Lecture 4: Association Measures and Variance Estimation

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Terminology

The following denotes a standard 2 x 2 table.

		Column		
		1	2	
Row	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$N = n_{\cdot\cdot}$

- $n_{1\cdot} = \sum_j n_{ij}$ represents the sum of row 1 *over* columns of J (1 and 2 in this example).
- This table can be generalized into an IxJ table where I represents the number of rows and J represents the number of columns.

Comparing Two Proportions

- Suppose you want to compare binary responses across a two factor group effect.
- Denote the response variable as Y and group variable as X.
- Let p_1 denote the probability of success given group 1 (i.e., $P(Y = 1|X = 1)$)
- Let p_2 denote the probability of success given group 2 (i.e., $P(Y = 1|X = 2)$)

  In terms of a (product binomial) contingency table,

		Binary Response (Y)		
		Success (Y=1)	Failure (Y=2)	
Group (X)	1	p_1	$1 - p_1$	1
	2	p_2	$1 - p_2$	1

- When $p_1 \neq p_2$, we want to quantify how the two probabilities are different or are associated.
- In other words, we want a single measure of how the treatments differ.

Quantifying treatment differences

There are two general classes of statistics that measure “association” of variables.

1. Absolute Measures
   - Measure the actual reduction in number of cases
   - Often used in a public health prevention study where the total number of cases reduced is of value
   - Absolute measures are relevant to the group as a whole (limits interpretation and application)
2. Relative Measures
   - Express how much more likely one group is to experience the outcome compared to another group
   - Relative measures can be interpreted at the individual level.

The research objective assists in the determination of which measure to use. Fortunately, both classes of measurement are obtainable from the same dataset.

Absolute Measures

Risk Difference

- Let Δ be defines as follows:

$$\Delta = p_1 - p_2, \qquad -1 \leq \Delta \leq 1,$$

- When the two rows are similar, $\Delta \rightarrow 0$ and indicates no group differences.
- Suppose $p_1 = .1$ and $p_2 = .2$ then $\Delta = .1 - .2 = -.1$

Number Needed to Treat (NNT)

- 'NNT' is defined as the inverse of the absolute risk reduction
- i.e., $NNT = 1/\Delta$

Example NNT

The results of the Diabetes Control and Complications Trial* into the effect of intensive diabetes therapy on the development and progression of neuropathy indicated that neuropathy occurred in 9.6% of patients randomized to usual care and 2.8% of patients randomized to intensive therapy. The NUMBER of patients we NEED TO TREAT with the intensive diabetes therapy to prevent one additional occurrence of neuropathy can be determined as follows:

RD = |9.6% - 2.8%| = 6.8%
NNT = 1/RD = 1/6.8

We therefore need to treat 15 diabetic patients with intensive therapy to prevent one from developing neuropathy.

*(Ann Intern Med 1995; 122:561-8)

NNT in details

Definitions

		TREATED	CONTROLS
ADVERSE EVENT	YES	a	b
	NO	c	d

LET:

pc = proportion of subjects in control group who suffer an event

pc = b / (b+d)

pt = proportion of subjects in treated group who suffer an event

pt = a / (a+c)

er = expected/baseline risk in untreated subjects

THEN:

Relative risk of event (RRe) = pt / pc

NNT

Relative risk of no event (RRne) = (1-pt) / (1-pc)

Odds ratio (OR) = (a*d) / (b*c)

Relative risk reduction (RRR) = (pc-pt) / pc = 1-RRe

Absolute risk reduction (ARR)/ risk difference (RD) = pc-pt

Number needed to treat (NNT):

NNT [risk difference] = 1 / RD

NNT [relative risk of event] = 1 / (pc*RRR)

NNT [relative risk of no event] = 1 / ((1-pc)*(RRne-1))

NNT [odds ratio] = (1-(pc*(1-OR)) / (pc*(1-pc)*(1-OR))

The most commonly quoted NNT statistic is NNT [risk difference] or the empirical NNT, which assumes a constant risk difference over different expected event rates.

Movement towards relative measures

- When p_1 or p_2 is close to 0 or 1, then Δ may have greater meaning.

Example:

Scenario A: Let $p_1 = 0.010$ and $p_2 = 0.001$, then $\Delta_a = 0.009$.

Scenario B: Let $p_1 = 0.410$ and $p_2 = 0.401$, then $\Delta_b = 0.009$.

Note that both $\Delta_a = \Delta_b = 0.009$, but that a 0.009 unit change in Scenario A seems more important than a 0.009 unit change in Scenario B.

This "importance" is quantified by Relative Measures of Association

Relative Risk or Risk Ratio

Define, Relative Risk (RR) as

$$RR = \frac{p_1}{p_2} \qquad 0 \leq RR \leq \infty,$$

A $RR = 1$ indicates independence (no association).

For the previous scenarios,

$$\begin{aligned} RR_a &= 0.010/0.001 \\ &= 10.0 \\ \\ RR_b &= 0.410/0.401 \\ &= 1.02 \end{aligned}$$

The estimate of RR dependent on the definition of the “success”. We will explore this concept later.

Log-Relative Risk

The log-relative risk is often used to alleviate the restrictions that the relative risk must be positive:

$$\log RR = \log \left(\frac{p_1}{p_2} \right) = \log(p_1) - \log(p_2)$$

where

$$-\infty \leq \log RR \leq \infty.$$

Log(RR) is also directly estimable using generalized linear models (GLM).

Relative Measures - Odds Ratio

- Recall the definition of odds,

$$\frac{p_t}{(1-p_t)} \quad = \quad \text{odds of success versus failure on group } t$$

- Then the ratio of the odds (odds ratio or OR) for group 1 to group 2 is

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} \qquad 0 \leq OR \leq \infty,$$

- Again, the log-odds ratio is often used to alleviate the restrictions that the odds ratio must be positive, i.e.,

$$\begin{aligned}\log OR &= \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)\\ &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right)\\ &= \text{logit}(p_1) - \text{logit}(p_2)\end{aligned}$$

  where $-\infty \leq \log OR \leq \infty$

- Note that the $\log(OR)$ is the difference in logits.

Additional Examination of OR

$$OR = \frac{\frac{P(Y=1|X=1)}{P(Y=2|X=1)}}{\frac{P(Y=1|X=2)}{P(Y=2|X=2)}}$$

Using Bayes's Law,

$$\begin{aligned}
P(Y=1|X=1) &= P(Y=1\bigcap X=1)/P(X=1)\\
P(Y=2|X=1) &= P(Y=2\bigcap X=1)/P(X=1)\\
P(Y=1|X=2) &= P(Y=1\bigcap X=2)/P(X=2)\\
P(Y=2|X=2) &= P(Y=2\bigcap X=2)/P(X=2)
\end{aligned}$$

SO

$$\begin{aligned}
OR &= \frac{P(Y=1\bigcap X=1)P(Y=2\bigcap X=2)}{P(Y=1\bigcap X=2)P(Y=2\bigcap X=1)}\\
&= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}
\end{aligned}$$

Estimating OR

We will see later that some study designs allow for you to estimate only one of the following

1. $P(Y = i \bigcap X = j) = \pi_{ij}$ (cross sectional data)
2. $P(Y = i | X = j)$ (prospective study stratified by row)
3. $P(X = j | Y = i)$ (retrospective (case-control) study stratified by column)

Regardless of the study design (or sampling mechanism), through the previous equalities, OR can be estimated by

$$\widehat{OR} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

since,

$$\widehat{\pi_{ij}} = n_{ij}/n$$

However, RR is only defined in studies where you can estimate item 2 above and item 2 occurs naturally in prospective studies.

Example

Suppose you observe the following:

		Outcome		
		Cold	No Cold	
Treatment	Vitamin C	17	122	139
	No Vitamin C	31	109	140
		48	231	279

We want to estimate RR, OR, log(RR) and log(OR).

$$\begin{aligned} RR &= p_1/p_2 \\ &= \frac{17/139}{31/140} \\ &= 0.5523 \end{aligned}$$

$$log(RR) = log(0.5523) = -0.5937$$

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \frac{17\times 109}{31\times 122} \\ &= 0.4900 \end{aligned}$$

$$log(OR) = log(0.4900) = -0.7133$$

Interpretation

In this example, our "success" was catching a cold. So, the following represent the correct interpretation of the estimates of RR and OR.

RR: The proportion of subjects likely to develop a cold who are under Vitamin C supplement is 0.5523 times the proportion to develop a cold NOT under Vitamin C supplement / Subjects taking Vitamin C supplements were about 46% (1 - 0.5523) less likely to develop a cold than subjects who did not take Vitamin C supplements

OR: The odds of catching a cold from those under Vitamin C supplement is 0.49 times the odds for those NOT under Vitamin C supplement/The odds of catching a cold for subjects taking Vitamin C supplements were 51% (1 - 0.49) less than subjects not taking Vitamin C.

Note: When you use RR, you can discuss likelihood (or probability of an outcome), but when you use OR, you can only draw inference on ODDS.

Properties of OR

Previously, we defined a "success" as catching a cold. It would seem reasonable that a successful treatment would prohibit a cold. Therefore, a success could have been defined as "no cold".

If we "flip" the columns, we get

		Outcome		
		No Cold	Cold	
Treatment	Vitamin C	122	17	139
	No Vitamin C	109	31	140
		231	48	279

and $OR_{\text{No Cold}} = (122 * 31)/(109 * 17) = 2.041$ and
$RR_{\text{No Cold}} = (122/139)/(109/140) = 1.127$.

Reciprocals of OR and RR

Note that

$$\begin{aligned} OR_{\text{No Cold}} &= 2.041 \\ &= 1/.4900 \\ &= 1/OR_{\text{Cold}} \end{aligned}$$

but that

$$\begin{aligned} RR_{\text{No Cold}} &= 1.127 \\ &\neq 1/.5523 \quad (1/.5523 = 1.81) \end{aligned}$$

Interpretation

Odds Ratio (for preventing a cold):

The odds of not catching a cold while taking vitamin c supplements is twice the odds of not catching a cold when not taking vitamin c.

Relative Risk (for preventing a cold):

An individual taking vitamin c supplements is about 12% more likely to avoid catching a cold than a person who does not take the vitamin c supplements.

Note: $OR \approx 2$; however, this does not mean that $p_1 \approx 2 \cdot p_2$ (that is a relative risk of 2).

Why the log?

It is not easy to see that 2.041 and 0.490 represent the same level of effect. However, in log terms
log(2.041) = 0.713

and

log(0.490) = -0.713

Now, you can see that both represent the same level of effect, just in different in direction.

Additional advantages of thinking in terms of logs is that log(ODDS) (or logits) are a special case of a generalized regression model we will discuss later.

Relationship of OR to RR

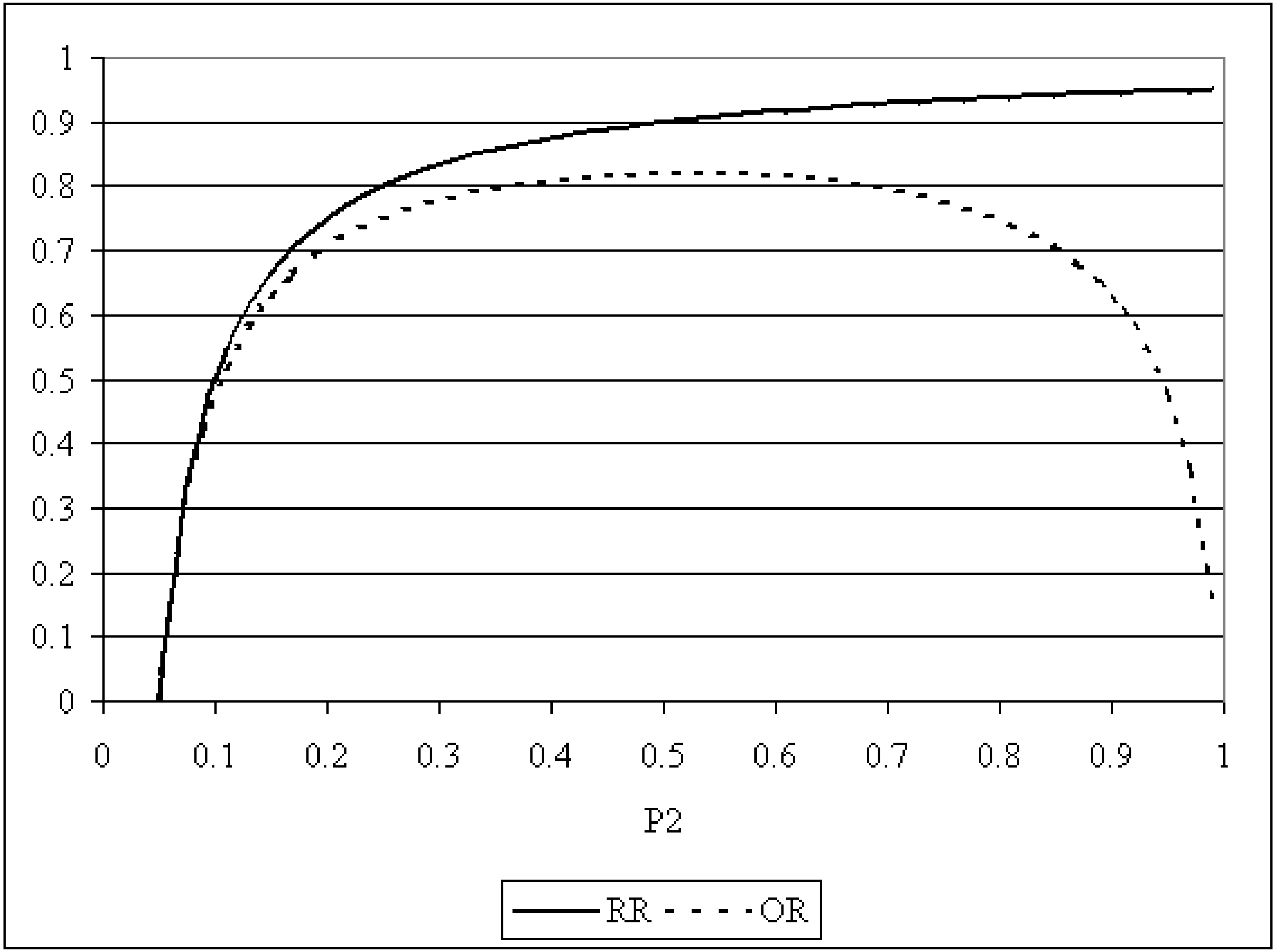
$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p2/(1-p_2)} \\ &= \frac{p_1}{p_2} \cdot \frac{1-p_2}{1-p_1} \\ &= RR \cdot \frac{1-p_2}{1-p_1} \end{aligned}$$

$\frac{1-p_2}{1-p_1}$ represents the bias when using OR as an estimate for RR.

When “p is small” for both the groups, $\frac{1-p_2}{1-p_1} \approx 1$, $OR \approx RR$. See section 2.2.5 of Agresti. Again, in this case, odds-ratio provides a rough indication of the relative risk (when it is not directly estimable in case of a case-control study).

However, when “p is large”, $\frac{1-p_2}{1-p_1} \neq 1$, so OR provides a poor estimate for RR.

Figure 1 Relative Risk and Odds Ratio for a fixed risk difference of $RD = P_1 - P_2 = -0.05$



Treatment Difference

Note Δ, $\log(RR)$ and $\log(OR)$ can be considered treatment differences on different scales, each can be written as

$$g(p_1) - g(p_2)$$

for the appropriate function $g(a)$:

TREATMENT DIFFERENCE	$g(a)$	$g(p_1) - g(p_2)$
RISK DIFF	a	$p_1 - p_2$
log (RR)	$\log(a)$	$\log(p_1) - \log(p_2)$
log (OR)	$\log\left(\frac{a}{1-a}\right)$	$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right)$

The function $g(\cdot)$ will surface later and be called the “link” function.

One sided Alternatives

Note, under the null H$_0$: $p_1 = p_2 = p$, the treatment difference on all scales equals 0, i.e.,

$$g(p_1) - g(p_2) = g(p) - g(p) = 0.$$

In general, we can form the following table:

	Null (H$_0$) NO ASSOCIATION	(H$_{A1}$) ALTERNATIVE 1	(H$_{A2}$) ALTERNATIVE 2
PROBS	$p_1 = p_2$	$p_1 > p_2$	$p_1 < p_2$
RISK DIFF	$\Delta = 0$	$\Delta > 0$	$\Delta < 0$
log (RR)	$\log(RR) = 0$	$\log(RR) > 0$	$\log(RR) < 0$
log (OR)	$\log(OR) = 0$	$\log(OR) > 0$	$\log(OR) < 0$
	All 3 = 0	All 3 > 0	All 3 < 0

All measures are in the same direction ($+$, $-$ or 0).

Motivation

- In categorical data analysis, we often take a function of a statistic
- For example,

$$\text{se}(p) = \sqrt{\frac{p(1-p)}{n}}$$

- As presented before, we may be interested in

$$\text{se}\left(\log\left(\frac{p}{1-p}\right)\right)$$

- That is, the standard error of the logit (p)
- Since p and $1-p$ are statistically dependent, this computation can be deceptively difficult

Delta Method

- The delta method is a useful method to derive the asymptotic variance of a test statistic
- Let $f(\theta)$ be a function of a statistic
- Then, according to the delta method, the standard error of $f(\theta)$ is

$$\text{se}\left(f(\theta)\right) = \left|\frac{d\ f(\theta)}{d\theta}\right| \text{se}(\theta)$$

Example - Sample logit

- Consider the following function of the binomial parameter

$$\log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

- Once again p and $1-p$ are statistically dependent, so the “variance of the sum is not the sum of the variances”
- We will apply the delta method. To do so we need to calculate

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

- Therefore,

$$\begin{aligned} \text{se}\left(\log(\tfrac{p}{1-p})\right) &= \left|\tfrac{1}{p(1-p)}\right| \sqrt{\tfrac{p(1-p)}{n}} \\ &= \tfrac{1}{\sqrt{np(1-p)}} \end{aligned}$$

Multivariate extension of the delta method

- Suppose, that $\theta = f(p_{11}, p_{12}, p_{21}, p_{22})$ where p_{ij} is defined as below

		Column		
		1	2	
Row	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	$N = n_{\cdot\cdot}$

- We want to derive the variance of

$$\theta = OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

- The multivariable version of the delta method is

$$\text{Var}\left(\widehat{\theta}\right) \approx \nabla f(p_{11}, p_{12}, p_{21}, p_{22}) \cdot Cov(p_{11}, p_{12}, p_{21}, p_{22}) \cdot \nabla f(p_{11}, p_{12}, p_{21}, p_{22})^T$$

- Where ∇ is the gradient vector. That is

$$\nabla f(p_{11}, p_{12}, p_{21}, p_{22}) = \left(\frac{\partial f}{\partial p_{11}}, \ldots, \frac{\partial f}{\partial p_{22}}\right)$$

Example - Variance of log odds ratio

- We want to estimate

$$Var(\log(OR)) = Var\left[\log\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right)\right]$$

- Let the function f be

$$f = (\log p_{11} - \log p_{12} - \log p_{21} + \log p_{22})$$

- Since these are not independent, we need to use the delta method
- Note that ∇f is

$$\nabla f = \left(\frac{1}{p_{11}}, \frac{-1}{p_{12}}, \frac{-1}{p_{21}}, \frac{1}{p_{22}}\right)$$

Variance Covariance Matrix

- The variance covariance matrix for a multinomial distribution with $c = 4$ categories

$$\Sigma = \frac{1}{n}\begin{bmatrix} p_{11}(1-p_{11}) & -p_{12}p_{11} & -p_{21}p_{11} & -p_{22}p_{11} \\ -p_{11}p_{12} & p_{12}(1-p_{12}) & -p_{21}p_{12} & -p_{22}p_{12} \\ -p_{11}p_{21} & -p_{12}p_{21} & p_{21}(1-p_{21}) & -p_{22}p_{21} \\ -p_{11}p_{22} & -p_{12}p_{22} & -p_{21}p_{22} & p_{22}(1-p_{22}) \end{bmatrix}$$

- Then $\nabla f \Sigma$ equals

$$\begin{aligned} \nabla f \Sigma &= \left(\frac{1}{p_{11}}, \frac{-1}{p_{12}}, \frac{-1}{p_{21}}, \frac{1}{p_{22}}\right) \times \\ & \quad n^{-1}\begin{bmatrix} p_{11}(1-p_{11}) & -p_{12}p_{11} & -p_{21}p_{11} & -p_{22}p_{11} \\ -p_{11}p_{12} & p_{12}(1-p_{12}) & -p_{21}p_{12} & -p_{22}p_{12} \\ -p_{11}p_{21} & -p_{12}p_{21} & p_{21}(1-p_{21}) & -p_{22}p_{21} \\ -p_{11}p_{22} & -p_{12}p_{22} & -p_{21}p_{22} & p_{22}(1-p_{22}) \end{bmatrix} \\ &= n^{-1}\left[(1-p_{11}+p_{11}+p_{11}-p_{11}), (-p_{12}-(1-p_{12})+p_{12}-p_{12}), \ldots\right] \\ &= n^{-1}\left[1, -1, -1, 1\right] \end{aligned}$$

- We now need $(\nabla f \Sigma) \times \nabla f^T$

- $(\nabla f \Sigma) \times \nabla f^T$ equals

$$\begin{aligned} &= n^{-1}\,[1,-1,-1,1] \times \begin{bmatrix} \frac{1}{p_{11}} \\ -\frac{1}{p_{12}} \\ -\frac{1}{p_{21}} \\ \frac{1}{p_{22}} \end{bmatrix} \\ &= n^{-1}\left[\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right] \end{aligned}$$

- Thus the variance of the log odds ratio is approximately

$$Var\widehat{(\log(OR))} = \frac{1}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right)$$

- substituting the MLEs for $\widehat{p_{ij}} = n_{ij}/n$ yields

$$Var\widehat{(\log(OR))} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

according to the delta method

- What about the variance of the odds ratio (instead of the log-odds)?
- We want the

$$Var\left(\widehat{\theta}\right)$$

- where $\theta = OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}$
- We could use the delta method to estimate this variance, but give it a try
- The partials in the gradient vector are rather unwieldily for matrix multiplication by hand
- So what do we do?
- We rely on another calculus “trick”
- That is, we will use the Taylor’s approximation of a function

- Suppose you know $E(X) = \mu$ and $Var(X) = \sigma^2$
- Let $Y = g(X)$ where g has the first two derivatives defined.
- That is, g' and g'' exist.
- Then, a second order Taylor Polynomial centered at μ is

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu) + \frac{1}{2}g''(\mu)(X - \mu)^2$$

- Then

$$\begin{aligned} E(Y) = E(g(X)) \quad &\approx \quad E(g(\mu)) + E(g'(\mu)(X - \mu) + E(\tfrac{1}{2}g''(\mu)(X - \mu)^2 \\ &= \quad g(\mu) + g'(\mu)(\mu - \mu) + \tfrac{1}{2}g''(\mu)E(X - \mu)^2 \\ &= \quad g(\mu) + \tfrac{1}{2}g''(\mu)\sigma^2 \end{aligned}$$

- A first order polynomial would yield

$$Y = g(X) \approx g(\mu) + g'(\mu)(X - \mu)$$

- We will use the zero order approximation for the variance estimation

$$
\begin{aligned}
Var(Y) &= Var(g(X)) \\
&= E\left[(g(x) - E(g(x)))^2\right] \\
&\approx E\left[(g(\mu) + g'(\mu)(X-\mu) - g(\mu))^2\right] \\
&= [g'(\mu)]^2 \, E\left[(X-\mu)^2\right] \\
&= [g'(\mu)]^2 \, Var(X)
\end{aligned}
$$

- Thus, for the variance of the odds ratio, consider the following function of the log-odds ratio

$$g(\log OR) = \exp(\log(OR)), X = \log(OR)$$

- Then by the Taylor expansion

$$Var(\exp(\log(OR))) \approx [OR]^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

- since

$$\frac{de^{\log(OR)}}{d\log(OR)} = e^{\log(OR)}$$

- and

$$e^{\log(OR)} = OR$$