Biostatistics: All About the Basics

Fern Jureidini Webb, Ph. D.
Associate Professor
University of Florida College of Medicine
Department of Community Health and Family Medicine
Department of Epidemiology

April 21 2016
Presented at NIH NIDDK NMRI’s 14th Annual Workshop
Speaker’s Disclosure

I, Fern Jureidini Webb have no conflict of interest to disclose.

Speaker’s Instruction for Interaction

Go to: http://www.fernjwebb.participoll.com/
Today’s Presentation

- Definition of Epidemiology
- Epidemiologic Research Cycle
- Study Factors/Variables
- Types of Data (Variables)
- Analysis and Evaluation
  - Measures of Frequency
    - Descriptive Analysis
  - Measures of Association
    - Statistical Analysis
    - Inferential Analysis
- Take Home Message
- Questions

http://www.fernjwebb.participoll.com/
Definitions of Epidemiology

- A branch of medical sciences involving the analysis of the incidence, distribution and control of disease and/or health in a population

- The study of the distribution and determinants of disease frequency [and health in the population with the premise that disease and/or health are not random]
  - Hennekens C, Buring J. Epidemiology in Medicine, 1st Edition. ISBN: 0-316-35636-0

Underlying assumption: disease or health distributions are not random events.

http://www.fernjwebb.participoll.com/
Epidemiologic Research Cycle

**Identify Question/Problem to Solve**
- Review current and seminal literature
- Create/establish team
- Form hypotheses
- Obtain institutional approval(s)

**Plan Protocol**
- Identify variables of interest
- Create analysis plan
- Determine research design
- Identify target/source population

**Conduct Study**
- Gather/collect data
- Analyze information
- Interpret findings

**Disseminate Findings**
- Share with key stakeholders
- Share with the science
  - Publications
  - Presentations

**Next Steps**
- Determine information learned
- Determine information still unknown
- Identify new/improved approaches to improve health

[http://www.fernjwebb.participoll.com/]
Epidemiologic Research Cycle

Plan Protocol
- Identify variables of interest
  - Create analysis plan
  - Determine research design
  - Identify target/source population

http://www.fernjwebb.participoll.com/
Study Factors/Variables

What is (are) the exposure(s) of prime interest?
- How is (are) the exposure(s) defined?
- How is (are) the exposure(s) measured?

What is (are) the outcome(s) of prime interest?
- How is (are) the outcome(s) defined?
- How is (are) the outcome(s) measured?

http://www.fernjwebb.participoll.com/
**Study Factors/Variables**

*We term the...*

<table>
<thead>
<tr>
<th>Exposure(s) of interest:</th>
<th>Outcome(s) of interest:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>Outcome</td>
</tr>
<tr>
<td>Treatment</td>
<td>Condition</td>
</tr>
<tr>
<td>Independent</td>
<td>Dependent</td>
</tr>
<tr>
<td>Antecedent</td>
<td>Consequent</td>
</tr>
<tr>
<td>Predictor</td>
<td>Response (or “Criterion”)</td>
</tr>
</tbody>
</table>

Types of Data (Variables)

- **Nominal data**
  - Unordered categories (i.e. ethnicity, gender, blood type)
  - No group/category is better/worse than the other

- **Ordinal data**
  - Ordered categories although distance between levels not exactly defined (i.e. excellent, very good, good, fair, poor)

- **Interval data**
  - Ordered and difference between points comparable
  - No ‘true’ zero (i.e. temperature)

- **Ratio data**
  - True zero point (i.e. cost, heart rate, blood pressure)
  - Defined difference/unit between values
  - Also called continuous

http://www.fernjwebb.participoll.com/
Let’s Practice ~ Which are nominal data?

A. Blood pressure, weight, age, income
B. Gender, race, hair color, religion/faith
C. Pain measures, education level, satisfaction
D. Temperature, money trends/stock market

http://www.fernjwebb.participoll.com/
Let’s Practice ~ Which are ordinal data?

A. Pain measures, education level, satisfaction
B. Blood pressure, weight, age, income
C. Gender, race, hair color, religion/faith
D. Temperature, money trends/stock market

http://www.fernjwebb.participoll.com/
Let’s Practice ~ Which are interval data?

A. Blood pressure, weight, age, income

B. Gender, race, hair color, religion/faith

C. Temperature, money trends/stock market

D. Pain measures, education level, satisfaction

http://www.fernjwebb.participoll.com/
Let’s Practice ~ Which are ratio data?

A. Gender, race, hair color, religion/faith

B. Blood pressure, weight, age, income

C. Pain measures, education level, satisfaction

D. Temperature, money trends/stock market

http://www.fernjwebb.participoll.com/
Epidemiologic Research Cycle

Plan Protocol

- Determine research design
- Identify target/source population
- Identify variables of interest
- Create analysis plan

http://www.fernjwebb.participoll.com/
Analysis and Evaluation

Measures of Frequency
• Descriptive Analysis

Measures of Association
• Statistical Analysis
• Inferential Analysis

http://www.fernjwebb.participoll.com/
Measures of Frequency

Review of Epidemiologic Measures
The “2 x 2” table

+-------+-------+-------+-------+
|       | YES   |       |       |       |
| YES   | a     | b     | a + b |       |
| NO    | c     | d     | c + d |       |
|       | a + c | b + d | a + b + c + d |   |
+-------+-------+-------+-------+

http://www.fernjwebb.participoll.com/
What are some of the uses of the 2x2 table?

A. Measures of Frequency
B. Measures of Association
C. Measures of Screening
D. Hypothesis Testing
E. All of the Above
Basic Measures of Frequency

• Counts - n

• Proportions – \( \frac{a}{a+b} \) (no time period) - i.e., percent

• Rates - \( \frac{a}{a+b} \) per 1,000, 10,000, 100,000 over a specified period of time

• Ratios - \( \frac{a}{b} \) (numerator and denominator are mutually exclusive)
Analysis

Descriptive Analysis
Used to describe information (measured by “variables”) or characteristics of those participating in our study

http://www.fernjwebb.participoll.com/
## Selected sociodemographic characteristics of participants (n=292)*

<table>
<thead>
<tr>
<th>Education</th>
<th>N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>2</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>GED/HS diploma</td>
<td>91</td>
<td>32</td>
</tr>
<tr>
<td>Some college/Associate’s Degree</td>
<td>40</td>
<td>14</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>95</td>
<td>33</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>49</td>
<td>17</td>
</tr>
<tr>
<td>Doctorate degree</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital Status</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>114</td>
<td>40</td>
</tr>
<tr>
<td>Other (including single, divorced,</td>
<td>173</td>
<td>60</td>
</tr>
<tr>
<td>widowed &amp; separated)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>35</td>
<td>18-73</td>
</tr>
<tr>
<td>Household income</td>
<td>$30,000-$49,000</td>
<td>&lt; $10K - &gt; $130K</td>
</tr>
</tbody>
</table>


http://www.fernjwebb.participoll.com/
Measures of Association

*Review of Epidemiologic Measures*

Used to describe how information (usually measured by variables) are associated or related to each other (variables)

Associations

Association:
The extent to which things occur together (non-directional)
OR
Statistical dependence between two variables:
(e.g., correlation between stages of change and weight)

\[ x \rightarrow y \rightarrow x \]

Independent, \( x \) (risk factor, exposure, treatment [clinical trials])
Dependent, \( y \) (Outcome, event)

\[ P(y) = x \]

In research, we would like to establish causal associations:
(uni-directional)

\[ x \rightarrow y \rightarrow x \]

http://www.fernjwebb.participoll.com/
Measures of Association

Relative Risk
Risk Ratio
Odds Ratio

Disease/Outcome

<table>
<thead>
<tr>
<th></th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>NO</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{a + c} & \quad \text{b + d} \\
\text{a + b} & \quad \text{c + d} \\
\text{a + b + c + d}
\end{align*}
\]
Choose the Appropriate Statistic to Measure the Association based on:

- Type and number of independent variables:
  - Nominal, ordinal, interval, continuous/ratio
  - One variable or multiple

- Type and number of dependent variables:
  - Nominal, ordinal, interval, continuous/ratio
  - One variable or multiple

- Same for any type of design or study

http://www.fernjwebb.participoll.com/
Let’s Practice!

Examining the association between stages of change and weight loss (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What is the primary independent variable?

A. Stages of Change

B. Weight Loss

C. Age, BMI, education, family health history, location, marital status, personal health

http://www.fernjwebb.participoll.com/
Let’s Practice!*

Examining the association between stages of change and weight loss (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What type of data is the primary independent variable?

A. Nominal
B. Ordinal
C. Interval
D. Ratio

http://www.fernjwebb.participoll.com/
Let’s Practice!*  

Examining the association between **stages of change** and **weight loss** (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What is the **dependent** variable?

A. Stages of Change

B. Age, BMI, education, family health history, location, marital status, personal health

C. Weight Loss

http://www.fernjwebb.participoll.com/
Let’s Practice!*

Examining the association between stages of change and weight loss (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What type of data is the dependent variable?

A. Nominal
B. Ordinal
C. Interval
D. Ratio

http://www.fernjwebb.participoll.com/
Let’s Practice!* 

Examining the association between *stages of change* and *weight loss* (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What type of data are the *other variables*?

A. Nominal

B. Ordinal

C. Ratio

D. Nominal, Ordinal & Ratio

http://www.fernjwebb.participoll.com/
Let’s Practice!*  
Examining the association between **stages of change** and **weight loss** (y/n). We will include age, body mass index [BMI], education, family health history (sum), healthy diet index (sum), location, marital status, personal health and life satisfaction in the model given their importance.

What statistic should we use to **measure this association**?  
A. Analysis of Variance (ANOVA)  
B. Multiple Regression  
C. Logistic Regression  
D. Chi-Square Test of Independence

Analysis

Inferential Analysis

- How do data from our sample reflect truth in the population?

http://www.fernjwebb.participoll.com/
Inferential Analysis: Chance

Statistical methods that evaluate the role of chance is the same for ANY/EVERY study

- Hypothesis testing
- Estimation/Confidence Intervals

http://ocw.tufts.edu/Content/1/lecturenotes/194194

http://www.fernjwebb.participoll.com/
Hypothesis Testing

\( H_0 \) = the null hypothesis. There is no association between stages of change and weight loss. Always start with the “null” ~ no difference!

\( H_A \) = the alternative hypothesis. There is an association between stages of change and weight loss.

There are four possible outcomes:

<table>
<thead>
<tr>
<th>Null Hypothesis is:</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject ( H_0 )</td>
<td>Type I error (P=( \alpha ))</td>
<td>Correct 1- ( \beta )</td>
</tr>
<tr>
<td>Fail to Reject ( H_0 )</td>
<td>Correct 1- ( \alpha )</td>
<td>Type II error (P=( \beta ))</td>
</tr>
</tbody>
</table>

“REALITY”

Usually, \( \alpha = .05 \)
\( \beta = .20 \) or .10
Power = 1- \( \beta \) = .80 or .90

http://www.fernjweebb.participoll.com/
Estimation: Confidence Intervals

0
1.0

• 1.0 = no association
If \( p \)-value > 0.05 (if set at 95%) or confidence interval (CI) includes 1
Fail to Reject \( H_0 \).

• > 1.0 = Those with exposure have dependent variable/outcome more than those without exposure
Relative Risk
Risk Ratio
Odds Ratio

• < 1.0 = Those with exposure have dependent variable/outcome less than those without exposure

If \( p \)-value < 0.05 (if set at 95%) or CI does not include 1
Statistically significant - Reject \( H_0 \).

\( p \)-value and CI should **ALWAYS*** give consistent findings!!
[* if based on same statistic *]

http://www.fernjwebb.participoll.com/
Let’s Practice!*

Weight loss modeled as the dependent variable

* These are fictitious data

Which associations are statistically significant?

A. Stages of change, family health, personal health

B. Age, BMI, education, location

C. Healthy diet index, marital status, life satisfaction

D. A and C above

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages of Change</td>
<td>1.50</td>
<td>1.10-2.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Age</td>
<td>1.00</td>
<td>0.97-1.04</td>
<td>0.08</td>
</tr>
<tr>
<td>BMI</td>
<td>1.01</td>
<td>0.96-1.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Education</td>
<td>1.34</td>
<td>0.65-2.74</td>
<td>0.11</td>
</tr>
<tr>
<td>Family Health</td>
<td>2.10</td>
<td>1.87-3.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Healthy Diet Index</td>
<td>3.10</td>
<td>2.81-4.27</td>
<td>0.001</td>
</tr>
<tr>
<td>Life Satisfaction</td>
<td>2.21</td>
<td>1.86-4.86</td>
<td>0.02</td>
</tr>
<tr>
<td>Location</td>
<td>1.01</td>
<td>0.78-1.32</td>
<td>0.21</td>
</tr>
<tr>
<td>Marital Status</td>
<td>0.76</td>
<td>0.36-0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>Personal Health</td>
<td>0.73</td>
<td>0.65-0.96</td>
<td>0.01</td>
</tr>
</tbody>
</table>

http://www.fernjwebb.participoll.com/
Let’s Practice!*
Weight loss modeled as the dependent variable
* These are fictitious data

Which associations are not statistically significant?

A. Stages of change, family health, personal health
B. Age, BMI, education, location
C. Healthy diet index, marital status, life satisfaction
D. A and C above

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages of Change</td>
<td>1.50</td>
<td>1.10-2.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Age</td>
<td>1.00</td>
<td>0.97-1.04</td>
<td>0.08</td>
</tr>
<tr>
<td>BMI</td>
<td>1.01</td>
<td>0.96-1.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Education</td>
<td>1.34</td>
<td>0.65-2.74</td>
<td>0.11</td>
</tr>
<tr>
<td>Family Health</td>
<td>2.10</td>
<td>1.87-3.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Healthy Diet Index</td>
<td>3.10</td>
<td>2.81-4.27</td>
<td>0.001</td>
</tr>
<tr>
<td>Life Satisfaction</td>
<td>2.21</td>
<td>1.86-4.86</td>
<td>0.02</td>
</tr>
<tr>
<td>Location</td>
<td>1.01</td>
<td>0.78-1.32</td>
<td>0.21</td>
</tr>
<tr>
<td>Marital Status</td>
<td>0.76</td>
<td>0.36-0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>Personal Health</td>
<td>0.73</td>
<td>0.65-0.96</td>
<td>0.01</td>
</tr>
</tbody>
</table>

http://www.fernjwebb.participoll.com/
Let’s Practice!*
If you saw this table, Which association might you question?

A. Stages of change
B. Education
C. Life Satisfaction
D. Personal Health

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages of Change</td>
<td>1.50</td>
<td>1.10-2.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Age</td>
<td>1.00</td>
<td>0.97-1.04</td>
<td>0.08</td>
</tr>
<tr>
<td>BMI</td>
<td>1.01</td>
<td>0.96-1.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Education</td>
<td>1.34</td>
<td>0.65-2.74</td>
<td>0.11</td>
</tr>
<tr>
<td>Family Health</td>
<td>2.10</td>
<td>1.87-3.39</td>
<td>0.03</td>
</tr>
<tr>
<td>Healthy Diet Index</td>
<td>3.10</td>
<td>2.81-4.27</td>
<td>0.001</td>
</tr>
<tr>
<td>Life Satisfaction</td>
<td>2.21</td>
<td>1.86-4.86</td>
<td>0.09</td>
</tr>
<tr>
<td>Location</td>
<td>1.01</td>
<td>0.78-1.32</td>
<td>0.21</td>
</tr>
<tr>
<td>Marital Status</td>
<td>0.76</td>
<td>0.36-0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>Personal Health</td>
<td>0.73</td>
<td>0.65-0.96</td>
<td>0.01</td>
</tr>
</tbody>
</table>

http://www.fernjwebb.participoll.com/
Take Home Message

- Choose a measure of association based on data/variable type for independent and dependent variables!

- Use your “cheat sheet” – no need to guess or memorize!

- Consult with a biostatistician/statistical expert during the planning phase of your study before you finalize design and begin conducting your study!!