Today’s lecture

Lecture 6: Dichotomous Variables & Chi-square tests

Sandy Eckel
seckel@jhsph.edu

29 April 2008

Proportions and $2 \times 2$ tables

<table>
<thead>
<tr>
<th>Population</th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 1</td>
<td>$x_1$</td>
<td>$n_1 - x_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Population 2</td>
<td>$x_2$</td>
<td>$n_2 - x_2$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$x_1 + x_2$</td>
<td>$n - (x_1 + x_2)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- Row 1 shows results of a binomial experiment with $n_1$ trials
- Row 2 shows results of a binomial experiment with $n_2$ trials

How do we compare these proportions

- Often, we want to compare $p_1$, the probability of success in population 1, to $p_2$, the probability of success in population 2
  - Usually: “Success” = Disease
  - Population 1 = Treatment 1
  - Population 2 = Treatment 2 (maybe placebo)

- How do we compare these proportions?
  - We’ve talked about comparing proportions by looking at their difference
  - But sometimes we want to look at one proportion ‘relative’ to the other
  - This approach depends on the type of study the data came from
Observational epidemiologic study designs

- Cross-sectional
- Cohort
- Case-control

Cross-sectional study design

- A ‘snapshot’ of a sample of the population
- Commonly a survey/questionnaire or one time visit to assess information at a single point in time
- Measures existing disease and exposure levels
- Difficult to argue for a causal relation between exposure and disease
- **Example:** A group of Finnish residents between the ages of 25 and 64 are mailed a questionnaire asking about their daily coffee consumption habits and their systolic BP from their last visit to the doctor.

Cohort study design

- Find a group of individuals without the disease and separate into those
  - with the exposure and
  - without the exposure
- Follow over time and measure the disease rates in both groups
- Compare the disease rate in the exposed and unexposed
- If the exposure is harmful and associated with the disease, we would expect to see higher rates of disease in the exposed group versus the unexposed group
- Allows us to estimate the incidence of the disease (rate at which new disease cases occur)
- **Example:** A group of Finnish residents between the ages of 25 and 64 and currently without hypertension are asked about their current levels of daily coffee consumption. These individuals are followed for 5 years at which point they are asked if they have been diagnosed with hypertension.

Case-control study design

- Identify individuals
  - with the disease of interest (case) and
  - those without the disease (control)
  and then look retrospectively (at prior records) to find what the exposure levels were in these two groups
- The goal is to compare the exposure levels in the case and control groups
- If the exposure is harmful and associated with the disease, we would expect to see higher levels of exposure in the cases than in the controls
- Very useful when the disease is rare
- **Example:** A group of Finnish residents with hypertension and a group of Finnish residents without hypertension are identified. These individuals are contacted and asked about their level of daily coffee consumption in the two years prior to diagnosis with hypertension.
(Almost) Cohort Study Example

Aceh Vitamin A Trial
Exposure levels (Vitamin A) assigned at baseline and then the children are followed to determine survival in the two groups.

- 25,939 pre-school children in 450 Indonesian villages in northern Sumatra
- 200,000 IU vitamin A given 1-3 months after the baseline census, and again at 6-8 months
- Consider 23,682 out of 25,939 who were visited on a pre-designed schedule

References:

Trial Outcome: 12 month mortality

<table>
<thead>
<tr>
<th>Vit A</th>
<th>Alive at 12 months?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>46</td>
</tr>
<tr>
<td>No</td>
<td>74</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
</tr>
</tbody>
</table>

Does Vitamin A reduce mortality?
- Calculate risk ratio or “relative risk”
  - Relative Risk abbreviated as RR
  - Could also compare difference in proportions: called “attributable risk”

Relative Risk Calculation

\[
\text{Relative Risk} = \frac{\text{Rate with Vitamin A}}{\text{Rate without Vitamin A}} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{46/12,048}{74/11,514} = 0.0038 / 0.0064 = 0.59
\]

- The death rate with vitamin A is 0.60 times (or 60% of) the death rate without vitamin A.
- Equivalent interpretation: Vitamin A group had 40% lower mortality than without vitamin A group!

Confidence interval for RR

- Step 1: Find the estimate of the log RR
  \[\log(\hat{p}_1/\hat{p}_2)\]
- Step 2: Estimate the variance of the log(RR) as:
  \[\frac{1 - \hat{p}_1}{n_1\hat{p}_1} + \frac{1 - \hat{p}_2}{n_2\hat{p}_2}\]
- Step 3: Find the 95% CI for log(RR):
  \[\log(RR) \pm 1.96 \cdot SD(\log RR) = (\text{lower}, \text{upper})\]
- Step 4: Exponentiate to get 95% CI for RR;
  \[e(\text{lower}, \text{upper})\]
Confidence interval for RR from Vitamin A Trial

95% CI for log relative risk is:

\[
\log(RR) \pm 1.96 \cdot SD(\log RR)
\]

\[
= \log(0.59) \pm 1.96 \cdot \sqrt{\frac{0.9962}{46} + \frac{0.9936}{74}}
\]

\[
= -0.53 \pm 0.37
\]

\[
= (-0.90, -0.16)
\]

95% CI for relative risk

\[
(e^{-0.90}, e^{-0.16}) = (0.41, 0.85)
\]

Does this confidence interval contain 1?

What if the data were from a case-control study?

- Recall: in case-control studies, individuals are selected by outcome status
- Disease (mortality) status defines the population, and exposure status defines the success
- \( p_1 \) and \( p_2 \) have a different interpretation in a case-control study than in a cohort study
- Cohort:
  - \( p_1 = P(\text{Disease} | \text{Exposure}) \)
  - \( p_2 = P(\text{Disease} | \text{No Exposure}) \)
- Case-Control:
  - \( p_1 = P(\text{Exposure} | \text{Disease}) \)
  - \( p_2 = P(\text{Exposure} | \text{No Disease}) \)

⇒ This is why we cannot estimate the relative risk from case-control data!

The Odds Ratio

- The odds ratio measures association in Case-Control studies
- Odds = \( \frac{P(\text{event occurs})}{P(\text{event does not occur})} = \frac{\hat{p}}{1-\hat{p}} \)
- Remember: the odds ratio is simply a ratio of odds!
- \( OR = \frac{\text{odds in group } 1}{\text{odds in group } 2} \)
- \( OR = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} \)

Which \( p_1 \) and \( p_2 \) do we use?

- We can actually calculate OR using either “case-control” or “cohort” set up
- Using “case-control” \( p_1 \) and \( p_2 \) where we condition on disease or no disease

\[
OR = \frac{(46/120)/(74/120)}{(12048/23562)/(11514/23562)} = \frac{46/74}{12048/11514} = 0.59
\]

- Using “cohort” \( p_1 \) and \( p_2 \) where we condition on exposure or no exposure

\[
OR = \frac{(46/12094)/(12048/12094)}{(74/11588)/(11514/11588)} = \frac{46/12048}{74/11514} = 0.59
\]

- We get the same answer either way!
The relative risk cannot be estimated from a case-control study
- The odds ratio can be estimated from a case-control study
- The OR estimates the RR when the disease is rare in both groups
- The OR is invariant to cohort or case-control designs, the RR is not
- We are introducing the OR now because it is an essential idea in logistic regression

Confidence interval for OR
- Step 1: Find the estimate of the log OR
  \[ \log \left( \frac{\hat{p}_1/(1 - \hat{p}_1)}{\hat{p}_2/(1 - \hat{p}_2)} \right) \]
- Step 2: Estimate the variance of the log(OR) as:
  \[ \frac{1}{n_1 p_1} + \frac{1}{n_1 q_1} + \frac{1}{n_2 p_2} + \frac{1}{n_2 q_2} \]
- Step 3: Find the 95% CI for log(OR):
  \[ \log(OR) \pm 1.96 \cdot SD(\log OR) = (\text{lower}, \text{upper}) \]
- Step 4: Exponentiate to get 95% CI for OR;
  \[ e^{(\text{lower}, \text{upper})} \]

The \( \chi^2 \) distribution
- Derived from the normal distribution
  \[ \chi_1^2 = \frac{(Y - \mu)^2}{\sigma^2} = Z^2 \]
  \[ \chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \]
  where \( Z_1, \ldots, Z_k \) are all standard normal random variables
- \( k \) denotes the degrees of freedom
- A \( \chi_k^2 \) random variable has
  - mean = \( k \)
  - variance = \( 2k \)
- Since a normal random variable can take on values in the interval \((-\infty, \infty)\), a chi-square random variable can take on values in the interval \((0, \infty)\)
**χ² Family of Distributions**

We generally use only a one-sided test for the χ² distribution.

Area under the curve to the right of the cutoff for each curve is 0.05.

Increasing critical value with increasing number of degrees of freedom.

**χ² Critical Values**

- Increasing critical value with increasing number of degrees of freedom.

**χ² Table**

For χ² random variables with degrees of freedom = df we’ll use

1. `pchisq(a, df, lower.tail=F)` to find $P(\chi_{df} \geq a) = ?$

2. `qchisq(b, df, lower.tail=F)` to find $P(\chi_{df} \geq ?) = b$
\( \chi^2 \) Goodness-of-Fit Test

Determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that the sample was drawn from a population with a specified distributional form, i.e.

- Normal
- Binomial
- Poisson
- etc...

Here, the expected cell counts would be derived from the distributional assumption under the null hypothesis.

The \( \chi^2 \) test statistic

\[
\chi^2 = \sum_{i=1}^{k} \left( \frac{(O_i - E_i)^2}{E_i} \right)
\]

where

- \( O_i \) = \( i^{th} \) observed frequency
- \( E_i \) = \( i^{th} \) expected frequency in the \( i^{th} \) cell of a table
- Degrees of freedom = (# categories − 1)

Note: This test is based on frequencies (cell counts) in a table, not proportions.

Example: Handgun survey I

- Survey 200 adults regarding handgun bill:
  - Statement: “I agree with a ban on handguns”
  - Four categories: Strongly agree, agree, disagree, strongly disagree
- Can one conclude that opinions are equally distributed over four responses?

Example: Handgun survey II

<table>
<thead>
<tr>
<th>Response (count)</th>
<th>1 Strongly agree</th>
<th>2 agree</th>
<th>3 disagree</th>
<th>4 Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responding ((O_i))</td>
<td>102</td>
<td>30</td>
<td>60</td>
<td>8</td>
</tr>
<tr>
<td>Expected ((E_i))</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum_{i=1}^{k} \left( \frac{(O_i - E_i)^2}{E_i} \right) = \left( \frac{(102 - 50)^2}{50} \right) + \left( \frac{(30 - 50)^2}{50} \right) + \left( \frac{(60 - 50)^2}{50} \right) + \left( \frac{(8 - 50)^2}{50} \right) = 99.36
\]

\( df = 4 - 1 = 3 \)
Example: Handgun survey III

- Critical value: \( \chi^2_{0.05} = \chi^2_{3.05} = 7.81 \)
- Since 99.36 > 7.81, we conclude that our observation was unlikely by chance alone (\( p < 0.05 \))
- Based on these data, opinions do not appear to be equally distributed among the four responses

\( \chi^2 \) Test of Independence I

- Test the null hypothesis that two criteria of classification are independent
- \( r \times c \) contingency table

\begin{tabular}{cccc|c}
 & 1 & 2 & 3 & \text{Total} \\
\hline
1 & \( n_{11} \) & \( n_{12} \) & \( n_{13} \) & \( \ldots \) & \( n_{1c} \) & \( n_1 \) \\
2 & \( n_{21} \) & \( n_{22} \) & \( n_{23} \) & \( \ldots \) & \( n_{2c} \) & \( n_2 \) \\
3 & \( n_{31} \) & \( n_{32} \) & \( n_{33} \) & \( \ldots \) & \( n_{3c} \) & \( n_3 \) \\
& \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r & \( n_{r1} \) & \( n_{r2} \) & \( n_{r3} \) & \( \ldots \) & \( n_{rc} \) & \( n_r \) \\
\hline
\text{Total} & \( n_1 \) & \( n_2 \) & \( n_3 \) & \( \ldots \) & \( n_c \) & \( n \) \\
\end{tabular}

\( \chi^2 \) Test of Independence II

- Test statistic:
  \[ \chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]
- Degrees of freedom = \((r - 1)(c - 1)\)
  where \( r \) is the number of rows and \( c \) is number of columns
- Assume the marginal totals are fixed

\( \chi^2 \) Test of Homogeneity (No association)

- Test the null hypothesis that the samples are drawn from populations that are homogenous with respect to some factor
  i.e. no association between group and factor
- Same test statistic as \( \chi^2 \) test of independence

- Test statistic:
  \[ \chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \]
- Degrees of freedom = \((r - 1)(c - 1)\)
  where \( r \) is the number of rows and \( c \) is number of columns
Example: Treatment response I

$\chi^2$ Test of Homogeneity (No association)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed Numbers</strong></td>
<td>A</td>
<td>37</td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>17</td>
<td>53</td>
<td>70</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>54</td>
<td>66</td>
<td>120</td>
</tr>
</tbody>
</table>

- Test $H_0$ that there is no association between the treatment and response
- Calculate what numbers of “Yes” and “No” would be expected assuming the probability of “Yes” was the same in both treatment groups
- Condition on total the number of “Yes” and “No” responses

Example: Treatment response II

- Expected proportion with “Yes” response = $\frac{54}{120} = 0.45$
- Expected proportion with “No” response = $\frac{66}{120} = 0.55$

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Yes (Expected)</th>
<th>No (Expected)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed</strong></td>
<td>A</td>
<td>37 (22.5)</td>
<td>13 (27.5)</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>17 (31.5)</td>
<td>53 (38.5)</td>
<td>70</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>54</td>
<td>66</td>
<td>120</td>
</tr>
</tbody>
</table>

- Get expected number of Yes responses on treatment A: $\frac{54}{120} \times 50 = 0.45 \times 50 = 22.5$
- Using a similar approach you get the other expected numbers

Example: Treatment response III

Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

\[
= \frac{(37 - 22.5)^2}{22.5} + \frac{(13 - 27.5)^2}{27.5} + \frac{(17 - 31.5)^2}{31.5} + \frac{(53 - 38.5)^2}{38.5}
\]

\[
= 29.1
\]

- Degrees of freedom = $(r-1)(c-1) = (2-1)(2-1) = 1$
- Critical value for $\alpha = 0.001$ is 10.82 so we see $p<0.001$
- Reject the null hypothesis, and conclude that the treatment groups are not homogenous (similar) with respect to response
- Response appears to be associated with treatment

Lecture 6 summary

- Dichotomous variables
  - cohort studies - relative risk
  - case-control - odds ratios
- Chi-square tests

Next time, we’ll talk about analysis of variance (ANOVA)