FLORIDA STATE UNIVERSITY COLLEGE OF MEDICINE

Research Workshop Series # 6
Hypothesis Testing

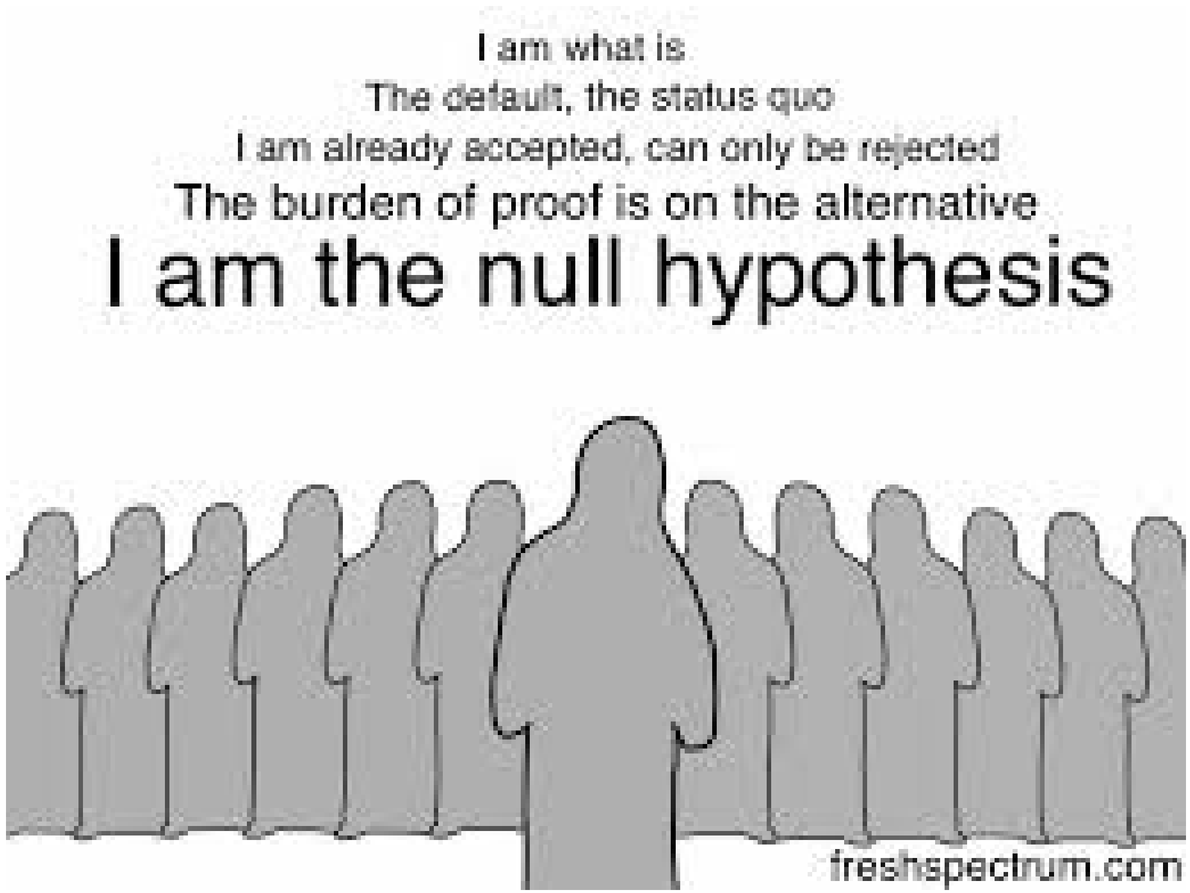
What is a hypothesis test?

Introduction

- When you think of statistical methods, you probably think of specific hypothesis tests
  - T-tests, ANOVA, etc.
- Do you know what they are and how to interpret them?
- We will cover with general concepts about hypothesis testing

Null Hypothesis

- What we assume is true at the beginning
- Evidence required to reject this



Alternative Hypotheses

- What we propose as an alternative explanation
- Usually we are hoping to be able to claim is true
  - Whether or not we can actually do this depends on the evidence from our experiment

Example: van Laar Clinical Trial

- H_0 : survival time is equal for patients randomized to the new drug versus a control regimen

- H_a : survival time is not equal for the two groups

$$H_0 : S_D(t) = S_C(t) \text{ for all } t$$
$$H_a : S_D(t) \neq S_C(t) \text{ for some } t$$

Example: Oliviera et al Denture RCT

- H_0 : mean masticatory performance is equal regardless of the denture adhesive used
- H_a : mean masticatory performance differs based on if a cream, powder, or no adhesive is used

$H_0 : \mu_c = \mu_p = \mu_{na}$

H_a : at least one of the means differs from the others

Pop Quiz

What is a p-value?

P-values

Definition:

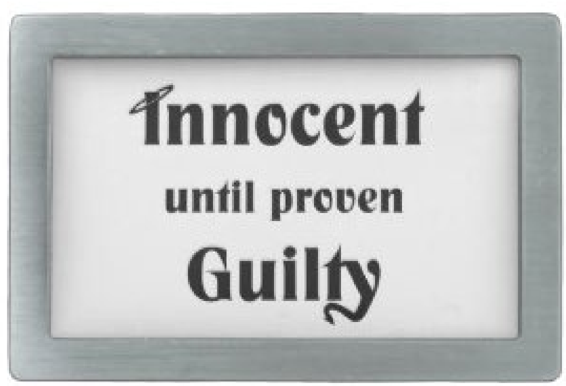
The p-value is the probability of observing data at least as extreme as the data in your experiment ***assuming that the null hypothesis is true***

P-values

- If the p-value is small, then it is unlikely to see data like ours simply as a result of chance.
- In that case, we may question whether our original assumption, the null hypothesis, was true.
- If desired, based on rules decided beforehand, we may be able to say that we ***reject the null hypothesis.***
- P-values may be reported *even if a strict binary decision is not required*.

Analogy: Trial By Jury

- Start by assuming the null.
- We need enough evidence to render the null questionable beyond a reasonable doubt.
  - P-values quantify the doubt
- Just as you never prove that someone is innocent, you never prove that the null is true.



Analogy: Trial By Jury

- You have to make a decision based on available information.
  - Assume the null hypothesis is true and require extensive evidence to reject it.

- You will never know for sure if you made the right decision.

Hypothesis Test Outcomes

Null Hypothesis	Reject	Do not reject
True	Type 1 error	Correct
False	Correct	Type II error

There are two different ways to be wrong in hypothesis testing. Each has a different consequence.

Type I Errors

- A Type 1 error is when we reject the null hypothesis when the null hypothesis is true.

- This may mean we conclude that new drugs are effective when they are really comparable to a placebo, that a gene is associated with a disease when in reality it is not, etc.

Significance Level

- The significance level, α, of a test is the probability of a type I error
- We MUST set this value before seeing any data
  - Typically, in a single experiment with only one hypothesis test, $\alpha = 0.05$ is used.
  - This tradition is arbitrary!
  - Other (usually smaller) values may be used as long as they are set ***a priori***.

Failing to Reject the Null

- If we fail to reject the null, then we simply do not have sufficient evidence to conclude make us question the null hypothesis
- We CANNOT prove that the null hypothesis is true
  - In fact it is usually false!
    - but it may be "just barely false"

Type II Errors

- A Type II error is when we fail to reject the null hypothesis when the null hypothesis is false
- This may mean we fail to conclude a new drug is effective when it may in fact be able to help patients
- Could easily happen if our sample size is too small

Trial by Jury Example

	Guilty	Not Guilty
Innocent	Type 1 error	Correct
Guilty	Correct	Type II error

Innocent until proven guilty: low type I error

Example: van Laar Clinical Trial

Either H_0 or H_a is true but we don't know which one. Let's look at both possibilities for the truth and both possible conclusions:

Example: van Laar Clinical Trial

Suppose H_0 is true

- Conclude that survival time is the same for patients on both regimens.
  - Fail to reject H_0: correct
    - We would not consider the new drug very helpful, and, for good reason, it would likely not go to market.

- Conclude that survival time differs among the groups.
  - Reject H_0: Type 1 error.
    - We might consider the new drug better for patient survival. This error could introduce an ineffective drug into the marketplace.
      - Placebo?

Example: van Laar Clinical Trial

Suppose H_a is true (and the drug was better)

- Conclude that survival time is the same for patients on both regimens
  - Fail to reject H_0 : Type II error
    - We would not consider the new drug very helpful, and it would likely not go to market.
      - Patients who could have benefited will not have access to the drug
      - Exceptions/modifications are often made for "orphan drugs"
- Conclude that survival time differs among the groups
  - Reject H_0 : Correct
    - We would consider the new drug better for patient survival.
      - The drug would have a chance of getting to market for patients

Connection between p-value and α

- If the p-value is less than α, then the test statistic is in the rejection region, and thus we reject the null hypothesis
- Otherwise, we do not reject the null hypothesis

Relationship between Hypothesis Testing and CIs

- If a $100(1 - \alpha)\%$ CI does not contain the null hypothesis parameter value, then the result must be statistically significant with $p < \alpha$

- If a $100(1 - \alpha)\%$ CI does contain the null hypothesis parameter value, then the result must not be statistically significant at the α significance level

Power

- Power is equal to the probability that we reject the null hypothesis when the null hypothesis is false.
- The probability of making a type II error is denoted by β.

$$\text{Power} = \text{P}(\text{reject } H_0 \mid H_a) = 1 - \beta$$

Factors Affecting Power

- Power increases
  - If the effect size increases
  - If the sample size increases
- Power decreases
  - If the significance level is reduced
  - If the standard deviation of the individual observations increases

Factors Affecting Sample Size

- Sample size increases as
  - the population variance increases
  - Significance level is reduced
  - Required power increases
  - Effect size decreases

Trade off between Type 1 and Type II errors

- Consider 2 extreme (unrealistic) procedures
  - Always reject the null: α will be large, $\beta = 0$
  - Never reject the null: β will be large, $\alpha = 0$
- In more practice, we do not have such drastic results, but we do observe something similar:
  - As α decreases, β increases and vice versa
- What do we do?
  - Fix α, then choose the test and sample size that minimizes β (maximizes power)

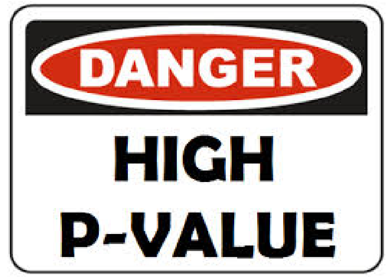
Importance of Power

- If a study is underpowered, then there is a low probability that we will reject the null hypothesis even if the null hypothesis is truly false.
  - Waste of money!
  - Unethical

Statistical Significance vs. Clinical Significance

- Statistical significance is not the same thing as clinical or practical significance
- Clinically important results can be missed, classified and statistically not significant, due to a small sample size or large variability.
- Clinically negligible can be statistically significant because the study was large.

Not Statistically Significant: what does that mean?



- It is possible that the null hypothesis really is true, but this is not guaranteed.
- It simply means that we didn't have enough evidence to reject the null hypothesis beyond a reasonable doubt.
  - We could have failed to recruit enough patients for the desired power level
  - we could have excessive variability.

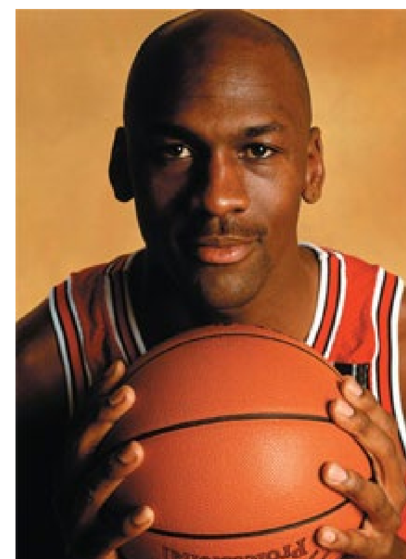
Example: non-significant result

- Consider this story from Vickers (2006a)
- A statistician shoots hoops with Michael Jordan.

  P-value = 0.07

- Should we really conclude that there is no difference in the proportion of times each one scores?

	Hits	Misses
Michel Jordan	7	0
Vickers (statistician)	3	4



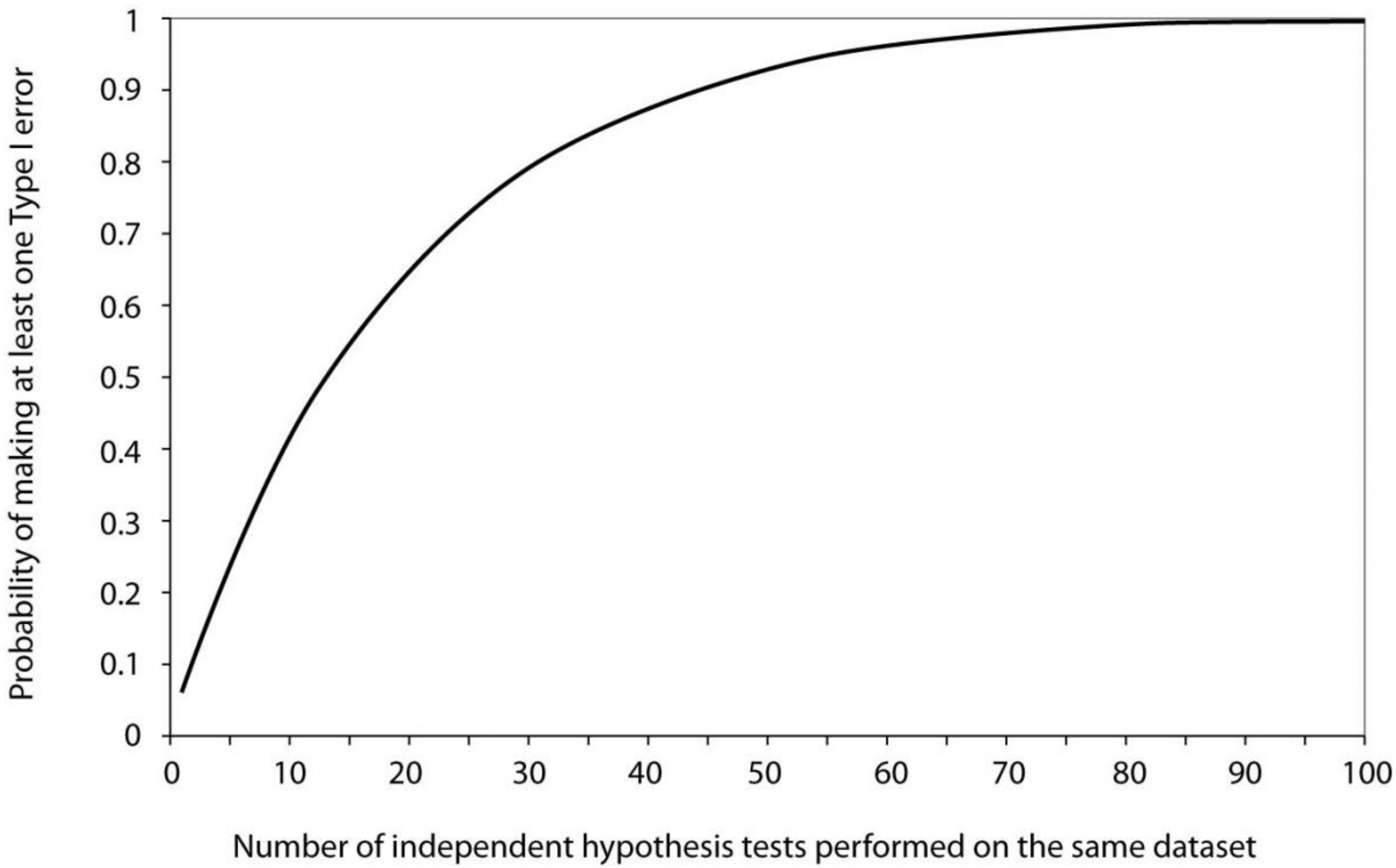
What is wrong with doing multiple hypothesis tests?

- Assume we are conducting *n* tests at 5% significance and all of the null hypotheses are true
- What is the overall type I error, called the family-wise error rate (FWER)?

P(at least one Type I error) = P(reject at least one test)

$= 1 - (1 - p)^n$

What is wrong with doing multiple hypothesis tests?



What is wrong with doing multiple hypothesis tests?

Assume we are conducting *n* tests at 5% significance and all of the null hypotheses are true

- 5 tests: we have a 22.6% chance of rejecting at least one null hypothesis (and making a type I error).
- 20 tests: we have a 66.2% chance.
- 100 tests: we have a 99.4% chance.
- 300 tests: we have a 99.99998% chance!

Ubiquity of Multiple Comparisons

- Multiple Subgroups
- Multiple Endpoints
- Multiple Research Questions
- Multiple Hypotheses

Hypothesis Testing is Frequentist

- Hypothesis testing answers questions about the probability of making a certain conclusion assuming that one of the hypotheses is true
  - P(reject $H_0 | H_0$) = α
  - P(do not reject $H_0 | H_a$) = β

Hypothesis Testing and Bayesian Inference

- Intuitively, we want to know the probability that one of the hypotheses is true based on the conclusion that we made
  - $P(H_0 \textit{ is true} \mid \text{reject } H_0) = ?$

- This requires the use of Bayesian inference

Discoveries

- Sometimes, rejecting the null hypothesis is called a *discovery*.
- We call it a discovery because we may (or may not) have discovered something scientifically interesting!
- Discoveries are valid in exploratory research. However, in confirmatory research, we need to beware of multiple comparisons.

False Discovery Rate

- Suppose you did a statistical test and rejected the null hypothesis.
- What is the probability that the rejection is a false positive?
- This question concerns FDR: the proportion of tests in which the null hypothesis is true out of all tests for which the null hypothesis is rejected.

$$FDR = \mathrm{P}(H_0 \text{ is true} \mid \text{reject } H_0)$$

- FDR is a Bayesian calculation.
- FDR is often used in studies with large numbers of tests, e.g. genomics, proteomics

Considering the FDR

- Note that in general $FDR \neq \alpha$ because

$$P(H_0 \text{ is true} \mid \text{reject } H_0) \neq P(\text{reject } H_0 \mid H_0 \text{ is true})$$

- This is called the posterior distribution.
- We can't know the truth for any individual study, so we must propose a distribution called a *prior*.
- Once we decide on a prior, we can calculate the posterior.

Considering the FDR

- We can explore the FDR in theory by considering 10,000 hypothetical studies in different scenarios.
  - For each one, we assume $\alpha = 0.05$ and $\beta = 0.2$.
  - We will look at three different priors:
    - $P(H_0) = 0.25$
    - $P(H_0) = 0.50$
    - $P(H_0) = 0.75$

FDR: $\alpha = 0.05$, $\beta = 0.2$, P(H_0 true)=0.25

- The FDR is 125/6125, or about 2%
- In this case, about 2% of rejected hypotheses will be false discoveries.

	Reject H_0	Do not reject H_0	Total
H_0 true	125	2375	2500
H_0 false	6000s	1500	7500
Total	6125	3875	10000

FDR: $\alpha = 0.05$, $\beta = 0.2$, $P(H_0 \text{ true}) = 0.5$

- The FDR is 250/4250, or about 6%
- In this case, about 65 of rejected hypotheses will be false discoveries.

	Reject H_0	**Do not reject H_0**	**Total**
H_0 true	250	4750	5000
H_0 false	4000	1000	5000
Total	4250	5750	10000

FDR: $\alpha = 0.05$, $\beta = 0.2$, $P(H_0 \text{ true}) = 0.75$

- The FDR is 375/2375, or about 16%
- In this case, about 16% of studies will be false discoveries.

	Reject H_0	**Do not reject H_0**	**Total**
H_0 true	375	7125	7500
H_0 false	2000	500	2500
Total	2375	7625	10000

Prior Distributions

- Sometimes we can use previous studies to generate reasonable prior distributions.
- Many times, we will not have previous data.
  - People often consider all possible values equally likely
    - This is called a non-informative prior
  - Results may be very sensitive to the choice of prior!

Reflections on Type 1 error, Power, and the FDR

- Always set the significance level of a test and set the sample size for desired power.

- Although the FDR is a quantity we would like to be able to calculate, doing so requires assumptions via Bayesian statistics.
  - How do you decide what prior to use?
  - Do you believe this type of inference is valid?

- Make sure you do not confuse Type I error with the false discovery rate!

More on the FDR

- Many scientists control the FDR as part of the experimental protocol. We will look at appropriate methods in a later lecture.

Proper Use of P-values

- P-values are just one piece of evidence
- ASA Statement on P-values
- Active area of research and discussion in statistics and science

Further Reading

- ASA Statement on P-values
- McShane and Gal (2017)
- https://www.tonyohagan.co.uk/academic/pdf/ExpertOpinion.pdf

References

- Vickers, Andrew J. “Michael Jordan Won't Accept the Null Hypothesis: Notes on Interpreting High P Values.” Medscape Business of Medicine: Stats for the Health Professional . May 15, 2006. Accessed August 13, 2014.

Thank you!

Questions & Discussion