Contingency Table Probability Structure

Dipankar Bandyopadhyay, Ph.D.

Department of Biostatistics,
Virginia Commonwealth University

Overview

- Over the next few lectures, we will examine the 2×2 contingency table
- Some authors refer to this as a "four fold table"
- We will consider various study designs and their impact on the summary measures of association

Rows Fixed: Product Binomial Case - [Prospective]

Question of interest: Does treatment affect outcome?

```
                        OUTCOME
                             NO
                  |COLD    | COLD   |  Total
T        ---------+--------+--------+
R        VITAMIN  |        |        |
E           C     |   17   |   122  |  139  <--|
A                 |        |        |          |  (fixed
T        ---------+--------+--------+          |-- by
M           NO    |        |        |          |  design)
E        VITAMIN  |   31   |   109  |  140  <--|
N           C     |        |        |
T        ---------+--------+--------+

         Total        48       231     279
```

Columns Fixed: Also Product Binomial - [Retrospective]

Question of interest: Does alcohol exposure vary among cases and controls?

```
                        STATUS
                |CASE     |CONTROL |   Total
A      ---------+---------+--------+
L      80+      |         |        |
C      (gm/day) |    96   |   109  |   205
O               |         |        |
H      ---------+---------+--------+
O      0-79     |         |        |
L      (gm/day) |   104   |   666  |   770
                |         |        |
       ---------+---------+--------+
       Total        200       775      975
                     ^         ^
                     |         |
```

N Fixed: Multinomial Case - [Cross-Sectional]

Question: Is there an association among cancer stage and smoking status?

		CANCER STAGE			
		SPREAD	NOT SPREAD	Total	
SMOKE	YES	300	600	900	
	NO	500	3000	3500	
	Total	800	3600	4400	<---(fixed by design)

Rows and Columns Fixed: Hypergeometric Case

Question of Interest: Is there gender bias in juror selection?

```
                     SELECTED
                     FOR JURY

                  |YES     |NO      |  Total
G        ---------+--------+--------+
E                 |        |        |
N        FEMALE   |    1   |    9   |  10
D                 |        |        |
E        ---------+--------+--------+
R                 |        |        |
         MALE     |   11   |    9   |  20
                  |        |        |
         ---------+--------+--------+
         Total        12       18     30
```

This distribution is will be used in Fisher's Exact testing.

Prospective Studies

We are going to begin examining contingency tables first by looking at prospective studies.

- Number on each treatment (or experimental) arm fixed by design.
- Rows are independent binomials.
- Question of interest: Does treatment affect outcome ?
- Usually the design for Experimental Studies, Clinical Trials.

In general, the 2 × 2 table is written as

		Outcome		
		1	2	
Treatment	1	Y_1	$n_1 - Y_1$	n_1
	2	Y_2	$n_2 - Y_2$	n_2

Facts about the distribution

- n_1 and n_2 are fixed by design
- Y_1 and Y_2 are independent with distributions:

$$Y_1 \sim Bin(n_1, p_1)$$

$$Y_2 \sim Bin(n_2, p_2)$$

- The distribution is the product of 2 independent binomials; often called the 'product binomial':

$$\begin{aligned} P(y_i, y_2|p_1, p_2) &= P(Y_1 = y_1|p_1)P(Y_2 = y_2|p_2) \\ &= \left(\begin{array}{c} n_1 \\ y_1 \end{array}\right)\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p_1^{y_1}(1-p_1)^{n_1-y_1} p_2^{y_2}(1-p_2)^{n_2-y_2} \end{aligned}$$

Question of interest (all the same)

- Does treatment affect outcome ?
- Are treatment and outcome associated ?
- Is the probability of success the same on both treatments ?
- How do we quantify treatment differences?
- Also, what test statistics can we use for

$$\mathrm{H}_0 : p_1 = p_2 = p$$

and the alternative is

$$\mathrm{H}_A : p_1 \neq p_2$$

MLE and estimated SEs of treatment differences

- To estimate these treatment differences, we must estimate the success probabilities p_1 and p_2.
- Intuitively, thinking of the two groups separately, the MLE's should be the proportion of successes in the two groups, i.e.,

$$\widehat{p}_1 = \frac{Y_1}{n_1}$$

and

$$\widehat{p}_2 = \frac{Y_2}{n_2}.$$

- However, we will derive these based on the likelihood of the product binomial.

The Likelihood for (p_1, p_2) is the product binomial distribution of (y_1, y_2, p_1, p_2).

$$\begin{aligned} L(p_1, p_2) &= P(Y_1 = y_1 | p_1) P(Y_2 = y_2 | p_2) \\ &= \left(\begin{array}{c} n_1 \\ y_1 \end{array} \right) \left(\begin{array}{c} n_2 \\ y_2 \end{array} \right) p_1^{y_1} (1 - p_1)^{n_1 - y_1} p_2^{y_2} (1 - p_2)^{n_2 - y_2} \end{aligned}$$

Then the log-likelihood is the sum of the two pieces,

$$\log L(p_1, p_2) = \\ \log\left[\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) p_1^{y_1}(1-p_1)^{n_1-y_1}\right] + \log\left[\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p_2^{y_2}(1-p_2)^{n_2-y_2}\right]$$

Similar to before, to find the MLE, we set the partial derivatives of $\log L(p_1, p_2)$ with respect to p_1 and p_2 to 0, and solve for $\widehat{p}_1$ and $\widehat{p}_2$:

Note: Agresti (and most statisticians) simply denote the natural logarithm as log instead of the ln as you would see in mathematics or physics. In this class, all references of log are consider the log to base e.

Now,

$$\log L(p_1, p_2) =$$

$$\log\left[\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) p_1^{y_1}(1-p_1)^{n_1-y_1}\right] + \log\left[\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p_2^{y_2}(1-p_2)^{n_2-y_2}\right]$$

The derivative of the log-likelihood with respect to p_1 is

$$\begin{aligned} \frac{d\log L(p_1,p_2)}{dp_1} &= \frac{d}{dp_1}\log\left[\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) p_1^{y_1}(1-p_1)^{n_1-y_1}\right] + \\ &\qquad \frac{d}{dp_1}\log\left[\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p_2^{y_2}(1-p_2)^{n_2-y_2}\right] \\ &= \frac{d}{dp_1}\log\left[\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) p_1^{y_1}(1-p_1)^{n_1-y_1}\right] + 0 \end{aligned}$$

since the the second part is not a function of p_1.

Note, though,

$$\frac{d \log L(p_1, p_2)}{dp_1} = \frac{d}{dp_1} \log \left[\begin{pmatrix} n_1 \\ y_1 \end{pmatrix} p_1^{y_1} (1 - p_1)^{n_1 - y_1} \right]$$

is just the derivative of a binomial log-likelihood with respect to its parameter p_1. From before, we have

$$\widehat{p}_1 = \frac{y_1}{n_1}$$

To explicitly show this, in the single binomial section, we showed that

$$\frac{d \log L(p_1, p_2)}{dp_1} = \frac{d}{dp_1} \log \left[\begin{pmatrix} n_1 \\ y_1 \end{pmatrix} p_1^{y_1} (1 - p_1)^{n_1 - y_1} \right] = \frac{y_1 - n_1 p_1}{p_1(1 - p_1)}$$

Similarly,

$$\frac{d \log L(p_1, p_2)}{dp_2} = \frac{y_2 - n_2 p_2}{p_2(1 - p_2)}$$

Then, the MLE's are found by simultaneously solving

$$\frac{d \log L(p_1, p_2)}{dp_1} = \frac{y_1 - n_1\widehat{p}_1}{\widehat{p}_1(1 - \widehat{p}_1)} = 0$$

and

$$\frac{d \log L(p_1, p_2)}{dp_2} = \frac{y_2 - n_2\widehat{p}_2}{\widehat{p}_2(1 - \widehat{p}_2)} = 0$$

which gives

$$\widehat{p}_1 = \frac{y_1}{n_1}$$

and

$$\widehat{p}_2 = \frac{y_2}{n_2}.$$

provided that $\widehat{p}_1, \widehat{p}_2 \neq 0, 1$

Since Y_1 and Y_2 are independent binomials we know that

$$Var(\widehat{p}_1) = \frac{p_1(1 - p_1)}{n_1}$$

and

$$Var(\widehat{p}_2) = \frac{p_2(1 - p_2)}{n_2}$$

Estimating treatment differences

To obtain the MLE of the log-odds ratio, we just plug $\widehat{p}_1$ and $\widehat{p}_2$ in to get

$$\begin{aligned} \log(\widehat{OR}) &= \log\left(\frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}\right) \\ &= \text{logit}(\widehat{p}_1) - \text{logit}(\widehat{p}_2) \end{aligned}$$

Now, suppose we want to estimate the variance of $\log(\widehat{OR})$.
Since the treatment groups are independent, $\text{logit}(\widehat{p}_1)$ and $\text{logit}(\widehat{p}_2)$ are independent, so that

$$Cov[\text{logit}(\widehat{p}_1), \text{logit}(\widehat{p}_2)] = 0,$$

The variance of differences of independent random variables is

$$\begin{aligned} Var[\log(\widehat{OR})] &= Var[\text{logit}(\widehat{p}_1) - \text{logit}(\widehat{p}_2)] \\ &= Var[\text{logit}(\widehat{p}_1)] + Var[\text{logit}(\widehat{p}_2)] \end{aligned}$$

Delta Method approximation

- The $Var[\log(\widehat{OR})]$ can be approximated by the delta method
- To do so we need to calculate

$$\begin{aligned}\frac{d}{d\,p}\left[\log(p) - \log(1-p)\right] &= \frac{1}{p} - \frac{-1}{1-p} \\ &= \frac{1}{p(1-p)}\end{aligned}$$

- Therefore,

$$\begin{aligned}\mathrm{Var}\left(\log(\tfrac{p}{1-p})\right) &= \left(\frac{1}{p(1-p)}\right)^2 \frac{p(1-p)}{n} \\ &= \frac{1}{np(1-p)} \\ &= \frac{1}{np} + \frac{1}{n(1-p)}\end{aligned}$$

- Using these results from the Delta Method, we have

$$Var[\mathrm{logit}(\widehat{p}_1)] = \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)}$$

and

$$Var[\mathrm{logit}(\widehat{p}_2)] = \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}$$

Then,

$$
\begin{aligned}
Var[\log(\widehat{OR})] &= Var[\text{logit}(\widehat{p}_1)] + Var[\text{logit}(\widehat{p}_2)] \\
&= \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}
\end{aligned}
$$

which we estimate by replacing p_1 and p_2 with $\widehat{p}_1$ and $\widehat{p}_2$,

$$
\begin{aligned}
\widehat{Var}[\log(\widehat{OR})] &= \frac{1}{n_1 \widehat{p}_1} + \frac{1}{n_1(1-\widehat{p}_1)} + \frac{1}{n_2 \widehat{p}_2} + \frac{1}{n_2(1-\widehat{p}_2)} \\
&= \frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}
\end{aligned}
$$

Note: This is the same result we obtained in the previous lecture; however, in this case we assumed two independent binomial distributions.

General formula for variance of treatment difference

The MLE of a treatment difference

$$\theta = g(p_1) - g(p_2)$$

is

$$\hat{\boldsymbol{\theta}} = g(\widehat{p}_1) - g(\widehat{p}_2)$$

Also, since $\widehat{p}_1$ and $\widehat{p}_2$ are independent, so $g(\widehat{p}_1)$ and $g(\widehat{p}_2)$ are independent.

Recall, the variance of a difference of two independent random variables is

$$Var[g(\widehat{p}_1) - g(\widehat{p}_2)] = Var[g(\widehat{p}_1)] + Var[g(\widehat{p}_2)]$$

Then, to obtain the large sample variance, we can apply the delta method to $g(\widehat{p}_1)$ to get $Var[g(\widehat{p}_1)]$ and to $g(\widehat{p}_2)$ to get $Var[g(\widehat{p}_2)]$ and then sum the two.

The results are summarized in the following table:

TREATMENT DIFFERENCE	ESTIMATE	Var(ESTIMATE)
RISK DIFF	$\widehat{p}_1 - \widehat{p}_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
log (RR)	$\log\left(\frac{\widehat{p}_1}{\widehat{p}_2}\right)$	$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$
log (OR)	$\log\left(\frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}\right)$	$\left[\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)}\right] + \left[\frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right]$

ESTIMATES of Standard Error, and LARGE SAMPLE CONFIDENCE INTERVALS

To estimate the variances, we can replace p_1 and p_2 with $\widehat{p}_1$ and $\widehat{p}_2$.

$$\widehat{Var}(\widehat{p}_1 - \widehat{p}_2) = \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2};$$

$$\widehat{Var}[\log(\widehat{RR})] = \frac{1-\widehat{p}_1}{n_1\widehat{p}_1} + \frac{1-\widehat{p}_2}{n_2\widehat{p}_2};$$

$$\begin{aligned}\widehat{Var}[\log(\widehat{OR})] \quad &= \quad \tfrac{1}{n_1\widehat{p}_1} + \tfrac{1}{n_1(1-\widehat{p}_1)} + \tfrac{1}{n_2\widehat{p}_2} + \tfrac{1}{n_2(1-\widehat{p}_2)} \\ &= \quad \tfrac{1}{y_1} + \tfrac{1}{n_1-y_1} + \tfrac{1}{y_2} + \tfrac{1}{n_2-y_2}\end{aligned}$$

Then, large sample 95% confidence interval for treatment differences can be obtained via

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

$$\log(\widehat{RR}) \pm 1.96\sqrt{\frac{1-\widehat{p}_1}{n_1\widehat{p}_1} + \frac{1-\widehat{p}_2}{n_2\widehat{p}_2}}$$

and

$$\log(\widehat{OR}) \pm 1.96\sqrt{\frac{1}{y_1} + \frac{1}{n_1-y_1} + \frac{1}{y_2} + \frac{1}{n_2-y_2}}$$

Confidence Interval for OR and RR

- You want a confidence interval for RR or OR that is assured to be in the interval $(0, \infty)$.
- Similar to what we did for a confidence interval for p, it is first better to get confidence intervals for $\log(RR)$ or $\log(OR)$, and to exponentiate the endpoints : i.e.,

$$\exp\{\log(\widehat{OR}) \pm 1.96\sqrt{\widehat{Var}[\log(\widehat{OR})]}\},$$

and

$$\exp\{\log(\widehat{RR}) \pm 1.96\sqrt{\widehat{Var}[\log(\widehat{RR})]}\},$$

Example: MI example

- Suppose clinical trial participants are randomized to either Placebo or Aspirin
- The subjects are followed prospectively for 5 years to determine whether or not an MI (or heart attack) occurs
- The following table summarizes the results

	Myocardial Infarction		
	Heart or Attack	No Attack	Total per Arm
Placebo	189	10845	11,034
Aspirin	104	10933	11,037

  - About 11035 randomized to each treatment
  - Overall probability of heart attack in Doctors is low

$$\frac{293}{22071} = 1.33\%$$

  The disease is ‘rare’.

Estimates and Test Statistics

The test statistics for $\text{H}_0 : p_1 = p_2$ versus $\text{H}_A : p_1 \neq p_2$

Parameter	Estimate	Estimated Standard Error	$Z-$Statistic (Est/SE)
RISK DIFF	.0077	.00154	5.00
log(RR)	.598 (RR=1.818)	.1212	4.934
log(OR)	.605 (OR=1.832)	.1228	4.927

- In each case, we reject the null, and the $Z-$statistic is about 5.
- The WALD test statistics using the Risk Difference, log OR, and log RR are slightly different.

Confidence Intervals Creation

- The following are the 95% confidence intervals

Parameter	Estimate	95% C.I.
RISK DIFF	.0077	[.0047,.0107]
RR	1.818	[1.433,2.306]
OR	1.832	[1.440,2.331]

- For the OR and RR, we exponentiated the 95% confidence intervals for the $\log(OR)$ and $\log(RR)$, respectively.
- None of the confidence intervals contain the null value for no association (0 for the RISK DIFFERENCE, 1 for the OR and RR).

Interpretation

- The **risk difference** has the interpretation that the 'Excess Risk' of a heart attack on Placebo is .0077. This 'fraction' is not very meaningful for rare diseases, but stated in terms of subjects, we can say that we would expect 77 more heart attacks in 10000 placebo subjects than in 10000 aspirin users.
- The **relative risk** has the interpretation that Individuals on Placebo have almost twice (1.8) the risk (or probability) of a heart attack than individuals on Aspirin
- The **odds ratio** has the interpretation that Individuals on Placebo have almost twice (1.8) the odds of a heart attack versus no heart attack than individuals on Aspirin

Relationship between OR and RR

- Recall, $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \left(\frac{p_1}{p_2}\right)\left[\frac{1-p_2}{1-p_1}\right] \\ &= RR\left[\frac{1-p_2}{1-p_1}\right] \end{aligned}$$

- When the disease is rare (in the example, $\widehat{p}_2 < \widehat{p}_1 < 2\%$),

$$\left[\frac{1-p_2}{1-p_1}\right] \approx \frac{1}{1} = 1; \quad and \quad OR \approx RR.$$

- In the example, $\widehat{OR} = 1.832$, $\widehat{RR} = 1.818$; i.e., they are almost identical.

LRT

- Now, we want to test the null hypothesis

$$\text{H}_0\text{:}p_1 = p_2 = p$$

  versus the alternative

$$\text{H}_A\text{:}p_1 \neq p_2$$

  with the likelihood ratio statistic (the likelihood ratio statistic generally has a two-sided alternative, i.e., it is χ^2 based).
- The general likelihood ratio statistic involves the estimate of $p_1 = p_2 = p$ under the null and (p_1, p_2) under the alternative.
- Thus, unlike the simple single binomial sample we discussed earlier, in which the null was $\text{H}_0\text{:}p = .5$, the parameters are not completely specified under the null. i.e., we must still estimate a common p under the null for the likelihood ratio.

General Likelihood Ratio Statistic

- The likelihood is a function of the parameter vector $\mathbf{p} = [p_1, p_2]'$.
- In large samples, it can be shown that

$$2\log\left\{\frac{L(\widehat{p}_1,\widehat{p}_2|\mathrm{H}_A)}{L(\tilde{p}_1,\tilde{p}_2|\mathrm{H}_0)}\right\} =$$

$$2[\log L(\widehat{p}_1,\widehat{p}_2|\mathrm{H}_A) - \log L(\tilde{p}_1,\tilde{p}_2|\mathrm{H}_0)] \sim \chi^2_{df}$$

- where $L(\widehat{p}_1,\widehat{p}_2|\mathrm{H}_A)$ is the likelihood after replacing $[p_1, p_2]$ by its estimate, $[\widehat{p}_1,\widehat{p}_2]$ under H_A, and

$$L(\tilde{p}_1,\tilde{p}_2|\mathrm{H}_0)$$

  is the likelihood after replacing $[p_1, p_2]$ by its estimate, $[\tilde{p}_1,\tilde{p}_2]$, under H_0. (In our case, $[\tilde{p}_1,\tilde{p}_2] = [\widehat{p},\widehat{p}]'$ since $p_1 = p_2 = p$ under the null).
- The degrees-of-freedom df is the difference in the number of parameters estimated under the alternative and null (In our example, $df = 2 - 1 = 1$).

MLE under the Null

- Thus, to use the likelihood ratio statistic, we need to estimate the common p under the null hypothesis.
- When $\mathrm{H}_0 : p_1 = p_2 = p$,
  $$E(Y_1) = n_1 p$$
  and
  $$E(Y_2) = n_2 p$$
- Then,
  $$E(Y_1 + Y_2) = E(Y_1) + E(Y_2) = n_1 p + n_2 p = (n_1 + n_2)p$$
- The 'pooled' estimate of p is
  $$\widehat{p} = \left(\frac{Y_1 + Y_2}{n_1 + n_2}\right) = \left(\frac{\text{total \# successes}}{\text{total sample size}}\right)$$
  which is unbiased and the MLE.
- Intuitively, when the probability of success is the same on both treatments, the best estimate (MLE) of p is obtained by pooling over the treatments.

Using the likelihood to obtain the MLE under the null

- Under the null H_0:$p_1 = p_2 = p$, the MLE of p is obtained from the likelihood

$$
\begin{aligned}
L(p) &= \left(\begin{array}{c} n_1 \\ y_1 \end{array}\right)\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p^{y_1}(1-p)^{n_1-y_1} p^{y_2}(1-p)^{n_2-y_2} \\
&= \left(\begin{array}{c} n_1 \\ y_1 \end{array}\right)\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) p^{y_1+y_2}(1-p)^{(n_1+n_2)-(y_1+y_2)},
\end{aligned}
$$

- Then,

$$
\begin{aligned}
\frac{d \log L(p)}{dp} &= \frac{d}{dp_1} \log\left[\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right)\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right)\right] \\
&\qquad + \frac{d}{dp_1} \log[p^{y_1+y_2}(1-p)^{(n_1+n_2)-(y_1+y_2)}] \\
&= \frac{y_1 + y_2 - (n_1+n_2)p}{p(1-p)}
\end{aligned}
$$

- This is the same first derivative as a single binomial sample, in fact, under the null,

$$Y_1 + Y_2 \sim Bin(n_1 + n_2, p),$$

and it is easily shown that the solution is

$$\widehat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$$

Using the Estimates to obtain the Likelihood Ratio Statistic

- Under the alternative,

$$\widehat{p}_1 = \frac{Y_1}{n_1} \qquad \text{and} \qquad \widehat{p}_2 = \frac{Y_2}{n_2},$$

and

$$\begin{array}{l} \log[L(\widehat{p}_1, \widehat{p}_2 | \mathrm{H}_A)] = \\ \quad \log\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) + y_1 \log(\widehat{p}_1) + (n_1 - y_1)\log(1-\widehat{p}_1) + \\ \qquad \log\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) + y_2 \log(\widehat{p}_2) + (n_2 - y_2)\log(1-\widehat{p}_2) \end{array}$$

Then,

$$\log[L(\widehat{p},\widehat{p}|\mathrm{H}_0)] = \\ \log\left(\begin{array}{c} n_1 \\ y_1 \end{array}\right) + y_1\log(\widehat{p}) + (n_1 - y_1)\log(1-\widehat{p}) + \\ \log\left(\begin{array}{c} n_2 \\ y_2 \end{array}\right) + y_2\log(\widehat{p}) + (n_2 - y_2)\log(1-\widehat{p})$$

- Under the alternative, we estimate 2 parameters, under the null, we estimated 1, so $df = 2 - 1 = 1$.
- Then, we take 2 times the differences in the log-likelihoods and compare it to a chi-square with 1 df.

Simplification of the LR statistic

Then, the likelihood ratio statistic equals 2 times the difference in the log-likelihoods under the alternative and null, or

$$
\begin{aligned}
G^2 &= 2[y_1 \log\left(\frac{\widehat{p}_1}{\widehat{p}}\right) + (n_1 - y_1)\log\left(\frac{(1-\widehat{p}_1)}{(1-\widehat{p})}\right) \\
&\quad + y_2 \log\left(\frac{\widehat{p}_2}{\widehat{p}}\right) + (n_2 - y_2)\log\left(\frac{(1-\widehat{p}_2)}{(1-\widehat{p})}\right)] \\
&= 2[y_1 \log\left(\frac{y_1}{n_1\widehat{p}}\right) + (n_1 - y_1)\log\left(\frac{n_1-y_1}{n_1(1-\widehat{p})}\right) + \\
&\quad + y_2 \log\left(\frac{y_2}{n_2\widehat{p}}\right) + (n_2 - y_2)\log\left(\frac{n_2-y_2}{n_2(1-\widehat{p})}\right)] \\
&\sim \chi_1^2
\end{aligned}
$$

under the null, in large samples

‘OBSERVED’ and ‘EXPECTED’ Cell Counts

- First, let’s look at the (2×2) table of ‘OBSERVED’ Cell Counts.

		OUTCOME		
		1	2	
TRT	1	Y_1	$(n_1 - Y_1)$	n_1
	2	Y_2	$(n_2 - Y_2)$	n_2
total		$Y_1 + Y_2$	$[(n_1 + n_2)$ $-(Y_1 + Y_1)]$	$(n_1 + n_2)$

- If we look at the likelihood ratio statistic,

$$\begin{aligned} G^2 &= 2[y_1 \log\left(\frac{y_1}{n_1\widehat{p}}\right) + (n_1 - y_1)\log\left(\frac{n_1-y_1}{n_1(1-\widehat{p})}\right) + \\ &\quad + y_2 \log\left(\frac{y_2}{n_2\widehat{p}}\right) + (n_2 - y_2)\log\left(\frac{n_2-y_2}{n_2(1-\widehat{p})}\right)] \end{aligned}$$

- In the numerator of the log's, we have the observed cell counts for the 4 cells in the table.
- Sometimes, statisticians let O_{ij} denote the observed count in row i, column j,

$$O_{11} = Y_1, \;\; O_{12} = n_1 - Y_1, \;\; O_{21} = Y_2, \;\; O_{22} = n_2 - Y_2$$

- Then, we can rewrite the observed table as

		OUTCOME		
		1	2	
TRT	1	O_{11}	O_{12}	$O_{11}+O_{12}$
	2	O_{21}	O_{22}	$O_{21}+O_{22}$
total		$O_{11}+O_{21}$	$O_{12}+O_{22}$	

- We will show that the likelihood ratio statistic is often written as

$$\begin{aligned} G^2 &= 2[y_1 \log\left(\frac{y_1}{n_1\widehat{p}}\right) + (n_1 - y_1)\log\left(\frac{n_1-y_1}{n_1(1-\widehat{p})}\right) + \\ &\quad + y_2 \log\left(\frac{y_2}{n_2\widehat{p}}\right) + (n_2 - y_2)\log\left(\frac{n_2-y_2}{n_2(1-\widehat{p})}\right)] \\ &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right), \end{aligned}$$

Simple Form of the Estimated Expected Counts

- First, suppose $p_1 \neq p_2$,
- Then, the (2×2) table of expected cell counts is

		OUTCOME		
		1	2	
TRT	1	$n_1 p_1$	$n_1(1-p_1)$	n_1
	2	$n_2 p_2$	$n_2(1-p_2)$	n_2
total		$n_1 p_1 + n_2 p_2$	$[(n_1+n_2)$ $-(n_1 p_1 + n_2 p_2)]$	(n_1+n_2)

- If we look at the n_1 subjects in the first row, we expect $n_1 p_1$ subjects to have outcome 1, and $n_1(1-p_1)$ of them to have outcome 2.
- Similarly, if we look at the n_2 subjects in the second row, we expect $n_2 p_2$ subjects to have outcome 1, and $n_2(1-p_2)$ of them to have outcome 2.

- Under the null, when the probability of success is the same on both treatments, $p_1 = p_2 = p$, the table of expected counts looks like

		OUTCOME		
		1	2	
TRT	1	$n_1 p$	$n_1(1-p)$	n_1
	2	$n_2 p$	$n_2(1-p)$	n_2
total		$(n_1+n_2)p$	$[(n_1+n_2)(1-p)]$	(n_1+n_2)

- Here, if we look at the n_1 subjects in the first row, we expect $n_1 p$ subjects to have outcome 1, and $n_1(1-p)$ of them to have outcome 2.
- Similarly, if we look at the n_2 subjects in the second row, we expect $n_2 p$ subjects to have outcome 1, and $n_2(1-p)$ of them to have outcome 2.

- Under $H_0 : p_1 = p_2 = p$, the table of estimated expected counts looks like

	OUTCOME 1	 2	
TRT 1	$n_1\widehat{p}$	$n_1(1-\widehat{p})$	n_1
TRT 2	$n_2\widehat{p}$	$n_2(1-\widehat{p})$	n_2
total	$(n_1+n_2)\widehat{p}$	$[(n_1+n_2)(1-\widehat{p})]$	(n_1+n_2)

- where, recall, $\widehat{p}$ is the 'pooled' estimate of p,

$$\widehat{p} = \left(\frac{Y_1 + Y_2}{n_1 + n_2}\right) = \left(\frac{\text{total \# successes}}{\text{total sample size}}\right).$$

- These estimated expected counts are denoted E_{ij}, (i^{th} row, j^{th} column), and are found in the denominator of the likelihood ratio statistic, with

$$E_{11} = n_1\widehat{p}, \quad E_{12} = n_1(1-\widehat{p}), \quad E_{21} = n_2\widehat{p}, \quad E_{22} = n_2(1-\widehat{p})$$

Simplification of Expected Cell Counts

- Substituting

$$\widehat{p} = \frac{Y_1 + Y_2}{n_1 + n_2},$$

and

$$1 - \widehat{p} = 1 - \frac{Y_1 + Y_2}{n_1 + n_2} = \frac{(n_1 + n_2) - (Y_1 + Y_2)}{n_1 + n_2},$$

in the table, we get the E_{ij}'s,

		OUTCOME		
		1	2	
TRT	1	$\frac{n_1(Y_1+Y_2)}{n_1+n_2}$	$\frac{n_1[(n_1+n_2)-(Y_1+Y_2)]}{n_1+n_2}$	n_1
	2	$\frac{n_2(Y_1+Y_2)}{n_1+n_2}$	$\frac{n_2[(n_1+n_2)-(Y_1+Y_2)]}{n_1+n_2}$	n_2
total		$(Y_1 + Y_2)$	$[(n_1 + n_2)$ $-(Y_1 + Y_2)]$	$(n_1 + n_2)$

- From this table, you can see that

$$E_{ij} = \frac{[i^{th} \text{ row total}] \cdot [j^{th} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]}$$

Summary

- We did all this to show that

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$$

- Note that, we can also write this as

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij}[\log(O_{ij}) - \log(E_{ij})]$$

- Writing it this way, we see that the likelihood ratio measures the discrepancy between the log of the observed counts, and the log of estimated expected counts under the null; if they are similar, you would expect the statistic to be small, and the null not to be rejected.