Probability and Mathematical Statistics in Data Science

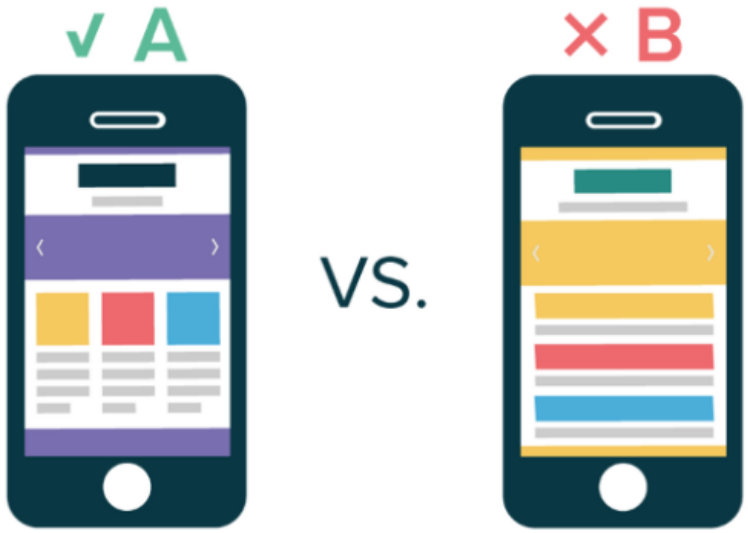
Lecture 24: Section 9.2: A/B Testing: Fisher's Exact Test

Data Science – A/B Testing
A Type of Randomized Experiment

- Let's say an online retailer like Walmart or Amazon wanted to decide which of two webpage layouts results in the most purchases for a certain product.

- For the next 200 visitors they randomly assign them to one of the two page layouts (A or B) and calculate the conversion rate – the number of customers that purchased a product.

Data Science – A/B Testing



SEGMENTING YOUR
A/B TEST RESULTS

	Layout A	Layout B
Visitors	200	200
Customers	20	15
Conversion %	10%	7.5%

https://productcoalition.com/

Data Science – A/B Testing
Null and Alternative Hypothesis

- **Null Hypothesis:** In the population, there is no difference in the conversion rate for Layout A compared to Layout B

- **Alternative Hypothesis:** In the population, there is no difference in the conversion rate for Layout A compared to Layout B

Fisher's Exact test

- Randomized controlled trial to see if botulinum toxin could help manage chronic pain
- 31 patients → 15 in treatment group, 16 in control group.
- Control group: 16 patients, 2 reported relief
- Treatment group: 15 patients, 9 reported relief
- H_0: The treatment has no effect (there would have been 11 patients reporting pain relief no matter what, and it just so happens that 9 of them were in the treatment group)
- H_1: The treatment has an effect

Fisher's Exact Test

	Treatment	Control	Total
No Relief	6	14	20
Relief	9	2	11
	15	16	31

The treatment group consists of a simple random sample of 15 of the 31 patients. Therefore under H_0, the distribution of X is hypergeometric with the following parameters:

- N = 31, the population size
- G = 11, the total number of "pain relief" patients
- n = 15, the size of the treatment group

Example: Gender bias?

- Rosen and Jerdee conducted several experiments using male bank supervisors (this was in 1974) who were given a personnel file and asked to decide whether to promote or hold the file. 24 were randomly assigned to a file labeled as that of a male employee and 24 to a female.

- 21 of the 24 males were promoted, and 14 of the females. Is there evidence of gender bias?

Example: Gender bias?

	Male	Female	Total
Promoted	21	14	35
Not Promoted	3	10	13
	24	24	48

Hypothesis Testing: Comparing Two Means

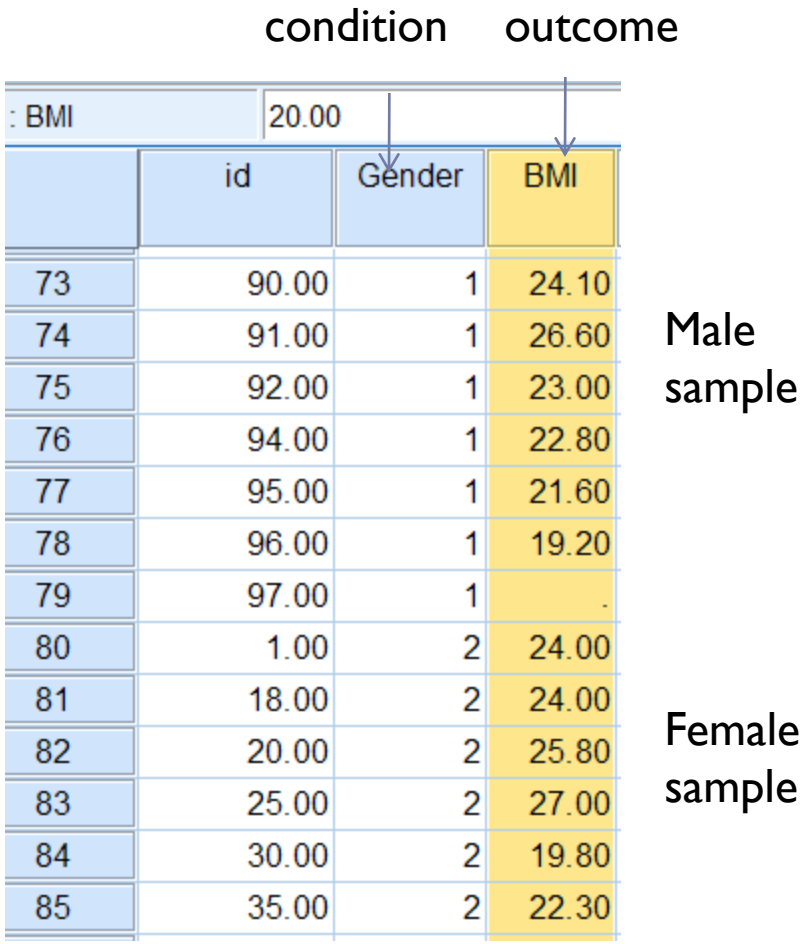
- So far the examples of hypothesis testing we have reasoned through are somewhat academic – reasoning through the steps for single group of sample data

- However, in reality the power of statistical analysis is in looking for relationships or making comparisons across (treatment) groups.

- The groups we compare do not have to be treatments. They could be simply males and females for example measured on some quantitative variable.

Example of *Independent samples*

- A quantitative variable as outcome and a binary variable as condition.
- The condition variable divides the outcome into two samples.
- Each sample represents a separate population.
- We assume two samples are *independent*.
  - The values in male sample do not affect the values in female sample.



Question: If some of the males and females are husband and wife, are the two samples independent?

Hypothesis Testing: Comparing Two Means

- In an randomized experiment patients are often randomized to two treatment groups and a quantitative variable is measured

- We compare groups by calculating the difference between the mean response across groups called the sample mean difference (often called the **sample effect size**)

- We calculate the standard error of the difference, test statistic and the p-value to decide whether the difference is real or due to chance as a result of sampling variation.

- The reasoning remains the same – only how we calculate the standard error and test statistic differs

The Standard Error of the Sample Mean Difference (or Sample Effect Size)

- The standard error measures how far on average the sample mean difference (or sample effect size) is expected to deviate from the unknown population mean difference (or population effect size). In theory, it is calculated as follows:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- In reality, it is calculated as follows:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example: Weight Loss for Diet versus Exercise program

- Let u1be the true population mean weight loss on Diet program
- Let u2 be the true population mean weight loss on Exercise program
-
- H0: There is no difference in true population mean weight loss on Diet or Exercise program
  - H0: u1=u2, or, u1-u2=0
- Ha: There is a difference in true population mean weight loss on Diet or Exercise program
  - Ha: u1 ≠ u2 or u1-u2 ≠ 0
- You decide beforehand the assignment of u1 and u2 to the particular weight loss program.
- The "sample mean difference" as an estimate of the "true mean difference" (whether a positive or negative value) should be thought of in terms the original assignment of u1 and u2 to the particular weight loss program.

Example: Weight Loss for Diet versus Exercise program

Diet Only:

sample mean = 7.2 kg
sample standard deviation = 4.1 kg
sample size = *n* = 42

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Exercise Only:

sample mean = 4.0 kg
sample standard deviation = 3.7 kg
sample size = *n* = 47

Standard Error of Difference = 0.83

Example: Weight Loss for Diet versus Exercise program

1. Determine the null and alternative hypotheses.

- ***Null hypothesis:*** No difference in average fat lost in population for two methods. Population mean difference is ***zero***.

- ***Alternative hypothesis:*** There is a difference in average fat lost in population for two methods. Population mean difference is not ***zero***.

Example: Weight Loss for Diet versus Exercise program

2. Collect and summarize data into a test statistic.

- The sample mean difference (Diet – Exercise) = 7.2 – 4.0 = 3.2 kg and the standard error of the difference is 0.83.

- **standardized score (test-statistic)**:

- = (sample mean difference – null value)/standard error of difference

standardized score (t-statistic) = *(3.2* – 0)/0.83 = 3.95

Example: Weight Loss for Diet versus Exercise program

3. Determine the *p*-value

- ***The alternative hypothesis was two-sided.***
- ***p*-value** = 2 × [proportion of bell-shaped curve above 3.95]
- ***p*-value** = 2 × [value <.00001] => p-value <0.0001

Example: Weight Loss for Diet versus Exercise program

4. Conclusion

- If there really no difference between dieting and exercise as fat loss methods, would see such an extreme result (difference in sample means) in less than 1 out of 10,000 samples of this size.

- For a 5% significance level, we reject the null hypothesis and conclude that there is a ***statistically significant difference between mean fat loss for the two methods***.