Statistical Inference: Type 1 and Type 2 errors, power and sample size

Fidel A Valea, MD

Professor and Chair

Department of Obstetrics and Gynecology

Virginia Tech Carilion School of Medicine

Conflicts and Credentials

- No conflict of interests
- 3 statistics classes in College
- 1 statistics class in Med School
- 1 statistics class in Fellowship
- Taught clinical research course for ObGyns for 3 years
- I am still learning and still asking questions…

Mark Twain…

"There are three kinds of lies: lies, damned lies, and **statistics**."

Terms

- Mean
- Null and alternative hypotheses
- Standard errors
- Standard Deviation
- p-value
- Alpha error (Type 1)
- Beta error (Type 2)
- Sample Size
- Power

Confusion

Mean (Average) μ

- Easy to calculate if one could sample the entire population
- When you can't sample the entire population you estimate the mean
- The more samples you take of the population, the better one can estimate the true mean

Mean: Exercise… Height of all 5th Graders in US

- Lets say we know that true average height is 60 inches
- We sample 100 students and the mean is 62
- We sample a different 100 and the mean is 59
- The more samples you take of the population, the better one can estimate the true mean
- Why is there a difference?

2 Types of Standard Errors

- Systematic error
  - The ruler is "off" (not accurate)
  - More Northern Europeans than expected
  - A bigger sample size in these cases will not help… Stats can't help much
- Random error
  - The variation in calculated means is real
  - Individual measurements are scattered around a true mean
  - Stats is really good at determining random error

Null Hypothesis H_0

- States that there is no association between the predictor and outcome variables in the population studied
- In other words… there is no difference between the groups
- How all statistical analyses start
- Ex. association of OCP's and breast cancer: there is no association between OCP use and breast cancer

Alternative Hypothesis H_1

- There is a difference between the study groups
- Cannot be tested directly, it is accepted when the null hypothesis is rejected
- Process of exclusion
- Initial studies on OCPs and breast cancer were negative
- More recent studies support a modest 20% increase in breast cancer therefore reject the null and accept the alternative: There is an association between OCP use and the incidence of breast Ca

Why is the OCP and Breast Cancer data conflicting?

- Is it a systematic error?
  - Maybe some of the studies are "stacked" with more high or low risk patients
- Is it a random error?
  - Maybe there is no difference and we are just seeing the distribution around the mean
- Are there other potential explanations for the differing results?
- How does one decide what study to believe?
  - Reputation: author / institution / journal
  - Trial design
  - Generalizability of the study population

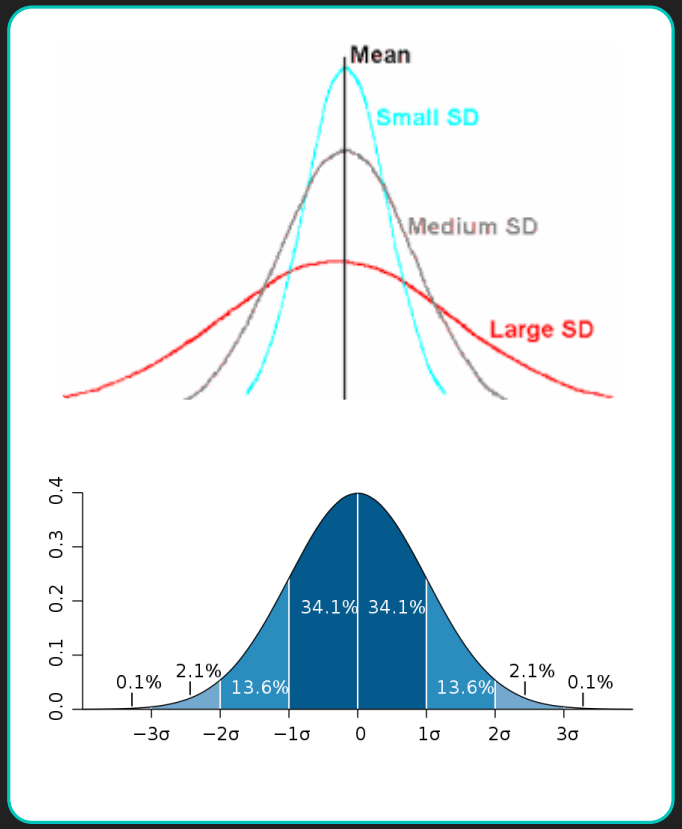
2 new concepts:

- Standard error of the means
- Standard Deviation

Standard deviation σ

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

- a measure of the amount of variation or dispersion of a set of values.
- A low standard deviation indicates that the values tend to be close to the mean.
- a high standard deviation indicates that the values are spread out over a wider range.



Standard Error of the means SE

$$\text{standard error} \quad SE = \frac{\sigma}{\sqrt{n}} \quad \text{sample size}$$

Example:

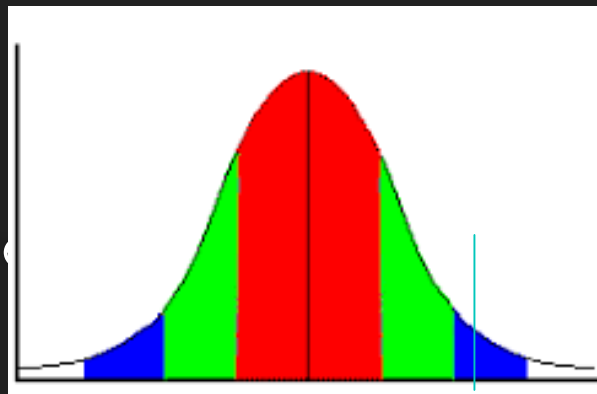
$n = 5$

$\sigma = 17$

$$= \frac{17}{\sqrt{5}}$$

$$SE = 7.6$$

- Recall the exercise of estimating the height of all 5th graders in the US
  - We took means from 50 capitol cities
  - When we plot results we are plotting means and not individual values
  - SE = Standard error or the means is the "standard deviation" for means
- It uses a Z test
- It allows you to estimate how likely a sample mean differs from the population mean
  - Lets say mean of all 5th graders is 150 cm
  - One mean in Madison was 167 cm
  - 17cm / SE 7.6 = 2.24 so we are 2.24 SE on a normal distribution curv
  - Less than a 2% chance



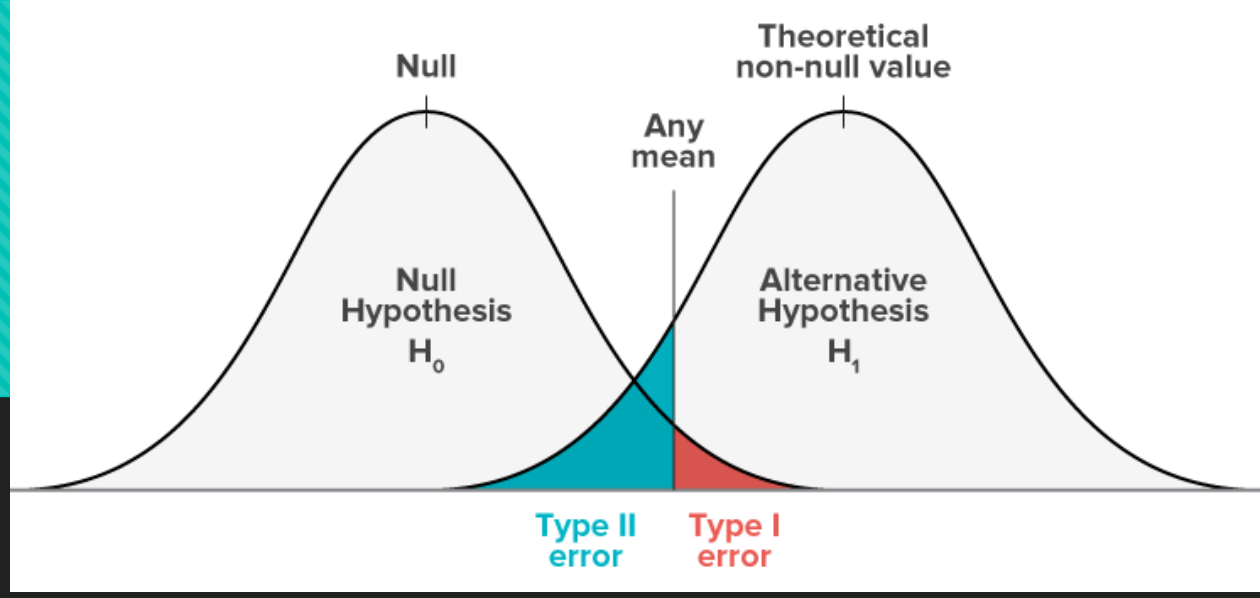
p-value

- Assumes the null hypothesis is correct
- It is the probability of obtaining a particular test result
- Compared to the results actually observed during the test
- When p-value is low… <5% the null hypothesis is rejected
- There is less than a 5% chance that the result is "similar" to the mean
- If the p-value is high (p=0.6) one accepts the null hypothesis… there is no difference between groups

What does a p-value of 0.05 mean?

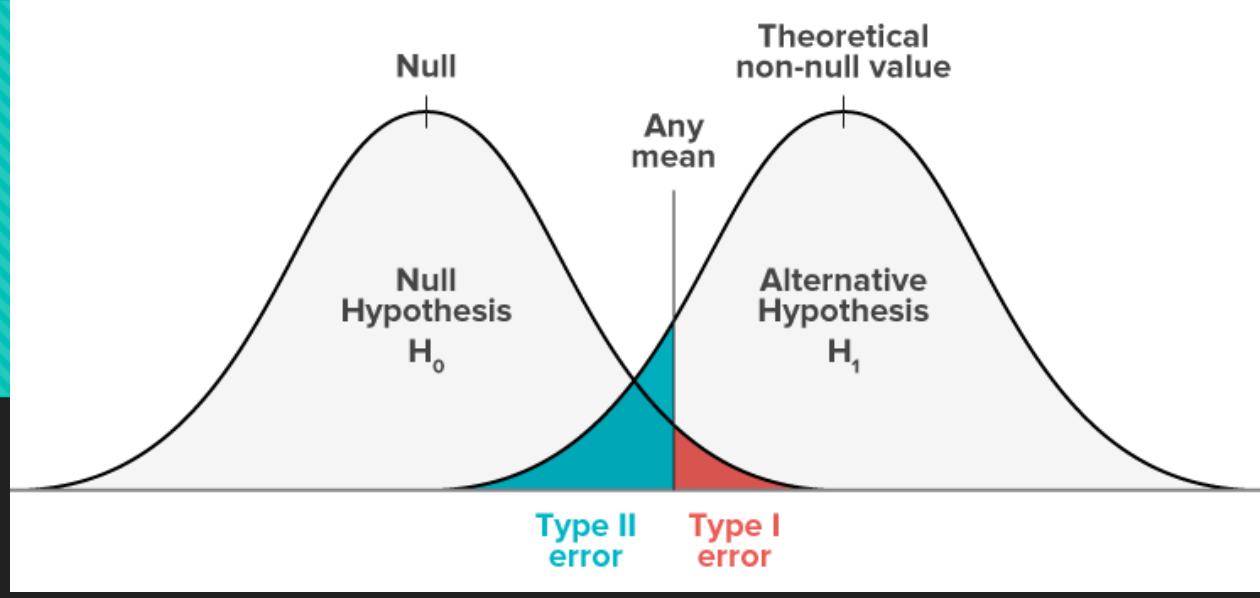
- 5% chance that there is no difference between groups
- 95% chance that there is a difference
- Remember… there is still a 5% chance (1/20) that the groups are the same
- Inherent in this principle is that 5% of "significant results" in the literature are actually not significantly different.
- Think of how this relates to the OCP and Breast Cancer studies
- Rejecting the null incorrectly is a Type-1(alpha) error

Type-1 alpha error



- Simplistically: it means incorrectly rejecting the null hypothesis
- In other words… saying there is a difference between two groups when indeed there is not
- The "red" area under the curve to the right of any mean is the alpha error
- Make sure you understand… QUESTIONS???
- This means that one in twenty significant studies are incorrect
- Remember Mark Twain

Type-2 Beta error



- The opposite of an alpha error… incorrectly accepting the null hypothesis
- In other words: saying there is no difference when there really is one
- The blue area to the left of a particular mean is the beta error
- Why does this happen?
  - Small sample size
  - Looking for very small differences between two groups
  - The study is underpowered

Let's conceptualize Beta error…

- Study: The effect of using IV antibiotics to treat PID vs observation
  - One would expect the difference to be obvious
  - Easy to tell
  - Probably will not need a lot of patients to confirm that IV antibiotics is superior to expectant management.
  - The results should be obvious and there should be a difference

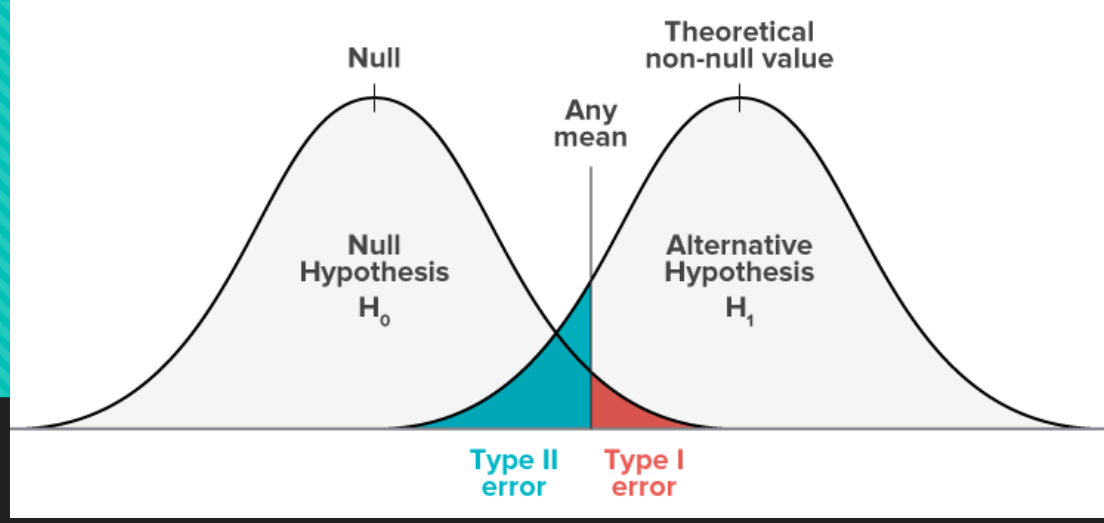
The other side of Beta error… the real side

- Study: 100 patients enrolled in RCT evaluating the use of LR vs NS for perioperative IVF looking at the incidence of hyponatremia
  - Rare event
  - Differences will likely be small
  - Would require a larger study to prove or disprove the hypothesis
  - Results: RR of using LR vs NS and the incidence of hyponatremia is 1.1, p=0.4
  - If you conclude that there is no difference you may have committed a beta error
  - Why???

Why would this be a beta error?

- We concluded that there was no difference as the p-value = 0.4
- The question that should be asked: What was the chance of finding a difference?
- What if I told you that there was only a 10% chance of finding a difference
- What if I told you that there was a 90% chance of finding a difference?
- This has to do with the concept of Power (1 – β)

What is Power? (1 – β)



- This is a concept that has to do with the ability to be able to find a difference if it is there
- Or, more pertinent to power, the confidence to say there truly is no difference if one is not found
- …Going back to the confidence in accepting the null hypothesis
- Closely related to sample size
  - Larger sample size has smaller standard errors
  - Less overlap of the two curves

Why is power important?

- Remember, statistics always start with the null hypothesis… there is no difference
- It is important to understand how many patients it will take to confidently say there is no difference… 10 patients, 100 patients, 1000 patients or more
- Why is that important?... Why do a study if it does not have a chance of giving you a reasonable answer?

Power and Study Design

- When designing a study:
  - Think of a hypothesis
  - Formulate the null and alternative hypotheses
  - Decide on study design: RCT, cohort, case-control etc
  - Set parameters: α, β, effect size, and n
  - Figure out how many subjects are needed
    - Depends on the study

Effect Size

- Very pertinent to Power
- The larger the effect size the easier it will be to detect an association
  - Ex. 90% reduction in Colon Cancer if one takes carotene supplements
  - Conversely, if you are looking for a 2% decrease in dementia if one takes 81mg ASA… it will be difficult to prove, need a lot of patients.
- In calculating sample size this entity is not known
  - It has to be estimated within reasonable clinical parameters (50% reduction is easier than 10%)

Sample Size Calculations

- State the null and alternative hypothesis
- Select the type of statistical test based on continuous or dichotomous variables
  - Ex t-test, Z-Statistic (chi-square), correlation coefficient
- Choose a reasonable effect size (variability can affect this)
- Set α and β
- Estimate incidence of the disease… very important

Sample Size Calculation

- Mostly done by computer program
- If done by hand using tables, must convert α and β into Z_{alpha} and Z_{beta}
- Look up in chart for specific statistical tests
  - Cohort and RCT use same formula
  - Case-Control uses: fraction of population at risk, Odds Ratio
  - Descriptive uses: prevalence (estimated), α and Z-value (corresponds to desired confidence level α

Final Thoughts on Sample Size/Power

- Remember… n is the number required in each arm
- Total subjects are 2n
- Once study complete and you know n, effect size, α and β…
- You can calculate Power to report in your manuscript
- Good general rules:
  - $\alpha = 0.05$
  - B = 0.2
- If you are missing a variable… look for comparable studies and estimate
- Ultimately… it is an educated guess as you have to start somewhere

Cohort or RCT

- Determine the P1 (incidence rate in the exposed)
- Determine (estimate) the p2 (incidence rate in unexposed)
- Set alpha error and look up corresponding Z_{alpha}
- Set beta error and look up corresponding Z_{beta}

WHI

- 8506 E/P 8102 Placebo
- Br Cancer: 166 vs 124 (1.95% vs 1.53%)
- What if in first 200 patients: 1 cancer in E/P and 4 cancers in placebo (Type 1 error)
- What if we sampled 200 pts: 1-2/100 in each arm… probably say no difference (Type 2 error)
- What if we sampled 2000 pts: 19-20 in E/P arm and 15-16 in Placebo… still not significantly different
- But with 16,000 patients… 24% difference, 9 extra cancers per 10,000 women

Questions