How to analyze and interpret binary data
Shah Islam, M.P.H., PStat®
Department of Biostatistics
Winthrop University Hospital
Lesson Overview

• Binary Data

• One sample z-test of proportion

• Two sample z-test for equality of proportion

• Chi-square test for comparisons between two categorical variables (test of independence)

• Relative Risk (RR) and Odds Ratio (OR)

• McNemar Chi-square test for paired data

• Binary Logistic Regression
Binary Data

- Developed CVD (Yes=1, No=0)
- Have hypertension (Yes=1, No=0)
- Readmission (Yes=1, No=0)
- Test positive (+ = 1) vs. Test negative (- = 0)
- Like this lame lecture? (Yes=1, No=0)
One sample test of proportion scenario

- Drug X is administered to 100 patients with a particular disease. 50 improve. Test whether this drug is **better** than drug Y, which is known to produce improvement in 45% of patients.

**Sample data**

<table>
<thead>
<tr>
<th>ID</th>
<th>DRUG</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>
One sample one-sided test of proportion

Step 1. Set up the hypothesis

H_0: P = P_0 = 0.45

[Proportion improved on Drug X = Proportion Improved on Drug Y]

H_1: P > 0.45

[Proportion improved on Drug X > Proportion Improved on Drug Y]

\(\alpha = 0.05\)

Step 2. Select appropriate test statistic

\[ z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \]
One sample test of proportion

Step 3. Set up decision rule
Reject $H_0$ if $p < 0.05$

Step 4. Compute test statistic and P-value
Where $p_0 = 0.45$
$p$-hat = 50/100 = 0.50
$n = 100$

$$Z = \frac{0.50 - 0.45}{\sqrt{0.45(1-0.45)/100}} = 1.005$$

Refer to excel spreadsheet for this calculation
One sample test of proportion

One sided p-value from excel = 1 - NORMSDIST(abs(Z))

= 1 - NORMSDIST(1.005)

= 0.157

Two sided p-value = 2 * (1 - NORMSDIST(1.005)) = 0.315

Step 5. **Conclusion**

Do not reject H₀ because one sided p-value **0.157 > 0.05**.

We do not have enough evidence to conclude that drug X is better than drug Y.
Chi-Square vs. Z-test for two proportions

When comparing only two proportions, such as in a 2x2 table where the columns represent counts of “diseased” and “non-diseased,” we can test

H0: \( p_1 = p_2 \) vs. Ha \( p_1 \neq p_2 \)

equally with a two-sided z test or with a chi-square test with 1 degree of freedom and get the same p-value. In fact, the two test statistics are related: \( X^2 = (z)^2 \).
Chi-Square vs. Z-test for two proportions

Difference is a matter of design. The difference between these two tests is subtle yet important.

- The two proportion z-test is used when the response variable has only two categories as outcomes and we are comparing two groups (2x2).

- In the test of proportions, the data are collected by randomly sampling from each sub-group separately. (Say, 100 blacks, 100 whites and so on.)

- In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit. This test can handle more than two categories. We are testing association between the variables.
A clinical trial is conducted to test the efficacy of a new drug for hypertension. The new drug is compared to a standard drug in a study involving n=80 participants. The primary outcome is systolic blood pressure measured after 4 weeks on the assigned drug. The table below shows characteristics of study participants measured at baseline.
# Data for Chi-Square Test and the Test of proportions

<table>
<thead>
<tr>
<th>Baseline Characteristic</th>
<th>Standard Drug (n=40)</th>
<th>New Drug (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD) Age</td>
<td>75.6 (4.8)</td>
<td>74.7 (5.6)</td>
</tr>
<tr>
<td>N (%) Male</td>
<td>13 (32.5%)</td>
<td>9 (22.5%)</td>
</tr>
<tr>
<td>Mean (SD) BMI</td>
<td>25.1 (3.0)</td>
<td>26.9 (4.1)</td>
</tr>
<tr>
<td>Mean (SD) SBP</td>
<td>150.4 (19.8)</td>
<td>144.5 (19.7)</td>
</tr>
<tr>
<td>N (%) Diabetic</td>
<td>11 (27.5%)</td>
<td>8 (20.0%)</td>
</tr>
<tr>
<td>N (%) Current Smokers</td>
<td>2 (5.0%)</td>
<td>1. (2.5%)</td>
</tr>
</tbody>
</table>

Two sample $z$-test for equality of proportion

Test if there is a significant difference in the proportion of diabetic participants in the standard as compared to the new drug group. Use a 5% level of significance.
Two sample test of proportions

Step 1. \( H_0: \ p_1=p_2 \)  
\( H_1: \ p_1\neq p_2 \)  
\( \alpha=0.05 \)

Step 2. Test statistic

\[
Z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}
\]

Step 3. Decision rule: reject \( H_0 \) if p-value<0.05
Two sample z-test for equality of proportion

Step 4. Compute the test statistic

The overall proportion is given as:

\[
\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{11 + 8}{80} = 0.24
\]

We now substitute to compute the test statistic

\[
Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.275 - 0.20}{\sqrt{0.24(1-0.24)\left(\frac{1}{40} + \frac{1}{40}\right)}} = 0.79
\]

Refer to excel spreadsheet for this calculation
Two sample z-test for equality of proportion

P-value=2*(1-NORMSDIST(0.79))=0.429

Step 5. Conclusion.
Do not reject $H_0$ because $p=0.429>0.05$. We do not have enough evidence at $\alpha=0.05$ to show that there is a difference in the proportions of diabetic participants in the Placebo as compared to the Standard Drug groups.
Cross-tabulated counts
Chi-Square test of independence

- Test if there is an association between gender and 2 treatments groups. (HINT: Are gender and treatment independent?) Use a 5% level of significance.
Chi-Square Test: Underlying Assumptions

✅ Frequency Data

✅ Adequate sample size

✅ Measures independent of each other

✅ Theoretical basis for the categorization of the variable

Cannot be used to analyze differences in scores or their means

Expected frequencies should not be less than 5

No subjects can be counted more than once

Categories should be defined prior to data collection and analysis
Chi-Square test of independence

Step 1. Set up hypotheses and determine level of significance.

$H_0$: Gender and treatment group are independent

$H_1$: $H_0$ is false.

$\alpha=0.05$.

Step 2. Select the appropriate test statistic.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$
Chi-Square test of independence

Step 3. Set up decision rule.
df=(2-1)(2-1)=1
Decision rule: Reject H₀ if p-value<0.05

Step 4. Compute the test statistic.
• Compute the expected count using the formula,
  – Expected count= (Row Total * Column Total)/N.

<table>
<thead>
<tr>
<th></th>
<th>Disease+</th>
<th>Disease-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Non Exposed</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

• Expected count for observed count ‘a’ = [(a+b)*(a+c)]/(a+b+c+d)
## Chi-Square test of independence

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>New</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>13</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(11)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>27</td>
<td>31</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>(29)</td>
<td>(29)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>40</td>
<td>80</td>
</tr>
</tbody>
</table>

The test statistic is computed as follows:

\[
\chi^2 = \frac{(13-11)^2}{11} + \frac{(9-11)^2}{11} + \frac{(27-29)^2}{29} + \frac{(31-29)^2}{29}
\]

\[
\chi^2 = 1.003
\]

p-value: =CHIDIST(1.003, 1)=0.317

Refer to Excel spreadsheet for this calculation.
Chi-Square test of independence

Step 5. Conclusion.

Do not reject $H_0$ because $p=0.317>0.05$. We do not have statistically significant evidence at $\alpha=0.05$ to show that $H_0$ is false or that gender and treatment group are not independent.
Exercise-two Sample Test of Proportions

Is the proportion of CVD different in smokers as compared to nonsmokers in the Framingham Offspring Study?

<table>
<thead>
<tr>
<th></th>
<th>Free of CVD</th>
<th>History of CVD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsmoker</td>
<td>2757</td>
<td>298</td>
<td>3055</td>
</tr>
<tr>
<td>Current smoker</td>
<td>663</td>
<td>81</td>
<td>744</td>
</tr>
<tr>
<td>Total</td>
<td>3420</td>
<td>379</td>
<td>3799</td>
</tr>
</tbody>
</table>

Step1.  $H_0: p_1=p_2$

$H_1: p_1 \neq p_2 \quad \alpha=0.05$

Exercise-Chi Square test of independence

The following data was collected in a clinical trial evaluating a new compound designed to improve wound healing in trauma patients. The new compound is compared against a placebo. After treatment for 5 days with the new compound or placebo the extent of wound healing is measured and the data are shown below.

Is there a difference in the extent of wound healing by treatment? (Hint: Are treatment and the percent wound healing independent?) Run the appropriate test at a 5% level of significance.

<table>
<thead>
<tr>
<th>Percent Wound Healing</th>
<th>0-25%</th>
<th>26-50%</th>
<th>51-75%</th>
<th>76-100%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Compound (n=125)</td>
<td>15</td>
<td>37</td>
<td>32</td>
<td>41</td>
<td>125</td>
</tr>
<tr>
<td>Placebo (n=125)</td>
<td>36</td>
<td>45</td>
<td>34</td>
<td>10</td>
<td>125</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>82</td>
<td>66</td>
<td>51</td>
<td></td>
</tr>
</tbody>
</table>

Step 1. Set up hypotheses and determine level of significance.

H₀: Treatment and percent wound healing are independent
H₁: H₀ is false. α=0.05

McNemar’s Test for Correlated (Dependent) Proportions

Basis/Rationale
• Two sample test of proportions and chi-square test are based on the assumptions that the samples are independent

However,
• What if we have a situation where that is not true.
  – proportion of subjects with the characteristic (or event) is the same before and after intervention.
  – for case-control pair matched data, the proportion of subjects exposed to the risk factor is equal for cases and controls.
Pre/Post Data Example

A randomly selected group of 120 students taking a standardized test for entrance into college shows a failure rate of 50%.

A company which specializes in coaching students on this type of test has indicated that it can significantly reduce failure rates through a four-hour seminar. The students are exposed to this coaching session, and re-take the test a few weeks later.

The school board is wondering if the results justify paying this firm to coach all of the students in the high school. Should they? Test at the 5% level.

Data source: https://www.msu.edu
## Sample data

### Raw data

<table>
<thead>
<tr>
<th>Student Id</th>
<th>Pass/Fail Before</th>
<th>Pass/Fail After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Summarized Data

<table>
<thead>
<tr>
<th></th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pass</td>
</tr>
<tr>
<td>Before</td>
<td>56</td>
</tr>
<tr>
<td>Fail</td>
<td>56</td>
</tr>
<tr>
<td>Total</td>
<td>112</td>
</tr>
</tbody>
</table>

Data source: [https://www.msu.edu](https://www.msu.edu)
McNemar’s Test for Correlated (Dependent) Proportions

Step 1. State the null hypothesis and determine level of significance

\[ H_0: \ b = c \text{ (no changes before vs. after)} \]
\[ H_1: \ b \neq c \]

Where \( b \) and \( c \) relate to the count of observation reflecting changes in status (Discordant Cells)

\[ \alpha = 0.05 \]
McNemar’s Test for Correlated (Dependent) Proportions

Step 2. Test Statistic
\[ \chi^2 = \frac{(b - c)^2}{b + c} \]

Where degrees of freedom = 1
Rule of thumb: (b+c) ≥ 20

Step 3. Decision rule: reject H_0 if p < 0.05
McNemar’s Test for Correlated (Dependent) Proportions

Step 4. Calculate test statistic and p-value (without correction)

\[ x^2 = \frac{(4-56)^2}{(4+56)} \text{ where d.f.}=1 \]

\[ X^2 = \frac{2704}{6045} = 45.07 \]

\[ P\text{-value}=\text{CHIDIST}(45.07, 1)=<.0001 \]

Step 5. **Reject H0 since } p<0.05. We can conclude that there is a significant change (increase passing rate, decrease failure rate) in standardized test result after the coaching session compared to pre-coaching.
McNemar-Matched Paired Scenario

• Breast cancer patients receiving mastectomy followed by chemotherapy were matched to each other on age and cancer stage.

• By random assignment, one patient in each matched pair received chemo perioperatively and for an additional 6 months, while the other patient in each matched pair received chemo perioperatively only.

Data source: http://www.biostat.umn.edu/
### McNemar-matched paired scenario

#### Chemo therapy data

<table>
<thead>
<tr>
<th></th>
<th>Periop. only</th>
<th>Periop. + 6 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survived 5 years</td>
<td>Died within 5 years</td>
</tr>
<tr>
<td>Periop. only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survived 5 years</td>
<td>510</td>
<td>17</td>
</tr>
<tr>
<td>Died within 5 years</td>
<td>5</td>
<td>90</td>
</tr>
</tbody>
</table>

Research Question: Does survival to 5 years differ by treatment groups?

Data source: http://www.biostat.umn.edu/
McNemar-matched paired scenario

Step 1. Set up hypothesis

$H_0$: $b = c$

$H_A$: $b \neq c$

Check rule of thumb: $b + c \geq 20$

Step 2. Test statistic with $df = 1$

$$
\chi^2 = \frac{(b - c)^2}{b + c}
$$

Step 3. Decision rule: reject $H_0$ if $p < 0.05$
McNemar-matched paired scenario

Step 4. Calculate test statistic and p-value (without correction)

\[ x^2 = \frac{(17-5)^2}{(17+5)} \text{ where df}=1 \]

\[ X^2 = \frac{144}{22}=6.55 \]

P-value = CHIDIST(6.55, 1) = 0.0105

<table>
<thead>
<tr>
<th></th>
<th>Periop. only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survived 5 years</td>
</tr>
<tr>
<td>Periop. only</td>
<td>510</td>
</tr>
<tr>
<td>Periop. + 6 months</td>
<td>Survived 5 years</td>
</tr>
<tr>
<td></td>
<td>Died within 5 years</td>
</tr>
</tbody>
</table>

Step 5.

Conclusion: reject H0 since \( p<0.05 \). Data provide evidence that an extra 6 months of chemo results in a different survival rate compared to treatment with perioperative chemo alone (\( p=0.0105 \))
A study is run to estimate the incidence of atrial fibrillation (AF) in men and women over the age of 60. Development of atrial fibrillation was monitored over a 10 year follow-up period. The data are summarized below.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Developed AF</th>
<th>Did not Develop AF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>120</td>
<td>6453</td>
<td>6573</td>
</tr>
<tr>
<td>Women</td>
<td>86</td>
<td>7074</td>
<td>7160</td>
</tr>
</tbody>
</table>

Questions:

1. Compute the relative risk of AF incidence comparing men to women
2. Compute the odds ratio of AF incidence comparing men and women

**Relative Risk (RR)**

<table>
<thead>
<tr>
<th></th>
<th>AF</th>
<th>No AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>men</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>women</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Developed AF</th>
<th>Did not Develop AF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>120</td>
<td>6453</td>
<td>6573</td>
</tr>
<tr>
<td>Women</td>
<td>86</td>
<td>7074</td>
<td>7160</td>
</tr>
</tbody>
</table>

**Relative Risk:**

\[
RR = \frac{a/(a + b)}{c/(c + d)}
\]

\[
RR = \frac{120/6573}{86/7160} = \frac{0.018}{0.012} = 1.50
\]

**Interpretation:** In this sample, men over the age of 60 has 50% higher risk of developing atrial fibrillation as compared to the women over 60.
Odds ratio (OR)

<table>
<thead>
<tr>
<th></th>
<th>AF</th>
<th>No AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>men</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>women</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Developed AF</th>
<th>Did not Develop AF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>120</td>
<td>6453</td>
<td>6573</td>
</tr>
<tr>
<td>Women</td>
<td>86</td>
<td>7074</td>
<td>7160</td>
</tr>
</tbody>
</table>

Odds Ratio: 

\[
\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}
\]

\[
\text{OR} = \frac{120/6453}{86/7074} = \frac{0.019}{0.012} = 1.58
\]

**Interpretation:** In this dataset, men over the age of 60 have 58% increased odds of having atrial fibrillation as compared with the women over 60.
Simple Logistic Regression Analysis

- Outcome is dichotomous (1=event, 0=non-event) and $p=P(\text{event})$
- Outcome is modeled as log odds

\[ \hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \]

\[
\log(\text{odds}) = \logit(p) = \ln\left( \frac{p}{1-p} \right) = b_0 + b_1 x
\]
Multiple Logistic Regression Analysis

- Outcome is dichotomous (1=event, 0=non-event) and \( p = P(\text{event}) \)
- Outcome is modeled as log odds

\[
\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p
\]
Example
Logistic Regression Analysis

Study to assess the relationship between obesity and incident CVD.

\[
\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -2.367 + 0.658 \text{ (Obesity)}
\]

\[
\hat{OR} = \exp(0.658) = 1.93
\]

\[
\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = -2.592 + 0.415 \text{ (Obesity)} + 0.655 \text{ (Age)}
\]

\[
\hat{OR} = \exp(0.415) = 1.52
\]
Estimation and Interpretation of Regression Coefficients

• Model parameters are estimated using maximum likelihood techniques
• $b_1$ is the log odds ratio
• $\exp(b_1)$ is the odds ratio estimate from a logistic regression model

Interpretation of Odds Ratio (OR):

**Unadjusted:**
Unadjusted OR for obesity is 1.93. Hence, we can conclude that the odds of having CVD for obese individuals is 93% greater as compared to non-obese.

**Adjusted:**
Age adjusted OR for obesity is 1.52. Hence, we can conclude that adjusted for age, odds of having CVD for obese individuals is 52% greater as compared to non-obese.

**Continuous predictor:**
OR from continuous predictor age is $\exp(0.655)= 1.93$. Hence, we can conclude that for every one year increase in age, the odds of having CVD increases by 93%
Tips to remember when interpreting regression coefficient...

\[ b_1 = 0 \implies \text{No association between Y and X} \]

\[ b_1 > 0 \implies \text{Probability of success increases as } X \text{ increases} \]

\[ b_1 < 0 \implies \text{Probability of success decreases as } X \text{ increases} \]
Example: Political party and alcohol...

Let’s say we are asked to evaluate the association between drinking alcohol and being a republican!

Outcome variable: Republican (yes/no)

Predictor Variable: Alcohol (continuous-number of alcoholic drinks/week)
Logistic regression

• Statistical question: Does alcohol drinking predict political party?
• What is the outcome variable? Political party
• What type of variable is it? Binary
• Are the observations correlated? No
• Are groups being compared? No, our independent variable is continuous
• → logistic regression
The logistic regression model...

\[ \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X \]

Logit function = log odds of the outcome
Fun Example: political party and drinking...

Model:
Log odds of being a Republican (outcome) =
Intercept + Weekly drinks (predictor)

Fit the data in logistic regression using a statistical software...
Fitted logistic model:

“Log Odds” of being a Republican = -.09 - 1.4 *(d/wk)

Slope for drinking can be directly translated into an odds ratio:

\[ e^{-1.4} = 0.25 \]

**Interpretation:** every 1 drink more per week decreases your odds of being a Republican by 75%

Data from web.stanford.edu