Statistical Test Selection in Epidemiologic Research

Vincent Lo Re, MD, MSCE, FISPE
Department of Medicine (Infectious Diseases)
Center for Pharmacoepidemiology Research and Training
Perelman School of Medicine
University of Pennsylvania

4th MURIA – June 18, 2018

CPeRT Center for Pharmacoepidemiology Research and Training

ISPE

"I was told there would be no math!"

- Chevy Chase
'Spies Like Us'

Learning Objectives

- Understand variable characteristics to guide statistical test selection
- Learn to use, interpret correlation coefficients
- Understand variables influencing sample size
- Gain familiarity with use, interpretation of linear and logistic regression

Outline

- Constructing a research project
- Correlation / regression
- Linear regression
- Logistic regression

Outline

- Constructing a research project
- Correlation / regression
- Linear regression
- Logistic regression

Constructing a Research Project

- Research question
- Variable characteristics
- Study design, sample size
- Statistical methods

Constructing a Research Project

- Research question
- Variable characteristics
- Study design, sample size
- Statistical methods

3 Types of Questions: Different Analyses

Research Question Type #1

- How much or common?
  - Design: Cross-sectional, cohort studies
    - Descriptive statistics:
      - Percentages, frequencies
      - Means (standard deviations)
      - Medians (interquartile ranges)
      - Prevalence (95% CIs)
      - Incidence (95% CIs)

Potential Analyses

Research Question Type #2

- Are these groups different?
  - Design: Case-control, cohort, RCTs
    - T-test: difference in means
    - Wilcoxon rank-sum: difference in medians
    - Chi square, Fisher's exact: diff. in frequencies
    - ANOVA, Kruskal-Wallis: diff. in means, medians among ≥3 groups
    - Odds ratios, hazard ratios, relative risks

Potential Analyses

Research Question Type #3

- Can certain variables predict outcome?
  - Design: Cohort study
    - Linear regression
    - Logistic regression
    - Survival analysis (Cox regression)

Potential Analyses

Constructing a Research Project

- Research question
- Variable characteristics
- Study design, sample size
- Statistical methods

Variable Characteristics to Consider

- Categorical or continuous?
  - Continuous: Normal or not?
- How many independent variables?
- How many groups?

Variable Characteristics to Consider

- Categorical or continuous?
  - Continuous: Normal or not?
- How many independent variables?
- How many groups?

Categorical or Continuous?

- Continuous: *Any value within a range*
  - Age (years)
  - Blood pressure (mm Hg)
  - Height (m)
  - Weight (kg)
  - CD4 cell count (cells/mm^3)

Categorical or Continuous?

- Categorical: *Discrete categories*

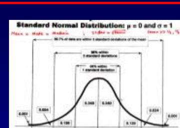
Nominal
Named categories with no order
- Blood type
- Medical specialty

Ordinal
Ordered categories, differences unequal
- NY Heart Class

Variable Characteristics to Consider

- Categorical or continuous?
  - Continuous: Normal or not?
- How many independent variables?
- How many groups?

Normal Distribution



- Continuous data
- Symmetrical, bell shaped
- Mean, median, mode all the same and located at the center
- Allows use of parametric tests (e.g., t-tests)
- If not, must use non-parametric tests

Variable Characteristics to Consider

- Categorical or continuous?
  - Continuous: Normal or not?
- How many independent variables?
- How many groups?

Depends on Clinical Question!

Constructing a Research Project

- Research question
- Variable characteristics
- Study design, sample size
- Statistical methods

Hypothesis Testing

- Develop hypothesis
- Test hypothesis:
  - Collect data, observe effect
    H_0 = No effect or difference
    H_a = Effect or difference
- How likely is it that effect occurs by chance
  - If very unlikely ($p < 0.05$), reject H_0

Effect Size, Significance, Power,

- Effect Size: Magnitude of effect being studied
  - Should represent clinically significant difference
- Significance: Probability of Type I error (α=0.05)
- Power: Probability of detecting difference (80%)
- Sample Size: n required to show a difference at set values of effect size, power, and significance

Determination of Sample Size

- Why is this necessary?
  - To detect effect size (OR, RR, HR) as significant
  - Avoid false-positive, false-negative conclusions
  - Avoid enrolling too many patients
- When to determine sample size?
  - During preparation of all protocols (perform early!)
- How to calculate?
  - Stata
  - Other programs: PS - Power / Sample Size, nQuery

Variables Used to Calculate Sample Size

- Detectable (clinically meaningful) difference (d*):
  - Magnitude of difference in proportions, means
- r: ratio of unexposed:exposed, controls:cases
- Power ($1 - \beta$):
  - Type II error (β) = prob. that there is no difference when one does exist (false-negative; set at 0.1, 0.2)
- Type I error (α):
  - Prob. of concluding that there is difference when one does not exist (false-positive; usually set at 0.05)

Variables Used to Calculate Sample Size

- p_1 (for proportions):
  - Proportion exposed who develop disease (cohort/ cross-sectional)
  - Proportion of cases exposed (case-control)
- p_0 (for proportions):
  - Proportion unexposed who develop disease (cohort/ cross-sectional)
  - Proportion of controls exposed (case-control)
- Standard deviation (σ) of continuous outcome

Calculation of Sample Size

- Primary outcome is variable for which you perform sample size calculation
  - If secondary outcomes important, ensure sample size is sufficient
- Typically, have more power to detect differences in continuous outcomes

Sample Size Calculation: Difference in Means

- Sample size for difference in means:

$$n = \frac{(Z_\beta + Z_{\alpha/2})^2 \, \sigma^2 \, (r + 1)}{(d^*)^2 r}$$

- Variables:
  - σ = standard deviation of outcome (σ^2=variance)
  - $Z_{\alpha/2}$ = type I error of 0.05; value=1.96
  - Z_β = type II error; for 0.2 [80% power], value=0.84
  - $(Z_\beta + Z_{\alpha/2})^2$ = 7.85 for 80% power
  - $(Z_\beta + Z_{\alpha/2})^2$ = 10.5 for 90% power
  - r = ratio; d* = detectable difference

Sample Size Calculation: Difference in Proportions

- Sample size for difference in proportions:

$$n = \frac{(Z_\beta + Z_{\alpha/2})^2 \, p_w (1 - p_w)(r + 1)}{(d^*)^2 r}$$

- Note: p_w = weighted average of p_1 and p_0

$$p_w = (p_1 + rp_0) / (1+r)$$

Variables Affecting Sample Size

- Detectable difference:
  - Smaller difference (effect size): ↑ sample size
- Power:
  - ↑ power (e.g., 80 → 90%): ↑ sample size
- Standard deviation of outcome (σ):
  - Smaller σ: ↓ sample size
- P_0:
  - Smaller p_0: ↑ sample size
- Significance level (α):
  - ↑ α → ↓ sample size

Sample Size Calculations

- You must increase sample size to reflect:
  - Loss to follow up
  - Expected response rate
  - Lack of adherence, etc.

  } ↓ Effective sample size
- Example:
  - Targeted number of exposed = 1,200 subjects
  - But only 70% expected to consent (30% refusal rate)
  - Adjust targeted number of exposed as follows:
    - 1,200 / 0.7 = 1,714 exposed
  - So if 1:1 ratio, need 1,714*2 = 3,428 total subjects

Constructing a Research Project

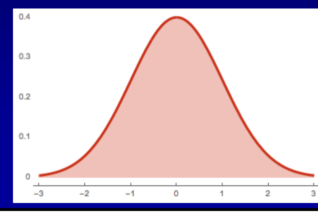
- Research question
- Variable characteristics
- Study design, sample size
- Statistical methods

Select Appropriate Statistical Test for Continuous Data

Purpose of Test	Normal Theory Test (Parametric)	Corresponding Non-Parametric Test
Compare paired data	Paired t-test	Wilcoxon signed-rank test
Compare 2 independent samples	Two-sample t-test	Wilcoxon rank-sum test (Mann-Whitney U test)
Compare ≥3 groups	One-way ANOVA	Kruskal-Wallis

Parametric Tests

- Continuous data
- Normally distributed



Non-Parametric Tests

- Not normally distributed continuous data
  - Small samples

    OR
  - Categorical data
    - Nominal, ordinal
    - Dichotomous (Outcome vs. no outcome)

Select Appropriate Statistical Test for Each Question

Purpose of Test	Normal Theory Test (Parametric)	Corresponding Non-Parametric Test
Compare paired data	Paired t-test	Wilcoxon signed-rank test
Compare 2 independent samples	Two-sample t-test	Wilcoxon rank-sum test (Mann-Whitney U test)
Compare ≥3 groups	One-way ANOVA	Kruskal-Wallis

Categorical Data Analysis: Chi Square Analysis

- Answers: “Are these groups different?”
- Contingency tables evaluate relation between values of ≥2 categorical variables
  - Rows, columns are independent

Categorical Data Analysis: Fisher’s and McNemar’s Tests

- Fisher’s Exact Test: Use if any value in a cell of table is <5
- McNemar’s Test: Use if data are from paired samples

Outline

- Constructing a research project
- Correlation / regression
- Linear regression
- Logistic regression

Correlation / Regression

- Examine relation between variables
- Correlation:
  - Tests significance of relation
- Regression:
  - Quantifies relationship, controlling for confounders

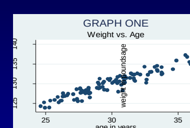
Correlation Coefficient

- Quantifies relationship between two variables
  - Correlation coefficient ("r") ranges -1 to +1

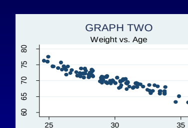
Value of r	Interpretation
r = 0	Two variables do not vary together at all
0 > r > 1	Two variables increase or decrease together
r = 1.0	Perfect correlation
-1 > r > 0	One variable increases as the other decreases
r = -1.0	Perfect negative or inverse correlation

- Often useful to graph data

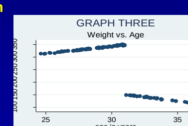
Correlations



Positive Correlation



Negative Correlation



Relation between weight and age is different for younger vs. older

Correlation Statistics

- Pearson's correlation (most widely used):
  - Assumes normally distributed data
  - Compute pairwise correlation ("r"), p values
- Spearman's rank-correlation coefficient:
  - One or more variables → not normally distributed
  - Less sensitive to effects of outlier data
  - Compute correlation ("r"), p values

Outline

- Constructing a research project
- Correlation / regression
- Linear regression
- Logistic regression

Linear Regression

- Allows you to relate <u>continuous</u> outcome (y) to one or more predictor variables (x_1, …, x_k)
  - Mean value of y is expressed as linear combination of x's (x's may be <u>continuous</u> or <u>categorical</u>)
  - Useful when have many potential confounders
- Have equation in the form:

$$y = \alpha + \beta * x$$

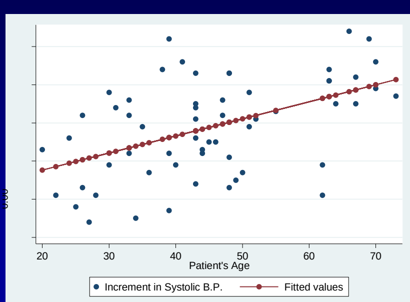
- Typically written as:

$$E(y) = \beta_0 + \beta_1 * x \rightarrow \text{fitted values}$$

Linear Regression

- Perform ordinary least squares regression of dependent variable y on independent variable x
- Estimates minimize squared distance between observed data and fitted values from model

Linear Regression: Determine Best Fit Line in Data



For every 1 unit (cm) ↑ in height, the pulmonary deadspace ↑ by 1.03 mL.

```
. regress deadspace height

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   32.81
       Model |  5607.43156     1  5607.43156           Prob > F      =  0.0001
    Residual |  2221.50178    13  170.884752           R-squared     =  0.7162
-------------+------------------------------           Adj R-squared =  0.6944
       Total |  7828.93333    14  559.209524           Root MSE      =  13.072

------------------------------------------------------------------------------
   deadspace |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.033323   .1803872     5.73   0.000     .6436202    1.423026
       _cons |   -82.4852   26.30147    -3.14   0.008    -139.3061   -25.66433
------------------------------------------------------------------------------
```

The constant is mean deadspace for a child without asthma.

Mean deadspace for child without asthma = 83 mL
Mean deadspace for child with asthma = 52.9 mL

```
.
. regress deadspace asthma

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =    9.92
       Model |  3388.05833     1  3388.05833           Prob > F      =  0.0077
    Residual |    4440.875    13  341.605769           R-squared     =  0.4328
-------------+------------------------------           Adj R-squared =  0.3891
       Total |  7828.93333    14  559.209524           Root MSE      =  18.483

------------------------------------------------------------------------------
   deadspace |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      asthma |    -30.125   9.565644    -3.15   0.008    -50.79032   -9.459683
       _cons |         83   6.985759    11.88   0.000     67.90819    98.09181
------------------------------------------------------------------------------
```

Multivariable Linear Regression

- Evaluate continuous outcome by linear relationship with independent variables
- Have >1 independent variable

Use of Multivariable Linear Regression

- Multiple factors may predict outcome
  - Blood pressure may be affected by weight, hormones, age, other factors
- Control for factors that can vary, but may confound, statistical analysis
  - E.g., age, sex, race, comorbidities
- Improve prediction

Outline

- Constructing a research project
- Correlation / regression
- Linear regression
- Logistic regression

Logistic Regression

- Alternate form of regression
  - Use when outcome is binary
    - Death versus survive
    - Acute MI versus no acute MI
  - One or more predictor variables

Logistic Regression

- Regression method for <u>binary</u> outcomes
- Useful for:
  - Continuous or discrete covariates
  - Adjusting for potential confounders
  - Evaluation of effect modifiers
- Provides OR (95% CI) of outcome for those with vs. without exposure of interest
-

Logistic Regression

- Determination of odds ratios (ORs) is based on maximum likelihood methods
  - Find coefficient values that maximize likelihood of obtaining observed data
- Output requested: β coefficients or ORs

Logistic Regression

```
. logit status admit_type

Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -93.176936
Iteration 2:   log likelihood = -92.528575
Iteration 3:   log likelihood = -92.524467
Iteration 4:   log likelihood = -92.524467

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =      15.11
                                                  Prob > chi2     =     0.0001
Log likelihood = -92.524467                       Pseudo R2       =     0.0755

------------------------------------------------------------------------------
      status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  admit_type |   2.184917   .7450489     2.93   0.003     .7246476    3.645186
       _cons |  -3.238678   .7208382    -4.49   0.000    -4.651495   -1.825861
------------------------------------------------------------------------------

. logistic status admit_type

Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =      15.11
                                                  Prob > chi2     =     0.0001
Log likelihood = -92.524467                       Pseudo R2       =     0.0755

------------------------------------------------------------------------------
      status | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  admit_type |   8.889907   6.623415     2.93   0.003     2.064004    38.28988
       _cons |   .0392157   .0282682    -4.49   0.000     .0095473    .1610788
------------------------------------------------------------------------------
```

Issues to Consider with Logistic Regression

- Which variables to include
- Which fitting method to use
- Collinearity
- Effect modifiers:
  - Does alcohol use level alter relation between drugs and acute liver injury?

Final Important Consideratons

- Know your data before analysis!
  - Look for missing data, develop plan to address
  - Graph variables, relationships between variables
- Collaboration with biostatistician is useful