\alpha + R(1-\beta))$. Thus, the PPV is the ratio of studies that successfully identify true relationships to the number of studies that achieve significance: $\boxed{\frac{(1-\beta) \cdot \frac{C \cdot R}{R+1}}{\frac{C}{R+1}(\alpha + R(1-\beta))}}$.

Q3. Evaluating type I errors Consider the following claim:

> If a null hypothesis is NOT rejected in multiple studies, then we have good evidence that the null is likely to be true.

Five datasets are created to study the same phenomenon. For each dataset there is a treatment group and a control group. Do the two groups differ, on average, with respect to the continuous response variable? There are pairs of columns representing each of the 5 studies. Your task is to:

1. Ascertain whether the response variable is different across the treatment and control. Look at p-values and confidence intervals.
2. Respond to the claim above.

You might consider R code like the following:

```
type1data <- read_delim("https://pomona.box.com/shared/static/acsduo30e4yzrc05mi2nnr56vj8qn3pb",
                        delim= "\t")
cbind("Test"=c("test1", "test2", "test3", "test4", "test5"),
rbind(t.test(resp1~group1, data=type1data) %>% tidy(),
t.test(resp2~group2, data=type1data) %>% tidy(),
t.test(resp3~group3, data=type1data) %>% tidy(),
t.test(resp4~group4, data=type1data) %>% tidy(),
t.test(resp5~group5, data=type1data) %>% tidy())) %>% kable()

response <- c(type1data$resp1, type1data$resp2, type1data$resp3, type1data$resp4, type1data$resp5)
group <- c(type1data$group1, type1data$group2, type1data$group3, type1data$group4, type1data$group5)
cbind("Test"="Pooled", t.test(response~group) %>% tidy()) %>% kable()
```

Solution:

I challenge the claim that a failure to reject a null hypothesis across multiple studies is tantamount to the null being true. For instance, in the above 5 studies, the result of the Welch Two Sample t-test for each study is non-significant: no study produced a p-value (much less an adjusted p-value) below 0.05. However, when the data from each of the 5 studies was combined and a single Welch's Two Sample t-test was run on the combined data, there was a significant difference observed between the two groups (p-val=0.0037).

The idea of combining studies to increase power and uncover latent true relationships underpins the whole field of metaanalysis. The first 5 studies were likely underpowered to detect a significant difference in group means, so the pooled analysis was required to obtain sufficient power to detect the difference.

Q4. Evaluating type II errors. Consider a large randomized controlled trial designed to investigate problem drinking in Australian university students (Kypri et al., *Randomized controlled trial of proactive web-based alcohol screening and brief intervention for university students.*, 2009). They specified 7 outcomes in advance, 3 were primary and 4 were secondary. No adjustments for multiple comparisons were made, and the p-values were reported to be 0.001, 0.02, 0.001 (primary endpoints), 0.59, 0.87, 0.22, 0.001 (secondary endpoints).

a. Adjust the p-values using Bonferroni, Holm, and Benjamini-Hochberg. Do all 3 methods give the same conclusions with respect to significance? Explain.

Adjustment using **Bonferroni** method: $p_{\text{Bonferonni}} = \min(p \cdot m, 1)$ where m is the number of tests. Thus, the adjusted bonferroni p-values are like so: $\boxed{p_{\text{Bonferonni}} = 0.007, 0.14, 0.007, 1, 1, 1, 0.007}$.

Adjustment using **Holm** method: $p_{\text{Holm}} = \max_{i \leq j}[\min((m - i + 1)p_i, 1)]$ where m is the number of tests. We begin by ranking the p-values in increasing order: 0.001, 0.001, 0.001, 0.02, 0.22, 0.59, 0.87. $p_{\text{Holm}} = \{0.007, \max(0.007, 6 \cdot 0.007) = 0.007, \max(0.007, 5 \cdot 0.007) = 0.007, \max(0.007, 4 \cdot 0.02) = 0.08, \max(0.08, 3 \cdot 0.22) = 0.66, \max(0.66, \min(2 \cdot 0.59, 1) = 1, 1\}$. So the final list of p-values, returned to original order is $\boxed{p_{\text{Holm}} = 0.007, 0.08, 0.007, 1, 1, 0.66, 0.007}$.

Adjustment using **Benjamini-Hochberg** method: $p_{\text{B-H}} = \min[\frac{m}{j} \cdot p_j, \tilde{p}_{j+1}]$. We begin by ordering the p-values in ascending order: 0.001, 0.001, 0.001, 0.02, 0.22, 0.59, 0.87. The $p_{\text{B-H}}$ are $\{\min(\frac{7}{1} \cdot 0.001, \tilde{p}_2), \min(\frac{7}{2} \cdot p_2, \tilde{p}_3), \min(\frac{7}{3} \cdot p_3, \tilde{p}_4), \min(\frac{7}{4} \cdot p_4, \tilde{p}_5), \min(\frac{7}{5} \cdot p_5, \tilde{p}_6), \min(\frac{7}{6} \cdot p_6, \tilde{p}_7), p_7\}$, which yielded $0.002, 0.002, 0.002, 0.035, 0.31, 0.89, 0.87$. Returning the p-values to their original order $\boxed{p_{\text{B-H}} = 0.002, 0.035, 0.002, 0.68, 0.87, 0.31, 0.002}$.

All 3 methods did not yield the same conclusions with respect to significance. Bonferroni and Holm both identified that the first and third primary endpoints and the fourth secondary endpoints were significant. Benjamini-Hochberg identified all three primary endpoints as significant as well as the fourth secondary endpoint.

b. Note that the Bonferroni and Holm adjusted p-values report the smallest familywise error under which each of the tests would reject the null hypothesis. Benjamini-Hochberg report the experiment wide FDR if all tests below a critical value are rejected. Explain why some of the adjusted p-values are repeated for Holm and BH.

Some adjusted p-values are repeated for Holm and BH because both methods involve taking the max (in the case of Holm) or min (in the case of BH) of neighboring adjusted p-values in the ordered list. If we're in a part of the p-value list where the p-values don't change much (i.e., all the p-values are very large or very small), then the adjusted p-values tend to get assigned the same values.

c. Explain how adding 100 more null tests would change each of the adjusted p-values (and corresponding conclusions).

For **Bonferroni**, adding 100 more null tests would force all the adjusted p-values to get approximately 100× larger, which would render none of the adjusted p-values significant. So none of the endpoints would be significant.

For **Holm**, adding 100 more null tests would do much the same: the m term in the formula would increase by 100. The first p-value (likely 0.001) would be adjusted to $\min(107 \cdot p_i, 1) = 0.107$ and every preceding adjusted p-value would be great than this number (because of the max term in the Holm formula). So none of the endpoints would achieve significance.

For **Benjamini-Hochberg**, adding 100 more null tests would cause many adjusted p-values to go up, but the smallest p-values (0.001) would remain significant. For example, I would anticipate 0 null p-values to be <0.001 (obtaining a p-value less than 0.001 using 100 random uniforms occurs approximately 9.5% of the time). Then, the smallest p-values would be $\frac{107}{3} \cdot 0.001 = 0.036$ (because 3 would be the highest index of 0.001 p-values in the ranked list). Suppose on average that there will be 2 null p-values ahead of 0.02 in the ranked list. Then the adjusted p-value would at minimum be $\frac{107}{6} \cdot 0.02 = 0.357$ which would be nonsignificant. So B-H would retain the p=0.001 values as significant, and all other tests would be rendered insignificant. So only the first and third primary endpoints and the fourth secondary endpoint would be significant.

```
praise()
```

```
## [1] "You are legendary!"
```