BIOS 312: Review of Key Concepts

Chris Slaughter

Department of Biostatistics, Vanderbilt University School of Medicine

January 3, 2013

Outline

Samples from populations

- Scientific hypotheses concern a population
  - Do *teenagers* learn math more quickly using the Singapore Method?
  - Will *people with heart disease* live longer if they are prescribed medication X?
- *Teenagers* and *people with heart disease*, are the populations of interest. However, it is rare that we can ever study the entire population
- The purpose of inferential statistics is to make valid inference about a populations based on a sample from that population
- We commonly make inference about population parameters

Inference and variability

- With biological questions, there is inevitably variation in the response across repetitions of the experiment
  - Exposure to a carcinogen increases your risk of cancer, but does not guarantee a cancer will develop
- Thus, biological questions must be phrased in a probabilistic (not deterministic) language
  - Deterministic: Does medication decrease blood pressure?
  - Probabilistic: Does medication **tend to** decrease blood pressure?

Choosing your summary measure

- The wording "tends to" is intentionally vague. There are many possible definitions
  - A lower average value (arithmetic mean)
  - A lower geometric mean (arithmetic mean on log scale)
  - A lower median
    - Median(Trt) - Median(Ctrl) < 0.0
    - Median(Trt - Ctrl) < 0.0
  - A lower proportion exceeding some threshold
  - A lower odds of exceeding some threshold
  - Pr(Trt $>$ Control) < 0.5
  - And many others...

Consider the science

- Defining "tends to" is primarily dictated by scientific considerations
  - You, not the data, get to choose which summary measure you care about
  - Which measure is most important to advance science?

Outline

Refining scientific questions

- To formally answer scientific questions, they must be refined into statistical questions
- Scientific question: Does aspirin prevent heart attacks?
  - An important question, but can't be addressed by statistics
  - Cause and effect dependent on study design

Refinement 1

Do people who take aspirin not have heart attacks?

- Problem: Deterministic

Refining scientific questions

- To formally answer scientific questions, they must be refined into statistical questions
- Scientific question: Does aspirin prevent heart attacks?
  - An important question, but can't be addressed by statistics
  - Cause and effect also depends on the study design

Refinement 2

Do people who take aspirin tend to have fewer heart attacks?

- Problem: No control group
  - We also need to know heart attack rate among individuals not taking aspirin

Refining scientific questions

- To formally answer scientific questions, they must be refined into statistical questions
- Scientific question: Does aspirin prevent heart attacks?
  - An important question, but can't be addressed by statistics
  - Cause and effect dependent on study design

Final Refinement

Is the incidence of heart attacks less in people who take aspirin than those who do not?

- Basic science: Is the incidence less by *any* amount?
- Clinical science: Is the incidence less by a *clinically relevant* amount?
- Note that we are addressing statistical association, not causation

Outline

Associations between variables

- An association exists between two variables if their probability distributions are not independent
  - For random variables X and Y with joint probabilty density function (pdf) $f(x, y)$, marginal pdfs $f_X(x)$ and $f_Y(y)$, and conditional pdfs $f(x|y)$ and $f(y|x)$
    - $f(x, y) = f_X(x)f_Y(y)$
    - $f(y|x) = f_Y(y)$
    - $f(x|y) = f_X(x)$
  - Independence means that there is no way that information about one variable could ever give any information at all about the probability that the other variable might take on a particular value
  - Association means that the distribution of one variable differs in some way (e.g. mean, median, variance, probability of being greater than 10) across at least two groups differing in their values of the other variable

Establishing independence

- Can we ever establish independence between two variables?
  - Very, very difficult
    - Two variables can be associated in many different ways
    - It is hard to examine every characteristic of a distribution across groups
- Conversely, we can show associations as soon as we establish some information that one variable provides about the other
- Negative studies (e.g. studies with $p > 0.05$ or CI that contains the null value)
  - Absence of evidence is not the same as evidence of absence
  - To make negative studies meaningful, we must...
    - Specify the type of association that we are looking for (e.g. mean, median)
    - Quantify the amount of uncertainty that might differ across groups

Example: Inference about an association between exposure (E) and disease (D)

- Data and results
  - Unexposed: 0 of 5 have disease D
    - Incidence rate: 0.00
    - 95% CI: [0.00, 0.52]
  - Exposed: 3 of 5 have disease D
    - Incidence rate: 0.60
    - 95% CI: [0.15, 0.95]
  - Fisher's test p-value: 0.17
- How would you summarize the results of the study examining the association between E and D? Critique the following

Answer 1

Since the p-value is greater than 0.05, we conclude that there is no association between exposure E and disease D.

Example: Inference about an association between exposure (E) and disease (D)

- Data and results
  - Unexposed: 0 of 5 have disease D
    - Incidence rate: 0.00
    - 95% CI: [0.00, 0.52]
  - Exposed: 3 of 5 have disease D
    - Incidence rate: 0.60
    - 95% CI: [0.15, 0.95]
  - Fisher's test p-value: 0.17
- How would you summarize the results of the study examining the association between E and D? Critique the following

Answer 2

Since the p-value is greater than 0.05, we lack evidence to conclude that there is an association between exposure E and disease D.

Example: Inference about an association between exposure (E) and disease (D)

- Data and results
  - Unexposed: 0 of 5 have disease D
    - Incidence rate: 0.00
    - 95% CI: [0.00, 0.52]
  - Exposed: 3 of 5 have disease D
    - Incidence rate: 0.60
    - 95% CI: [0.15, 0.95]
  - Fisher's test p-value: 0.17
- How would you summarize the results of the study examining the association between E and D? Critique the following

Answer 3

We observed incidence rates of 60% in the exposed (95% CI: [15%, 95%]) and 0% in the unexposed (95% CI: [0%, 52%]). The precision of the study was not adequate to demonstrate that such a large difference in incidence rates would be unlikely in the absence of a true association.

Outline

The multiple comparisons problem

- When you perform many hypothesis tests, your chance of making a type 1 error increases
  - Type 1 error: Probability of rejecting the null hypothesis when the null hypothesis is true
- Consider the follow hypothetical example
  - An investigator decides to examine an association between eating red meat and cancer. He collects clinical data on a cohort of individuals who eat red meat and a cohort who does not eat red meat.
  - In the analysis, the investigator compares incidence rates between the two groups. Also makes comparisons stratified by gender, race, and lifestyle factors.

Investigator's summary of findings

"The research study uncovered a significant association between consuming red meat and the incidence of lung cancer in non-smoking males ($p < 0.05$)" (No significant associations were found in any other subgroup.)

Two possible conclusions

Conclusion 1

There is an association between consuming red meat and cancer in non-smoking men

Conclusion 2

This finding is a type 1 error

Which do you suspect is the truth?

http://xkcd.com/882/

Probability calculation for multiple comparisons

- For a null hypothesis H_0 that is true, and a test performed at significance level α
  - $\Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$
  - $\Pr(\text{do not reject } H_0 | H_0 \text{ is true}) = 1 - \alpha$
- Next suppose that n independent hypothesis tests $(H1_0, H2_0, \ldots, Hn_0)$ are performed at level α and all n null hypotheses are true
  - $\Pr(\text{do not reject } H1_0, H2_0, \ldots, Hn_0 | \text{ all } Hi_0 \text{ are true}) = (1 - \alpha)^n$

n	1	2	4	8	12	16	20	30
$(1 - .05)^n$	0.95	0.90	0.81	0.66	0.54	0.44	0.36	0.22
$(1 - .01)^n$	0.99	0.98	0.96	0.92	0.89	0.85	0.82	0.74

- If 30 independent tests are performed at $\alpha = 0.05$ and all null hypotheses are true, the probability of falsely rejecting at least one null hypothesis is 78%! We are very likely to make a mistake.
- Choosing a smaller $\alpha = 0.01$ helps, but the probability of an error is still 26% for 30 tests

Outline

Regression models

- This semester, I will introduce several regression models
- Which regression model you choose to use is based on the parameter being compared across groups

Means	$\rightarrow$ Linear regression
Odds	$\rightarrow$ Logistic regression
Rates	$\rightarrow$ Poisson regression
Hazards	$\rightarrow$ Proportional Hazards (Cox) regression

Regression models

- In Bios 311, we discussed how to examine association between an outcome and a predictor of interest (POI)
- Regression models generalize two sample tests

2-sample t-test	$\rightarrow$ Linear regression
Pearson chi-squared test	$\rightarrow$ Logistic regression
Wilcoxon signed-rank test	$\rightarrow$ Proportional odds regression

Multivariable regression

- We will learn how to adjust for additional variables using a regression model
  - Confounders
  - Effect modifiers
  - Precision variables
- When building a multivariable regression model, many possible choices along the way
  - Which covariates to include?
  - What model provides the best fit to the data?
  - Does the model satisfy underlying assumptions?
  - etc.
- Theme of this course: When building a multivariable model, data driven decisions lead to multiple comparisons problem

Multivariable regression

- Instead, I will emphasize
  - Putting science before statistics
  - Regression models that are robust to distributional assumptions (relieving the need for extensive model checking)
  - Detailed, pre-specified analysis plans