Unit 9: Inference for Categorical Data

Statistics S-100 Teaching Team

Summer 2024

Introduction

Inference for binomial proportions

Inference for two-way tables

Measures of effect size in two-by-two tables

Introduction

Tools for assessing association

We have covered methods for numerical outcomes:

- numerical outcome with a categorical predictor
  - ◇ two-sample t-tests and ANOVA
  - ◇ simple/multiple linear regression
- numerical outcome with a numerical predictor
  - ◇ simple linear regression
- numerical outcome with several predictors, numerical or categorical
  - ◇ multiple linear regression

Next, methods for categorical outcomes:

- categorical outcome with a categorical predictor
  - ◇ χ^2 test of independence
  - ◇ Fisher's exact test
  - ◇ simple/multiple logistic regression
- binary outcome with a numerical predictor
  - ◇ simple logistic regression
- binary outcome with several predictors, numerical or categorical
  - ◇ multiple logistic regression

Inference for binomial proportions

Fatal vehicle collisions

According to the National Highway Traffic Safety Administration (NHTSA) there were 33,949 fatal vehicle collisions across the US in 2018.

The cause of each accident is reported (e.g., distraction, drowsiness, alcohol consumption, etc.) as well as the location.

In Massachusetts, 136 out of 341 fatal collisions involved alcohol consumption.

Questions that can be addressed with inference. . .

- What is the estimated **population proportion** of alcohol-related fatal collisions in MA?
- What is the 95% confidence interval for the estimated population proportion of alcohol-related fatal collisions in MA?
- The nationwide proportion of alcohol-related fatal collisions is thought to be 0.33. Do the observed data for MA suggest that the probability of alcohol being involved in a fatal collision is greater in MA than nationwide?

Inference for binomial proportions

The collision data are binomial data; "success" can be defined as alcohol being involved.

Suppose X is a binomial random variable with parameters n and p, where n is the number of trials and p is the probability of success.

- The parameter of interest is p, the population probability of success; i.e., population probability of a fatal collision involving alcohol consumption.
- The estimate of p from the observed sample is $\hat{p} = x/n$, where x is the observed number of successes.

Inference for p can be done using the normal approximation to the binomial, or directly using the binomial distribution.

- The normal approximation approach relies on the **Central Limit Theorem**.
- The binomial approach is an example of an **exact** test, in which it is not necessary to approximate the sampling distribution of the test statistic.

Normal theory approach (CLT for the sample proportion)

The sampling distribution of $\hat{p}$ is approximately normal when

1. The sample observations are independent, and
2. At least 10 successes and 10 failures are expected in the sample: $np \geq 10$ and $n(1-p) \geq 10$.[1]

Under these conditions, $\hat{p}$ is approximately normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Since p is unknown, it is necessary to substitute either $\hat{p}$ or p_0 for p in the standard error term when computing confidence intervals and test statistics.

[1] This condition is commonly referred to as the success-failure condition.

Inference with the normal approximation

In the context of calculating CIs, substitute $\hat{p}$ for p.

An approximate two-sided 95% confidence interval for p is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

With 95% confidence, the interval (0.347, 0.453) captures the population proportion of fatal collisions in MA that involved alcohol consumption.

```r
#calculate confidence interval
prop.test(x = 136, n = 341,
          conf.level = 0.95)$conf.int
```

```
## [1] 0.3468427 0.4531286
## attr(,"conf.level")
## [1] 0.95
```

Inference with the normal approximation...

In the testing context, substitute p_0 for p.

The test statistic z for the null hypothesis $H_0 : p = p_0$ is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(1 - p_0)}{n}}}$$

- If the proportion of fatal collisions in MA that involved alcohol consumption were actually 0.33, there would be only a 0.0041 probability of observing a sample proportion of alcohol-related fatal collisions equal to 0.399 or larger.
- Thus, these data suggest that the proportion of alcohol-related fatal collisions in MA is higher than the nationwide proportion of 0.33.

```r
#conduct hypothesis test
prop.test(x = 136, n = 341, p = 0.33,
          alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  136 out of 341, null probability 0.33
## X-squared = 6.9981, df = 1, p-value = 0.00408
## alternative hypothesis: true p is greater than 0.33
## 95 percent confidence interval:
##  0.3547446 1.0000000
## sample estimates:
##        p
## 0.398827
```

Exact inference for binomial data

Definition of the p-value: the probability of observing 136 or more successes out of 341 trials if the null hypothesis $H_0 : p = 0.33$ were true.

```r
#use pbinom( )
pbinom(135, 341, p = 0.33, lower.tail = FALSE)
```

```
## [1] 0.004507281
```

```r
#use binom.test( )
binom.test(x = 136, n = 341, p = 0.33, alternative = "greater")
```

```
## 
##  Exact binomial test
## 
## data:  136 and 341
## number of successes = 136, number of trials = 341, p-value = 0.004507
## alternative hypothesis: true probability of success is greater than 0.33
## 95 percent confidence interval:
##  0.3545283 1.0000000
## sample estimates:
## probability of success
##               0.398827
```

Inference for the difference of two proportions

The one-sample z-test for a population proportion is analogous to the one-sample t-test for a population mean:

- Sample statistic: $\hat{p}$, Parameter: p, Null hypothesis: $H_0 : p = p_0$
- Sample statistic: $\overline{x}$, Parameter: μ, Null hypothesis: $H_0 : \mu = \mu_0$

Similarly, there exists a two-sample z-test for the difference of population proportions that is analogous to the two-sample t-test for the difference of population means:

- Sample statistic: $\hat{p}_1 - \hat{p}_2$, Parameter: $p_1 - p_2$, Null hypothesis: $H_0 : p_1 - p_2 = 0$
- Sample statistic: $\overline{x}_1 - \overline{x}_2$, Parameter: $\mu_1 - \mu_2$, Null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$

For completeness, slides 13-14 show the details of the two-sample proportions test. We will focus on learning a more flexible approach for analyzing the association between two categorical variables.

Inference for the difference of two proportions...

The normal model can be applied to $\hat{p}_1 - \hat{p}_2$ if

1. The two samples are independent, the observations in each sample are independent, and
2. At least 10 successes and 10 failures are expected in each sample.

The standard error of the difference in sample proportions is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

In hypothesis testing, the following estimate of p is used to compute the standard error:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

Fatal vehicle collisions. . .

Does the population proportion of alcohol-related fatal collisions differ between MA and UT?

```
##      Cause
## State Alcohol Not Alcohol Sum
##   MA      136         205 341
##   UT       58         179 237
##   Sum     194         384 578
```

```r
#analyze the data
prop.test(x = c(136, 58), n = c(341, 237))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(136, 58) out of c(341, 237)
## X-squared = 14.207, df = 1, p-value = 0.0001637
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07504718 0.23315531
## sample estimates:
##    prop 1    prop 2
## 0.3988270 0.2447257
```

Inference for two-way tables

Inference for two-way tables

A two-way table summarizes information about the relationship between two categorical variables.

Testing for a difference between p_1 and p_2 is equivalent to testing for association in a two-way table that has two rows and two columns.

	Outcome: Success	**Outcome: Failure**	**Total**
Group 1	x_1	$n_1 - x_1$	n_1
Group 2	x_2	$n_2 - x_2$	n_2
Total	$x_1 + x_2$	$(n_1 - x_1) + (n_2 - x_2)$	$n_1 + n_2$

Treating HIV$^+$ infants

In resource-limited settings, single-dose nevirapine is given to an HIV$^+$ woman during birth to prevent mother-to-child transmission of the virus.

- Exposure of the infant to nevirapine (NVP) may foster the growth of resistant strains of the virus in the child.
- If the child is HIV$^+$, should they be treated with nevirapine or a more expensive drug, lopinarvir (LPV)?

In this setting, the possible outcomes are virologic failure (the virus becomes resistant) versus stable disease (virus growth is prevented).

The following table summarizes the results of a 2012 study comparing NVP versus LPV in treatment of HIV-infected infants.[a] Children were randomized to receive either NVP or LPV.

	Stable Disease	**Virologic Failure**	**Total**
NVP	87	60	147
LPV	113	27	140
Total	200	87	287

[a] Violari, et al. *NEJM* 2012; 366: 2380-2389.

Formulating hypotheses in a two-way table

The main question of interest:

- Do the data support the claim of a difference in outcome by treatment?

If there is no difference in outcome by treatment, then knowing treatment provides no information about outcome; treatment assignment and outcome are *independent* (i.e., *not associated*).

- H_0: Treatment and outcome are not associated.
- H_A: Treatment and outcome are associated.
  - ◇ This is inherently a two-sided alternative.

The χ^2 test of independence

In the χ^2 test, the observed number of cell counts are compared to the number of **expected** cell counts, where the expected counts are calculated under the null hypothesis.

- The test statistic quantifies how far the observed results deviate from what is expected under the null hypothesis.
- A larger test statistic represents stronger evidence against the null hypothesis of independence.

Expected cell counts

If treatment had no effect on outcome, what would we expect to see?

- Let $A = \{\text{assignment to NVP}\}$
- Let $B = \{\text{virologic failure}\}$

Under the hypothesis of independence,

$$P(A \text{ and } B) = P(A) \times P(B) = \left(\frac{147}{287}\right)\left(\frac{87}{287}\right)$$

The expected cell count in the upper right corner would be

$$(287)\left(\frac{147}{287}\right)\left(\frac{87}{287}\right) = 44.56$$

What about the other cells?

Formula for expected cell counts

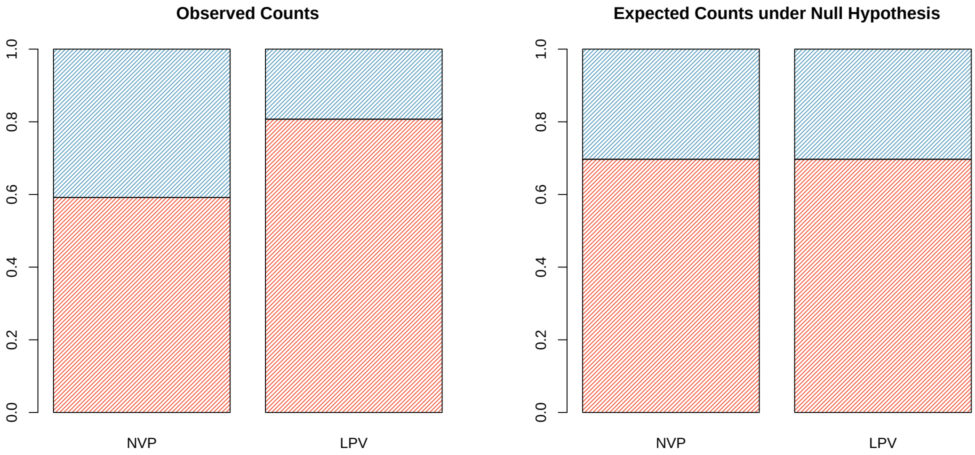
The expected count for the i^{th} row and j^{th} column is

$$E_{i,j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n},$$

where n is the total number of observations.

	Stable Disease	**Virologic Failure**	**Total**
NVP	87 (102.44)	60 (44.56)	147
LPV	113 (97.56)	27 (42.44)	140
Total	200	87	287

Visual comparison of observed versus expected



Observed Counts

Expected Counts under Null Hypothesis

The χ^2 test statistic

The χ^2 **test statistic** is calculated as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

and is approximately distributed χ^2 with degrees of freedom $(r-1)(c-1)$, where r is the number of rows and c is the number of columns.

- $O_{i,j}$ represents the observed count in row i, column j.
- $E_{i,j}$ represents the expected count in row i, column j.

Assumptions for the χ^2 test:

- *Independence*. Each case that contributes a count to the table must be independent of all other cases in the table.
- *Sample size*. Each expected cell count must be greater than or equal to 10.[a]
  - ◇ For tables larger than 2×2, it is appropriate to use the test if no more than $1/5$ of the expected counts are less than 5, and all expected counts are greater than 1.

These assumptions must be met for the test statistic to be approximately distributed χ^2.

[a] Some sources use a less strict sample size condition. For example, the `chisq.test()` function only shows a warning if one of the expected counts is smaller than 5.

The χ^2 test in R

```
hiv.table <- matrix(c(87, 113, 60, 27), nrow = 2, ncol = 2, byrow = F)
dimnames(hiv.table) <- list("Drug" = c("NVP", "LPV"),
                           "Outcome" = c("Stable Disease", "V. Failure"))
chisq.test(hiv.table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  hiv.table
## X-squared = 14.733, df = 1, p-value = 0.0001238
```

```
chisq.test(hiv.table)$expected
```

```
##      Outcome
## Drug   Stable Disease V. Failure
##   NVP      102.43902   44.56098
##   LPV       97.56098   42.43902
```

Residuals in the χ^2 test

For each cell in a table, the **residual** equals

$$\frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}.$$

Residuals with a large magnitude contribute the most to the χ^2 statistic.

- If a residual is positive, the observed value is greater than the expected value.
- If a residual is negative, the observed value is less than the expected.

Residuals in the χ^2 test...

```
chisq.test(hiv.table)$residuals
```

```
##      Outcome
## Drug  Stable Disease V. Failure
##   NVP      -1.525412   2.312824
##   LPV       1.563082  -2.369939
```

Examining the residuals can be informative for understanding direction of association.

- Which drug is associated with stable disease; i.e., which drug should be recommended for treatment of HIV-infected infants?

Treating *C. difficile* infection

Clostridium difficile is a bacterium that causes inflammation of the colon. Antibiotic treatment is typically not effective. Infusion of feces from healthy donors has been reported as an effective treatment.

A randomized trial was conducted to compare the efficacy of donor-feces infusion versus vancomycin, the antibiotic typically prescribed to treat *C. difficile* infection.

	Cured	Uncured	Sum
Fecal Infusion	13	3	16
Vancomycin	4	9	13
Sum	17	12	29

Table 1: Fecal Infusion Study Results

Can a χ^2 test be used to analyze these results?

Fisher's exact test

Fisher's exact test works even when sample sizes are small.[2]

In this particular experiment, we observed 17 cured individuals (out of 29 total) when 16 were assigned to fecal infusion and 13 to vancomycin.

- Under $H_0 : p_1 = p_2$, individuals in one treatment group are just as likely to be cured as individuals in the other group.
- If H_0 is true (and the study had the same setup):
  - ◇ What is the probability that of the 17 cured individuals, 13 were in the fecal infusion group?
  - ◇ What are the possible sets of results that indicate stronger evidence in favor of fecal infusion?
  - ◇ What is the probability of seeing even stronger evidence in favor of fecal infusion as an effective treatment?

The p-value for Fisher's exact test is calculated by adding together the individual conditional probabilities of obtaining each table that is **as extreme or more extreme than the one observed**, under the null hypothesis and given that the marginal totals are considered fixed.

[2] In this course, Fisher's exact test is only discussed in the context of 2×2 tables.

The hypergeometric distribution

Let X represent the number of successes in a series of repeated Bernoulli trials, where sampling is done without replacement.

- In a population of size N, there are m total successes.
- What is the probability of observing exactly k successes when drawing a sample of size n?

For example, imagine an urn with m white balls and $N - m$ red balls. Draw n balls without replacement. What is the probability of observing k white balls in the sample?

	White Ball	**Red Ball**	**Total**
Sampled	k	$n - k$	n
Not Sampled	$m - k$	$N - n - (m - k)$	$N - n$
Total	m	$N - m$	N

The hypergeometric distribution. . .

To calculate $P(X = k)$ where $X \sim \text{HGeom}(m, N - m, n)$, use `dhyper( )`:

```r
dhyper(k, m, N - m, n)
```

Suppose the urn contains 10 white balls, 15 red balls, and a sample of size 8 is drawn. What is the probability of observing 5 white balls in the sample?

```r
dhyper(5, 10, 25 - 10, 8)
```

```
## [1] 0.1060121
```

Treating *C. difficile* infection. . .

Given that 17 individuals out of 29 were cured and that 16 individuals were in the fecal infusion group (and that H_0 is true), what is the probability that 13 of the cured individuals were in the fecal infusion group?

- $N = 29$, $m = 17$, and $n = 16$
- Calculate $P(X = 13)$.

```r
#probability of observed results
dhyper(13, 17, 29 - 17, 16)
```

```
## [1] 0.007715441
```

Fisher's exact test…

For a one-sided *p*-value…

- Sum the probabilities of the results as or more extreme than those observed; that is, the probability of the observed table and that of all tables that are more extreme in the direction specified by the alternative hypothesis.

```r
#one-sided p-value
phyper(12, 17, 29 - 17, 16, lower.tail = FALSE)
```

```
## [1] 0.008401063
```

For a two-sided *p*-value…

- Consider extreme tables to be all tables with probabilities less than that of the observed; sum the probabilities of tables representing results as or more extreme than those observed.

Treating *C. difficile* infection...

```r
#enter the data
infusion.table = matrix(c(13, 3, 4, 9), nrow = 2, ncol = 2, byrow = T)
dimnames(infusion.table) = list("Outcome" = c("Cured", "Uncured"),
                                "Treatment" = c("Fecal Infusion",
                                                "Vancomycin"))

fisher.test(infusion.table, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  infusion.table
## p-value = 0.008401
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.735233      Inf
## sample estimates:
## odds ratio
##   8.848725
```

Treating *C. difficile* infection. . .

```
##      k     prob
## 1    0 0.000000
## 2    1 0.000000
## 3    2 0.000000
## 4    3 0.000000
## 5    4 0.000035
## 6    5 0.001094
## 7    6 0.012036
## 8    7 0.063046
## 9    8 0.177317
## 10   9 0.283708
## 11  10 0.264794
## 12  11 0.144433
## 13  12 0.045135
## 14  13 0.007715
## 15  14 0.000661
## 16  15 0.000024
## 17  16 0.000000
```

```
#P(X leq 5) + P(X geq 13)
phyper(5, 17, 29 - 17, 16) +
  phyper(12, 17, 29 - 17, 16, lower.tail = F)
```

```
## [1] 0.009530323
```

```
#two-sided p-value
fisher.test(infusion.table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  infusion.table
## p-value = 0.00953
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.373866 78.811505
## sample estimates:
## odds ratio
##   8.848725
```

Measures of effect size in two-by-two tables

Measures of effect size for categorical outcomes

Recent article: "Aspartame Is a Possible Cause of Cancer in Humans, a W.H.O. Agency Says"

- Many studies have investigated potential links between artificial sweeteners and cancer.
- “The highest category of aspartame intake ($\geq$ 143 mg/day) was associated with elevated relative risk of non-Hodgkin lymphoma (RR = 1.64, 95% CI 1.17 - 2.29) in men.”[3]
- “High consumption of aspartame was associated with stomach cancer (OR = 2.27, 95% CI 0.99 - 5.44), while a lower risk was observed for breast cancer (OR = 0.28, 95% CI 0.08 - 0.83).”[4]

Results from studies done to investigate the effect of a risk factor on an outcome of interest are often reported as relative risks (RRs) or odds ratios (ORs).

- Important caveat: RR and OR should always be examined in the context of the **absolute risk**; i.e., estimate of risk in the baseline group.

[3] Study referenced in 2015 dietary guidelines advisory report

[4] Study by Palomar-Cros, et al.

Relative risk in a 2×2 table

The **relative risk (RR)** is a measure of the risk of a certain event occurring in one group relative to the risk of the event occuring in another group.

The risk of virologic failure among the NVP group is

$$\frac{\#\text{ in NVP group and had virologic failure}}{\text{total }\#\text{ in NVP group}} = \frac{60}{147} = 0.408$$

The risk of virologic failure among the LPV group is

$$\frac{\#\text{ in LPV group and had virologic failure}}{\text{total }\#\text{ in LPV group}} = \frac{27}{140} = 0.193$$

Thus, the relative risk of virologic failure comparing NVP to LPV is $0.408/0.193 = 2.11$.

- Children treated with NVP are estimated to be more than twice as likely to experience virologic failure.

Confidence interval for relative risk

Let y_1 and y_2 represent the observed number of successes in two groups of size n_1 and n_2. Let the risk (of the event defined as success) in each group be represented as $\hat{p}_1 = y_1/n_1$ and $\hat{p}_2 = y_2/n_2$ and the estimated relative risk be $\widehat{RR} = \hat{p}_1/\hat{p}_2$.

$$\text{SE}_{\log(\widehat{RR})} = \sqrt{\frac{1-\hat{p}_1}{y_1} + \frac{1-\hat{p}_2}{y_2}}$$

A $100(1-\alpha)\%$ confidence interval for log(RR)[5] is given by

$$\log(\widehat{RR}) \pm \left(z^{\star} \times \text{SE}_{\log(\widehat{RR})}\right)$$

To obtain the confidence interval for RR, exponentiate the bounds of the CI for log(RR).

[5] This CI is valid when all expected cell counts ≥ 10.

Confidence interval for relative risk...

Compute a 95% CI for the relative risk of virologic failure comparing NPV to LPV.

$$\text{SE}_{\log(\widehat{RR})} = \sqrt{\frac{1-\hat{p}_1}{y_1} + \frac{1-\hat{p}_2}{y_2}} = \sqrt{\frac{1-0.408}{60} + \frac{1-0.193}{27}} = 0.199$$

95% CI for log(RR):

$$\log(2.11) \pm (1.96)(0.199) \rightarrow (0.358, 1.139)$$

95% CI for RR:

$$(e^{0.358}, e^{1.139}) \rightarrow (1.430, 3.125)$$

```r
library(epitools)
riskratio(hiv.table, rev = "rows")$measure
```

```
##       risk ratio with 95% C.I.
## Drug  estimate   lower    upper
##   LPV 1.000000      NA       NA
##   NVP 2.116402 1.43177 3.128405
```

Odds and the odds ratio in a 2×2 table

The **odds** of an event E are $\frac{P(E)}{1-P(E)}$.

The **odds ratio (OR)** is a measure of the odds of a certain event occurring in one group relative to the odds of the event occurring in another group.

The odds of virologic failure among the NVP group is

$$\frac{\text{\# in NVP group and had virologic failure}}{\text{\# in NVP group and did not have virologic failure}} = \frac{60}{87} = 0.690$$

The odds of virologic failure among the LPV group is

$$\frac{\text{\# in LPV group and had virologic failure}}{\text{\# in LPV group and did not have virologic failure}} = \frac{27}{113} = 0.239$$

Thus, the odds ratio of virologic failure comparing NVP to LPV is $0.690/0.239 = 2.89$.

- The odds of virologic failure when treated with NVP are almost three times as large as the odds of virologic failure when treated with LPV.

Odds and probabilities

With some algebra, it is possible to show the following relationship:

$$\text{odds} = \frac{p}{1-p} \qquad p = \frac{\text{odds}}{1+\text{odds}}$$

Probabilities and odds increase or decrease together.

- Note that while probabilities always have values between 0 and 1 (inclusive), odds can be much larger than 1.

Probability	**Odds** $= p/(1-p)$	**Odds**
0	$0/1 = 0$	0
$1/100 = 0.01$	$1/99 = 0.0101$	$1 : 99$
$1/10 = 0.10$	$1/9 = 0.11$	$1 : 9$
$1/4$	$1/3$	$1 : 3$
$1/3$	$1/2$	$1 : 2$
$1/2$	$(\frac{1}{2})/(\frac{1}{2}) = 1$	$1 : 1$
$2/3$	$(2/3)/(1/3) = 2$	$2 : 1$
$3/4$	3	$3 : 1$
1	$1/0 \approx \infty$	∞

Confidence interval for odds ratio

Let a, b, c, and d represent the four cell counts in a 2×2 table.

$$\mathrm{SE}_{\log(\widehat{OR})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

A $100(1-\alpha)\%$ confidence interval for log(OR)[6] is given by

$$\log(\widehat{OR}) \pm \left(z^{\star} \times \mathrm{SE}_{\log(\widehat{OR})}\right)$$

To obtain the confidence interval for OR, exponentiate the bounds of the CI for log(OR).

[6] This CI is valid when all expected cell counts ≥ 10.

Confidence interval for odds ratio...

Compute a 95% CI for the odds ratio of virologic failure comparing NPV to LPV.

$$\text{SE}_{\log(\widehat{OR})} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{87} + \frac{1}{60} + \frac{1}{113} + \frac{1}{27}} = 0.272$$

95% CI for log(OR):

$$\log(2.89) \pm (1.96)(0.272) \rightarrow (0.578, 1.595)$$

95% CI for OR:

$$(e^{0.578}, e^{1.595}) \rightarrow (1.693, 4.920)$$

```r
oddsratio(hiv.table, rev = "rows", method = "wald")$measure
```

```
##      odds ratio with 95% C.I.
## Drug  estimate    lower    upper
##   LPV 1.000000       NA       NA
##   NVP 2.886335 1.693248 4.920088
```

Relative risk versus odds ratio

The relative risk cannot be used in studies that use **outcome-dependent sampling**, such as a case-control study:

- Suppose in the HIV study, researchers had identified 100 HIV-positive infants who had experienced virologic failure (cases) and 100 who had stable disease (controls), then recorded the number in each group who had been treated with NVP or LPV.
- With this design, the sample proportion of infants with virologic failure no longer estimates the population proportion.
    - ◇ Similarly, the sample proportion of infants with virologic failure in a treatment group no longer estimates the proportion of infants who would experience virologic failure in a hypothetical population treated with that drug.

The odds ratio remains valid even when it is not possible to estimate incidence of an outcome from sample data.