Multiple comparisons

Patrick Breheny

April 19

Multiple comparisons

- So far in this class, I've painted a picture of research in which investigators set out with one specific hypothesis in mind, collect a random sample, then perform a hypothesis test
- Real life is a lot messier
- Investigators often test dozens of hypotheses, and don't always decide on those hypotheses before they have looked at their data
- Hypothesis tests and p-values are much harder to interpret when multiple comparisons have been made

Environmental health emergency . . .

- As an example, suppose we see five cases of a certain type of cancer in the same town
- Suppose also that the probability of seeing a single case in a town this size is 1 in 10
- If the cases arose independently (our null hypothesis), then the probability of seeing five cases in the town in a single year is $\left(\frac{1}{10}\right)^5 = .00001$
- This looks like pretty convincing evidence that chance alone is an unlikely explanation for the outbreak, and that we should look for a common cause
- This type of scenario occurs all the time, and suspicion is usually cast on a local industry and their waste disposal practices, which may be contaminating the air, ground, or water

. . . or coincidence?

- But there are a lot of towns and a lot of types of cancer
- Suppose we were to carry out this kind of investigation for 10,000 different towns and 100 different diseases
- Then we would expect $(10,000)(100)(.00001) = 10$ of these tests to have p-values below .00001 just by random chance
- As a result, further investigations by epidemiologists and other public health officials rarely succeed in finding a common cause in these sorts of situations
- The lesson: if you keep testing null hypotheses, sooner or later, you'll find "significant" differences regardless of whether or not one exists

Other examples

The issue of multiple testing comes up a lot – for example,

- Subgroup analyses: separate analyses of the subjects by sex or by age group or patients with severe disease/mild disease
- Multiple outcomes: we might collect data on whether the patients died, how long the patients were in the intensive care unit, how long they required mechanical ventilation, how many days they required treatment with vasopressors, etc.
- Multiple risk factors for a single outcome

Breast cancer study

- If an investigator begins with a clear set of hypotheses in mind, however, then there are methods for carrying out tests while adjusting for multiple comparisons
- For example, consider a study of hereditary breast cancer carried out at the National Institutes of Health
- Many cases of hereditary breast cancer are due to mutations in either the BRCA1 or the BRCA2 gene; in this study, the goal of the researchers was to compare gene expression profiles in these two types of tumors
- They looked at 3,226 genes, carrying out a two-sample t-test for each gene to see if the expression level of the gene differed between BRCA1 tumors and BRCA2 tumors (*i.e.*, they got 3,226 p-values)

Probability of a single mistake

- If we accepted $p < .05$ as convincing evidence, what is the probability that we would reject at least one null hypothesis, even if all null hypotheses are true (assuming the tests are independent)?

$$\begin{aligned} P(\text{At least one rejection}) &= 1 - P(\text{No rejections}) \\ &\approx 1 - .95^{3,226} \\ &> 0.9999999999999999 \end{aligned}$$

- Once again, if we test a large number of null hypotheses, we are guaranteed to find "significant" results, even if no real differences exist

The Bonferroni correction

- All this suggests that we should modify our idea of significance in light of multiple testing
- If we want to keep our overall probability of making a type I error at 5%, we must require that the p-value of an individual test is much lower than 5%
- In particular, consider testing each individual hypothesis by comparing our p-values to a new, lower value α^*, where

$$\alpha^* = \frac{\alpha}{h},$$

and h is the number of hypothesis tests that we are conducting

The Bonferroni correction (cont'd)

- This approach to multiple comparison adjustment is called the *Bonferroni correction*
- The appeal of the Bonferroni correction is that (a) it is very simple to implement and (b) it is easy to show that:

$$P(\text{Reject any true null hypothesis}) < \alpha,$$

regardless of any dependencies among the tests

Bonferroni correction applied to the breast cancer study

- For the breast cancer study, $\alpha^* = 0.05/3226 = 0.000015$
- Note that it is still possible to find significant evidence of a gene-cancer association, but we require much more evidence to be convincing in light of the multiple testing issue
- In the breast cancer study:
  - 545 genes had p-values below 0.05
  - 4 genes had p-values below 0.000015

Family-wise error rates

- The probability we considered on the previous slide,

$$P(\text{Reject any true null hypothesis}),$$

  is referred to as the *family-wise error rate*, or FWER
- A variety of approaches have been proposed to control the FWER
- For example, an alternative procedure is to set

$$\alpha^* = 1 - (1 - \alpha)^{1/h};$$

  this is known as the *Šidák procedure*, and controls the FWER under the assumption that the tests are independent (slide 6 shows the basic idea)

False discovery rate

- An alternative strategy for dealing with multiple testing is to estimate *false discovery rates*
- Instead of trying to control the overall probability of a type I error, the false discovery rate estimates the proportion of significant findings that are type I errors
- If a cutoff of α for the individual hypothesis tests results in s significant findings, then the false discovery rate is:

$$\text{FDR} = \frac{h\alpha}{s}$$

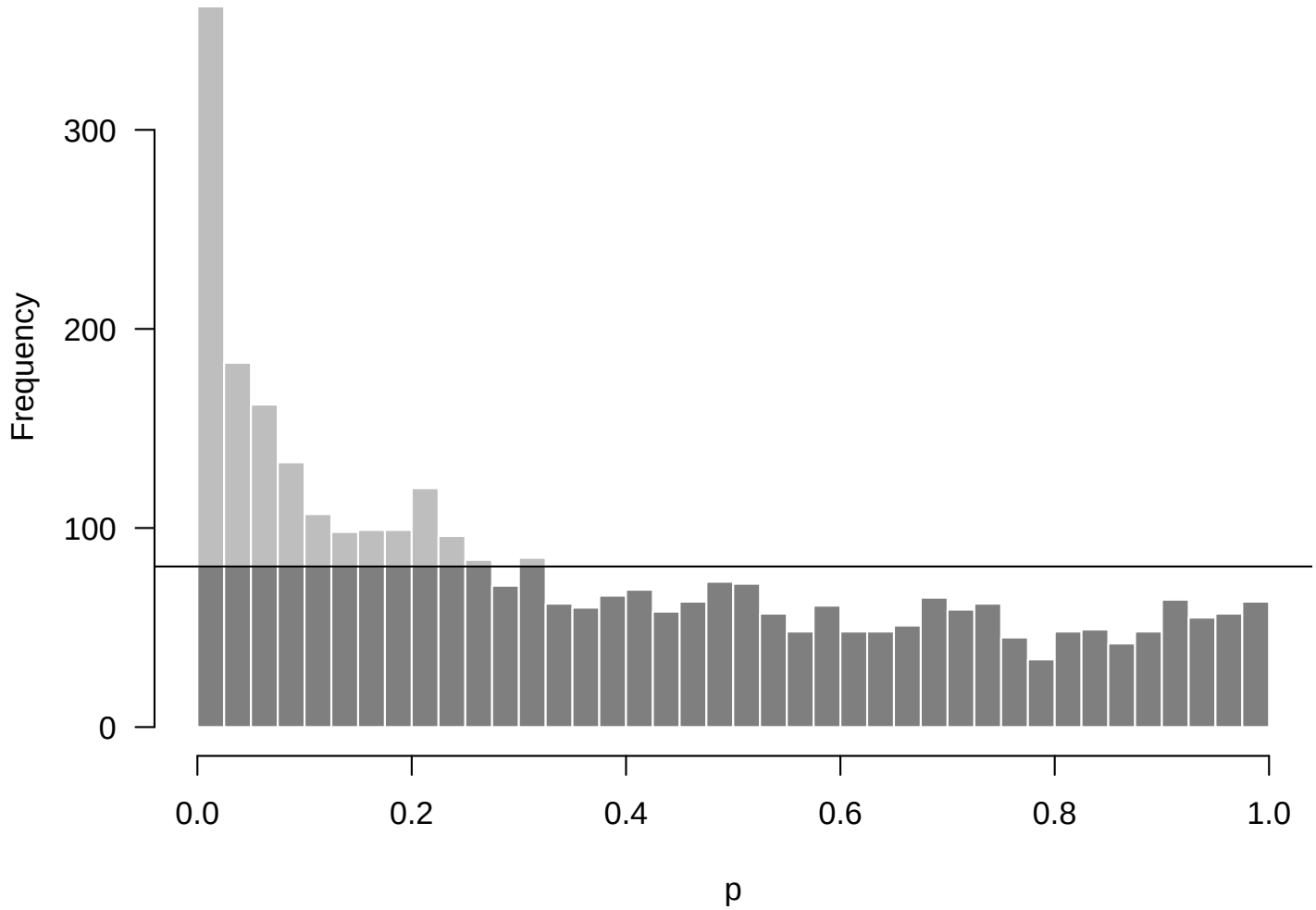
False discovery rate: breast cancer study

- So for example, in the breast cancer study, $p < .01$ for 207 of the hypothesis tests
- By chance, we would have expected $3226(.01) = 32.26$ significant findings by chance alone
- Thus, the false discovery rate for this p-value cutoff is

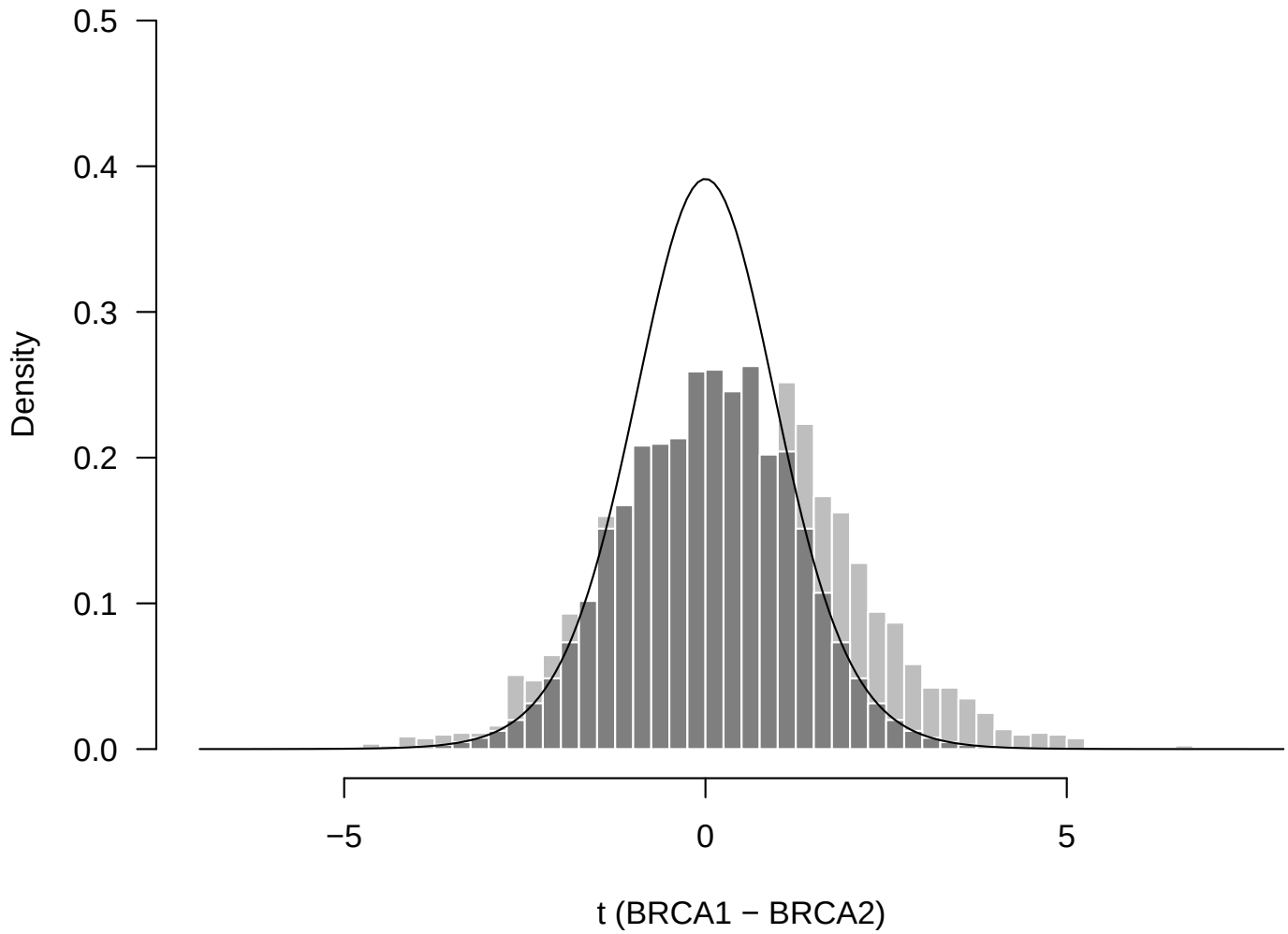
$$\text{FDR} = \frac{32.26}{207} = 15.6\%$$

- We can expect roughly 15.6% of these 207 genes to be spurious results, linked to breast cancer only by chance variability

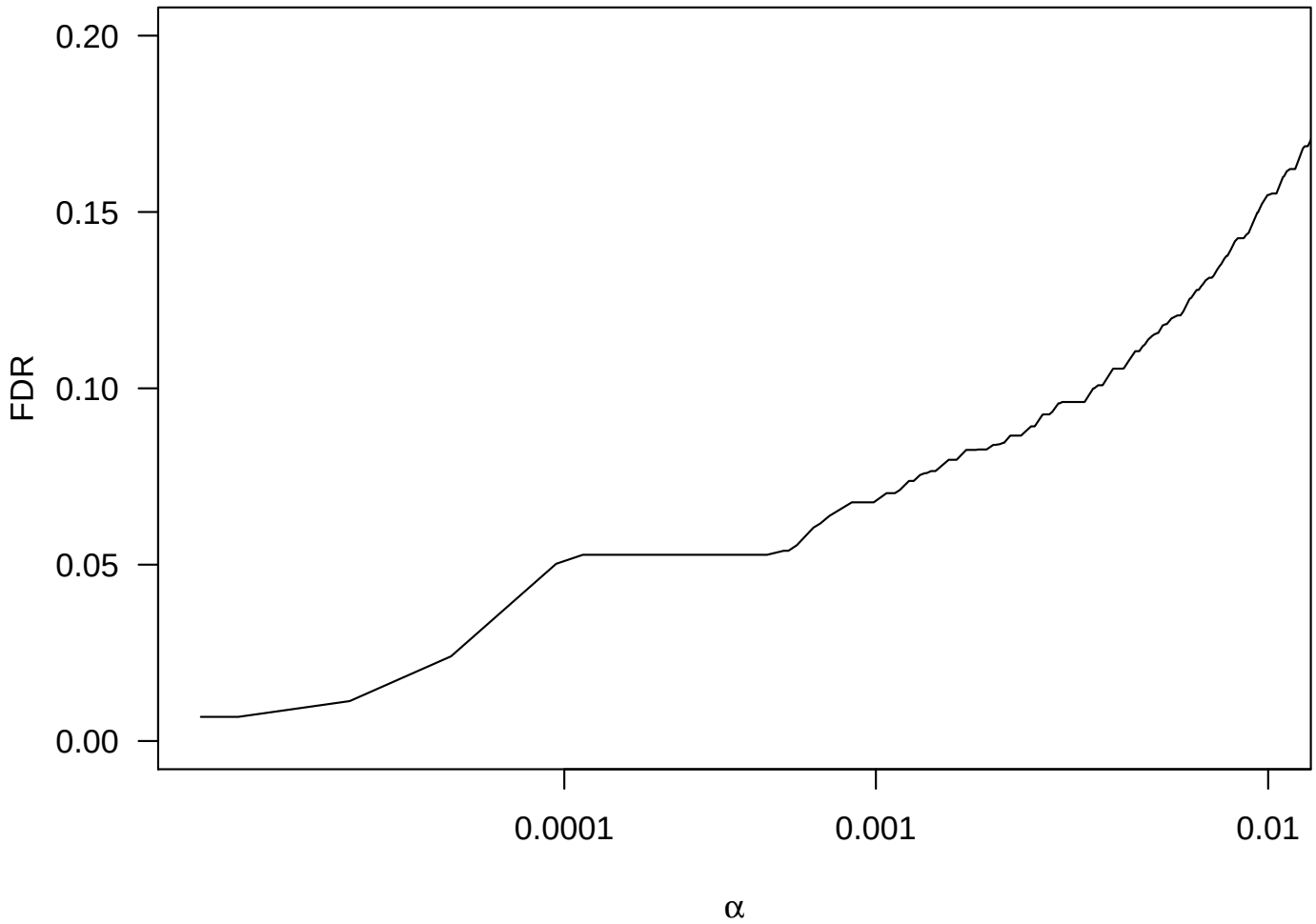
Breast cancer study: Visual idea of FDR



Breast cancer study: FDR by t values



Breast cancer study: FDR vs. α



FDR vs. FWER

- To see the difference between what FDR and FWER mean, let's compare the two at a value of 0.1:
- Bonferroni approach:
  - $0.1/3226 = 0.00003$;
  - 4 genes have p-values smaller than 0.00003
- False discovery rates:
  - 124 genes have p-values less than 0.0038
  - $3226(0.0038) = 12.4$
  - Thus, $\text{FDR} = 12.4/124 = 0.1 \implies$ we can select 124 genes with a false discovery rate of 10%

FDR vs. FWER (cont'd)

- With FWER, we want to limit the probability of making *even a single mistake*
- This is a pretty severe restriction, and we are only able to select 4 genes before our probability of committing a single type I error exceeds 10%
- On the other hand, FDR explicitly allows us to make mistakes – indeed, on the previous slide we estimate that we have made 12.4 mistakes
- The restriction is instead that these false discoveries can only make up a prespecified percentage of the total discoveries; keep in mind that this is a more liberal goal than the goal of FWER

To adjust or not to adjust?

- A common question that arises in reporting data analyses is whether tests should be adjusted for multiple comparisons or not
- Opinions differ on this matter
- Clearly, multiple comparisons matter – however, this does not necessarily imply that the author has to be the one who makes these adjustments
- A quite reasonable argument may be made that if the author reports all of the analyses they performed, then the reader can make whatever sort of multiple comparison adjustment he/she feels is appropriate
- Others argue that this approach can lead to misunderstandings if the audience is unfamiliar with multiple testing issues

Example: Drop in Childhood Obesity in Toddlers

- A good example of this issue occurred in a 2014 study published in *JAMA* looking at U.S. obesity trends
- The authors found no evidence of changes in obesity overall, but they also performed a subgroup analysis looking at changes by age group
- They looked at 6 different age groups and found that in one age group (2-5 year olds), there has been a significant reduction in obesity over the past decade ($p = 0.03$)

Example: Drop in Childhood Obesity in Toddlers (cont'd)

- The authors did not perform any explicit multiple comparison adjustments, but did informally take the multiple comparisons into account in their discussion: "Because these age subgroup analyses and tests for significance did not adjust for multiple comparisons, these results should be interpreted with caution."
- Furthermore, the article's conclusion was, "Overall, there have been no significant changes in obesity prevalence in youth or adults between 2003-2004 and 2011-2012."
- However, the "significant" drop in childhood obesity was widely picked up by media outlets, and in those news stories the reader is given no indication of the multiple testing issues or the borderline nature of how convincing this evidence was

Large-scale studies

- Of course, sometimes the number of hypotheses is simply far too large to report them all
- The breast cancer gene expression study is a good example of this – the authors are not going to publish a table containing all 3,226 p-values
- In cases like this, multiple comparison adjustments are essential and the use of FDR and FWER calculations is widespread and has become widely recognized as essential

Summary

- If you keep testing null hypotheses you are guaranteed to arrive at "significant" differences even if no real differences exist
- This problem can be avoided by accounting for the multiple tests that have been performed:
  - Bonferroni correction
  - False discovery rates
- Understand the conceptual difference between what the Bonferroni and FDR procedures are trying to accomplish with respect to multiple comparisons