Week 1 – Thursday session

Statistical inference and Crude Analysis

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Week 1: Discussion Topics

1. Statistical Inference
   - Hypothesis testing
     - Null and alternative hypothesis
     - Form of the test statistic
     - P-values
     - Neyman-Pearson hypothesis tests
     - Limitations
   - Confidence Intervals
     - Confidence, not probability
     - Relationship to Neyman-Pearson hypothesis testing
   - Statistical versus biological/public health significance
   - Sources of random variability
2. Crude Analysis
   - Person-time data - testing, point estimates, confidence intervals
   - Case-control data - testing, point estimates, confidence intervals
   - Binomial data (cumulative incidence; prevalence)
3. The EPI 202 Road map

Statistical Significance
Influence of the Size of the Numerator and Denominator

- Recall the form of the statistic used for hypothesis testing:

$$Z^2 = \frac{\left[X - E\left(X \mid H_0\right)\right]^2}{Var\left(X \mid H_0\right)} \sim \chi_1^2$$

- The evidence against the null hypothesis increases as Z^2 gets larger as either the numerator gets larger or the denominator gets smaller

Statistical Significance
Influence of Numerator Size

- Z^2 increases as its numerator, $[X - E(X|H_0)]^2$, increases
  - The value of the numerator is determined by the underlying phenomenon under investigation and cannot be altered by the researcher.
  - It is a function of the magnitude and direction of the exposure-disease association.
  - In a study of incidence rates in an exposed and unexposed group, X = $\hat{RD}$, the estimated incidence rate difference. $E(X|H_0) = 0$. The numerator of the test statistic, $[X - E(X|H_0)]^2$, increases as the estimated rate difference becomes large (i.e. >> 0) or small (i.e. <<0).

Statistical Significance
Influence of Denominator Size

- Z^2 increases as the denominator, $Var(X|H_0)$, decreases

  N increases, more statistical information and greater precision, Var(X|H_0) decreases; Z^2 increases, p gets smaller.

  - $Var(X|H_0)$ depends on both the value of X specified by H_0 and the sample size of the investigation.
  - All else being equal, the larger the sample size, the smaller the ***variance*** because there is more ***statistical information*** and the estimate of the exposure-outcome relationship can be estimated with greater ***precision***.
  - Sample size is an arbitrary quantity; it is purely a function of study design and bears no relationship to any scientific quantity of interest.

Z^2, Sample Size and Statistical Significance

- By making a study sufficiently large and thereby making the denominator of Z^2 sufficiently small, one can guarantee a "statistically significant" result no matter how small the difference between the groups are.

All else being equal, the value of the Z-squared test statistic gets larger when

$$Z^2 = \frac{[X - E(X \mid H_0)]^2}{Var(X \mid H_0)} \sim \chi_1^2$$

- The estimate of the association gets closer to the null
- The estimate of the association gets further from the null
- The variance gets larger
- The variance gets smaller
- Do not know

Total Results: 0

All else being equal, the Z^2 test statistic gets larger when either the numerator gets larger or the denominator gets smaller.

The nominator gets larger when the point estimate gets further from the null.

The denominator gets smaller when the variance under the null gets smaller. (More sample size)

Assuming the null is true and there are no structural or other sources of bias, there will be a higher proportion of statistically significant findings at the alpha=0.05 level in

"I have a positive attitude…
but it might be a false positive."

- A large study
- A small study
- The same in large and small studies
- Not sure

The probability of incorrectly rejecting the null when it is true is the same whether the study is large or small (because we have defined it as the alpha).

Unlike the p-value, alpha can be thought of as a long-term error risk.

Assuming the null is true and there are no structural or other sources of bias, the false positive findings (incorrectly rejecting the null), will on average be further from the null in

"I have a positive attitude…
but it might be a false positive."

- Large studies
- Small studies
- The same in large and small studies
- Not sure

All else being equal, the variance of the estimator from a small study will be larger. Hence, the denominator of the test statistic will be larger in a small study.

In order for the test statistic to exceed the critical value, the numerator of the test statistic must be larger in a small study (on average be further from the null).

Random Variability in Randomized Trials (1)

- The source of random variability in a randomized trial is well understood. Exposure (treatment) is assigned at random to each study participant.

- On average, both measured and unmeasured confounders are randomly distributed between the treatment and control groups. (Exchangeability is assumed)

- The source of random variability in this setting is identified as the random treatment assignment.

Random Variability in Randomized Trials (2)

- The p-value has the interpretation that under the null, i.e. when there is no difference between the treated and untreated groups with respect to the outcome, the observed association is due to a random imbalance between the treated and untreated groups with respect to
  - unmeasured confounders or
  - confounders that were measured but not adjusted for statistically

Random Variability in Observational Studies

- In an observational study, exposure is not randomly assigned. What, then, is the source of random variability to which the p-value refers in this setting?
- If exposure is not assigned by a chance mechanism, what is the meaning of the expression "These results could have been due to chance"?
- Because there is no clear physical source of randomness introduced into the design, the meaning of p-values and confidence intervals is in question.
- Additional assumptions are needed. These assumptions are largely not verifiable from observed data.

Methods for Person-time Data
Doll and Hill NCI Monograph 1966

- A classic prospective cohort study of cigarette smoking. The subjects are British male doctors. Here we investigate the relationship between cigarette smoking and coronary heart disease mortality (CHD).

Notation

	E	$\bar{E}$	
Cases	a	b	M_1
Person-Time	N_1	N_0	T

Smoking

	Yes	No	
Deaths	630	101	731
Person-years	142,247	39,220	181,467

Methods for Person-time Data
Hypothesis Test – Choice of X

- The test statistic is:

$$Z^2 = \frac{\left[X - E\left(X \mid H_0\right)\right]^2}{Var\left(X \mid H_0\right)} \sim \chi_1^2$$

- X can be defined in several ways:
  - X = number of exposed cases = a = 630
  - $E(X|H_0) = E(a|H_0)$
    - Total number of cases * Pr(E)
    - $M_1(N_1/T)$
    - 731 * (142,247/181,467)
    - 731 * 0.784
    - 573.0

Methods for Person-time Data

Point estimate of the rate difference ($\widehat{IRD}$)

$$\widehat{IRD} = \frac{a}{N_1} - \frac{b}{N_0} = \frac{630}{142{,}247\ PY} - \frac{101}{39{,}220\ PY}$$

$$= \frac{4.43}{1{,}000\ PY} - \frac{2.58}{1{,}000\ PY} = 1.85/1{,}000\ PY$$

- These data indicate that the rate of CHD deaths associated with smoking among British male doctors was 1.85 per 1,000 person-years higher among those who smoked compared with those who did not (assuming no confounding, selection bias, information bias or any other source of bias).

Methods for Person-time Data
Variance of the Incidence Rate Difference ($\widehat{IRD}$)

95% Confidence Interval $\quad \mathrm{X} \pm 1.96\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\mathrm{X})}$

$$X = \hat{IRD} = 0.00185$$

$$\hat{Var}(X) = \hat{Var}(\hat{IRD}) = \frac{\hat{I}_1}{N_1} + \frac{\hat{I}_0}{N_0} = \frac{a}{N_1^2} + \frac{b}{N_0^2}$$

$$= \frac{630}{142{,}247^2} + \frac{101}{39{,}220^2} = \frac{9.680}{10^8\ PY^2}$$

Note: The variance of the IRD at it's observed value is not equal to the variance of the IRD under the null hypothesis when IRD=0.

Methods for Person-time Data

Point estimate of the incidence rate ratio ($\widehat{IRR}$)

- $\widehat{IRR} = \frac{a}{N_1} \Big/ \frac{b}{N_0} = \frac{630}{142{,}247\ PY} \Big/ \frac{101}{39{,}220\ PY} = 1.72$

- The CHD mortality rate was 72% higher among smokers compared with non-smokers (assuming no confounding or any other source of bias)

Methods for Person-time Data

Variance of the ℓn(Incidence Rate Ratio) ($\ell n(\widehat{IRR})$)

- 95% Confidence interval: $X \pm 1.96\sqrt{\widehat{Var}\ (X)}$
- Let X = $\ell n(\widehat{IRR})$ = $\ell n(1.72)$ = 0.5422
- $\widehat{Var}$(X) = $\widehat{Var}(\ell n(\widehat{IRR}))$ = $\frac{1}{a} + \frac{1}{b}$

$$= \frac{1}{630} + \frac{1}{101} = 0.01149$$

Methods for Person-time Data
Confidence Interval for the Rate Ratio ($\widehat{IRR}$)

- 95% Confidence interval for $\ln(IRR)$:

$$0.5422 \pm 1.96\sqrt{0.01149}$$
$$= (0.3321, 0.7523)$$

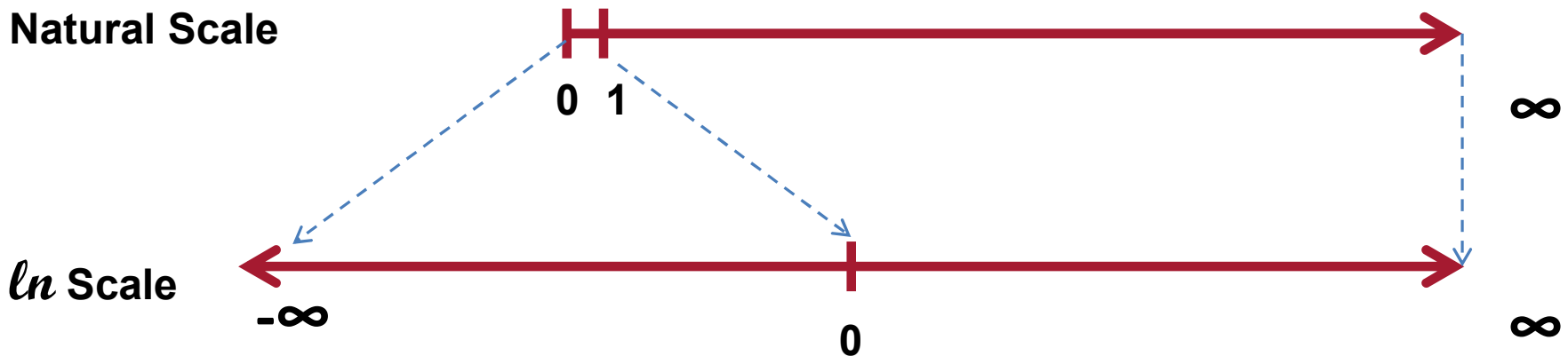
- 95% Confidence interval for (IRR):

- $e^{(X \pm 1.96\sqrt{\widehat{Var}(X)})} = e^{(0.3321, 0.7522)}$

= (1.39, 2.12)

- These data are consistent with rate ratios ranging from 1.4 to 2.1 with 95% confidence (assuming no confounding, selection bias or any other source of bias)

Ratios on the *ln* Scale



	IRR	*ln*IRR
Bike helmet non-use vs. use	IRR=10	*ln*(10)=2.3
Bike helmet use vs. non-use	IRR=1/10=0.1	*ln*(1/10)=-2.3

Brief Review of Logs and Exponents

- For the purpose of this course and in nearly all work in epidemiology, biostatistics, and statistics, *log* refers to the *natural log (ln)*
  - $\exp\{X\} = e^X$
  - $\exp\{0\} = 1$
  - $\exp\{1\} = 2.718281828$
  - $ln[\exp(X)] = X$
  - $ln[1] = 0$
  - $ln[2.718281828] = 1$
  - $ln[2.718281828^2] = ln\ [7.389056] = 2$
  - $ln[0] = -\infty$
  - $ln[\infty] = \infty$
  - $\exp[ln(X)] = X$
  - $\exp\{A\}*\exp\{B\} = \exp\{A+B\}$
  - $\exp\{A\}/\exp\{B\} = \exp\{A-B\}$
  - $ln\{AB\} = ln\ \{A\} + ln\ \{B\}$
  - $ln\{A/B\} = ln\ \{A\} - ln\ \{B\}$
  - $ln[1/X] = -ln[X]$

Test of No Exposure-Disease Association

$$Z^2 = \frac{[X - \hat{E}(X \mid H_0)]^2}{\hat{Var}(X \mid H_0)} \sim X_1^2$$

	H_0	X	$\hat{E}(X \mid H_0)$	$\hat{Var}(\hat{X} \mid H_0)$	
Methods for Count Data (Closed cohort and cross-sectional Studies)					
E $\bar{E}$ cases a b M_1 Non-cases N_1-a N_0-b M_0 N_1 N_0 T	$C_1=C_o$ $C_1/C_o=1$ $C_1-C_o=0$	a	$\frac{N_1 M_1}{T}$	$\frac{M_1 M_0 N_1 N_0}{T^3}$	Unstratified
		$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^3}$	Stratified
Methods for Person-time Data (Open Cohort and Closed Cohort studies)					
E $\bar{E}$ cases a b M_1 PT N_1 N_0 T	$I_1=I_o$ $I_1/I_o=1$ $I_1-I_o=0$	a	$\frac{N_1 M_1}{T}$	$\frac{N_1 N_0 M_1}{T^2}$	Unstratified
		$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{N_{1i} N_{0i} M_{1i}}{T_i^2}$	Stratified
Methods for Case-control Data					
E $\bar{E}$ cases a b M_1 controls c d M_0 N_1 N_0 T	OR=1 $I_1/I_o=1$	a	$\frac{N_1 M_1}{T}$	$\frac{M_1 M_0 N_1 N_0}{T^2(T-1)}$	Unstratified
		$\sum a_i$	$\sum \frac{N_{1i} M_{1i}}{T_i}$	$\sum \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^2(T_i-1)}$	Stratified

Confidence Intervals
Count Data

$$X \pm Z_{1-\alpha/2}\sqrt{\hat{Var}(X)}$$

	X	w_i	$\hat{Var}(X)$
Methods for Count Data (closed cohort and cross-sectional studies)			
Cumulative incidence difference	$\frac{a}{N_1} - \frac{b}{N_0}$	--	$\frac{ac}{N_1^3} + \frac{bd}{N_0^3}$
Summary cumulative incidence difference	$\frac{\sum w_i\left[\frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}}\right]}{\sum w_i}$	$\frac{N_{1i}N_{0i}}{T_i}$	$\frac{\sum\left(\frac{a_i c_i N_{0i}^2}{T_i^2(N_{1i}-1)} + \frac{b_i d_i N_{1i}^2}{T_i^2(N_{0i}-1)}\right)}{\left(\sum \frac{N_{1i}N_{0i}}{T_i}\right)^2}$
Cumulative incidence ratio (ℓn)	$\ell n\left\{\frac{a}{N_1} \Big/ \frac{b}{N_0}\right\}$	--	$\frac{c}{aN_1} + \frac{d}{bN_0}$
Summary cumulative incidence ratio (ℓn)	$\ell n\left\{\frac{\sum w_i\left[\frac{a_i}{N_{1i}} \Big/ \frac{b_i}{N_{0i}}\right]}{\sum w_i}\right\}$	$\frac{b_i N_{1i}}{T_i}$	$\frac{\sum\left(M_{1i}N_{1i}N_{0i} - a_i b_i T_i\right)/T_i^2}{\left[\sum \frac{a_i N_{0i}}{T_i}\right]\left[\sum \frac{b_i N_{1i}}{T_i}\right]}$

Confidence Intervals
Person-time Data

$$X \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(X)}$$

	X	w_i	$\widehat{Var}(X)$
Methods for Person-Time Data (open cohort and closed cohort studies)			
Rate difference	$\frac{a}{N_1} - \frac{b}{N_0}$	--	$\frac{a}{N_1^2} + \frac{b}{N_0^2}$
Summary rate difference	$\frac{\sum w_i\left[\frac{a_i}{N_{1i}} - \frac{b_i}{N_{0i}}\right]}{\sum w_i}$	$\frac{N_{1i}N_{0i}}{T_i}$	$\frac{\sum\left(\frac{a_i N_{0i}^2 + b_i N_{1i}^2}{T_i^2}\right)}{\left(\sum \frac{N_{1i}N_{0i}}{T_i}\right)^2}$
Rate ratio (ℓn)	$\ell n\left\{\frac{a}{N_1} \Big/ \frac{b}{N_0}\right\}$	--	$\frac{1}{a} + \frac{1}{b}$
Summary rate ratio (ℓn)	$\ell n\left\{\frac{\sum w_i\left[\frac{a_i}{N_{1i}} \Big/ \frac{b_i}{N_{0i}}\right]}{\sum w_i}\right\}$	$\frac{b_i N_{1i}}{T_i}$	$\frac{\sum\left(M_{1i}N_{1i}N_{0i}\right)/T_i^2}{\left[\sum \frac{a_i N_{0i}}{T_i}\right]\left[\sum \frac{b_i N_{1i}}{T_i}\right]}$

Confidence Intervals
Case-control Data

$$X \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(X)}$$

	X	w_i	$\widehat{Var}(X)$
Methods for case-control Data			
Odds ratio (ℓn)	$\ell n\left\{\frac{ad}{bc}\right\}$	--	$\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$
Summary odds ratio (ℓn)	$\ell n\left\{\frac{\sum w_i \frac{a_i d_i}{b_i c_i}}{\sum w_i}\right\}$	$\frac{b_i c_i}{T_i}$	RGB variance

RGB variance:

$$\frac{1}{2}\left[\frac{\sum\left(\frac{a_i d_i}{T_i}\right)\left(\frac{a_i+d_i}{T_i}\right)}{\left(\sum\frac{a_i d_i}{T_i}\right)^2}+\frac{\sum\left(\frac{a_i d_i}{T_i}\right)\left(\frac{c_i+b_i}{T_i}\right)+\sum\left(\frac{b_i c_i}{T_i}\right)\left(\frac{a_i+d_i}{T_i}\right)}{\left(\sum\frac{a_i d_i}{T_i}\right)\left(\sum\frac{b_i c_i}{T_i}\right)}+\frac{\sum\left(\frac{b_i c_i}{T_i}\right)\left(\frac{c_i+b_i}{T_i}\right)}{\left(\sum\frac{b_i c_i}{T_i}\right)^2}\right]$$

Computational formulas for point estimates

Computational Form of the Mantel Haenszel Estimators

Methods for Count Data (closed cohort and cross-sectional studies)

Summary cumulative incidence ratio

$$\hat{CIR}_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i N_{0i}}{T_i}}{\sum_{i=1}^{I} \frac{b_i N_{1i}}{T_i}}$$

Summary cumulative incidence difference

$$\frac{\sum \left(\frac{a_i N_{0i} - b_i N_{1i}}{T_i} \right)}{\sum \frac{N_{1i} N_{0i}}{T_i}}$$

Methods for Person-Time Data (open cohort and closed cohort studies)

Summary rate ratio

$$\hat{IRR}_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i N_{0i}}{T_i}}{\sum_{i=1}^{I} \frac{b_i N_{1i}}{T_i}}$$

Summary rate difference

$$\frac{\sum \left(\frac{a_i N_{0i} - b_i N_{1i}}{T_i} \right)}{\sum \frac{N_{1i} N_{0i}}{T_i}}$$

Methods for Case-control Data

Summary odds ratio

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{T_i}}$$

Test of Homogeneity of Effect Measures

$$H = \sum_{i=1}^{I} \frac{[\hat{X}_i - \hat{X}_{summary}]^2}{\hat{Var}_i[\hat{X}_i]} \sim \chi^2_{I-1}$$

	H_0	$\hat{X}_i$	$\hat{X}_{summary}$	$\hat{Var}_i(\hat{X}_i)$
Methods for Count Data (closed cohort and cross-sectional studies)				
Difference measure	$CID_1=CID_2=\ldots=CID_i$ $CID_i=CID_j$ for all i,j	CID_i	$CID_{summary}$	$\frac{a_i c_i}{N_{1i}^3} + \frac{b_i d_i}{N_{0i}^3}$
Ratio measure	$CIR_1=CIR_2=\ldots=CIR_i$ $CIR_i=CIR_j$ for all i,j	$\ell n\,(CIR_i)$	$\ell n(CIR_{MH})$	$\frac{c_i}{a_i N_{1i}} + \frac{d_i}{b_i N_{0i}}$
Methods for Person-Times Data (open cohort and closed cohort studies)				
Difference measure	$IRD_1=IRD_2=\ldots=IRD_i$ $IRD_i=IRD_j$ for all i,j	IRD_i	$IRD_{summary}$	$\frac{a_i}{N_{1i}^2} + \frac{b_i}{N_{0i}^2}$
Ratio measure	$IRR_1=IRR_2=\ldots=IRR_i$ $IRR_i=IRR_j$ for all i,j	$\ell n(IRR_i)$	$\ell n(IRR_{MH})$	$\frac{1}{a_i} + \frac{1}{b_i}$
Methods for Case-control Data				
Ratio measure	$OR_1=OR_2=\ldots=OR_i$ $OR_i=OR_j$ for all i,j	$\ell n(OR_i)$	$\ell n(OR_{MH})$	$\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$

Confidence Intervals
Matched Case-control Data

		Controls	
		E	$\bar{E}$
Case	E	f_{11}	f_{10}
	$\bar{E}$	f_{01}	f_{00}

$OR_{MH} = f_{10} / f_{01}$

Test statistic:
Ho: $OR_{MH} = 1$
Ha: $OR_{MH} \neq 1$

$$Z^2 = \frac{(f_{10} - f_{01})^2}{f_{10} + f_{01}} \sim \chi_1^2$$

Variance formula for confidence interval:

$$\text{Var}(\ln (OR_{MH})) = \frac{1}{f_{10}} + \frac{1}{f_{01}}$$

HAVE A GOOD WEEKEND