2-Way Tables

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Fall 2018

Outline

- Analyzing associations
- Notation & Probability Structure
  - Joint
  - Marginal
  - Conditional
- Sampling Models
- No Association (Independence)
- Measuring Association in 2×2 tables – general comments
  - Differences of proportions
  - Relative risk
  - **Odds ratios** — very important
  - Relationship between Odds ratios and relative risk
  - Correlation (phi coefficient)
  - Comparison of odds ratios and correlation
  - Others
- Types of experimental designs

Analyzing Associations

Two aspects:

Description

- How strong is the relationship?
- What is the most appropriate way to measure it?
- What is the nature of the relationship?

Inference

- Hypothesis testing.
- Confidence intervals for parameters.

How we do this depends on the substantive problem of the study and How the data were obtained.

Notation & Probability

General "rules"

- Greek letters refer to population quantities.
- Upper case Roman letters refer to random variables.
- Lower case Roman letters refer to sample values.

Unless otherwise stated,

- X and Y are categorical variables.
- X has I levels.
- Y has J levels.
- There are IJ cells in a cross–classification of X and Y.
- X is the row variable, which is indexed by i.
- Y is the column variable, which indexed by j.

A **2–way contingency table** is a cross–classification of observations by the levels of 2 discrete variables. The cells of the table contain frequency counts.

More Notation & Definitions

- The number of variables is often referred to as the "**dimension of the table**".
- The "**size**" of the table often refers to the number of cells.
- The size of (for example) a 2–way table is $I \times J$.
- An Example of a "2×2" table: In a French study, a double blind experiment on the therapeutic value of vitamin C (ascorbic acid) for treating the common cold was conducted during 2 periods of 5–7 days (Fienberg, 1985). The subjects were 279 skiers, 139 of whom received 1 gram of vitamin C and the other 140 skiers received a placebo.

		Outcome		
		Cold	No Cold	
Treatment	vitamin C	17	122	139
Group	placebo	31	109	140
		48	231	279

- n_{ij} equals the frequency in the i^{th} level of X and the j^{th} level

Probability Distributions for Contingency Tables

$$\begin{aligned} \pi_{ij} &= P(X = i, Y = j) \\ &= \text{probability that a randomly selected individual falls into the } (i,j)^{\text{th}} \text{ cell of the contingency table.} \end{aligned}$$

Probability Distributions:
Joint Distribution of X and Y consists of the set of the π_{ij}'s:

X \ Y		
	π_{11}	π_{12}
	π_{21}	π_{22}
		1.00

Marginal Distributions

Marginal Distribution of X and Y are the sums of cell probabilities across of the columns and rows, respectively:

π_{11}	π_{12}	$\boldsymbol{\pi_{1+}}$
π_{21}	π_{22}	$\boldsymbol{\pi_{2+}}$
$\boldsymbol{\pi_{+1}}$	$\boldsymbol{\pi_{+2}}$	1.00

Conditional Distributions

For observed data, we use p instead of π and

$$\begin{aligned} p_{ij} &= \text{proportion of observations in the } (i,j)\text{cell} \\ &= \frac{n_{ij}}{n} \end{aligned}$$

Conditional Distribution

When one variable is a "response" and the other is an "explanatory" variable, we focus on the distribution of the response variable conditional on the explanatory variable.

	Outcome		
	Cold	No Cold	
vitamin C	$17/139 = \mathbf{.12}$	$122/139 = \mathbf{.88}$	$.12 + .88 = 1.00$
placebo	$31/140 = \mathbf{.22}$	$109/140 = \mathbf{.78}$	$.22 + .78 = 1.00$

Conditional Distributions (continued)

and in general notation,

$n_{11}/n_{1+} = p_1$	$n_{12}/n_{1+} = (1 - p_1)$	1.00
$n_{21}/n_{2+} = p_2$	$n_{22}/n_{2+} = (1 - p_2)$	1.00

Sampling Models

These are extensions of the Poisson and Binomial models that we discussed for 1 variable.

Poisson Sampling No margins of a table are fixed by design. Each cell is considered an independent Poisson random variable.

The following data are the number of game–related concussions of players on 49 college football teams between 1975–1982 (Buckley, 1988; Agresti, 1990).

		Activity		
		Block	Tackle	
Situation	Passing	47	147	194
	Rushing	190	341	531
		237	488	725

Poisson Sampling

Example 2: These data are from records of accidents in 1988 compiled by the Department of Highway Safety and Motor Vehicles in Florida (Agresti, 1990, 1996).

		Injury	
		Fatal	Nonfatal
Seat Belt	No	1601	162,527
Use	Yes	510	412,368

Poisson Sampling

Example 3: (Approximate) 2008 admissions data at UIUC reported by Chicago Tribune and former Illinois President J.B. White.

	Admitted		
Category	Yes	No	
"I list"	37	123	160
General	8,000	18,000	26,000
	8,037	18,123	26,160

Multinomial Sampling

Multinomial sampling Only the total number of observations, n, is fixed by design. The margins are free to vary.

Example: Job satisfaction (Andersen, 1985). These data are from a large scale investigation of blue collar workers in Denmark (1968).

Supervisor Satisfaction	Low		High		
Worker satisfaction	Low	High	Low	High	
	162	196	107	247	712

But usually we write this as

		Worker satisfaction		
		Low	High	
Supervisor satisfaction	Low	162	196	
	High	107	247	
				712

Independent Binomial Sampling

Independent Binomial Sampling One margin is fixed by design while the other(s) is free to vary.

Example: The study on the effectiveness of vitamin C on preventing colds.

		Outcome		
		Cold	No Cold	
Treatment	vitamin C	17	122	139
Group	placebo	31	109	140
		48	231	279

Independent Multinomial Sampling The "response" variable has more than two categories.

Pseudo-Independent Binomial Sampling

Pseudo-Independent Binomial Sampling When one variable is considered the response and the other variable is considered the explanatory variable, but only the total n is fixed by design. We may want to treat the data as if it were independent binomial samples.

- What sampling model did the data come from?
- Consider job satisfaction example where worker's satisfaction is the response variable and their supervisor's satisfaction is an explanatory variable.
- Different sampling models usually lead to the same inferential methods.
- Importance of Considering Sampling Design: **sampling and design do make a difference regarding conclusions that can be made.**

Other Sampling Models

Both margins fixed by design.

Example 1: Data from Kramer on acceptance of new sibling by 30 firstborn 3–5 year old children. The variables age and sibling acceptance were created by taking median splits:

		Sibling Acceptance		
		Lower	Higher	
Age	Younger	9	6	15
	Older	6	9	15
		15	15	30

Other Sampling Models

Both margins fixed by design.

Example 2 : 1970 draft lottery of 19–26 year olds. Each day of the year (including Feb 29) was typed on a slip of paper and inserted into a capsule. The capsules were mixed and were assigned a "drawing number" according to their position in the sequence of capsules picked from a bowl. Below is a cross-classification of months by drawing number where drawing numbers are grouped into thirds:

Drawing numbers	Months												
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Totals
1–122	9	7	5	8	9	11	12	13	10	9	12	17	122
123–244	12	12	10	8	7	7	7	7	15	15	12	10	122
245–366	10	10	16	14	15	12	12	11	5	7	6	4	122
Totals	31	29	31	30	31	30	31	31	30	31	30	31	366

Other Designs for high–way tables same as the ones we've talked about and extensions of these.

No Association (Statistical Independence)

Situation: One response variable and the other is an explanatory variable. Example:

	Outcome		
	Cold	No Cold	
vitamin C	$17/139 = .12$	$122/139 = .88$	$.12 + .88 = 1.00$
placebo	$31/140 = .22$	$109/140 = .78$	$.22 + .78 = 1.00$
	$48/279 = .17$	$231/279 = .83$	$.17 + .83 = 1.00$

If response and explanatory variables are independent, then

- The conditional probabilities of responses given levels of the explanatory variable should be equal, and
- They should equal the marginal probabilities over levels of the explanatory variable.

$\Longrightarrow$ **Homogeneous Distributions**

Homogeneous Distributions

Situation: One response variable and the other is an explanatory variable.

Example 2: Height of Presidential candidates data (excluding ties):

	The winner was		
	Taller	Shorter	
1856–1928	$4/12 = .33$	$8/12 = .67$	$.33 + .67 = 1.00$
1932–1992	$13/15 = .87$	$2/15 = .13$	$.87 + .13 = 1.00$
	$17/27 = .71$	$10/27 = .37$	$.71 + .37 = 1.00$

Two Response Variables

Two variables are **Statistically Independent** if the

$$\begin{aligned} \text{Joint probabilities} &= \text{Product of the marginal probabilities} \\ \pi_{ij} &= \pi_{i+}\pi_{+j} \end{aligned}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Responses to the two items for the GSS (1994):

- Item 1: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
- Item 2: Working women should have paid maternity leave.

Two items from the GSS (1994)

- Item 1: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
- Item 2: Working women should have paid maternity leave.

	Item2					
Item 1	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree	
Strongly Agree	97	96	22	17	2	234
Agree	102	199	48	38	5	392
Disagree	42	102	25	36	7	212
Strongly Disagree	9	18	7	10	2	46
	250	415	102	101	16	884

For "Strongly Agree" on Item 1 and "Disagree" on Item 2,

$$p_{ij} = 17/884 = .019 \neq p_{i+}p_{+j} = (234/884)(101/884) = (.265)(.114) = .030$$

Measuring Association in 2×2 tables

Ways to study and analyze the relationship between two variables.

Multiple ways to do measure association:

- Differences of Proportions
- Relative risk
- Odds Ratios
- Correlation (phi coefficient)

Differences of Proportions

in the population,

$$\begin{aligned} \pi_1 &= \text{probability of ``success'' given row 1} \\ 1-\pi_1 &= \text{probability of ``failure'' given row 1} \end{aligned}$$

and

$$\begin{aligned} \pi_2 &= \text{probability of ``success'' given row 2} \\ 1-\pi_2 &= \text{probability of ``failure'' given row 2} \end{aligned}$$

These are conditional probabilities.

		Outcome	
		Cold	No Cold
Treatment	vitamin C	π_1	$1-\pi_1$
Group	placebo	π_2	$1-\pi_2$

Definition: Difference of proportions equals $(\pi_1 - \pi_2)$.

Differences of Proportions

Definition: Difference of proportions equals $(\pi_1 - \pi_2)$.

- **Properties:**
  - $-1 \leq (\pi_1 - \pi_2) \leq 1$
  - If variables are independent, then $(\pi_1 - \pi_2) = 0$
- **Estimation:** $(p_1 - p_2)$ estimates $(\pi_1 - \pi_2)$
- Example:

		Outcome	
		Cold	No Cold
Treatment Group	vitamin C	.12	.88
	placebo	.22	.78

$$(p_1 - p_2) = (.12 - .22) = -.10$$

or

$$(1 - p_1) - (1 - p_2) = .10$$

Is $(p_1 - p_2) = -.10$ "big" or "small"?

Confidence Interval for Differences of Proportion

Using Agresti's notation, $n_{1+} = N_1$ and $n_{2+} = N_2$.

- An Estimate of the standard error of $(p_1 - p_2)$

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$$

(i.e., standard error of difference between two means)

- Example:

$$\begin{aligned} \hat{\sigma}(p_1 - p_2) &= \sqrt{\frac{(.12)(.88)}{139} + \frac{(.22)(.78)}{140}} \\ &= \sqrt{.00077 + .00123} = .0455 \end{aligned}$$

- So, a large sample $(1-\alpha) \times 100\%$ confidence interval for $(\pi_i - \pi_2)$ is

$$(p_1 - p_2) \pm z_{\alpha/2}\hat{\sigma}(p_1 - p_2)$$

- 95% CI for our example:

$$-.10 \pm 1.96(.045) \quad \longrightarrow \quad (-.19, -.01)$$

Problem with Difference of Proportions

A (fixed) difference maybe more important when both p's are close to 0 or 1 than when both p's are close to .5.

e.g., $(p_1 - p_2) = .09$ can get this from

$$(.10 - .01) = .09 \qquad \textit{or} \qquad (.50 - .41) = .09$$

On the left: 1st p is 10 times larger than 2nd

On the right: 1st p is 1.2 times larger than 2nd.

Another measure for proportions . . .

Relative Risk

Relative Risk

The Relative Risk of a "success" is the ratio of the probabilities,

$$\frac{\pi_1}{\pi_2}$$

Note: $0 \leq \pi_1/\pi_2$.

For observed data, estimate using observed proportions. For the vitamin C and cold study, the relative risk of getting a cold given "exposure" to vitamin C equals

$$\frac{p_1}{p_2} = \frac{.12}{.22} = .552$$

Note: $p_2/p_1 = .22/.12 = 1.8$

The distribution of p_1/p_2 is highly skewed, unless N_1 and N_2 are large.

We'll just let SAS/PROC FREQ or R compute confidence intervals....

Odds Ratios

Fundamental measure (definition) of association for 2×2 tables.

Row 1: Odds of a "success" equals

$$\text{odds}_1 = \frac{\text{P(success} \mid \text{row 1)}}{\text{P(failure} \mid \text{row 1)}} = \frac{\pi_1}{1-\pi_1}$$

$$\begin{aligned} \text{odds} &= 1 \Longrightarrow \text{success \& failure equally likely} \\ \text{odds} &> 1 \Longrightarrow \text{success more likely than failure} \\ \text{odds} &< 1 \Longrightarrow \text{failure more likely than success} \end{aligned}$$

Row 2: Odds of a "success" equals

$$\text{odds}_2 = \frac{P(\text{success} \mid \text{row 2})}{P(\text{failure} \mid \text{row 2})} = \frac{\pi_2}{1-\pi_2}$$

Example: Odds Ratios

	Outcome		
	Cold	No Cold	
vitamin C	$17/139 = .12$	$122/139 = .88$	$.12 + .88 = 1.00$
placebo	$31/140 = .22$	$109/140 = .78$	$.22 + .78 = 1.00$
	$48/279 = .17$	$231/279 = .83$	$.17 + .83 = 1.00$

$$\text{odds(cold given vitamin C)} = \frac{.12}{.88} = .14$$

$$\text{odds(cold given a placebo)} = \frac{.22}{.78} = .28$$

The odds ratio equals the ratio of two odds

$$\text{odds ratio} = \theta = \frac{\text{odds}_1}{\text{odds}_2} = 0.49$$

Odds and Probabilities

There is a one-to-one relationship between odds and probabilities:

$$\begin{aligned}
\frac{\pi_1}{1-\pi_1} &= \text{odds}_1 \\
\frac{1}{\pi_1} - 1 &= \frac{1}{\text{odds}_1} \\
\frac{1}{\pi_1} &= 1 + \frac{1}{\text{odds}_1} \\
\pi_1 &= \frac{\text{odds}_1}{1+\text{odds}_1}
\end{aligned}$$

Odds Ratios and Statistical Independence

If the distributions of the column variable conditional on the rows are the same, then the two variables are statistically independent and

$$\begin{aligned} \pi_1 &= \pi_2 \\ 1-\pi_1 &= 1-\pi_2 \\ \frac{\pi_1}{1-\pi_1} = \text{odds}_1 &= \text{odds}_2 = \frac{\pi_2}{1-\pi_2} \end{aligned}$$

The "**odds ratio**", which is the ratio of two odds, equals 1

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = 1$$

Example:

$$\frac{.12/.88}{.22/.78} = \frac{.14}{.28} = .49$$ Note: without roundoff error

Note: $\text{odds}_1 = \theta \text{odds}_2$, so $(.14) = .49(.28)$ or $(1/.49)(.14) = .28$.

Properties of Odds Ratios

Possible values and their meaning

- Independence when $\pi_1 = \pi_2$ (*conditional probabilities*).

$$\text{odds}_1 = \text{odds}_2 \Longrightarrow \theta = \frac{\text{odds}_1}{\text{odds}_2} = 1$$

- Dependence when $\pi_1 > \pi_2$.

$$\text{odds}_1 > \text{odds}_2 \Longrightarrow 1 < \theta < \infty$$

If $\theta = 5$, individuals in row 1 are more likely to have a "success" than those in row 2.

If $\theta = 5$, the odds of a "success" in row 1 are 5 times the odds in row 2.

Possible values and their meaning

Dependence when $\pi_1 < \pi_2$:

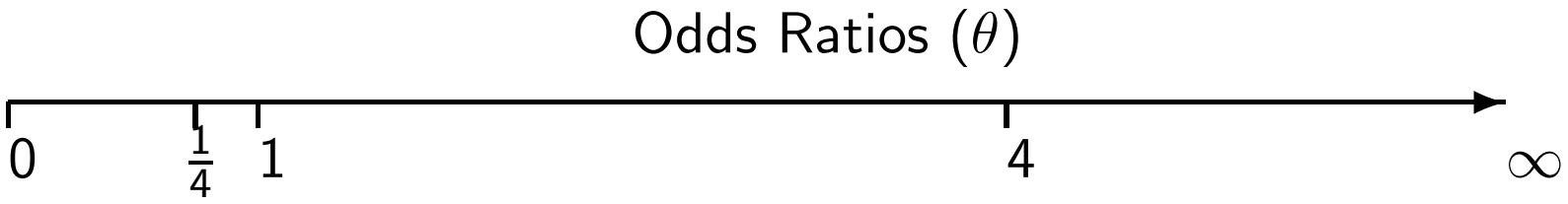
$$\text{odds}_1 < \text{odds}_2 \Longrightarrow 0 < \theta < 1$$

- If $\theta = 0.3$, individuals in row 2 are more likely to have a "success" than those in row 1.
- If $\theta = 0.3$, the odds of a "success" in row 1 are .3 times the odds in row 2.
- If $\theta = 0.3$, the odds of a "success" in row 2 are $(1/.3) = 3.33$ times the odds in row 1.
- In our example, $\hat{\theta} = 0.49$ means that
  - The odds of getting a cold given vitamin C are 0.49 times the odds of getting a cold given a placebo.
  - The odds of getting a cold given a placebo are $(1/.49) = 2.04$ times the odds given vitamin C.
  - Getting a cold is less likely given vitamin C than given a placebo.

Multiplicative Symmetry of θ

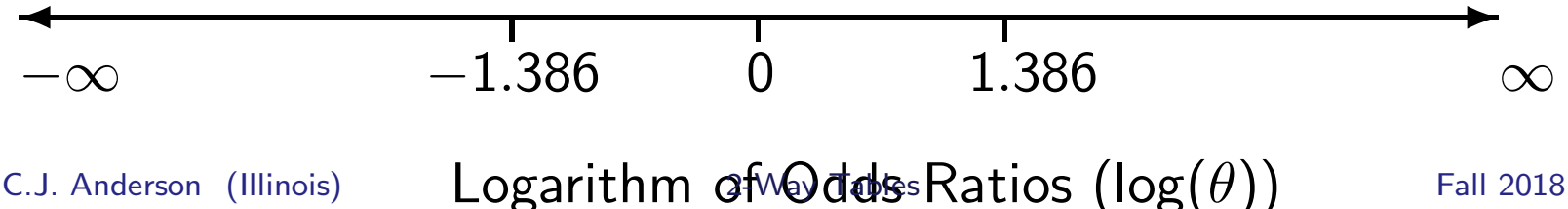
Odds ratios are multiplicatives symmetric around 1.

An association with odds ratio of $\theta = 4$ is of the same strength as one with odds ratio equal to $(1/4) = .25$



The logarithm of odds ratios, and these are *additively symmetric around* 0.

$$\begin{aligned} \log(.25) &= -1.386 \\ \log(1) &= 0 \\ \log(4) &= 1.386 \end{aligned}$$



Invariance of θ Interchanging Categories

Odds ratios are invariant with respect to interchanging the categories of both variables.

	No Cold	Cold
placebo	109	31
vitamin C	122	17

$\longrightarrow$

	No cold	Cold
placebo	.78	.22
vitamin C	.88	.12

$$\hat{\theta} = \frac{.78/.22}{.88/.12} = .490$$

If the categories of just 1 variable are switched, the odds ratio in the re-arranged table will equal $1/\theta$.

	Cold	No Cold
placebo	.22	.78
vitamin C	.12	.88

$$\longrightarrow \hat{\theta} = \frac{.22/.78}{.12/.88} = 1/.490 = 2.07$$

Invariance of θ Interchanging Variables

Odds ratios are invariant with respect to interchanging variables (i.e., odds ratios are symmetric with respect to variables).

	Vitamin C	Vitamin Placebo			Vitamin C	Vitamin Placebo
Cold	17	31	48	$\longrightarrow$	.35	.65
No Cold	122	109	231		.53	.47

$$\begin{aligned} \text{odds}_1 &= .35/.65 = .548 \\ \text{odds}_2 &= .53/.47 = 1.119 \end{aligned}$$

$$\hat{\theta} = \frac{\text{odds}_1}{\text{odds}_2} = \frac{.548}{1.119} = .490$$

θ is the Cross-product Ratio

When both variables are "response" variables. (e.g., concussions playing football crossed by situation and activity).

$$\begin{aligned}\theta &= \frac{(\pi_{11}/\pi_{1+})/(\pi_{12}/\pi_{1+})}{(\pi_{21}/\pi_{2+})/(\pi_{22}/\pi_{2+})} \\ &= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \\ &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\end{aligned}$$

"**Cross–product ratio**" in the 2×2 table:

π_{11}	π_{12}
π_{21}	π_{22}

For sample data,

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

θ and Marginal Distributions

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

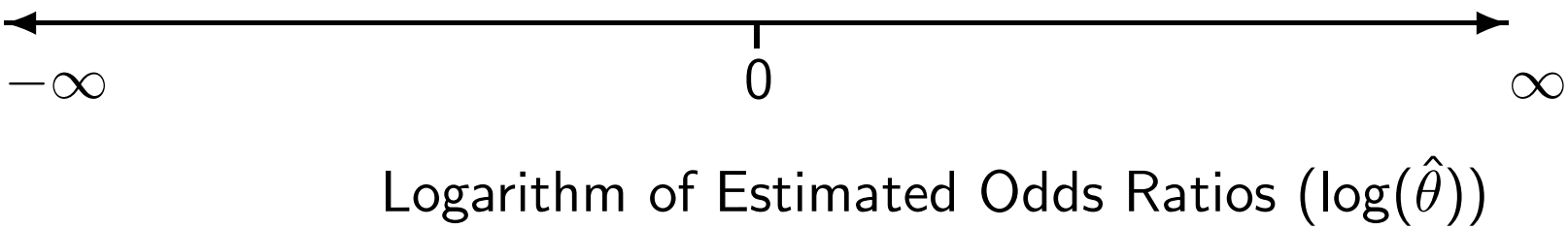
Odds ratios do not depend on the marginal distributions of either variable.

Odds ratios only depend on cell probabilities (proportions or counts) and not on marginal values.

Odds Ratios and Inference

Problem: The sampling distribution of $\hat{\theta}$ can be very skewed. Suppose that $\theta = 1$, $\hat{\theta}$ can't be much smaller than θ but it could be much larger than θ.

Solution: Use $\log(\theta)$ and $\log(\hat{\theta})$ for inference.



Sampling Distribution of $\log(\hat{\theta})$

Asymptotic Standard Error (ASE) of $\log(\hat{\theta})$ is

$$ASE(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

As n_{ij}'s get bigger

- $ASE(\log\hat{\theta})$ gets smaller.
- The sampling distribution of $\log\hat{\theta} \longrightarrow \mathcal{N}(\log(\theta), \sigma^2)$

So, a $(1-\alpha)\times 100\%$ large sample confidence interval for $\log(\theta)$ is

$$\log(\hat{\theta}) \pm z_{\alpha/2} ASE(\log\hat{\theta})$$

Confidence Interval for θ

Example: A 95% CI for $\log(\theta)$ for our vitamin C/cold data is

$$\begin{aligned} \log(.490) &\pm 1.96\sqrt{\frac{1}{17}+\frac{1}{109}+\frac{1}{122}+\frac{1}{31}} \\ -0.7134 &\pm 1.96(.32932) \longrightarrow (-1.3589, -.067977) \end{aligned}$$

And to get the CI for θ, just take exp:

$$(.257, .934)$$

INPUT SAS and Measures of Association

```
DATA skiers;
INPUT treat $ outcome $ count ;
LABEL treat ='Treatment Group';
DATALINES;
     placebo          cold        31
     placebo          no_cold    109
     ascorbic_acid    cold        17
     ascorbic_acid    no_cold    122
;
PROC FREQ;
    WEIGHT count;
    TABLES treat*outcome / nopercent nocol chisq measures;
RUN;
```

Produced by "TABLES"

Table of treat by outcome

Frequency
Row Pct

	Outcome		
treat(Treatment Group)	cold	no_cold	Total
ascorbic	17	122	139
	12.23	87.77	
placebo	31	109	140
	22.14	77.86	
Total	48	231	279

Produced by "CHISQ"

Statistics for Table of treat by outcome

Statistic	DF	Value	Prob
Chi-Square	1	4.8114	0.0283
Likelihood Ratio Chi-Square	1	4.8717	0.0273
Continuity Adj. Chi-Square	1	4.1407	0.0419
Mantel-Haenszel Chi-Square	1	4.7942	0.0286
Phi Coefficient		-0.1313	
Contingency Coefficient		0.1302	
Cramer's V		-0.1313	

Produced by "CHISQ"

Fisher's Exact Test	
Cell (1,1) Frequency (F)	17
Left-sided Pr <= F	0.0205
Right-sided Pr >= F	0.9910
Table Probability (P)	0.0115
Two-sided Pr <= P	0.0385

Produced by "MEASURES"

Statistics for Table of treat by outcome

Statistic	Value	ASE
Gamma	-0.3423	0.1454
Kendall's Tau-b	-0.1313	0.0581
Stuart's Tau-c	-0.0991	0.0448
Somers' D C|R	-0.0991	0.0448
Somers' D R|C	-0.1740	0.0764
Pearson Correlation	-0.1313	0.0581
Spearman Correlation	-0.1313	0.0581
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0935	0.1041
Lambda Symmetric	0.0695	0.0784
Uncertainty Coefficient C|R	0.0190	0.0169
Uncertainty Coefficient R|C	0.0126	0.0113
Uncertainty Coefficient Symmetric	0.0152	0.0135

Also Produced by "MEASURES"

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.4900	0.2569	0.9343
Cohort (Col1 Risk)	0.5523	0.3209	0.9506
Cohort (Col2 Risk)	1.1273	1.0120	1.2558

Sample Size = 279

Note: If a cell equals 0, then you can add .5 to all cells and compute statistics.

R: Data input & Basic Measures

```
# Give names to variable and levels
var.levels ← expand.grid(cold=c('yes','no'),
                         treatment=c('vitamin C','placebo'))
# Create a data frame with 2 x 2 = 4 cases
colds.freq ← data.frame(var.levels,
                        count=c(17,122,31,109))
# Put data in tabular form
( colds.tab ← xtabs(count ~ treatment + cold,
                    data=colds.freq) )
# Compute various measures:  Note:  includes
correction for continuity.
library(Epi)
colds.mat ← as.matrix(colds.tab)
twoby2(colds.mat)
```

R Output

	yes	no	P(yes)	95% conf.	interval
vitamin C	17.00	122.00	0.12	0.08	0.19
placebo	31.00	109.00	0.22	0.16	0.30

		95% conf.	interval
Relative Risk:	0.55	0.32	0.95
Sample Odds Ratio:	0.49	0.26	0.93
Conditional MLE Odds Ratio:	0.49	0.24	0.97
Probability difference:	-0.10	-0.19	-0.01

Note: I used the `xtable` command to get LaTex code for above tables.

Alternative (better?) R Script

The correction for continuity isn't all that common any more, to get G^2 and X^2 without it, the easiest ways is to use

```
# Put data in tabular form
( colds.tab ← xtabs(count ~ treatment + cold,
                    data=colds.freq) )
# Compute various measures:
colds.mat ← as.matrix(colds.tab)
chisq.test(colds.mat,correct=FALSE)
```

which gives just X^2, or

```
library(MASS)
( model.loglm ← loglm(~ 1 + 2, colds.tab) )
```

which gives X^2 and G^2.

Odds Ratio, Relative Risk & Confusion

"The Odds of Execution" (1994), *Technology Review*, p 42–43.
1987 U.S. Supreme Court McClesky vs Kemp ruling dealing with Georgia death sentencing and race of victim

> *"...even after taking account of 39 nonracial variables, defendants charged with killing white victims were* 4.3 *time more likely to receive a death sentence as defendants charged with killing blacks."*

- $P(D|W)$ = probability death sentence given victim is white.
- $P(D|B)$ = probability death sentence given victim is black.
- Incorrect interpretation: $P(D|W) = 4.3P(D|B)$.
- Correct interpretation

$$\frac{P(D|W)}{1-P(D|W)} = 4.3\left(\frac{P(D|B)}{1-P(D|B)}\right)$$

- Suppose that $P(D|W) = .99$,

$$\frac{P(D|B)}{1-P(D|B)} = \frac{1}{4.3}\left(\frac{.99}{.01}\right) = 23 \quad \text{and} \quad P(D|B) = .96$$

Relationship between Odds Ratios & Relative Risk

Odds ratios and relative risk are related:

$$\begin{aligned} \text{odds ratio} &= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \\ &= \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} \\ &= \text{relative risk}\frac{(1-\pi_2)}{(1-\pi_1)} \end{aligned}$$

If both π_1 and π_2 are small, then

- $(1-\pi_1)$ and $(1-\pi_2)$ are large.
- $(1-\pi_2)/(1-\pi_1)$ is close to 1.
- odds ratio $\approx$ relative risk.

Example: Odds Ratios & Relative Risk

Back to our vitamin C and cold example, where

$p_1 = .12$ (cold given vitamin C),
$p_2 = .22$ (cold given placebo), and

$$\frac{(1-p_2)}{(1-p_1)} = .887$$

which is "close" to 1, and

$$\begin{aligned} \text{odds ratio} &= .490 \\ \text{relative risk} &= .552 \end{aligned}$$

Using Odds Ratios/Relative Risk Relationship

To compute a rough estimate relative risk when relative risk cannot be estimated — case control studies.

Oral contraceptive use and heart attack (Agresti, 1990): This is a **Retrospective, case control** study.

Oral	Heart Attack				Heart Attack	
Contraceptive	Yes	No			Yes	No
Used	23	34	57	$\longrightarrow$	.40	.60
Never	35	132	167		.21	.79
	58	160	224			

Heart Attack is the response variable and we want to condition on oral contraceptive use (the explanatory variable).

But this does **not** make sense here, because the column margin is fixed by design.

Estimating Relative Risk via Odds Ratio

- The each **column** is a separate binomial distribution.
- The column marginal proportions do not reflect population marginal probabilities.

Assuming the π_1 and π_2 (conditional probabilities of heart attack given oral contraceptive use/no use) are small in the population, then

$$\text{odds ratio} = \hat{\theta} = \frac{(23)(132)}{(35)(34)} = 2.55$$

and this is a rough indication of the relative risk π_1/π_2.

Correlation (phi coefficient)

Pearson product moment correlation.

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

(can also use proportions or probabilities).

Properties:

- Depends on margins as well as cells.
- $-1 \leq r \leq 1$.
- r can only equal 1 if the margins are equal.
- $r = \sqrt{X^2/n}$.

Example: Correlation

	Outcome		
	Cold	No Cold	
vitamin C	17	122	139
placebo	31	109	140
	48	231	

$$
\begin{aligned}
r &= \frac{(17)(109) - (31)(122)}{\sqrt{(139)(140)(48)(231)}} \\
&= -1929/146892 \\
&= -.131
\end{aligned}
$$

Also, $r = \pm\sqrt{\chi^2/N} = \pm\sqrt{4.8114/279} = -.131$

Association versus Marginal Distribution

The information in 2–way tables can be separated into 2 parts:

- 2 Marginal distributions.
- Association between the 2 variables.

With contingency tables,

- There are many differ possible measures for a 2×2 table.
- More that one measure (statistic) is needed to fully describe association in tables larger than 2×2.

With 2 continuous (numerical) variables that have a (joint) bivariate normal distribution

- Marginal distributions described by mean and variance of each variables.
- Association described by the correlation (covariance) coefficient.

Characterizing Association in Table w/ θ

For a 2–way, $I \times J$ table, you need a set of $(I-1)(J-1)$ odds ratios to completely characterize the association between the 2 variables.

Start with	IJ	cells
Take out	-1	because $\sum_i \sum_j n_{ij} = n$ (or $\sum_i \sum_j p_{ij} = 1$)
marginal	$-(I-1)$	because $\sum_i n_{i+} = n$
information	$-(J-1)$	because $\sum_j n_{+j} = n$
which leaves	$(I-1)(J-1)$	pieces of information

Characterizing Association in Table w/ θ

Gender by political party identification (from GSS, 1991)

	Party Identification			
Gender	Democrat	Independent	Republican	
Females	279	73	225	577
Males	165	47	191	403
	444	120	416	980

Need 2 odds ratios:

	Democrat	Independent
Females	279	73
Males	165	47

$\theta = (279)(47)/(165)(73)$
$= 1.089$

	Independent	Republican
Females	73	225
Males	47	191

$\theta = (73)(191)/(47)(225)$
$= 1.318$

The third possible odds ratio is

$$\begin{aligned}\theta_{(FM:DR)} &= \theta_{(FM:DI)}\theta_{(FM:IR)} \\ &= (1.089)(1.318) \\ &= 1.435\end{aligned}$$

and as a check

$$\theta_{(FM:DR)} = (279)(191)/(165)(225) = 1.435$$

Summary Comments on Measures

Goodman & Kruskal papers (1979):

- (1954) Criteria for judging measures, some new ones for specific contexts.
- (1959) Supplement to '54 paper, historical & bibliographic material.
- (1963) Large sample standard errors for sample inference.
- (1972) Unified way to derive asymptotic variances.

Four general classes of measures:

- Measures based on cross-product ratio for 2×2 tables
- Measures based on X^2 used to test independence in 2–way table (i.e., correlation).
- "Proportional reduction of error" measure that indicates the relative value of using row categories to predict column categories.
- "Proportion of explained variance" measure which provides a close analogy for categorical data to the squared correlation coefficient for continuous data.

Comments on Measures of Association

- **Normed** measures such that they lie between 0 & 1 or -1 & 1.
- **Symmetric vs Asymmetric** measures focus on joint or conditional distributions.
- For more information (and more references), see Agresti (2012) or Wickens (1989).

Types of Designs:

- Retrospective
  - Case-controls
- Prospective
  - Clinical trials (experiments)
  - Cohort studies
- Cross-Sectional

Retrospective

Retrospective or "look into the past".

- Sample those with and those without attribute of interest.
- Used to ensure that you have enough cases for events that are relatively rare in the population.
- Example: Oral contraceptive use and heart attacks (Agresti, 1990)

  Subjects: 58 women under age 45 treated for heart attack in two hospital regions in England and Wales during 1968–1972. Each case was matched with 3 control patients in the same hospital who were not being treated for heart attack.

		Heat Attack	
		Yes	No
Oral Contraceptive Use	Used	23 (.397)	34 (.205)
	Never Used	35 (.603)	132 (.795)
	total	58 (1.00)	166 (1.00)

Prospective

Prospective or "look into the future". Take a sample, wait some period of time, then count the number of outcomes/events/attributes of interest.

There are 2 kinds of prospective studies:

Clinical trials & Cohort Studies

Clinical trials (experiments): Subjects are randomly assigned to groups. Examples:

- The French study on Vitamin C and colds.
- Study by Alper & Raymond (1995) Imposing Views, Imposing Shoes: A Statistician as a Sole Model. *American Statistician, 49*, 317–319.
  - Classes assigned randomly to one of two groups — control in which professor wore ordinary shoes and treatment group in which professor wore Nikes.
  - After 3 times/week for 14 weeks, checked to see if students bought Nikes or not.

Cohort Study

Cohort study: Subjects make their own choice as to which group they belong or "come as they are".

Examples:

- Women decide whether they will use oral contraceptive.
- Kramer data on sibling acceptance by gender.
- Sample students at different ages (e.g., accelerated longitudinal design)

Cross-sectional:

Take a sample from the population of interest and record which group a person falls into and the outcome of interest.

Examples:

- Take a large sample of women and record whether they have heart disease (or high blood pressure) and whether or not they have used oral contraceptives.
- Job Satisfaction (from Andersen) — part of a large scale investigation of blue collar workers in Denmark (1968).

3 variables:	own satisfaction	(low, high)
	supervisor's satisfaction	(low, high)
	management quality	(bad, good)

Observational and Experimental:

- Observational:
  - Case-control.
  - Cohort.
  - Cross-sectional.
- Experimental:
  - Clinical trials.
- Longitudinal or repeated measures.

. . . Next: Testing relationships