Principles of Hypothesis Testing for Public Health

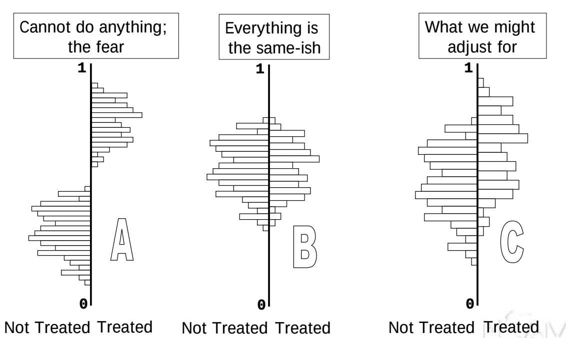
Laura Lee Johnson, Ph.D.
Statistician
National Center for Complementary and Alternative Medicine

johnslau@mail.nih.gov
Fall 2008

Questions I Usually Get

- ITT is like generalizing to real life
- I am not a fan of stratification
  - Except by clinic/site
  - Not everyone agrees with me
- OK to adjust for (some) variables
  - Baseline covariates
    - Cannot stratify a continuous variable
      - At least rarely can you do it well
  - Some variables are not ok, or you just upgraded to a fancy model!

Remember: How Much Overlap Do We Want?



Objectives

- Discuss commonly used terms
  - P-value
  - Power
  - Type I and Type II errors
- Present a few commonly used statistical tests for comparing two groups

Outline

- ➢ Estimation and Hypotheses
- How to Test Hypotheses
- Confidence Intervals
- Regression
- Error
- Diagnostic Testing
- Misconceptions

Estimation and Hypotheses

- ➢ Inference
- ➢ How we use Hypothesis Testing
- Estimation
- Distributions
- Hypothesis testing
- Sides and Tails

Statistical Inference

- Inferences about a population are made on the basis of results obtained from a sample drawn from that population
- Want to talk about the larger population from which the subjects are drawn, not the particular subjects!

You Use Hypothesis Testing

- Designing your study
- Reviewing the design of other studies
  - Grant or application review (e.g. NIH study section, IRB)
- Interpreting your study results
- Interpreting other's study results
  - Reviewing a manuscript or journal
  - Interpreting the news

I Use Hypothesis Testing

- Do all you do
- Analyze the data to find the results
  - Program formulas not presented here in detail
- You can analyze the data, too, but be careful

Analysis Follows Design

Questions → Hypotheses → Experimental Design → Samples → Data → Analyses →Conclusions

What Do We Test

- **Effect or *Difference* we are interested in**
  - **Difference in Means or Proportions**
  - **Odds Ratio (OR)**
  - **Relative Risk (RR)**
  - **Correlation Coefficient**
- **Clinically important difference**
  - **Smallest difference considered biologically or clinically relevant**
- **Medicine: usually 2 group comparison of population means**

Estimation and Hypotheses

- ✓**Inference**
- ✓**How we use Hypothesis Testing**
- ➢**Estimation**
- **Distributions**
- **Hypothesis testing**
- **Sides and Tails**

Estimation: From the Sample

- Point estimation
  - Mean
  - Median
  - Change in mean/median
- Interval estimation
  - Variation (e.g. range, σ^2, σ, $\sigma/\sqrt{n}$)
  - 95% Confidence interval

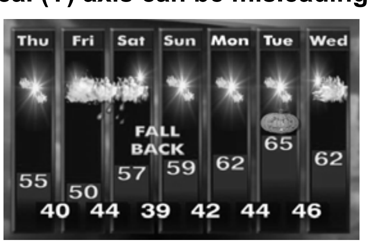
Pictures, Not Numbers

- Scatter plots
- Bar plots (use a table)
- Histograms
- Box plots

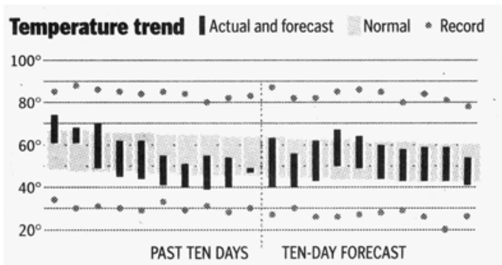
- Not Estimation
  - See the data and check assumptions

Graphs and Tables

- A picture is worth a thousand t-tests
- Vertical (Y) axis can be misleading



Like the Washington Post Weather, Though



Estimation and Hypotheses

- ✓Inference
- ✓How we use Hypothesis Testing
- ✓Estimation
- ➢Distributions
- Hypothesis testing
- Sides and Tails

Distributions

- Parametric tests are based on distributions
  - Normal Distribution (standard normal, bell curve, Z distribution)
- Non-parametric tests still have assumptions, but not based on distributions

2 of the Continuous Distributions

- Normal/Gaussian distribution: N(μ, σ^2)
  - μ = mean, σ^2 = variance
  - Z or standard normal = N(0,1)
- t distribution: t_ω
  - ω = degrees of freedom (df)
    - Usually a function of sample size
  - Mean = $\bar{X}$ (sample mean)
  - Variance = s^2 (sample variance)

Binary Distribution

- Binomial distribution: B (n, p)
  - Sample size = n
  - Proportion ‘yes’ = p
  - Mean = np
  - Variance = np(1-p)
- Can do exact or use Normal

Many More Distributions

- Not going to cover
- Poisson
- Log normal
- Gamma
- Beta
- Weibull
- Many more

Estimation and Hypotheses

✓Inference
✓How we use Hypothesis Testing
✓Estimation
✓Distributions
➢Hypothesis Testing
➢Sides and Tails

Hypothesis Testing

- Null hypothesis (H_0)
- Alternative hypothesis (H_1 or H_a)

Null Hypothesis

- Usually that there is no effect
  - Mean = 0
  - OR = 1
  - RR = 1
  - Correlation Coefficient = 0
- Generally fixed value: mean = 4
- If an equivalence trial, look at NEJM paper or other specific resources

Alternative Hypothesis

- Contradicts the null
- There *is* an effect
- What you want to prove
- If equivalence trial, special way to do this

Example Hypotheses

- H_0: $\mu_1 = \mu_2$
- H_A: $\mu_1 \neq \mu_2$
  - Two-sided test
- H_A: $\mu_1 > \mu_2$
  - One-sided test

1 vs. 2 Sided Tests

- Two-sided test
  - No *a priori* reason 1 group should have stronger effect
  - Used for most tests
- One-sided test
  - Specific interest in only one direction
  - Not scientifically relevant/interesting if reverse situation true

Use a 2-Sided Test

- Almost always
- If you use a one-sided test
  - Explain yourself
  - Penalize yourself on the alpha
    - 0.05 2-sided test becomes a 0.025 1-sided test

Take Home: Hypothesis Testing

- Null hypothesis (H_0)
- Alternative hypothesis (H_1 or H_a)
- What do you expect to happen?
- Never "accept" anything
  - Reject the null hypothesis
  - Fail to reject the null hypothesis

Outline

✓Estimation and Hypotheses
➢How to Test Hypotheses
- Confidence Intervals
- Regression
- Error
- Diagnostic Testing
- Misconceptions

Experiment

- ➢ Develop hypotheses
- ➢ Collect sample/Conduct experiment
- Calculate test statistic
- Compare test statistic with what is expected when H_0 is true

Information at Hand

- 1 or 2 sample test?
- Outcome variable
  - Binary, Categorical, Ordered, Continuous, Survival
- Population
- Numbers (e.g. mean, standard deviation)

Example: Hypertension/Cholesterol

- Mean cholesterol hypertensive men
- Mean cholesterol in male general (normotensive) population (20-74 years old)
- In the 20-74 year old male population the mean serum cholesterol is 211 mg/ml with a standard deviation of 46 mg/ml

One Sample: Cholesterol Sample Data

- Have data on 25 hypertensive men
- Mean serum cholesterol level is 220mg/ml ($\bar{X}$ = 220 mg/ml)
  - Point estimate of the mean
- Sample standard deviation: s = 38.6 mg/ml
  - Point estimate of the variance = s^2

Compare Sample to Population

- Is 25 enough?
  - Next lecture we will discuss
- What difference in cholesterol is clinically or biologically meaningful?
- Have an available sample and want to know if hypertensives are different than general population

Situation

- May be you are reading another person's work
- May be already collected data

- If you were designing up front you would calculate the sample size
  - But for now, we have 25 people

Cholesterol Hypotheses

- H_0: $\mu_1 = \mu_2$
- H_0: μ = 211 mg/ml
  - μ = POPULATION mean serum cholesterol for male hypertensives
  - Mean cholesterol for hypertensive men = mean for general male population
- H_A: $\mu_1 \neq \mu_2$
- H_A: $\mu \neq$ 211 mg/ml

Cholesterol Sample Data

- Population information (general)
  - μ = 211 mg/ml
  - σ = 46 mg/ml (σ^2 = 2116)
- Sample information (hypertensives)
  - $\bar{X}$ = 220 mg/ml
  - s = 38.6 mg/ml (s^2 = 1489.96)
  - N = 25

Experiment

✓ Develop hypotheses
✓ Collect sample/Conduct experiment
➢ Calculate test statistic
- Compare test statistic with what is expected when H_0 is true

Test Statistic

- Basic test statistic for a mean

$$\text{test statistic} = \frac{\text{point estimate of } \mu - \text{target value of } \mu}{\sigma_{\text{point estimate of } \mu}}$$

- σ = standard deviation (sometimes use σ/√n)
- For 2-sided test: Reject H_0 when the test statistic is in the upper or lower 100*α/2% of the reference distribution
- What is α?

Vocabulary

- Types of errors
  - Type I (α) (false positives)
  - Type II (β) (false negatives)
- Related words
  - Significance Level: α level
  - Power: 1- β

Unknown Truth and the Data

Truth / Data	H_0 Correct	H_A Correct
Decide H_0 "fail to reject H_0"	1- α True Negative	β False Negative
Decide H_A "reject H_0"	α False Positive	1- β True Positive

α = significance level

1- β = power

Type I Error

- α = P(reject H_0 | H_0 true)
- Probability reject the null hypothesis given the null is true
- False positive
- Probability reject that hypertensives' μ=211mg/ml when in truth the mean cholesterol for hypertensives is 211

Type II Error (or, 1- Power)

- β = P(do not reject H_0 | H_1 true)
- False Negative
- Probability we NOT reject that male hypertensives' cholesterol is that of the general population when in *truth* the mean cholesterol for hypertensives *is different* than the general male population

Power

- Power = 1-β = P(reject H_0 | H_1 true)
- Everyone wants high power, and therefore low Type II error

Cholesterol Sample Data

- N = 25
- $\bar{X}$ = 220 mg/ml
- μ = 211 mg/ml
- s = 38.6 mg/ml (s^2 = 1489.96)
- σ = 46 mg/ml (σ^2 = 2116)
- α = 0.05
- Power? Next lecture!

Z Test Statistic and N(0,1)

- Want to test continuous outcome
- Known variance
- Under H_0 $\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$
- Therefore,

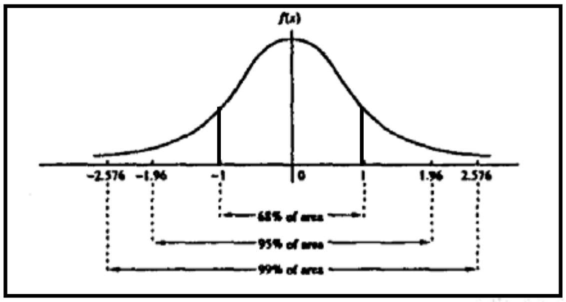
$$\text{Reject } H_0 \text{ if } \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > 1.96 \text{ (gives a 2-sided } \alpha = 0.05 \text{ test)}$$

$$\text{Reject } H_0 \text{ if } \bar{X} > \mu_0 + 1.96 \frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - 1.96 \frac{\sigma}{\sqrt{n}}$$

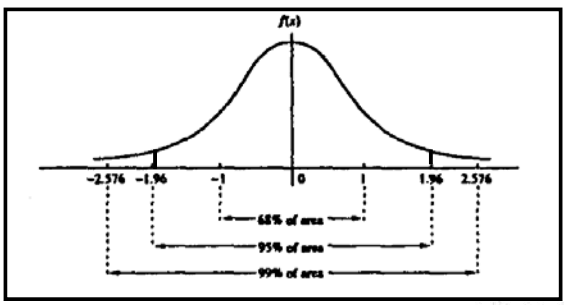
Experiment

- ✓ Develop hypotheses
- ✓ Collect sample/Conduct experiment
- ✓ Calculate test statistic
- ➢ Compare test statistic with what is expected when H_0 is true
  - Reference distribution
  - Assumptions about distribution of outcome variable

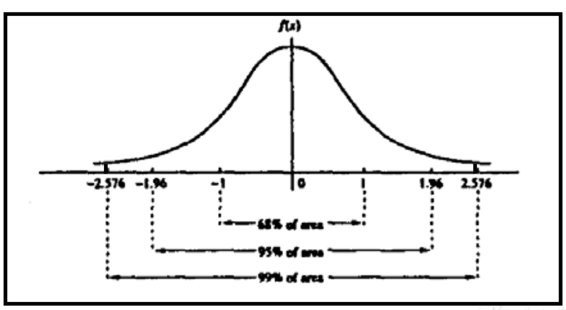
Z or Standard Normal Distribution



Z or Standard Normal Distribution



Z or Standard Normal Distribution



How to test?

- Rejection interval
  - Like a confidence interval but centered on the null mean
- Z test or Critical Value
  - N(0,1) distribution and alpha
- *t* test or Critical Value
  - *t* distribution and alpha
- P-value
- Confidence interval

General Formula (1-α)% Rejection Region for Mean Point Estimate

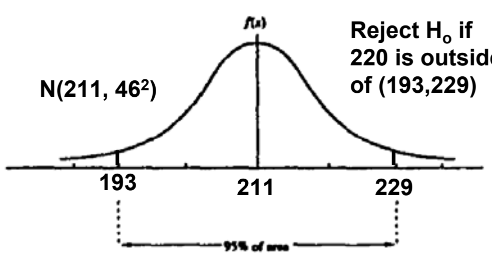
$$\left(\mu - \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}}, \mu + \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)$$

- Note that $+Z_{(\alpha/2)} = -Z_{(1-\alpha/2)}$
- 90% CI : Z = 1.645
- 95% CI : Z = 1.96
- 99% CI : Z = 2.58

Cholesterol Rejection Interval Using H_0 Population Information

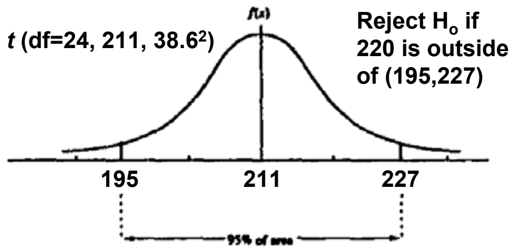
N(211, 46^2)

Reject H_o if 220 is outside of (193,229)



Normal Distribution

Cholesterol Rejection Interval Using H_0 Sample Information



t Distribution (df = 24)

Side Note on *t* vs. *Z*

- If s = σ then the *t* value will be larger than the Z value
- BUT, here our sample standard deviation (38.6) was quite a bit smaller than the population sd (46)
  - HERE intervals using *t* look smaller than Z intervals BUT
  - Because of sd, not distribution

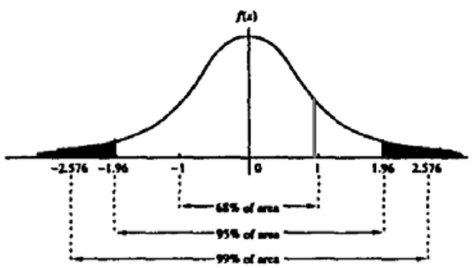
How to test?

- ✓ Rejection interval
  - Like a confidence interval but centered on the null mean
- ➢ Z test or Critical Value
  - N(0,1) distribution and alpha
- ➢ *t* test or Critical Value
  - *t* distribution and alpha
- P-value
- Confidence interval

Z-test: Do Not Reject H_0

$$|Z| = \left|\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}\right| = \left|\frac{220 - 211}{46 / \sqrt{25}}\right| = 0.98 < 1.96$$

Z or Standard Normal Distribution



Determining Statistical Significance: Critical Value Method

- **Compute the test statistic Z (0.98)**
- **Compare to the critical value**
  - **Standard Normal value at α-level (1.96)**
- **If |test statistic| > critical value**
  - **Reject H_0**
  - **Results are *statistically significant***
- **If |test statistic| < critical value**
  - **Do not reject H_0**
  - **Results are *not statistically significant***

T-Test Statistic

- Want to test continuous outcome
- Unknown variance (s, not σ)
- Under H_0 $\frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t_{(n-1)}$
- Critical values: statistics books or computer
- t-distribution approximately normal for degrees of freedom (df) >30

Cholesterol: t-statistic

- Using data
- For α = 0.05, two-sided test from t(24) distribution the critical value = 2.064
- | T | = 1.17 < 2.064
- The difference is not statistically significant at the α = 0.05 level
- Fail to reject H_0

Almost all 'Critical Value' Tests: Exact Same Idea

- Paired tests
- 2-sample tests
- Continuous data
- Binary data
- See appendix at end of slides

How to test?

- ✓ Rejection interval
  - Like a confidence interval but centered on the null mean
- ✓ Z test or Critical Value
  - N(0,1) distribution and alpha
- ✓ *t* test or Critical Value
  - *t* distribution and alpha
- ➢ P-value
- Confidence interval

P-value

- Smallest α the observed sample would reject H_0
- Given H_0 is true, probability of obtaining a result as extreme or more extreme than the actual sample
- MUST be based on a model
  - Normal, t, binomial, etc.

Cholesterol Example

- P-value for two sided test
- = 220 mg/ml, σ = 46 mg/ml
- n = 25
- H_0: μ = 211 mg/ml
- H_A: μ ≠ 211 mg/ml

Determining Statistical Significance: P-Value Method

- Compute the exact p-value (0.33)
- Compare to the predetermined α-level (0.05)
- If p-value < predetermined α-level
  - Reject H_0
  - Results are *statistically significant*
- If p-value > predetermined α-level
  - Do not reject H_0
  - Results are *not statistically significant*

P-value Interpretation Reminders

- Measure of the strength of evidence in the data that the null is not true
- A random variable whose value lies between 0 and 1
- NOT the probability that the null hypothesis is true.

How to test?

- ✓ Rejection interval
  - Like a confidence interval but centered on the null mean
- ✓ Z test or Critical Value
  - N(0,1) distribution and alpha
- ✓ *t* test or Critical Value
  - *t* distribution and alpha
- ✓ P-value
- Confidence interval

Outline

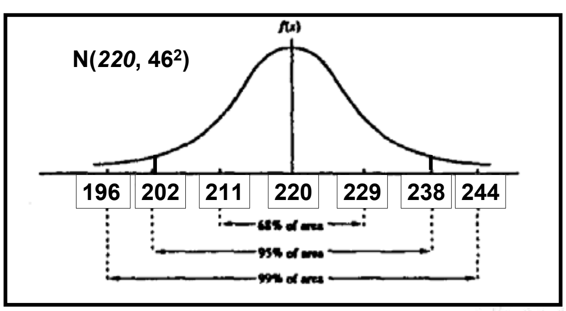
✓Estimation and Hypotheses
✓How to Test Hypotheses
➢Confidence Intervals
- Regression
- Error
- Diagnostic Testing
- Misconceptions

General Formula (1-α)% CI for μ

$$\left(\bar{X} - \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)$$

- Construct an interval around the point estimate
- Look to see if the population/null mean is inside

Cholesterol Confidence Interval Using Population Variance (*Z*)



CI for the Mean, Unknown Variance

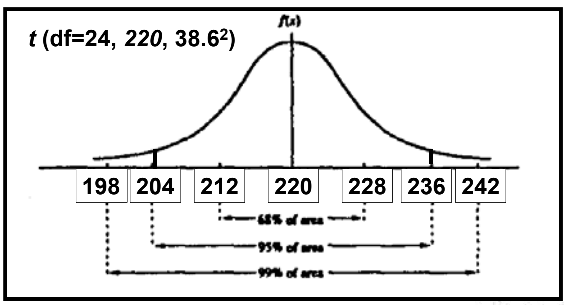
- Pretty common
- Uses the t distribution
- Degrees of freedom

$$\left(\bar{X} - \frac{t_{n-1,1-\alpha/2}s}{\sqrt{n}}, \bar{X} + \frac{t_{n-1,1-\alpha/2}s}{\sqrt{n}}\right)$$

$$= \left(220 - \frac{2.064*38.6}{\sqrt{25}}, 220 + \frac{2.064*38.6}{\sqrt{25}}\right)$$

$$= (204.06, 235.93)$$

Cholesterol Confidence Interval Using Sample Data (*t*)



But I Have All Zeros! Calculate 95% upper bound

- Known # of trials without an event (2.11 van Belle 2002, Louis 1981)
- Given no observed events in *n* trials, 95% upper bound on rate of occurrence is 3 / (*n* + 1)
  - No fatal outcomes in 20 operations
  - 95% upper bound on rate of occurrence = 3 / (20 + 1) = 0.143, so the rate of occurrence of fatalities could be as high as 14.3%

Hypothesis Testing and Confidence Intervals

- Hypothesis testing focuses on where the sample mean is located
- Confidence intervals focus on plausible values for the population mean

CI Interpretation

- Cannot determine if a particular interval does/does not contain true mean
- Can say in the long run
  - Take many samples
  - Same sample size
  - From the same population
  - 95% of similarly constructed confidence intervals will contain true mean

Interpret a 95% Confidence Interval (CI) for the population mean, μ

- "If we were to find many such intervals, each from a different random sample but in exactly the same fashion, then, in the long run, about 95% of our intervals would include the population mean, μ, and 5% would not."

Do NOT interpret a 95% CI…

- "There is a 95% probability that the true mean lies between the two confidence values we obtained from a particular sample"
- "We can say that we are 95% confident that the true mean does lie between these two values."
- Overlapping CIs do NOT imply non-significance

Take Home: Hypothesis Testing

- Many ways to test
  - Rejection interval
  - Z test, t test, or Critical Value
  - P-value
  - Confidence interval
- For this, all ways will agree
  - If not: math wrong, rounding errors
- Make sure interpret correctly

Take Home Hypothesis Testing

- How to turn questions into hypotheses
- Failing to reject the null hypothesis DOES NOT mean that the null is true
- Every test has assumptions
  - A statistician can check all the assumptions
  - If the data does not meet the assumptions there are non-parametric versions of tests (see text)

Take Home: CI

- Meaning/interpretation of the CI
- How to compute a CI for the true mean when variance is known (normal model)
- How to compute a CI for the true mean when the variance is NOT known (t distribution)

Take Home: Vocabulary

- Null Hypothesis: H_0
- Alternative Hypothesis: H_1 or H_a or H_A
- Significance Level: α level
- Acceptance/Rejection Region
- Statistically Significant
- Test Statistic
- Critical Value
- P-value, Confidence Interval

Outline

✓Estimation and Hypotheses
✓How to Test Hypotheses
✓Confidence Intervals
➢Regression
- Error
- Diagnostic Testing
- Misconceptions

Regression

- Continuous outcome
  - Linear
- Binary outcome
  - Logistic
- Many other types

Linear regression

- Model for simple linear regression
  - $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
  - β_0 = intercept
  - β_1 = slope
- Assumptions
  - Observations are independent
  - Normally distributed with constant variance
- Hypothesis testing
  - H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$

In Order of Importance

1. Independence
2. Equal variance
3. Normality

(for ANOVA and linear regression)

More Than One Covariate

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$
- *SBP* =
  $\beta_0 + \beta_1 \textit{Drug} + \beta_2 \textit{Male} + \beta_3 \textit{Age}$
- β_1
  - **Association between Drug and SBP**
  - **Average difference in SBP between the Drug and Control groups, given sex and age**

Testing?

- **Each *β* has a *p*-value associated with it**
- **Each model will have an F-test**
- **Other methods to determine fit**
  - **Residuals**
- **See a statistician and/or take a biostatistics class. Or 3.**

Repeated Measures (3 or more time points)

- **Do NOT use repeated measures AN(C)OVA**
  - **Assumptions quite stringent**
- **Talk to a statistician**
  - **Mixed model**
  - **Generalized estimating equations**
  - **Other**

An Aside: Correlation

- Range: -1 to 1
- Test is correlation is ≠ 0
- With N=1000, easy to have highly significant ($p<0.001$) correlation = 0.05
  - Statistically significant that is
  - No where CLOSE to meaningfully different from 0
- Partial Correlation Coefficient

Do Not Use Correlation. Use Regression

- Some fields: Correlation still popular
  - Partial regression coefficients
- High correlation is > 0.8 (in absolute value). Maybe 0.7
- Never believe a *p*-value from a correlation test
- Regression coefficients are more meaningful

Analysis Follows Design

Questions → Hypotheses → Experimental Design → Samples → Data → Analyses →Conclusions

Outline

✓Estimation and Hypotheses
✓How to Test Hypotheses
✓Confidence Intervals
✓Regression
➢Error
- Diagnostic Testing
- Misconceptions

Is α or β more important ?

- Depends on the question
- Most will say protect against Type I error
- Need to think about individual and population health implications and costs

Microarray / Gene Chip

- False negative (Type II error)
  - Miss what could be important
  - Are these samples going to be looked at again?
- False positive (Type I error)
  - Waste resources following dead ends

HIV Screening

- False positive
  - Needless worry
  - Stigma
- False negative
  - Thinks everything is ok
  - Continues to spread disease
- For cholesterol example?

What do you need to think about?

- Is it worse to treat those who truly are not ill or to not treat those who are ill?
- That answer will help guide you as to what amount of error you are willing to tolerate in your trial design

Outline

✓Estimation and Hypotheses
✓How to Test Hypotheses
✓Confidence Intervals
✓Regression
✓Error
➢Diagnostic Testing
• Misconceptions

Little Diagnostic Testing Lingo

- False Positive/False Negative (α, β)
- Positive Predictive Value (PPV)
  - Probability diseased given POSITIVE test result
- Negative Predictive Value (NPV)
  - Probability NOT diseased given NEGATIVE test result
- Predictive values depend on disease prevalence

Sensitivity, Specificity

- Sensitivity: how good is a test at correctly IDing people who have disease
  - Can be 100% if you say everyone is ill (all have positive result)
  - Useless test with bad Specificity
- Specificity: how good is the test at correctly IDing people who are well

Example: Western vs. ELISA

- 1 million people
- ELISA Sensitivity = 99.9%
- ELISA Specificity = 99.9%
- 1% prevalence of infection
  - 10,000 positive by Western (gold standard)
  - 9990 true positives (TP) by ELISA
  - 10 false negatives (FN) by ELISA

1% Prevalence

- 990,000 not infected
  - 989,010 True Negatives (TN)
  - 990 False Positives (FP)
- Without confirmatory test
  - Tell 990 or ~0.1% of the population they are infected when in reality they are not
  - PPV = 91%, NPV = 99.999%

1% Prevalence

- 10980 total test positive by ELISA
  - 9990 true positive
  - 990 false positive
- 9990/10980 = probability diseased GIVEN positive by ELISA = PPV = 0.91 = 91%
- 989,020 total test negatives by ELISA
  - 989,010 true negatives
  - 10 false negatives
- 989010/989020 = NPV = 99.999%

0.1% Prevalence

- 1,000 infected – ELISA picks up 999
  - 1 FN
- 999,000 not infected
  - 989,001 True Negatives (TN)
  - 999 False Positives (FP)
- Positive predictive value = 50%
- Negative predictive value = 99.999%

10% Prevalence

- 99% PPV
- 99.99% NPV

Prevalence Matters

(Population You Sample to Estimate Prevalence, too)

- Numbers look "good" with high prevalence
  - Testing at STD clinic in high risk populations
- Low prevalence means even very high sensitivity and specificity will result in middling PPV
- Calculate PPV and NPV for 0.01% prevalence found in female blood donors

Prevalence Matters

- PPV and NPV tend to come from good cohort data
- Can estimate PPV/NPV from case control studies but the formulas are hard and you need to be REALLY sure about the prevalence
  - Triple sure

A Little More Testing

High OR Does Not a Good Test Make

- Diagnostic tests need separation
- ROC curves
  - Not logistic regression with high OR
- Strong association between 2 variables does not mean good prediction of separation

What do you need to think about?

- How good does the test need to be?
  - 96% sensitivity and 10% specificity?
  - 66% AUC? (What is that?)
- Guide you as to what amount of differentiation, levels of sensitivity, specificity, PPV and NPV you are willing to tolerate in your trial design

Outline

- ✓Estimation and Hypotheses
- ✓How to Test Hypotheses
- ✓Confidence Intervals
- ✓Regression
- ✓Error
- ✓Diagnostic Testing
- ➢Misconceptions

Avoid Common Mistakes: Hypothesis Testing

- If you have paired data, use a paired test
  - If you don't then you can lose power
- If you do NOT have paired data, do NOT use a paired test
  - You can have the wrong inference

Avoid Common Mistakes: Hypothesis Testing

- These tests have assumptions of independence
  - Taking multiple samples per subject ? Statistician MUST know
  - Different statistical analyses MUST be used and they can be difficult!
- Distribution of the observations
  - Histogram of the observations
  - Highly skewed data - t test - incorrect results

Avoid Common Mistakes: Hypothesis Testing

- Assume equal variances and the variances are not equal
  - Did not show variance test
  - Not that good of a test
  - ALWAYS graph your data first to assess symmetry and variance
- Not talking to a statistician

Estimates and P-Values

- Study 1: 25±9
  - Stat sig at the 1% level
- Study 2: 10±9
  - Not statistically significant (ns)
- 25 vs. 10 wow a big difference between these studies!
  - Um, no. 15±12.7

Comparing A to B

- Appropriate
  - Statistical properties of A-B
  - Statistical properties of A/B
- NOT Appropriate
  - Statistical properties of A
  - Statistical properties of B
  - Look they are different!

Not a big difference? 15?!?

- Distribution of the difference
  - 15±12.7
  - Not statistically significant
  - Standard deviations! Important.
- Study 3 has much larger sample size!
  - 2.5±0.9

3 Studies. 3 Answers, Maybe

- Study # 3 is statistically significant
- Difference between study 3 and the other studies
  - Statistical
  - Different magnitudes
- Does study 3 replicate study 1?
- Is it all sample size?

Misconceptions

- P-value = inferential tool
  - Helps demonstrate that population means in two groups are not equal
- Smaller p-value → larger effect
  - Effect size is determined by the difference in the sample mean or proportion between 2 groups

Misconceptions

- A small p-value means the difference is statistically significant, not that the difference is clinically significant
  - A large sample size can help get a small p-value
- Failing to reject H_0
  - There is not enough evidence to reject H_0
  - Does NOT mean H_0 is true

Analysis Follows Design

Questions → Hypotheses → Experimental Design → Samples → Data → Analyses →Conclusions

Questions?

Appendix

- Formulas for Critical Values
- Layouts for how to choose a test

Do Not Reject H_0

$$220=\bar{X} > \mu_0 + 1.96\frac{\sigma}{\sqrt{n}} = 211+1.96*\frac{46}{\sqrt{25}} = 228.03 \text{ NO!}$$

$$220=\bar{X} < \mu_0 - 1.96\frac{\sigma}{\sqrt{n}} = 211\text{-}1.96*\frac{46}{\sqrt{25}} = 192.97 \text{ NO!}$$

Paired Tests: Difference Two Continuous Outcomes

- Exact same idea
- Known variance: Z test statistic
- Unknown variance: t test statistic
- H_0: $\mu_d = 0$ vs. H_A: $\mu_d \neq 0$
- Paired Z-test or Paired t-test

$$Z = \frac{\bar{d}}{\sigma / \sqrt{n}} \quad or \quad T = \frac{\bar{d}}{s / \sqrt{n}}$$

2 Samples: Same Variance

+ Sample Size Calculation Basis

- Unpaired - Same idea as paired
- Known variance: Z test statistic
- Unknown variance: t test statistic
- H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 \neq \mu_2$
  - H_0: $\mu_1 - \mu_2 = 0$ vs. H_A: $\mu_1 - \mu_2 \neq 0$
- Assume common variance

$$Z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{1/n + 1/m}} \quad or \quad T = \frac{\bar{x} - \bar{y}}{s\sqrt{1/n + 1/m}}$$

2 Sample Unpaired Tests: 2 Different Variances

- Same idea
- Known variance: Z test statistic
- Unknown variance: t test statistic
- H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 \neq \mu_2$
- H_0: $\mu_1 - \mu_2 = 0$ vs. H_A: $\mu_1 - \mu_2 \neq 0$

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}} \quad or \quad T = \frac{\bar{x} - \bar{y}}{\sqrt{s_1^2 / n + s_2^2 / m}}$$

One Sample Binary Outcomes

- Exact same idea
- For large samples
  - Use Z test statistic
  - Set up in terms of proportions, not means

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Two Population Proportions

- Exact same idea
- For large samples use Z test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$$