Variance Derivations 1

Basic Rules

Epidemiologic Methods 2
Murray A. Mittleman and Elizabeth Mostofsky
Department of Epidemiology, Harvard School of Public Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Introduction

- All variances that we'll use can be derived using a combination of 10 different properties and what we know about some basic distributions
- Some of these are are more intuitive than others, depending on your calculus background

10 Basic Properties

Take any constants r and s; any random variables X and Y

1. $E(r) = r$
2. $E(r + s) = E(r) + E(s) = r + s$
3. $E(sX) = sE(X)$
   - $\Rightarrow E(r + sX) = r + sE(X)$
4. $Var(r) = 0$
5. $Var(rX) = r^2 Var(X)$

10 Basic Properties

Take any constants r and s; any random variables X and Y

6. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
   - If $X \amalg Y$, $Var(X + Y) = Var(X) + Var(Y)$
7. $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$
   - If $X \amalg Y$, $Var(X - Y) = Var(X) + Var(Y)$
8. For a fn $f(\cdot)$, $Var(f(X)) = [f'(X)]^2 Var(X)$
   - Delta method
9. $\ln(\frac{a}{b}) = \ln(a) - \ln(b)$
10. $f'(\ln(x)) = \frac{1}{x}$

Properties of Important Distributions

- Binomial Distribution, $X \sim Bin(n, p)$
  - $E(X) = np$
  - $Var(X) = np(1-p) = npq$
  - $\Rightarrow Var(\frac{X}{n}) = (\frac{1}{n})^2 Var(X) = \frac{p(1-p)}{n} = \frac{pq}{n}$
- Poisson Distribution, $X \sim Pois(\lambda)$
  - $E(X) = \lambda$
  - $Var(X) = \lambda$
- Hypergeometric Distribution, $X \sim Hypergeo(n, m, T)$
  - $E(X) = \frac{nm}{T}$
  - $Var(X) = n\frac{m}{T}\frac{T-m}{T}\frac{T-n}{T-1}$

Variance Derivations 2

Closed Cohort Data

Epidemiologic Methods 2
Murray A. Mittleman and Elizabeth Mostofsky
Department of Epidemiology, Harvard School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Closed Cohort

	E	$\overline{E}$	
D	a	b	M_1
$\overline{D}$	c	d	M_0
	N_1	N_0	T

Confidence Intervals for Risk Ratio and Difference Measures

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(X)}$
- We will derive variance estimates for the risk difference and the ln risk ratio in order to calculate these confidence intervals
  - $\hat{RD} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(RD)}$
  - $ln(\hat{RR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(ln(RR))}$ (and exponentiate)

Variance for the Risk Difference

$$X = RD = \left(\frac{a}{N_1} - \frac{b}{N_0}\right)$$

$Var(X) = Var\left(\frac{a}{N_1} - \frac{b}{N_0}\right)$	*Definition*
$= Var\left(\frac{a}{N_1}\right) + Var\left(\frac{b}{N_0}\right)$	*Rule 7*
$= \frac{1}{N_1^2} Var(a) + \frac{1}{N_0^2} Var(b)$	*Rule 5*
$= \frac{N_1 p_1 q_1}{N_1^2} + \frac{N_0 p_0 q_0}{N_0^2}$	*Binomial distribution*
$= \frac{N_1 \frac{a}{N_1} \frac{c}{N_1}}{N_1^2} + \frac{N_0 \frac{b}{N_0} \frac{d}{N_0}}{N_0^2}$	*Substitute for p,q*
$= \frac{ac}{N_1^3} + \frac{bd}{N_0^3}$	*Reduce*

Variance for the Risk Ratio

$$X = RR = \left(\frac{a}{N_1} / \frac{b}{N_0}\right)$$

$Var(ln(X)) = Var\left(ln\left(\frac{a}{N_1}\right) - ln\left(\frac{b}{N_0}\right)\right)$	*Rule 9*
$= Var\left(ln\left(\frac{a}{N_1}\right)\right) + Var\left(ln\left(\frac{b}{N_0}\right)\right)$	*Rule 7*
$= \left(\frac{1}{\frac{a}{N_1}}\right)^2 Var\left(\frac{a}{N_1}\right) + \left(\frac{1}{\frac{b}{N_0}}\right)^2 Var\left(\frac{b}{N_0}\right)$	*Rules 8 and 10*
$= \left(\frac{1}{\frac{a}{N_1}}\right)^2 \left(\frac{\frac{a}{N_1}\frac{c}{N_1}}{N_1}\right) + \left(\frac{1}{\frac{b}{N_0}}\right)^2 \left(\frac{\frac{b}{N_0}\frac{d}{N_0}}{N_0}\right)$	*Binomial distribution*
$= \frac{N_1^2}{a^2}\frac{a}{N_1}\frac{c}{N_1}\frac{1}{N_1} + \frac{N_0^2}{b^2}\frac{b}{N_0}\frac{d}{N_0}\frac{1}{N_0}$	*Reduce*
$= \frac{c}{aN_1} + \frac{d}{bN_0}$	*Reduce*

Confidence Intervals for Risk Ratio and Difference Measures

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(X)}$
- Plug in ratio or difference variance estimates that we just derived
  - $\hat{RD} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(RD)}$
  - $ln(\hat{RR}) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(ln(RR))}$ (and exponentiate)

Test for No Exposure-Disease Association

$$Z^2 = \frac{(X - E(X|H_0))^2}{Var(X|H_0)} \sim \chi^2_{df=1}$$

- $X = RD = \frac{a}{N_1} - \frac{b}{N_0}$
- $E(X|H_0) = 0$
- From before, generally $Var(X) = \frac{ac}{N_1^3} + \frac{bd}{N_0^3} = \frac{a}{N_1}\frac{c}{N_1}\frac{1}{N_1} + \frac{b}{N_0}\frac{d}{N_0}\frac{1}{N_0}$
  - Under the null, $\frac{a}{N_1} = \frac{b}{N_0} = \frac{M_1}{T}$ and $\frac{c}{N_1} = \frac{d}{N_0} = \frac{M_0}{T}$
  - $\Rightarrow Var(X|H_0) = \frac{M_1}{T}\frac{M_0}{T}\frac{1}{N_1} + \frac{M_1}{T}\frac{M_0}{T}\frac{1}{N_0} = \frac{M_1 M_0 (N_1 + N_0)}{T^2 N_1 N_0} = \frac{M_1 M_0}{T N_1 N_0}$

Variance Derivations 3

Person-time Data

Epidemiologic Methods 2
Murray A. Mittleman and Elizabeth Mostofsky
Department of Epidemiology, Harvard School of Public Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Open Cohort

- Or a closed cohort with person-time data

	E	$\overline{E}$	
Cases	a	b	M_1
PT	N_1	N_0	T

Confidence Intervals for Rate Ratio and Difference Measures

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(X)}$
- We will derive variance estimates for the rate difference and the ln rate ratio in order to calculate these confidence intervals
  - $\hat{IRD} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(IRD)}$
  - $ln(\hat{IRR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(ln(IRR))}$ (and exponentiate)

Variance for the Rate Difference

$$X = IRD = \left(\frac{a}{N_1} - \frac{b}{N_0}\right)$$

$Var(X) = Var(\frac{a}{N_1} - \frac{b}{N_0})$	*Definition*
$= Var(\frac{a}{N_1}) + Var(\frac{b}{N_0})$	*Rule 7*
$= \frac{1}{N_1^2} Var(a) + \frac{1}{N_0^2} Var(b)$	*Rule 5*
$= \frac{a}{N_1^2} + \frac{b}{N_0^2}$	*Poisson distribution*

Variance for the log Rate Ratio

$$X = IRR = (\tfrac{a}{N_1} / \tfrac{b}{N_0})$$

$Var(ln(X)) = Var(ln(\frac{a}{N_1}) - ln(\frac{b}{N_0}))$	*Rule 9*
$= Var(ln(\frac{a}{N_1})) + Var(ln(\frac{b}{N_0}))$	*Rule 7*
$= \left(\frac{1}{\frac{a}{N_1}}\right)^2 Var(\frac{a}{N_1}) + \left(\frac{1}{\frac{b}{N_0}}\right)^2 Var(\frac{b}{N_0})$	*Rules 8 and 10*
$= (\frac{N_1}{a})^2(\frac{1}{N_1})^2 Var(a) + (\frac{N_0}{b})^2(\frac{1}{N_0})^2 Var(b)$	*Rule 5 and Reduce*
$= (\frac{N_1}{a})^2(\frac{1}{N_1})^2 a + (\frac{N_0}{b})^2(\frac{1}{N_0})^2 b$	*Poisson Distribution*
$= \frac{1}{a} + \frac{1}{b}$	*Reduce*

Confidence Intervals for Rate Ratio and Difference Measures

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(X)}$
- Plug in ratio or difference variance estimates that we just derived
  - $\hat{IRD} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(IRD)}$
  - $ln(\hat{IRR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(ln(IRR))}$ (and exponentiate)

Test for No Disease-Exposure Association

$$Z^2 = \frac{(X - E(X|H_0))^2}{Var(X|H_0)} \sim \chi^2_{df=1}$$

- $X = a$
- We note that a follows a binomial distribution
  - $a \sim bin(n_1 = M_1, p_1 = \frac{N_1}{T}, q_1 = \frac{N_0}{T})$
- $E(X|H_0) = np = M_1(\frac{N_1}{T})$
- $Var(X|H_0) = npq = M_1(\frac{N_1}{T})(\frac{N_0}{T}) = \frac{M_1 N_1 N_0}{T^2}$

Variance Derivations 4

Case-Control Data

Epidemiologic Methods 2
Murray A. Mittleman and Elizabeth Mostofsky
Department of Epidemiology, Harvard School of Public Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Case-Control Study Data

	E	$\bar{E}$	
Case	a	b	M_1
Control	c	d	M_0
	N_1	N_0	T

- Let $p_1 = \frac{a}{M_1}$
- Let $p_0 = \frac{c}{M_0}$

Confidence Interval for the Odds Ratio

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(X)}$
- We will derive the variance estimate for the ln odds ratio in order to calculate this confidence interval
  - $ln(\hat{OR}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(ln(OR))}$ (and exponentiate)

Variance for the log Odds Ratio

$$X = OR = (\tfrac{a}{b}/\tfrac{c}{d}) = (\tfrac{p_1}{1-p_1}/\tfrac{p_0}{1-p_0})$$

$Var(ln(OR)) = Var(ln(\frac{p_1}{1-p_1}) - ln(\frac{p_0}{1-p_0}))$	*Rule 9*
$= Var(ln(\frac{p_1}{1-p_1})) + Var(ln(\frac{p_0}{1-p_0}))$	*Rule 7*
$= Var(ln(p_1) - ln(1-p_1)) + Var(ln(p_0) - ln(1-p_0))$	*Rule 9*
$= Var(p_1)(\frac{1}{p_1} + \frac{1}{(1-p_1)})^2 + Var(p_0)(\frac{1}{p_0} + \frac{1}{(1-p_0)})^2$	*Rules 8 and 10*
$= Var(\frac{a}{M_1})(\frac{1-p_1+p_1}{p_1(1-p_1)})^2 + Var(\frac{c}{M_0})(\frac{1-p_0+p_0}{p_0(1-p_0)})^2$	*Substitute and Rearrange*
$= \frac{1}{M_1^2}Var(a)(\frac{1}{p_1(1-p_1)})^2 + \frac{1}{M_0^2}Var(c)(\frac{1}{p_0(1-p_0)})^2$	*Rule 5*

Variance for the log Odds Ratio

$$X = OR = (\tfrac{a}{b}/\tfrac{c}{d}) = (\tfrac{p_1}{1-p_1}/\tfrac{p_0}{1-p_0})$$

$= \frac{1}{M_1^2}M_1\frac{a}{M_1}(1-\frac{a}{M_1})(\frac{1}{p_1(1-p_1)})^2 + \frac{1}{M_0^2}M_0\frac{c}{M_0}(1-\frac{c}{M_0})(\frac{1}{p_0(1-p_0)})^2$	*Binomial distribution*
$= \frac{ab}{M_1^3}(\frac{M_1^2}{ab})^2 + \frac{cd}{M_0^3}(\frac{M_0^2}{cd})^2$	*Substitute and Reduce*
$= \frac{ab}{M_1^3}\frac{M_1^4}{a^2b^2} + \frac{cd}{M_0^3}\frac{M_0^4}{c^2d^2} = \frac{M_1}{ab} + \frac{M_0}{cd}$	*Reduce*
$= \frac{a+b}{ab} + \frac{c+d}{cd} = (\frac{a}{ab} + \frac{b}{ab}) + (\frac{c}{cd} + \frac{d}{cd})$	*Substitute and reduce*
$= \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$	*Reduce*

Confidence Interval for the Odds Ratio

- General form: $\hat{X} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(X)}$
- Plug in odds ratio variance estimate that we just derived
  - $ln(\hat{OR}) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(ln(OR))}$ (and exponentiate)

Test for No Disease-Exposure Association

$$Z^2 = \frac{(X - E(X|H_0))^2}{Var(X|H_0)} \sim \chi^2_{df=1}$$

- $X = a$
- We assume under the null that a follows a hypergeometric distribution
  - $a \sim Hypergeo(N_1, M_1, T)$
- $E(X|H_0) = \frac{M_1 N_1}{T}$
- $Var(X|H_0) = M_1(\frac{N_1}{T})(\frac{T-N_1}{T})(\frac{T-M_1}{T-1}) = M_1(\frac{N_1}{T})(\frac{N_0}{T})(\frac{M_0}{T-1}) = \frac{M_1 M_0 N_1 N_0}{T^2(T-1)}$