Data Science Foundations of Decision Making

Multiple hypothesis testing

PURDUE UNIVERSITY | College of Science

Recall: hypothesis testing errors

	H_0 true	H_1 true
Accept H_0	Correct $1-\alpha$	Type II error β
Reject H_0	Type I error α	Correct $1-\beta$

- α = Type I error; β = Type II error

Testing more than one hypothesis

Purdue University | College of Science

Testing more than one hypothesis

- If we perform one hypothesis tests, what is the probability of a false positive? (i.e., reject the null when it is true)
  - P(Making an error) = α *(typically 0.05)*
  - P(Not making an error) = 1 - α *(1-0.05 = 0.95)*

Testing more than one hypothesis

- If we perform one hypothesis tests, what is the probability of a false positive? (i.e., reject the null when it is true)
  - P(Making an error) = α *(typically 0.05)*
  - P(Not making an error) = $1 - \alpha$ *(1-0.05 = 0.95)*
- In general, if we perform m hypothesis tests, what is the probability of at least one false positive?
  - P(Not making an error in m tests) = $(1 - \alpha)^m$

    when m=2, $0.95^2=0.90$
  - P(Making at least one error in m tests) = $1 - (1 - \alpha)^m$

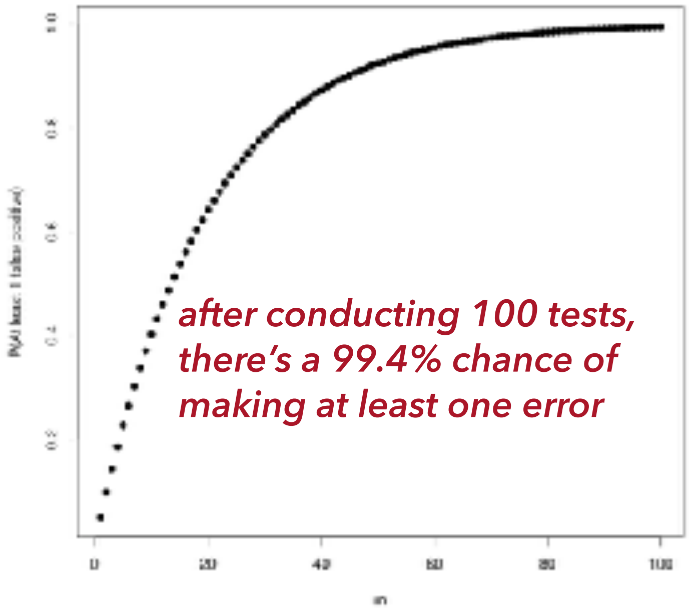
    $1-0.95^2=0.10$

Testing more than one hypothesis

- If we perform one hypothesis tests, what is the probability of a false positive? (i.e., reject the null when it is true)
  - P(Making an error) = α *(typically 0.05)*
  - P(Not making an error) = 1 - α *(1-0.05 = 0.95)*
- In general, if we perform m hypothesis tests, what is the probability of at least one false positive?
  - P(Not making an error in m tests) = $(1 - \alpha)^m$

    when m=2, $0.95^2=0.90$
  - P(Making at least one error in m tests) = $1 - (1 - \alpha)^m$

    $1-0.95^2=0.10$



after conducting 100 tests, there’s a 99.4% chance of making at least one error

The # of test conducted increase , the chance to make error↑

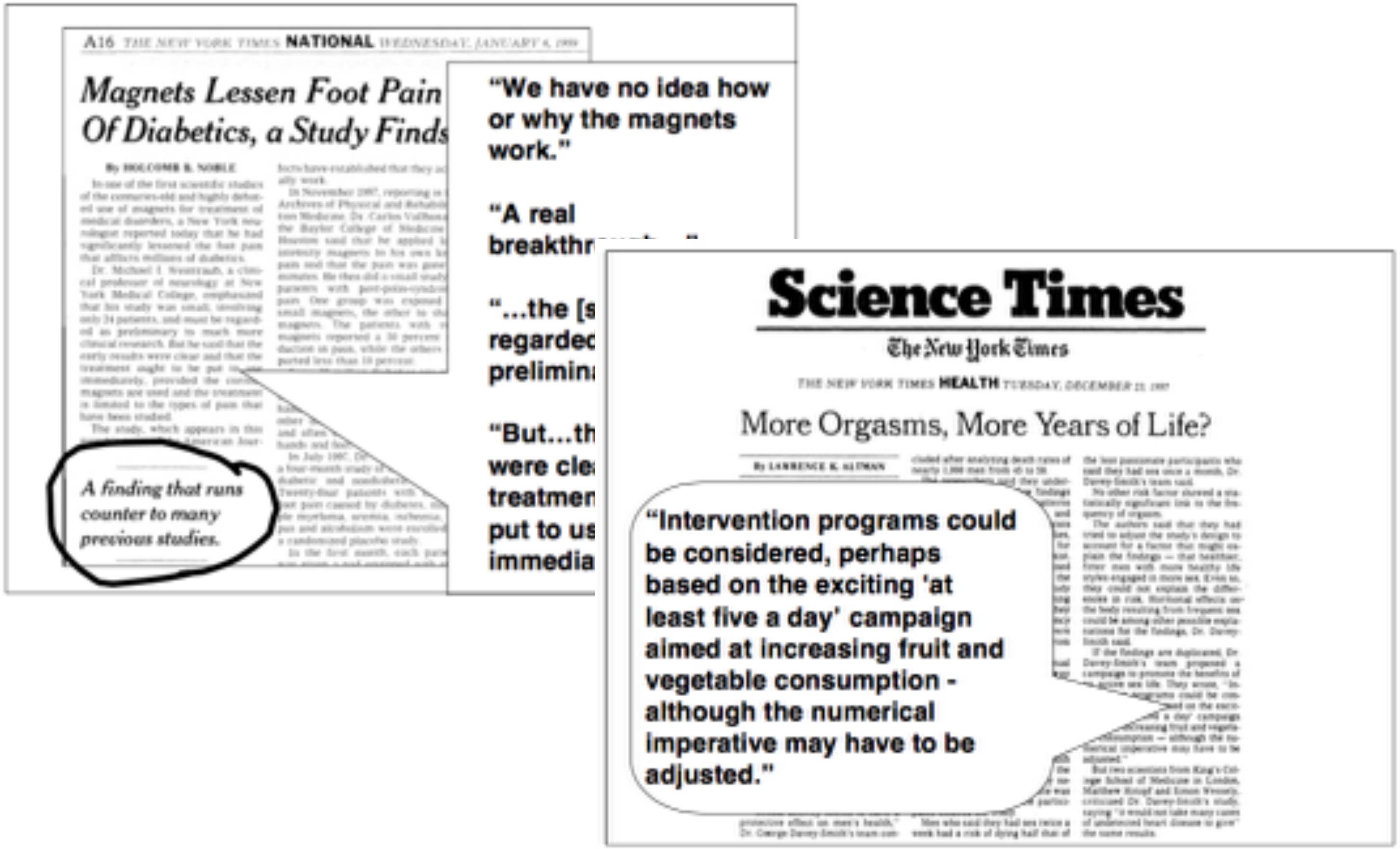
What happens when you test multiple hypotheses?

Things identified as cancer risks (Altman and Simon 1992)

- Electric razors
- Fluorescent lights
- Allergies
- Being a waiter
- Owning a pet bird
- Eating hot dogs
- Being short
- Being tall
- Having a refrigerator
- …

Things identified as cancer risks (Altman and Simon 1992)

- Electric razors
- Fluorescent lights
- Allergies
- Being a waiter
- Owning a pet bird
- Eating hot dogs
- Being short
- Being tall
- Having a refrigerator
- …



A16 THE NEW YORK TIMES **NATIONAL** WEDNESDAY, JANUARY 6, 1999

Magnets Lessen Foot Pain Of Diabetics, a Study Finds

By HOLCOMB B. NOBLE

In one of the first scientific studies of the centuries-old and highly debated use of magnets for treatment of medical disorders, a New York neurologist reported today that he had significantly lessened the foot pain that afflicts millions of diabetics.

Dr. Michael I. Weintraub, a clinical professor of neurology at New York Medical College, emphasized that his study was small, involving only 24 patients, and must be regarded as preliminary to much more clinical research. But he said that the early results were clear and that the treatment ought to be put in use immediately, provided the correct magnets are used and the treatment is limited to the types of pain that have been studied.

The study, which appears in this [illegible] American Jour-

A finding that runs counter to many previous studies.

"We have no idea how or why the magnets work."

"A real breakthr

"…the [s regarded prelimina

"But…th were clea treatmen put to us immedia

Science Times

The New York Times

THE NEW YORK TIMES **HEALTH** TUESDAY, DECEMBER 23, 1997

More Orgasms, More Years of Life?

By LAWRENCE K. ALTMAN

"Intervention programs could be considered, perhaps based on the exciting 'at least five a day' campaign aimed at increasing fruit and vegetable consumption - although the numerical imperative may have to be adjusted."

Multiple comparisons

- 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

Adjusting for multiple hypotheses

- When people say “adjusting p-values for the number of hypothesis tests performed” what they mean is controlling the Type I error rate α
- This is a very active area of statistics - many different methods have been developed
- Simple corrections adjust the level of significance to ensure that experiment-wide Type I error rates are controlled

Bonferonni correction

- Very simple method for ensuring that the overall Type I error rate of α is maintained when performing m independent hypothesis tests
- Approach: reject any hypothesis with p-value $\leq \alpha/m$
- For example, to ensure an experiment-wide Type I error rate of $\alpha=0.05$ when 10,000 hypothesis tests are performed, Bonferroni correction uses a p-value threshold of $\alpha^* = 0.05/10000 = 5 \times 10^{-6}$ to declare significance

Example

- A study published in the New England Journal of Medicine investigated whether vitamin intake affected the risk of breast cancer
- The study carried out hypothesis tests concerning:
  - vitamin C, vitamin E, and vitamin A
- And reported three separate p-values:
  - 0.67 for vitamin C, 0.07 for vitamin E, and 0.001 for vitamin A

Example

- Interpreting each p-value is straightforward, but what do they mean together?
- Suppose we set the type I error rate at $\alpha = 0.05$ for each test; what is the probability of committing at least one type I error?
- P (At least one error | All three H_0 are true) = 1 − 0.953 = 14.3%

Example

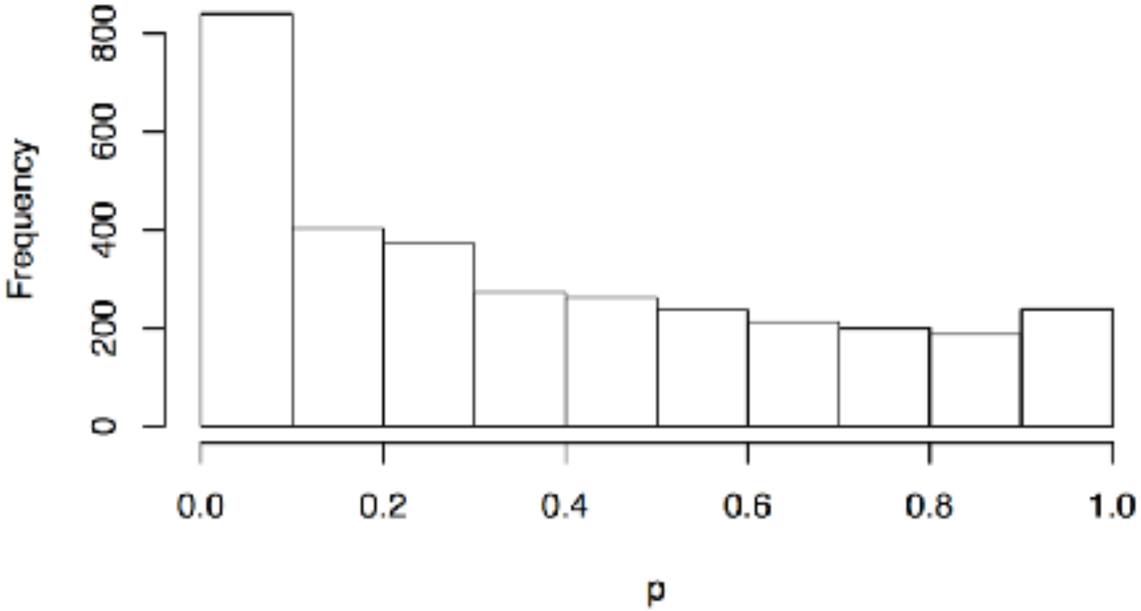
- Instead of testing each individual hypothesis at $\alpha = 0.05$, we use a Bonferroni corrected $\alpha^* = 0.05/m$
- For the vitamin study, $m = 3$ so $\alpha^* = 0.017$
- Thus, the vitamin A finding is still significant even in light of the multiple comparisons that were made (recall that its p-value was 0.001)

Bonferonni limitations

- The Bonferroni approach works well when the number of hypotheses is fairly small and making a single type I error is costly
- However, studies often test large numbers of hypotheses, expecting to find dozens of significant results, and if 3 or 4 type I errors were introduced, no great harm would be done
- Examples: AB testing, genome-wide studies

Breast cancer gene study

- For example, a landmark study by researchers at the National Institutes of Health tried to find genes associated with breast cancer
- They looked at 3,226 genes and found 207 genes with a pvalue less than α = 0.01



False discovery rate

False discovery rate

- If they had used the Bonferroni correction, they would have had to test each gene using a significance level of $\alpha^* = 1.0 \times 10^{-5}$
  - This is quite strict, only four of pvalues are below this threshold

False discovery rate

- If they had used the Bonferroni correction, they would have had to test each gene using a significance level of $\alpha^* = 1.0 \times 10^{-5}$
  - This is quite strict, only four of pvalues are below this threshold
- An alternative to the Bonferroni correction that is as strict is controlling the false discovery rate (FDR)

False discovery rate

- If they had used the Bonferroni correction, they would have had to test each gene using a significance level of $\alpha^* = 1.0 \times 10^{-5}$
  - This is quite strict, only four of pvalues are below this threshold
- An alternative to the Bonferroni correction that is as strict is controlling the false discovery rate (FDR)
- Instead of trying to control the overall probability of a type I error, the FDR controls 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 among m tests, then the false discovery rate is:

$$FDR = \frac{m\alpha}{s}$$

FDR for breast cancer gene study

FDR for breast cancer gene study

- In the breast cancer study, there were 207 p-values below α = 0.01 so

$$FDR = \frac{m\alpha}{s} = \frac{3226 \cdot 0.01}{207} = 0.156$$

FDR for breast cancer gene study

- In the breast cancer study, there were 207 p-values below α = 0.01 so

$$FDR = \frac{m\alpha}{s} = \frac{3226 \cdot 0.01}{207} = 0.156$$

- This means about 15.6% of our 207 significant findings are expected to be type I errors

FDR for breast cancer gene study

- In the breast cancer study, there were 207 p-values below α = 0.01 so

$$FDR = \frac{m\alpha}{s} = \frac{3226 \cdot 0.01}{207} = 0.156$$

- This means about 15.6% of our 207 significant findings are expected to be type I errors

$$FDR = \frac{3226 \cdot 0.0038}{122} = 0.10$$

FDR for breast cancer gene study

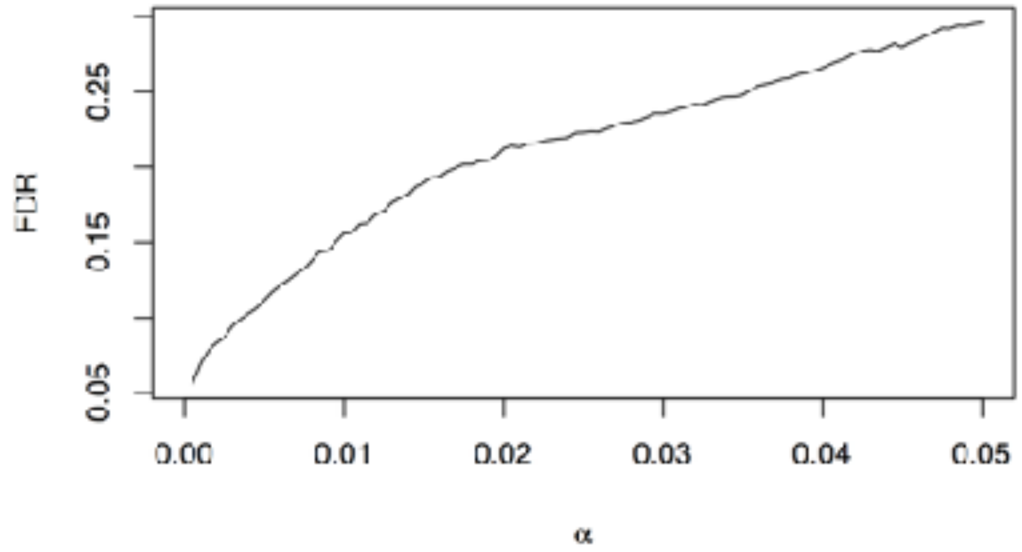
- In the breast cancer study, there were 207 p-values below α = 0.01 so

$$FDR = \frac{m\alpha}{s} = \frac{3226 \cdot 0.01}{207} = 0.156$$

- This means about 15.6% of our 207 significant findings are expected to be type I errors

- If we choose an FDR of 0.10, then we'd set the pvalue threshold to α* = 0.0038 because there are 122 genes with p-vals less than 0.0038

$$FDR = \frac{3226 \cdot 0.0038}{122} = 0.10$$



Takeaways

- When you're the investigator, you can account for multiple comparisons because you can keep track of all the comparisons/tests that are made
- Beware of other investigators that make many comparisons but only publish the few that were significant
  - Exploratory analyses can easily generate hundreds of p-values, many of which will be significant
- The FDA regulates this for clinical trials by requiring investigators to:
  - Plan all analyses before the data are collected
  - Complete and report all planned analyses