CFAR BIOS Core

Michael Hudgens, PhD
Core Director

Katie Mollan, MS
Biostatistician

Bios PhD Student, GRA

Stephen R. Cole, PhD
Professor of Epidemiology
CNICS Epi and Bios Core Director

Sonia Napravnik, PhD
Assistant Professor of Medicine
CFAR BIOS Core

Michael Hudgens, PhD
Core Director

Stephen R. Cole, PhD
Professor of Epidemiology
CNICS Epi and Bios Core Director

Katie Mollan, MS
Biostatistician

Joe Rigdon
Bios PhD Student, GRA

Sonia Napravnik, PhD
Assistant Professor of Medicine
An Overview of Biostatistics

• Know Your Data
• Statistical Graphics

• Statistical Inference
• Sample Size/Power
• Resources/References

• Pitfalls along the way
An Overview of Biostatistics

• Not an exhaustive list or catalog of which statistical methods to employ in various settings
• (Almost) no math/formulas
• No software training

• General principles and guidelines
• Mistakes to avoid
• Additional resources
“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”
-H.G. Wells
Moving from data to results

A) Merging databases to make analysis datasets
   • Major part of analyzing big studies & clinical trials

B) Describe the data
   • N, mean, median, Q1, Q3, standard deviation, frequency (%)
   • Review data, visualize with graphics

C) Clean & query the data as needed
   • Intersection of statistics and data management

D) Statistical analysis

E) Interpret and report results
Look at your data

• Review data records for oddities
  • Out of range values
  • Missing values
  • Pitfall: logical inconsistencies (e.g. adults getting shorter, dates out of order, dead but still on-study)

• Find out why data are missing

• Large studies
  • Review a subset of records, use frequency listings to review data combinations and missing data patterns
  • Include logic checks in your database
### Dataset Example (Rectangle, i.e. Matrix)

<table>
<thead>
<tr>
<th>ID</th>
<th>AGE</th>
<th>SEX</th>
<th>HDL</th>
<th>CHOL</th>
<th>CD4PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>M</td>
<td>59</td>
<td>178</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>M</td>
<td>65</td>
<td>220</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>M</td>
<td>45</td>
<td>213</td>
<td>187</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>F</td>
<td>55</td>
<td>195</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>M</td>
<td>60</td>
<td>187</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>M</td>
<td>38</td>
<td>205</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td></td>
<td>55</td>
<td>180</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>F</td>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>F</td>
<td>45</td>
<td>195</td>
<td>57</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>M</td>
<td>198</td>
<td>46</td>
<td>32</td>
</tr>
</tbody>
</table>

Hypothetical (not real) data set
### Dataset Example (Potential Problems)

<table>
<thead>
<tr>
<th>ID</th>
<th>AGE</th>
<th>SEX</th>
<th>HDL</th>
<th>CHOL</th>
<th>CD4PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>M</td>
<td>59</td>
<td>178</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>M</td>
<td>65</td>
<td>220</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>M</td>
<td>45</td>
<td>213</td>
<td>187</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>F</td>
<td>55</td>
<td>195</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>M</td>
<td>60</td>
<td>187</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>M</td>
<td>38</td>
<td>205</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>M</td>
<td>55</td>
<td>180</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>F</td>
<td>55</td>
<td>180</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>F</td>
<td>45</td>
<td>195</td>
<td>57</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>M</td>
<td>198</td>
<td>46</td>
<td>32</td>
</tr>
</tbody>
</table>
Corrections were made.

Now it’s looking better..

But do you notice anything odd?
The updated dataset has a duplicate row!

<table>
<thead>
<tr>
<th>ID</th>
<th>AGE</th>
<th>SEX</th>
<th>HDL</th>
<th>CHOL</th>
<th>CD4PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>M</td>
<td>59</td>
<td>178</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>M</td>
<td>65</td>
<td>220</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>M</td>
<td>45</td>
<td>213</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>F</td>
<td>55</td>
<td>195</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>F</td>
<td>55</td>
<td>195</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>M</td>
<td>60</td>
<td>187</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>M</td>
<td>38</td>
<td>205</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>F</td>
<td>55</td>
<td>180</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>F</td>
<td></td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>F</td>
<td>45</td>
<td>195</td>
<td>57</td>
</tr>
<tr>
<td>10</td>
<td>41</td>
<td>M</td>
<td>46</td>
<td>198</td>
<td>32</td>
</tr>
</tbody>
</table>
Data Example 2

Imagine this dataset continues for n=50 patient IDs...

Notice anything strange?

<table>
<thead>
<tr>
<th>ID</th>
<th>BIOMARKER</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IFNg</td>
<td>1.50</td>
</tr>
<tr>
<td>1</td>
<td>TGFa</td>
<td>2.46</td>
</tr>
<tr>
<td>1</td>
<td>IL6</td>
<td>5.87</td>
</tr>
<tr>
<td>1</td>
<td>IL7</td>
<td>1.50</td>
</tr>
<tr>
<td>1</td>
<td>IL10</td>
<td>5.12</td>
</tr>
<tr>
<td>2</td>
<td>IFNg</td>
<td>2.31</td>
</tr>
<tr>
<td>2</td>
<td>TGFa</td>
<td>10.23</td>
</tr>
<tr>
<td>2</td>
<td>IL6</td>
<td>5.15</td>
</tr>
<tr>
<td>2</td>
<td>IL7</td>
<td>1.50</td>
</tr>
<tr>
<td>2</td>
<td>IL10</td>
<td>6.24</td>
</tr>
<tr>
<td>3</td>
<td>IFNg</td>
<td>9.60</td>
</tr>
<tr>
<td>3</td>
<td>TGFa</td>
<td>1.50</td>
</tr>
<tr>
<td>3</td>
<td>IL6</td>
<td>1.50</td>
</tr>
<tr>
<td>3</td>
<td>IL7</td>
<td>8.42</td>
</tr>
<tr>
<td>3</td>
<td>IL10</td>
<td>7.46</td>
</tr>
</tbody>
</table>
Data Example 2

Some results are censored at the limit of detection.

This is important to know for choosing an analysis method

(Probably cannot assume normality)

<table>
<thead>
<tr>
<th>ID</th>
<th>BIOMARKER</th>
<th>RESULT</th>
<th>CENSOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IFNg</td>
<td>1.50</td>
<td>L</td>
</tr>
<tr>
<td>1</td>
<td>TGFa</td>
<td>2.46</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IL6</td>
<td>5.87</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IL7</td>
<td>1.50</td>
<td>L</td>
</tr>
<tr>
<td>1</td>
<td>IL10</td>
<td>5.12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IFNg</td>
<td>2.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TGFa</td>
<td>10.23</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IL6</td>
<td>5.15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IL7</td>
<td>1.50</td>
<td>L</td>
</tr>
<tr>
<td>2</td>
<td>IL10</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>IFNg</td>
<td>9.60</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>TGFa</td>
<td>1.50</td>
<td>L</td>
</tr>
<tr>
<td>3</td>
<td>IL6</td>
<td>1.50</td>
<td>L</td>
</tr>
<tr>
<td>3</td>
<td>IL7</td>
<td>8.42</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>IL10</td>
<td>7.46</td>
<td></td>
</tr>
</tbody>
</table>
Know your data summary

- Look at records from the original data, intermediate data sets, and analysis data
- Use cross-frequency listings, plots, range checks, and logic checks to “look at the data”
- Be careful with coded variables, be sure you know what 0/1 represent, 1, 2, 3.. and so on
Know your data. Pitfalls + Scandal

Misconduct in science
An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | From the print edition

ANIL POTTI, Joseph Nevins and their colleagues at Duke University in Durham, North Carolina, garnered widespread attention in 2006. They reported in the New England Journal of Medicine that they could predict the course of a patient’s lung cancer using devices called expression arrays, which log the activity patterns of thousands of genes in a sample of tissue as a colourful picture (see above). A few months later, they wrote in Nature Medicine that they had developed a similar technique which used gene expression in laboratory cultures of cancer cells, known as cell lines, to predict which chemotherapy would be most effective for an individual patient suffering from lung, breast or ovarian cancer.

How Bright Promise in Cancer Testing Fell Apart

Errors.. “Some seemed careless — moving a row or a column over by one in a giant spreadsheet — while others seemed inexplicable.” -NYTimes
Useful Reference:


STATISTICAL GRAPHICS
Pitfalls: What not to plot

- Don’t use a 3D chart unless the 3rd dimension displays another variable
- Don’t use 2 dimensions (area) to convey one-dimensional data
- Pie charts are rarely ideal, bar charts provide an axis for orientation

Friends don’t let friends plot imaginary dimensions

-Tufte, 1983 pg. 69

How to Lie with Statistics – Huff 1954
http://en.wikipedia.org/wiki/Misleading_graph
Statistical Graphics

• Show the data, even a simple plot will do
  • Scatter Plot (association)
  • Spaghetti Plot (trajectories, paired data)
  • Histogram

• Summary Statistics
  • Box plot
  • Bar chart

• Estimation and inference
  • Kaplan-Meier Curves
  • Forest Plot
Aim for high data-ink ratio

\[
data\text{-ink ratio} = \frac{\text{‘ink’ used to display the data}}{\text{total ‘ink’ used to display the graphic}}
\]

Low ratio
Data are lost in a sea of gridlines and labels

http://en.wikipedia.org/wiki/Misleading_graph
http://www.davidgiard.com/2011/05/12/DataVisualizationPart5DataInk.aspx
Aim for high data-ink ratio

\[
data-ink \text{ ratio} = \frac{\text{‘ink’ used to display the data}}{\text{total ‘ink’ used to display the graphic}}
\]

Low ratio
Data are lost in a sea of gridlines and labels

High ratio 😊
Ah.. There we go

http://en.wikipedia.org/wiki/Misleading_graph
http://www.davidgiard.com/2011/05/12/DataVisualizationPart5DataInk.aspx
Graphics rules of thumb

- Label the axes with reasonable size font
- Always start axis at zero when applicable
- Be thoughtful and fair with graph dimensions
- Use axis breaks sparingly
- Use gridlines and boxes sparingly

Manhattan plot ↯ skyscraper


Sophisticated graphic from *Science*
Scatter Plot of Pharmacy Refill Adherence Levels at 12 Months after Starting cART for Patients with and without Virologic Failure

http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0050109
Figure 1. The relationship between T-cell activation and plasma HIV RNA levels (log_{10} transformed) in 153 individuals recently diagnosed with HIV infection. A smooth line was generated by linear regression with quadratic equations.

Chomont, Nicolas, et al. "HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation."
Individual subject concentration-time profiles over the 24-h study period for African-American REAL cohort participants (n = 38)

Interquartile Range 25th to 75th percentile
Box Plot REAL Study Example

Plotted Log10 scale

C12
p=0.41

Wilcoxon Rank-Sum nonparametric test to compare 2 independent groups (QDMRK vs. REAL)

Section of Figure 2:
Many Types of Error Bars

• Descriptive
  • ± 1 standard deviation (SD)
  • Interquartile Range (IQR), 25\textsuperscript{th} to 75\textsuperscript{th} percentile
  • Range, smallest to largest

• Statistical Inference
  • ± 1 standard error (SE or SEM),
  • Confidence interval (CI), such as 95% CI

• If the type of error bar is not specified, that is a problem
Error Bar Example
REAL Study C12 for Raltegravir (n=38)

Data from Table 2:

SEM: Standard Error of Mean
Change in Y-axis Scale

Plotted log10 Scale

Not ideal here due to censored and skewed data

Use a Boxplot to show this and more in one plot

10 ng/mL is 50% of the Lower limit of quantification

Data from Table 2:
Kaplan Meier Survival Curves

Figure 2. Kaplan–Meier Estimates of the Cumulative Risk of Infant HIV-1 Infection or a Composite of HIV-1 Infection or Death by 28 Weeks.

Shown is the probability of HIV-1 infection or a composite of HIV-1 infection or death among infants who tested negative for HIV-1 infection at 2 weeks (Panels A and B) and among all infants who underwent randomization (Panels C and D) in three groups of mother–infants pairs: women who received an antiretroviral (ARV) regimen, infants who received nevirapine (NVP) prophylaxis, and control subjects. Rates were compared with the use of the log-rank test.

KM Plot with Confidence Interval

Figure 1 Kaplan–Meier curve for overall survival (interrupted line 95% confidence interval).

**Figure 2.** Estimated Effect of Abacavir–Lamivudine (ABC-3TC) versus Tenofovir DF–Emtricitabine (TDF-FTC) on the Hazard of Virologic Failure, According to Baseline Characteristics.

In the univariate analysis, modeling for treatment-effect modification used the continuous form for age, CD4 cell count, and HIV-1 RNA level. Estimated treatment effects are shown as example values for these continuous variables. Multivariate analysis showed similar results.

Graphics Conclusion

• A good picture is worth a 1000 words
• Encourage your analyst to provide graphical displays of the data
• These were published examples, don’t worry if your plots are simple or not as elegant
• Some journals have a graphics department that redraw graphics for a consistent look
STATISTICAL INFERENCE

Confidence is what you have before you understand the problem – Woody Allen
Statistical Inference

• Drawing inference or conclusions about a population of interest,
• based on data sampled from the population,
• while appropriately accounting for uncertainty due to not having data from the entire population
Population & Sample

inference
Statistical Inference

Many statistical methods available
Choice driven by several factors, including

1. Type of study being analyzed
   - Case-control study
   - Cohort study
   - Randomized clinical trial
   - Meta-analysis
   - Complex survey sample
Statistical Inference

1. Type of study
2. Type of endpoint being analyzed
   - Binary, ordinal, or continuous endpoint
   - Repeated/longitudinal measurements
Pitfall: Ignoring the independence assumption

**Independent**
- Single measurements from separate individuals (or animals)

**Not Independent**
- Repeated measure on same subject or animal
  - Before and after
  - Longitudinal
  - Assay A and B
- Measurements from separate individuals in the same household
Pitfall: Ignoring the independence assumption

- Ignoring possible dependence between observations essentially assumes the sample size is larger than it is
- Leading to anti-conservative inference, i.e., confidence intervals that are too narrow, p-values too small, and increasing the likelihood of a false-positive result
Statistical Inference

1. Type of study
2. Type of endpoint being analyzed
   - Binary, ordinal, or continuous endpoint
   - Repeated/longitudinal measurements
   - Time to event endpoint (subject to censoring) or assay (w/ limit of detection)
   - More generally, is it missing or incomplete? Eg, due to missed visits, lost specimen, non-response, lost-to-follow-up
Pitfall: Ignoring missing data

- The default of most statistical software programs is to use only observations with complete data ("complete case" analysis)
- May be reasonable if missing data is uncommon
- Otherwise can yield invalid and/or inefficient results
- Certain types of missing data (e.g., right censoring in survival analysis) can be handled by standard software
- For other types of missing data, more sophisticated methods (imputation, weighting) may be required
Statistical Inference

• 1. Type of study
• 2. Type of endpoint being analyzed
• 3. Sample size
  • Asymptotic or exact statistical methods
Pitfall: Large sample methods with small n

• The justification for most statistical methods relies on the sample size being large
• AKA asymptotic approximations
• These approximations fail when the sample size is small (or possibly when the sample size is not small, but the endpoint is rare)
• Methods appropriate for small sample sizes (aka “exact” methods) should be used in this case
Statistical Inference

1. Type of study
2. Type of endpoint being analyzed
3. Sample size
4. Additional considerations
   - Efficiency – get the most out of the data
   - Robustness – to outliers, violations to assumptions
   - Assumptions underlying the method
Pitfall: Failure to Check Assumptions

• All statistical methods rely on certain assumptions
• If these assumptions are not true the resulting inference may be incorrect
• Thus it is critical to understand these underlying assumptions and, if possible, assess veracity for a given study/data set
Choopanya et al. Primary analysis based on Cox proportional hazards model
What types of assumptions?

- Parametric methods make very strong assumptions (e.g., log viral load is Normally distributed, time until infection follows an Exponential distribution), but are not robust.
- Non-parametric methods (e.g., Kaplan-Meier estimator) make very few assumptions, so tend to be valid in a wide range of settings but not efficient.
- Semi-parametric and rank-based type methods are popular:
  - Wilcoxon ranksum test
  - Generalized estimating equations (GEE)
  - Cox proportional hazards (PH) model
- Tend to be more robust than parametric methods (i.e., make weaker assumptions) and often almost as efficient.
Statistical Inference

• Typically one of two forms

• Estimation: Compute an estimate and associated confidence interval

• Testing: Conduct a statistical test of a null hypothesis, p-value
Estimation

• Ideally, we’d like to know some characteristic of the population of interest (*parameter*)
• Eg, prevalence of HIV in women of child-bearing age in Malawi
• Typically we can’t measure the whole population. So we sample from the population and compute an *estimate* of the parameter from the sample
• Estimate is *random* in the sense that if we were to sample from the population repeatedly, each estimate would potentially be different
• In contrast, the *parameter* is viewed as *fixed*
Population & Sample

Target Population
μ: mean

Sample
\( \bar{x}: \text{mean} \)

True population parameter
Sample estimate

inference
Pitfalls: Over-Reliance on Point Estimate

• Estimates alone do not describe uncertainty associated with inference; need to report confidence intervals also (more on this in a moment)
• Point estimate represents a single summary measure of the data and typically will not tell the whole story and may even mislead
• Eg, odds ratios
• Eg, next slide shows various examples where the Pearson correlation coefficient r, a measure of linear association between two variables, may or may not be misleading (know your data!)
Pitfall: Confusing SD and SE

- SD: Standard Deviation
- SE: Standard Error

- Sample SD: describes the spread of the data, how much do the observations tend to deviate from the mean
- Population SD: analog of sample SD for the population; unknown; fixed; a parameter
Population & Sample

Target Population
- μ: mean
- σ: standard deviation

Sample
- \( \bar{x} \): mean
- s: standard deviation

True population parameters
Sample estimates

Inference
Pitfall: Confusing SD and SE

- SD: Standard Deviation
- SE: Standard Error

- Sample SD: describes the spread of the data, how much do the observations tend to deviate from the mean
- Population SD: analog of sample SD for the population; unknown; fixed; a parameter

- SE refers to the distribution of the estimator
- Imagine computing the estimator (e.g., sample mean) repeatedly; SE is the standard deviation of the different estimators
What is a confidence interval (CI)?

- Provides an indication of the uncertainty associated with an observed estimate.

- How to interpret? 95% CI: Imagine conducting 100 studies, 95 of the 100 intervals should contain the true parameter of interest.

- To achieve narrower confidence intervals, decrease confidence (90% CI will be narrower than a 95% CI) or increase sample size.
Reporting Results

• General recommendation is to report estimate as well as confidence interval

• Cohen et al (2011) “a total of 39 HIV-1 transmissions were observed (incidence rate, 1.2 per 100 person-years; 95% confidence interval [CI], 0.9 to 1.7)”
Statistical Inference

• Typically one of two forms

• Estimation: Compute an estimate and associated confidence interval

• Testing: Conduct a statistical test of a null hypothesis, p-value
Hypothesis Testing

- Stipulate a null hypothesis

- Construct a test statistic based on data from sample, typically defined in such a way to help distinguish whether or not the null might be true

- Assess likelihood of obtaining a statistic as or more extreme than the observed statistic under the null: p-value
  - If the p-value is small (e.g., <0.05), one of two possibilities:
    - Just witnessed something unusual
    - The null hypothesis is not true

- Reject the null if likelihood is very small (say <.05)

- Threshold for rejecting null: significance level $\alpha$
Example: BAN Study (Chasela et al NEJM 2010)

- Evaluate efficacy of a maternal triple-drug antiretroviral regimen or infant nevirapine prophylaxis for 28 weeks during breast-feeding to reduce postnatal HIV transmission in Malawi
- Two primary null hypotheses of interest:
  1. the cumulative risk of HIV by 28 weeks will be the same in the infant intervention arm as in the no-ARV arm
  2. the cumulative risk of HIV by 28 weeks will be the same in the maternal intervention arm as in the no-ARV arm
- Log-rank test used to test each null hypothesis
- P<0.001 and P=0.02; reject both nulls
Confidence interval (CI) and p-value

• Generally recommended to not report p-values alone; reduces entire data set to a single number

• Accompany with parameter estimate and CI

• Cohen et al (NEJM 2011) “Of the 28 linked transmissions, only 1 occurred in the early-therapy group (hazard ratio, 0.04; 95% CI, 0.01 to 0.27; P<0.001)”

• Jamieson et al (Lancet 2012) “The cumulative risk of HIV-1 transmission by 48 weeks was significantly higher in the control group (7%, 95% CI 5–9) than in the maternal-antiretroviral (4%, 3–6; p=0.0273) or the infant-nevirapine (4%, 2–5; p=0.0027) groups.”
Pitfall: Over-emphasis on “significance”

- P<0.05 usually accompanied by phrase “statistically significant”

- Large effects in a small sample may be scientifically important but not statistically significant (underpowered)

- Small effects in a large sample can be statistically significant (high power to detect small effects), but **not** scientifically/clinically meaningful
Pitfall: Significant association implies causality
STATISTICAL INFERENDE

A Simple Example
Comparing Proportions

Estimated Prevalence of “+” (95% confidence interval)

A: 57.6% (49.7 to 65.5%)
B: 45.8% (37.9 to 53.7%)

Prevalence is 11.8% higher in group A.

Is the difference statistically significant?
Pitfall: Confidence Interval Overlap

• For the two sample problem, CIs overlapping does not indicate a lack of significant difference

• Need to conduct a hypothesis test or compute CI for difference
Comparing Proportions

Group A: 57.6%
Group B: 45.8%

The estimated difference is 11.8% with a 95% confidence interval from 0.7% to 23.0%

Chi-square p=0.04

At the 0.05 level of significance, there is a significant difference in prevalence between the two groups.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Group A</th>
<th>Group B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>87</td>
<td>70</td>
<td>157</td>
</tr>
<tr>
<td>-</td>
<td>64</td>
<td>83</td>
<td>147</td>
</tr>
<tr>
<td>Total</td>
<td>151</td>
<td>153</td>
<td>304</td>
</tr>
</tbody>
</table>
STATISTICAL INFERENCE

Multiple Comparisons
Multiple Comparisons

• Suppose we conduct many hypothesis tests, using data from the same study or individual

• Eg, compare group A versus group B with respect to several different assays, several different output measurements from the same assay (eg, flow cytometry), large panel of lab tests

• What is the probability of making at least one type I error (false positive) ?

• Eg, suppose we conduct 10 tests, each with $\alpha=0.05$, and in truth all nulls are true. What is the chance of making at least one type I error (falsely rejecting the null) ?
**Multiple Comparisons**

**Familywise Error Rate:** probability of getting a significant result ($p<0.05$) by chance when there are multiple independent tests and no correction for the number of tests carried out.

10 independent tests: 40% probability of a type I error (false positive).

Reference: Intuitive Biostatistics
Multiple Comparisons

• Methods are available to adjust for MC
  • Bonferroni adjustment
    • divide $\alpha$ by the number of tests to be performed
    • appropriate for independent tests, conservative otherwise
  • Hochberg, Holm, ...
  • False Discovery Rate (FDR) adjustment
    • Less stringent, useful when a lot of tests are done (genomics)

• When to adjust for MC?

• In settings such as genomics where a large number of tests are performed, MC adjustment essential
Multiple Comparisons

- In randomized studies with more than two arms, MC adjustment standard
  - BAN: A log-rank test procedure of size 0.025 used to test each null hypothesis
- Analyzing the same outcome multiple ways to assess the sensitivity of the primary approach can be done without MC adjustment
- Conducting multiple tests without MC adjustment and reporting only “significant” results not appropriate
- Minimal approach: report all test conducted and state the results are “without correction for multiple comparisons” and make clear how many comparisons were carried out. Informally, the reader can expect 1 in 20 p<0.05 to be from chance
STATISTICAL INFERENCE

Choosing an Endpoint
Choosing an Endpoint

• To avoid post-hoc data dredging, important to select primary endpoint(s) prior to analysis

• Considerations when selecting an endpoint:
  • Relevance
    • What is the most important question to answer?
  • Reliable (precise) and valid (accurate)
    • Substantial measurement error or bias?
  • Rate
    • How often does the endpoint occur (e.g., incident HIV infection) in the population of interest? This will affect power (more to come)
Types of Endpoints

• **Clinical endpoints**
  • HIV infection, AIDS-defining event, death
  • Eg, HIV vaccine efficacy studies use HIV infection as endpoint

• **Surrogate endpoints**
  • Not always feasible or ethical to wait for clinical endpoints
  • Laboratory: HIV-1 RNA, CD4, lipids, biomarkers...
  • Eg, early phase HIV vaccine studies rely on immunogenicity endpoints
  • Potentially misleading, ie, effect on surrogate may not imply effect on clinical endpoint

• **Composite endpoints**
  • Eg, time until HIV infection or death (BAN)
  • More relevant from public health or policy standpoint?
  • Can simplify analysis and increase power, but results may be difficult to interpret
SAMPLE SIZE AND POWER
Sample Size and Power Calculation

• Most grants/protocols include sample size justification

• Sample size too small: may miss scientifically meaningful differences

• Sample size too big: waste of resources

• Sample size justification typically in terms of power
Type I and Type II Error: Hypothesis Testing

<table>
<thead>
<tr>
<th>Population (Truth)</th>
<th>DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis...</td>
<td>Reject the null</td>
</tr>
<tr>
<td>Is True</td>
<td>Type I error (α) “false positive”</td>
</tr>
<tr>
<td>Is False</td>
<td>Correct “power”</td>
</tr>
</tbody>
</table>

Typically in study design α set to 5% and power (1-β) set to ≥80%
Sample Size/Power Calculation

• Based on endpoints associated with primary aims
• Requires prior knowledge (or a guess) regarding the endpoints (eg, mean, rate, SD) in each arm of the study
• Power is a function of exact specification of the alternative hypothesis; sometimes called the “effect size”
• Effect size chosen based on scientific or clinical relevance, not by the statistician
Sample Size/Power Calculation

• Takes one of two forms:
  1. For fixed sample size, what is the power to detect a particular difference between the arms of the study (i.e., effect size)?
  2. Given a particular difference (effect size) that we would like to detect, what sample size is required to insure adequate power?

Generally three approaches

• 1. Mathematical formula in a book/article
• 2. Statistical software (which often relies on 1)
• 3. Simulation study (modeling)
Sample Size Formula: Comparing Two Means

- Continuous outcome, two groups, alpha = 0.05, power = 80%
- Sample size per group

\[ n = \frac{16}{\Delta^2} \]

where \( \Delta = (\mu_0 - \mu_1)/\sigma \) is the difference in means between the two groups (effect size) divided by the population SD, i.e., standardized difference

- Note for continuous outcome need to know SD in addition to effect size
- For 90% power, replace 16 with 21

Gerald van Belle. Statistical Rules of Thumb. Chapter 2
What impacts $n$?

• **Effect size of interest or # of events**
  • The smaller the effect size (# of events), the larger $n$ will need to be

• **Variability of outcome measure**
  • The greater the variability, the larger the $n$ will need to be

• **Alpha and power (e.g. 5% and 80%)**
  • The smaller alpha or the greater the power, the larger $n$ will need to be

• **Number of groups and multiple comparisons**
  • The greater the number of MC, the larger $n$ will need to be
RESOURCES & REFERENCES
Statistical Software

• SAS: industry standard, legacy, learning curve
  • Useful for data management and statistics

• R: free, flexible, cutting edge methods, learning curve, [http://www.r-project.org/](http://www.r-project.org/)
  • R Packages: akin to the peer-reviewed Wiki world of statistics
  • R Commander, Graphical User Interface (GUI) [http://www.rcommander.com/](http://www.rcommander.com/)
  • S-plus: proprietary version, very similar code

• STATA: GUI or syntax, more tailored to epidemiology and biostatistics
Statistical Software

• GraphPad Prism: appears user-friendly and has some built in statistical guidance
  • QuickCalcs: free quick analysis on the fly
  • “Intuitive Biostatistics” –Harvey Motulsky

• Many, many more: StatXact, SigmaPlot, SPSS, Minitab, SAS JMP...

• Microsoft Excel: spreadsheet software
  • Quick graphs (use chart layout tools carefully for best results)
  • Reasonable for simple graphics and analyses
Statistical Software

• For sample size/power
  • nQuery software available for free at UNC
  • SAS PROC POWER

• Software available for free or discount at UNC
  http://software/sites.unc.edu/software-category/science-and-statistics/
  • SAS, SAS JMP, and R (R is freeware)
  • Discounts for several software packages (Stata, SPSS,..)
  • Sample size calculation: nQuery Advisor
# Available Software

## Filter Selection...

<table>
<thead>
<tr>
<th>$</th>
<th>Title</th>
<th>Category</th>
<th>Audience</th>
<th>Platform</th>
<th>GET SOFTWARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>Amos</td>
<td>Science and Statistics</td>
<td>Faculty, Staff</td>
<td>Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>ChemDraw Ultra</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Mac, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>dof Power DBMS/Copy</td>
<td>Science and Statistics</td>
<td>Faculty, Staff</td>
<td>Linux, Unix, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>JMP</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Linux, Mac, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>JMP Genomics</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>nQuery Advisor</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>NVivo</td>
<td>Science and Statistics</td>
<td>Faculty, Staff</td>
<td>Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>R</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Linux, Unix, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>SAS 9.3</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Linux, Unix, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>SPSS</td>
<td>Science and Statistics</td>
<td>Faculty, Staff</td>
<td>Mac, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>Stata GradPlan</td>
<td>Science and Statistics</td>
<td>Faculty, Student</td>
<td>Linux, Mac, Windows</td>
<td>GET SOFTWARE</td>
</tr>
<tr>
<td>$</td>
<td>SUDAAN</td>
<td>Science and Statistics</td>
<td>Faculty, Staff, Student</td>
<td>Windows</td>
<td>GET SOFTWARE</td>
</tr>
</tbody>
</table>
References and Resources

- Intuitive Biostatistics – Motulsky
- Principles of Biostatistics – Pagano
- Modern Epidemiology – Rothman
- Statistics notes in BMJ
  http://www-users.york.ac.uk/~mb55/pubs/pbstnote.htm
- Bios 662
- CFAR Biostatistics Core (including these slides)
  http://cfar.med.unc.edu/content/biostatistics-core
Biostatistics Training at UNC

Odum Institute
http://www.irss.unc.edu/odum/home2.jsp

Biostatistics Department
http://sph.unc.edu/department-pages/biostatistics/

BIOS 600 PRINCIPLES OF STATISTICAL INFERENCE (3). Prerequisite, knowledge of basic descriptive statistics. Major topics include elementary probability theory, probability distributions, estimation, tests of hypotheses, chi-squared procedures, regression, and correlation.

BIOS 610 BIOSTATISTICS FOR LABORATORY SCIENTISTS (3). Prerequisite, elementary calculus. Introduces the basic concepts and methods of statistics, focusing on applications in the experimental biological sciences.
NC TraCS Biostatistics Training

Annually: Introduction to Study Design and Strategies for Data Analysis (1 week summer course)

This Fall: Biostatistics Seminar Series for Clinical and Translational Scientists begins August 29!

A seven part series running Aug-Nov

tracs.unc.edu/biostats-seminar-fall2013
NC TraCS Biostatistics Training

The **goal** of this series is to provide clinical and translational researchers who have basic quantitative training in biostatistical methods with a more in depth understanding of selected topics and to introduce them to more advanced methods.

The **target audience** for the seminar series includes junior clinical and translational scientists who currently serve as or plan to serve as Principal Investigators leading interdisciplinary research teams that include biostatisticians. For example, this includes junior faculty in the Schools of Medicine, Public Health, Nursing, Pharmacy and Dentistry.
# Fall 2013 NC Tracs Biostatistics Seminars

To receive information about future sessions, contact [NCTraCS_BiosSeminar@unc.edu](mailto:NCTraCS_BiosSeminar@unc.edu).

## 2013 - 2014 Sessions

<table>
<thead>
<tr>
<th>Date / Time</th>
<th>Topic / Location</th>
<th>Presenter</th>
</tr>
</thead>
</table>
| Thu. 8/29/13 1:00 - 2:30 | Some Essentials of Randomized Controlled Trials | Mark Weaver, PhD  
UNC School of Medicine |
| Wed. 9/12/13 1:00 - 2:30 | Data Reliability and Validity | Kant Bangdiwala, PhD  
UNC Gillings School of Global Public Health |
| Thu. 9/26/13 1:00 - 2:30 | DSMBs and Interim Analyses | Sonia Davis, DrPH  
UNC Gillings School of Global Public Health, CSCC |

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu. 10/10/13</td>
<td>1:00 - 2:30</td>
<td>Cluster-Randomized Trials</td>
<td>John Preissner, PhD</td>
<td>UNC Gillings School of Global Public Health</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brinkhous-Bullitt 219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu. 10/24/13</td>
<td>1:00 - 2:30</td>
<td>Issues in Non-inferiority Trials</td>
<td>Rosalie Dominik, DrPH</td>
<td>UNC Gillings School of Global Public Health, CSCCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bondurant Hall, G074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu. 11/7/13</td>
<td>1:00 - 2:30</td>
<td>Survey Sampling and Working with the CSRL</td>
<td>Donglin Zeng, PhD, and Robert Agans, PhD</td>
<td>Co-Directors of Carolina Survey Research Laboratory (CSRL)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brinkhous-Bullitt 219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu. 11/21/13</td>
<td>1:00 - 2:30</td>
<td>Design and Analysis of SMARTs</td>
<td>Michael Kosorok, PhD</td>
<td>UNC Gillings School of Global Public Health</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Sequential Multiple Assignment Randomized Trials)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brinkhous-Bullitt 219</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank you for coming and staying!
Thanks to Joe Rigdon, Ali Fokar, David Rosen, Catherine Grodensky, Carol Golin for helpful feedback!
Next Week:

"An Overview of Data Management"

Ali Foker
UNC CFAR Clinical Core

and

"Survey Development"

Carol Golin and Catherine Grodensky
UNC CFAR Social and Behavioral Science Core

August 30, 2013
8:30 - 10:00 a.m. (1.5 hours)
1131 Bioinformatics
I USED TO THINK CORRELATION IMPLIED CAUSATION.

THEN I TOOK A STATISTICS CLASS. NOW I DON'T.

SOUNDS LIKE THE CLASS HELPED. WELL, MAYBE.