Contingency Tables

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

September 28–October 5, 2010

Case Study

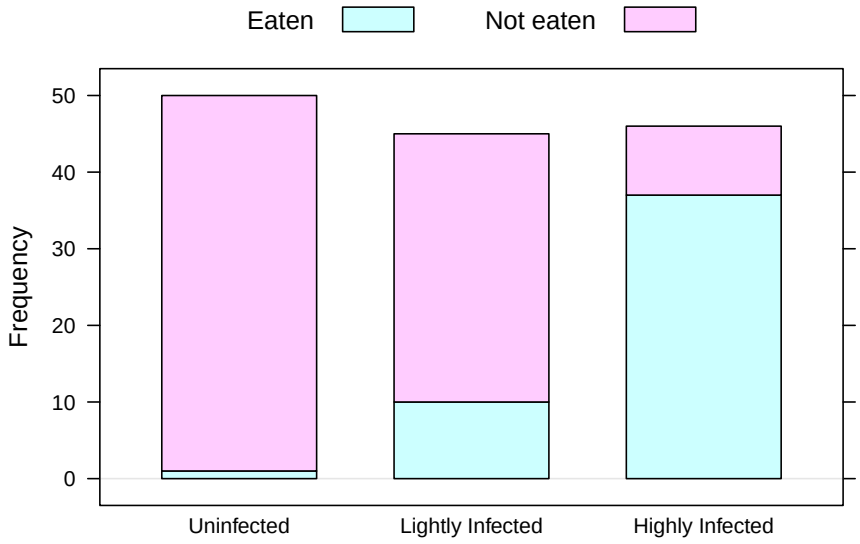
Example

Example 9.3 beginning on page 213 of the text describes an experiment in which fish are placed in a large tank for a period of time and some are eaten by large birds of prey. The fish are categorized by their level of parasitic infection, either uninfected, lightly infected, or highly infected. It is to the parasites' advantage to be in a fish that is eaten, as this provides an opportunity to infect the bird in the parasites' next stage of life. The observed proportions of fish eaten are quite different among the categories.

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

The proportions of eaten fish are, respectively, $1/50 = 0.02$, $10/45 = 0.222$, and $37/46 = 0.804$.

Stacked Bar Graph



Graphing Tabled Counts

- A stacked bar graph shows:
  - the sample sizes in each sample; and
  - the number of observations of each type within each sample.
- This plot makes it easy to compare sample sizes among samples and counts within samples, but the comparison of estimates of conditional probabilities among samples is less clear.

Mosaic Plot

Mosaic Plot

- A *mosaic plot* replaces absolute frequencies (counts) with relative frequencies within each sample.
- This plot makes comparisons of estimated conditional probabilities very clear.
- The cost is that the sample size information is lost.

Estimating Differences between Proportions

- In the setting of the experiment, we observe a difference between the proportions of eaten fish in the lightly and highly infected fish.
- A point estimate of this difference is

$$\frac{37}{46} - \frac{10}{45} = 0.804 - 0.222 = 0.582$$

- How can we quantify uncertainty in this estimate?

Confidence Intervals

- A confidence interval for a difference in proportions $p_1 - p_2$ is based on the sampling distribution of the difference in sample proportions.
- If the two samples are independent,

$$\mathrm{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- If both samples are large enough (depending on how close the proportions are to 0 or 1), this sampling distribution is approximately normal.

Confidence Interval

95%Confidence Interval for $p_1 - p_2$

A 95% confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - 1.96\text{SE}\Big(\hat{p}_1 - \hat{p}_2\Big) < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + 1.96\text{SE}\Big(\hat{p}_1 - \hat{p}_2\Big)$$

where $\hat{p}_i = x_i/n_i$ for $i = 1, 2$ and

$$\text{SE}\Big(\hat{p}_1 - \hat{p}_2\Big) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- This formula will be more accurate for large n_1 and n_2.
- A rough rule of thumb is that each sample should have at least five observations of each type.
- Maybe this method can be improved by adding fake observations like the one sample case?

Application

- For the infected fish case study, a confidence interval for the difference in probabilities of being eaten between highly and lightly infected fish is

$$0.415 < p_{\text{high}} - p_{\text{light}} < 0.749$$

(Show calculations on the board.)

In the settings of the experiment, we are 95% confident that the probability a highly infected fish is eaten is greater than the corresponding probability for a lightly infected fish by an amount between 0.415 and 0.749.

Odds Ratios

- Odds ratios are an alternative way to think about probabilities.

Definition

- The *odds* in favor of an event with probability p are $p/(1-p)$.
- The *odds ratio* in favor of an event between two groups is the odds in favor for the first group divided by the odds in favor for the second group.

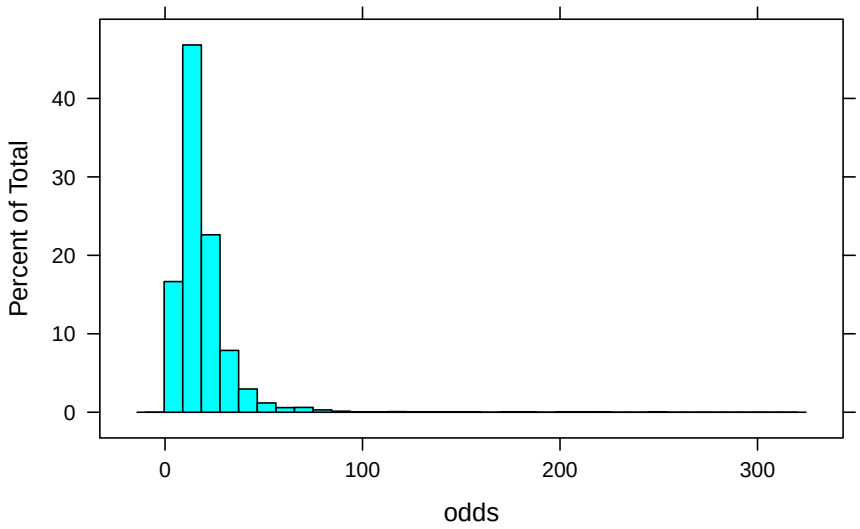
$$\text{odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- Odds ratios are estimated by plugging in sample proportions.

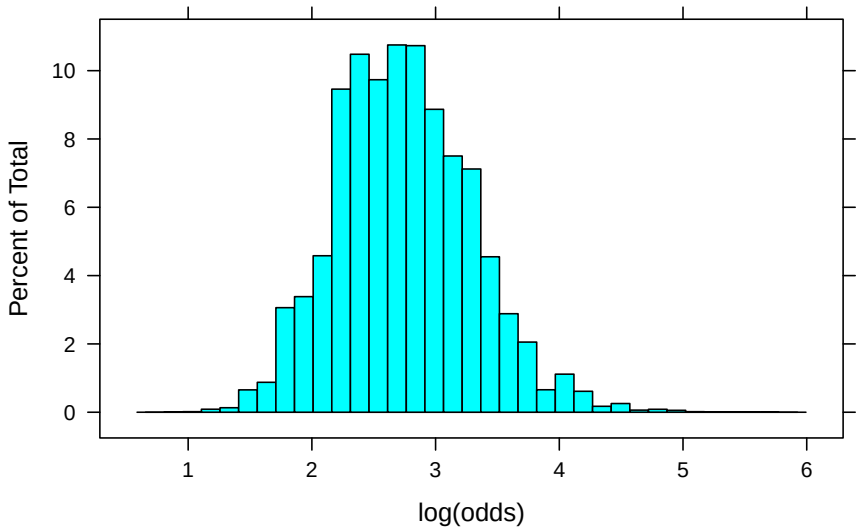
Sampling Distribution of the Odds Ratio

- We explore the sampling distribution of the odds ratio when $n_1 = 46$, $p_1 = 0.8$, $n_2 = 45$, and $p_2 = 0.22$ which are estimates from the case study.
- We simulate 100,000 odds ratios from independent samples and graph the results.

Graph of Odds Ratio



Graph of Log of Odds Ratio



Comparison

- The sampling distribution of the odds ratio is very *skewed to the right*.
- The sampling distribution of the log odds ratio is *fairly symmetric and bell-shaped*.
- We will use the normal approximation for the log odds ratio and then translate back.
- The standard error of the odds ratio can be estimated as

$$\text{SE}\big(\ln(\text{odds ratio})\big) = \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}}$$

Confidence Interval for Odds Ratio

95%Confidence Interval for $\dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$

A 95% confidence interval for the odds ratio is

$$\exp\Big(\ln\widehat{\text{OR}} - 1.96\text{SE}\big(\ln\widehat{\text{OR}}\big)\Big) < \frac{p_1/(1-p_1)}{p_2/(1-p_2)} < \exp\Big(\ln\widehat{\text{OR}} + 1.96\text{SE}\big(\ln\widehat{\text{OR}}\big)\Big)$$

where $\widehat{\text{OR}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$ and

$$\text{SE}\big(\ln\widehat{\text{OR}}\big) = \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_2} + \frac{1}{n_2 - x_2}}$$

Note the equivalent expression:

$$\widehat{\text{OR}}\exp\Big(-1.96\text{SE}\big(\ln\widehat{\text{OR}}\big)\Big) < \frac{p_1/(1-p_1)}{p_2/(1-p_2)} < \widehat{\text{OR}}\exp\Big(+1.96\text{SE}\big(\ln\widehat{\text{OR}}\big)\Big)$$

Application to the Case Study

- The estimated odds for being eaten in the highly infected group is $37/9 = 4.111$.
- The estimated odds for being eaten in the lightly infected group is $10/35 = 0.286$.
- The estimated odds ratio is 14.389 and its natural logarithm is 2.666.
- The estimated SE of the log odds ratio is

$$\sqrt{\frac{1}{37} + \frac{1}{9} + \frac{1}{10} + \frac{1}{35}} = 0.516$$

- $e^{2.666-1.96(0.516)} \doteq 5.229$ and $e^{2.666+1.96(0.516)} \doteq 39.594$.
- The 95% confidence interval is $5.229 < \text{OR} < 39.594$.

Interpretation

In the experimental setting of the infected fish case study, we are 95% confident that the odds of being eaten in the highly infected group are between 5.2 and 39.6 times higher than in the lightly infected group.

A Second Case Study

Example

Example 9.4 on page 220 describes an experiment. In Costa Rica, the vampire bat *Desmodus rotundus* feeds on the blood of domestic cattle. If the bats respond to a hormonal signal, cows in estrous (in heat) may be bitten with a different probability than cows not in estrous. (The researcher could tell the difference by harnessing painted sponges to the undersides of bulls who would leave their mark during the night.)

	In estrous	Not in estrous	Total
Bitten by a bat	15	6	21
Not bitten by a bat	7	322	329
Total	22	328	350

The proportion of bitten cows among those in estrous is $15/22 = 0.682$ while the proportion of bitten cows among those not in estrous is $6/328 = 0.018$.

Estimating Differences in Proportions

- Find a 95% confidence interval for the difference in probabilities of being bitten by a vampire bat between cows in estrous and those not.

$$0.682 - 0.018 \pm 1.96\sqrt{\frac{0.682(1-0.682)}{22} + \frac{0.018(1-0.018)}{328}}$$

$$0.468 < p_1 - p_2 < 0.859$$

In the study setting in Costa Rica, we are 95% confident that the probability that a cow in estrous is bitten by a vampire bat is larger than the probability of cow not in estrous being bitten by an amount between 0.468 and 0.859.

Odds Ratio for Vampire Bats

- The estimated odds for being bitten for a cow in estrous are $15/7 = 2.143$.
- The estimated odds for being bitten for a cow not in estrous are $6/322 = 0.019$.
- The estimated odds ratio is 115 and its natural logarithm is 4.745.
- The estimated SE of the log odds ratio is

$$\sqrt{\frac{1}{15} + \frac{1}{7} + \frac{1}{6} + \frac{1}{322}} = 0.616$$

- $e^{4.745-1.96(0.616)} \doteq 34.392$ and $e^{4.745+1.96(0.616)} \doteq 384.536$.
- The 95% confidence interval is $34.392 < \text{OR} < 384.536$.

Under the study conditions in Costa Rica, we are 95% confident that the odds that a cow in estrous is bitten by a vampire bat are between 34.392 and 384.536 times higher than for cows not in estrous.

Hypothesis Tests

- Chapter 9 describes three methods for testing independence between two categorical variables.
  - χ^2 test of independence;
  - G-test;
  - Fisher's Exact Test.
- If we think of one of these variable as grouping observations into populations and the other as a response, then each test is equivalent to a test with the null hypothesis that all population proportions (or conditional probabilities) are equal.
- For example,

$$H_0\text{: infection level and being eaten are independent}$$

is equivalent to

$$H_0\text{: } p_1 = p_2 = p_3$$

where p_1, p_2, and p_3 are probabilities of being eaten for the three groups.

Comparisons Among Tests

- The χ^2 test of independence and G-test can be applied to tables of any size.
- Fisher's Exact Test is only defined for 2×2 tables.
- The χ^2 test of independence and G-test compute test statistics for which the true sampling distributions are approximated by χ^2 distributions.
- Fisher's Exact Test computes p-values on the basis of sampling without replacement and the p-value is exact.
- The χ^2 test of independence and G-test p-value calculations are only accurate if sample sizes are large enough (due to the approximation).
- The G-test is based on likelihood ratios and may be more accurate than the χ^2 test which approximates likelihood ratios.

The χ^2 Test of Independence

- The χ^2 test of independence compares *the observed counts in the table* with *the expected values of those counts under the null distribution*.
- The test statistic measures *discrepancy between observed and expected counts*.
- If the discrepancy is larger than expected (from a random chance model), then there is evidence against the null hypothesis of independence.

The Test Statistic

Test Statistic for χ^2 Test of Independence

$$X^2 = \sum_{i \in \text{rows}} \sum_{j \in \text{columns}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

- O_{ij} is the observed count in row i and column j;
- $E_{ij} = \frac{(\text{row sum } i)(\text{column sum } j)}{(\text{table sum})}$ is the expected count in row i and column j;

Expected Counts

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

Explain expected counts in reference to the example:

- Calculations and estimates assume independence (the null hypothesis).
- The observed proportion getting eaten is 48/141.
- The observed proportion that are uninfected is 50/141.
- Probabilities of these events are estimated by their observed proportions.
- Under independence,

$$\text{P(eaten\&uninfected)} = \text{P(eaten)P(uninfected)}$$

Expected Counts (cont.)

- Plugging in observed proportions as estimates, this is

$$\text{P(eaten\&uninfected)} \approx \frac{48}{141} \times \frac{50}{141}$$

- Under the null hypothesis, the observed count in each cell is a *binomial random variable* with $n = 141$ and p estimated as above as a product of marginal proportions.

$$O_{ij} \sim \text{Binomial}(n, p_{ij})$$

  where n is the total number of observations in the table and p_{ij} is the estimated probability for cell in row r and column c.
- The expected value of this random variable is $E_{ij} = np_{ij}$, or

$$E_{ij} = 141 \times \frac{48}{141} \times \frac{50}{141} = \frac{48 \times 50}{141}$$

- In general,

$$E_{ij} = \frac{(\text{row sum } r)(\text{column sum } c)}{(\text{table sum})}$$

Expected Counts in Example

Observed Counts:

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

Expected Counts:

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	17	15.3	15.7	48
Not eaten	33	29.7	30.3	93
Total	50	45	46	141

Calculating the test statistic

$$\begin{aligned} X^2 &= \sum_{r \in \text{rows}} \sum_{c \in \text{columns}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(1-17)^2}{17} + \cdots + \frac{(9-30.3)^2}{30.3} \\ &= 69.8 \end{aligned}$$

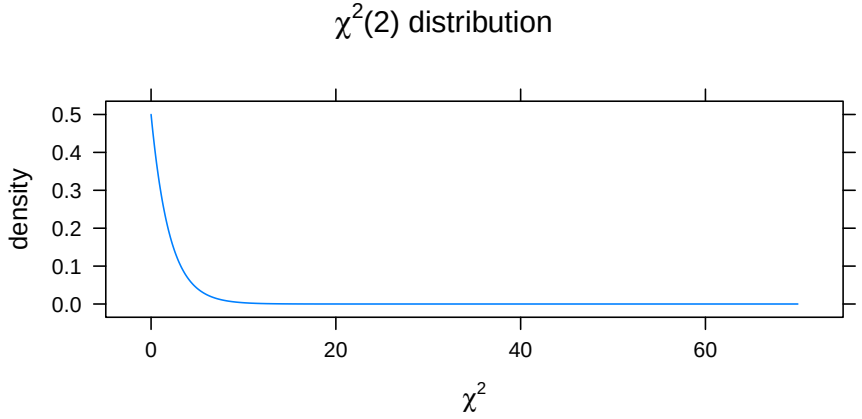
- The sum is over all cells in the table.
- If there are some cells where the observed counts and expected counts differ by a lot, the test statistic will be large.
- If all observed counts are close to expected counts, then the test statistic will be small.

Sampling Distribution

- The sampling distribution of the test statistic under the null hypothesis of independence can be estimated using simulation.
- For large enough samples (no more than 20% of expected counts $<$ 5), the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom is a good approximation.
- This is the distribution of a sum of $(r-1)(c-1)$ squared independent standard normal random variables (which we will see next week).
- The expected value of the test statistic is $(r-1)(c-1)$.
- The p-value is the area to the right of the test statistic under a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Application to Example

- In the example, $r = 2$ and $c = 3$ so there are $(2-1)(3-1) = 2$ degrees of freedom.
- The test statistic of 69.8 is much larger than 2.
- The p-value is about 6.6×10^{-16}.



Interpretation

There is overwhelming evidence ($X^2 = 69.8$, $n = 141$, $df = 2$,$p < 10^{-15}$, χ^2 test of independence) that infection status is not independent of the probability of being eaten for fish under these experimental conditions.

The G-test

- The G-test is nearly identical to the χ^2 test in that the test statistic is compared to a χ^2 distribution.
- The difference is that the test statistic is computed on the basis of likelihood.
- The G-test is an example of a *likelihood ratio test*.

The next several slides contain the background of likelihood ratio tests as they apply to contingency tables. Feel free to skim them without worrying about details. You will not be tested on the details! Just try to capture a rough idea for where the test statistic of the G-test arises. (Bret, you may want to skim these slides to get to the punch line.)

Likelihood Ratio Tests

Definition

In a *likelihood ratio test*, the null hypothesis assumes a likelihood model with k_0 free parameters which is a special case of the alternative hypothesis likelihood model with k_1 free parameters. The two likelihood models are maximized with likelihoods L_0 and L_1 respectively. The test statistic is $G = 2(\ln L_1 - \ln L_0)$ which, for large enough samples, has approximately a $\chi^2(k_1 - k_0)$ distribution when the null hypothesis is true.

The Multinomial Distribution

Definition

The *multinomial distribution* is a generalization of the binomial distribution where there is an independent sample of size n and each outcome is in one of k categories with probabilities p_i for the ith category ($\sum_{i=1}^{k} p_i = 1$). The probability that there are x_i outcomes of type i for $i = 1, 2, \ldots, k$ is

$$\binom{n}{x_1, \ldots, x_k} \left(p_1^{x_1} \cdots p_k^{x_k}\right)$$

where

$$\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! \cdots x_k!}$$

is called a *multinomial coefficient* and $x_1 + \cdots + x_k = n$.

Likelihood for Multinomial Distribution

- In the binomial distribution, we can rethink the parameters for fixed n with probabilities p_1 and p_2 for the two categories with $p_1 + p_2 = 1$, so there is only one free parameter. (If you know p_1, you also know p_2.)
- The maximum likelihood estimates are $\hat{p}_1 = x_1/n$ and $\hat{p}_2 = x_2/n = 1 - \hat{p}_1 = (n - x_1)/n$.
- For more categories, the maximum likelihood estimates are $\hat{p}_i = x_i/n$ for $i = 1, \ldots, k$.
- The maximum likelihood is then

$$L = \binom{n}{x_1, \ldots, x_k} \left(\left(\frac{x_1}{n}\right)^{x_1} \times \cdots \times \left(\frac{x_k}{n}\right)^{x_k}\right)$$

and the maximum log-likelihood is

$$\ln L = \ln \binom{n}{x_1, \ldots, x_k} + \sum_{i=1}^{k} x_i \ln \left(\frac{x_i}{n}\right)$$

Contingency Tables

- The observed outcomes $\{O_{ij}\}$ in a contingency table with r rows and c columns are jointly are modeled with a multinomial distribution with parameters $\{p_{ij}\}$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$.
- There are rc probabilities.

Contingency Tables: Alternative Model

- Under the alternative hypothesis of no independence, the only restriction on the probabilities is that they sum to one, so there are $k_1 = rc - 1$ free parameters.
- The maximum likelihood estimates are

$$p_{ij} = \frac{O_{ij}}{n}$$

- The maximum log-likelihood is

$$\ln L_1 = \ln \binom{n}{O_{11}, \ldots, O_{rc}} + \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \ln \left(\frac{O_{ij}}{n} \right)$$

Contingency Tables: Null Model

- Under the null hypothesis of independence, $p_{ij} = p_{i\cdot} \times p_{\cdot j}$ for all i and j where there are $r - 1$ free parameters for the row factor and $c - 1$ free parameters for the column factor, for a total of $k_0 = r + c - 2$.
- The maximum likelihood estimates are

$$\hat{p}_{i\cdot} = \frac{\text{sum of observations in row } i}{n}$$

for the row probabilities and

$$\hat{p}_{\cdot j} = \frac{\text{sum of observations in column } j}{n}$$

for the column probabilities.
- The maximum likelihood estimate for p_{ij} is $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{E_{ij}}{n}$.
- The maximum log-likelihood is

$$\ln L_0 = \ln \binom{n}{O_{11}, \ldots, O_{rc}} + \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \ln \left(\frac{E_{ij}}{n} \right)$$

Test Statistic

- The test statistic is

$$G = 2(\ln L_1 - \ln L_0)$$

which equals

$$G = 2\left(\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \left(\ln \left(\frac{O_{ij}}{n} \right) - \ln \left(\frac{E_{ij}}{n} \right) \right) \right)$$

which can be simplified to

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

- The difference in the number of free parameters is

$$(rc - 1) - (r + c - 2) = rc - r - c + 1 = (r - 1)(c - 1)$$

Now, back to our regularly scheduled lecture.

The G-Test

The G-Test

The test statistic

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

where

- O_{ij} is the observed count in row r and column c;
- $E_{ij} = \frac{(\text{row sum } r)(\text{column sum } c)}{(\text{table sum})}$ is the expected count in row r and column c;

has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom when the null hypothesis is true and n is large enough.

Expected Counts Again

Observed Counts:

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not eaten	49	35	9	93
Total	50	45	46	141

Expected Counts:

	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	17	15.3	15.7	48
Not eaten	33	29.7	30.3	93
Total	50	45	46	141

Case Study: G-Test

$$\begin{aligned} G &= 2\left(\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)\right) \\ &= 2\left(O_{11}\ln\left(\frac{O_{11}}{E_{11}}\right) + \cdots + O_{rc}\ln\left(\frac{O_{rc}}{E_{rc}}\right)\right) \\ &= 77.9 \end{aligned}$$

- The p-value is approximately 1.2×10^{-17}.
- Compare G to the χ^2 test of independence test statistic value of 69.8.

Interpretation

There is overwhelming evidence ($G = 77.9$, $n = 141$, $df = 2$,$p < 10^{-16}$, G-test) that infection status and is not independent of the probability of being eaten for fish under these experimental conditions.

Fisher's Exact Test

- Fisher's exact test is based on an alternative probability model for 2×2 tables.
- Think of one factor as an outcome and the other as designating groups.
- Fisher's exact test imagines the 2×2 tables if the groups of the same size had been randomly created with sampling without replacement rather than using the factor to form the groups.
- The p-value is the probability of selecting any table at least as extreme as the actual table.
- Sampling without replacement is described by the *hypergeometric distribution*.

Vampire Bats revisited

	In estrous	Not in estrous	Total
Bitten by a bat	15	6	21
Not bitten by a bat	7	322	329
Total	22	328	350

Here are other tables with even more extreme differences in proportions of being bitten, but with the same marginal totals.

16	5
6	323

17	4
5	324

18	3
4	325

19	2
3	326

20	1
2	327

21	0
1	328

P-value Calculation

	In estrous	Not in estrous	Total
Bitten by a bat	X	$21 - X$	21
Not bitten by a bat	$22 - X$	$307 + X$	329
Total	22	328	350

- The p-value calculation focuses on any single cell; here the top left.
- Imagine the 21 bitten cows as red balls and the 329 cows not bitten as white balls.
- Sample 22 without replacement at random, and let X be the number of red balls in the sample.
- The probability of having exactly x red balls in the sample is

$$\frac{\binom{21}{x}\binom{329}{22-x}}{\binom{350}{22}}$$

as there are $\binom{21}{x}$ ways to pick which x red balls are sampled, $\binom{329}{22-x}$ ways to pick which $22 - x$ white balls are sampled, and $\binom{350}{22}$ total ways to choose 22 balls from 350.

P-value Calculation

- The actual grouping of cows by estrous status has $X = 15$.
- The p-value is the probability $X \geq 15$.

$$P = \sum_{x=15}^{21} \frac{\binom{21}{x}\binom{329}{22-x}}{\binom{350}{22}}$$

- This calculation is tedious by hand, but can be done in R using the `dhyper()` function.

```
> sum(dhyper(15:21, 21, 329, 22))
[1] 1.004713e-16
```

Interpretation

There is overwhelming evidence ($p \doteq 10^{-16}$, Fisher's one-sided exact test) that the probability a cow in estrous will be bitten by a vampire bat is larger than that for a cow not in estrous in a setting similar to the study in Costa Rica.

R for Fisher

- Even easier, there is a built-in function `fisher.test()`.
- The following example shows how.

```
> x = matrix(c(15, 7, 6, 322), nrow = 2, ncol = 2)
> fisher.test(x, alternative = "greater")

Fisher's Exact Test for Count Data

data:  x
p-value < 2.2e-16
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 35.49817      Inf
sample estimates:
odds ratio
  108.3894
```


What you should know

You should know:

- how to find a confidence interval for a difference in proportions;
- how to find a confidence interval for an odds ratio;
- how to test for independence in contingency tables using:
  - the χ^2 test of independence;
  - the G-test;
  - Fisher's exact test
- how to determine which tests are appropriate in which situations.

R Details

- The R function `chisq.test()` can be used to automate calculations for the χ^2 test of independence.
- The R function `fisher.test()` can be used to automate calculations for Fisher's exact test.
- There is no built-in function in R for the G-test, but a new R appendix will contain one.