Week 1 – Tuesday 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

EPI201/202 Sequence

- **EPI201**
  - Exchangeability
  - Measures of Disease Frequency, Association and Effect
  - Directed Acyclic Graphs
  - Study Design
  - Target Trial Emulation
  - Confounding and Effect Measure Modification
  - Causal Inference

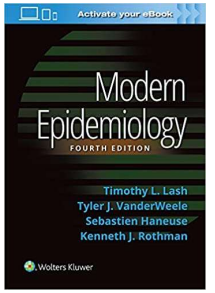
- **EPI202**
  - Statistical Inference in Epidemiology
  - Crude and Stratified Tabular Analysis
  - Matching in Design and Analysis
  - Maximum Bias Attributable to Uncontrolled Confounding
  - Misclassification and Measurement Error
  - Introduction to Regression Models (Linear and Logistic)
  - Inverse Probability of Treatment Weighting / Standardization

- Students enrolled in EPI201 are expected to enroll in EPI202

Texts and Readings

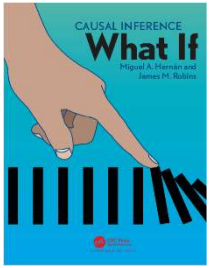
- **Modern Epidemiology**

  Lash TL, VanderWeele TJ, Haneuse S, Rothamn KJ. Modern Epidemiology (4th ed.) Philadelphia, PA: Wolters-Kluwer, 2021 (ISBN-13: 978-1-4511-9328-2)
  http://id.lib.harvard.edu/alma/99155293309703941/catalog



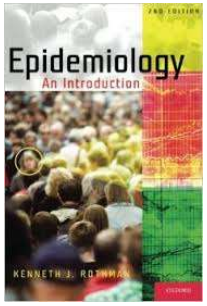
- **Causal Inference**

  Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
  http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/



- **Optional**

  Rothman KJ. Epidemiology: An Introduction (2nd ed.) New York, NY: Oxford University Press, 2012 (ISBN-10 0199754551)
  http://ezp-prod1.hul.harvard.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2114235&site=ehost-live&scope=site



- Additional materials posted on course website

Logistics

- **Lecture Material**
  - Modules comprised of short videos covering the week's material divided into small chunks are posted on the Canvas site together with self-assessment questions. Modules are to be completed prior to Thursday at 9:30 AM. Completion of the weekly self-assessment questions on time count for 10% of the final grade in the course.
  - Lecture slides are available from the course's web site
- **Weekly Discussion sessions**
  - The class will meet on Tuesday and Thursday mornings (9:45-11:15 AM and 11:30-1:00 PM in Kresge G1). During interactive class sessions, we will start with a review of the key concepts from the weeks videos and then will work on problems that apply and extend the material in the pre-recorded videos. During these sessions, there will be ample opportunity to ask questions and discuss course material. Students are expected to participate in these sessions actively. On the week of the midterm and final examinations, the Tuesday sessions will be devoted to review of course content covered on the upcoming exam. We will record class sessions; however, since the interactive sessions include some small group work, the recordings may be of limited value. Since some material will be covered only during these sessions, attendance is strongly recommended.
- **Lab**
  - One weekly 90-minute session led by a teaching fellow to review the week's material, answer questions and work through problem sets that will be introduced during lab. Students are required to sign up for a lab session and should attend that session.

- **Office Hours**
  - Attend any and as many as you find helpful

Assignments

- **4 Homework Assignments (30%)**
  - Discuss together
  - Submit your own work online by Thursday at **9:30 AM ET**
  - Review and discuss in lab

- **Weekly Self assessment Questions (10%)**
  - Open book/notes…
  - To be completed before Thursday at **9:30 AM ET**

- **Midterm Exam (25%)**
  - In class, closed book exam – Thursday **November 17, 2022.** You may take the exam at any of the offered times.

- **Final Exam (35%)**
  - In class, closed book exam – Will be offered on **Wednesday December 14 from 5:30 – 7:00 PM or during class on Thursday December 15, 2022.** You may take the final exam at any of the offered times.

Policies and Expectations

- **Academic Integrity**
  - Collaboration is encouraged
  - Submit your own work
  - https://www.hsph.harvard.edu/office-of-education/education-policies/academic-support/
- **Ask questions!**
  - Interactive class sessions
  - Labs with Teaching Fellows
  - Office Hours

What introductory epidemiology course did you complete?

- EPI 201
- EPI 500
- EPI 505
- EPI 208
- ID 200
- Other

Total Results: 0

What is your preferred statistical software program(s)? Choose one or more.

- SAS
- Stata
- R
- Other

Total Results: 0

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

Hypothesis Testing
Steps

- Specify the null hypothesis and alternative hypothesis
- Collect data that can be used to test the hypothesis
- Form a *test statistic* that has a known probability distribution under the assumptions implied by the null hypothesis

Hypothesis Testing
Form of the Test Statistics

- If the sample size is sufficiently large, we use a test statistic of the following form:

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

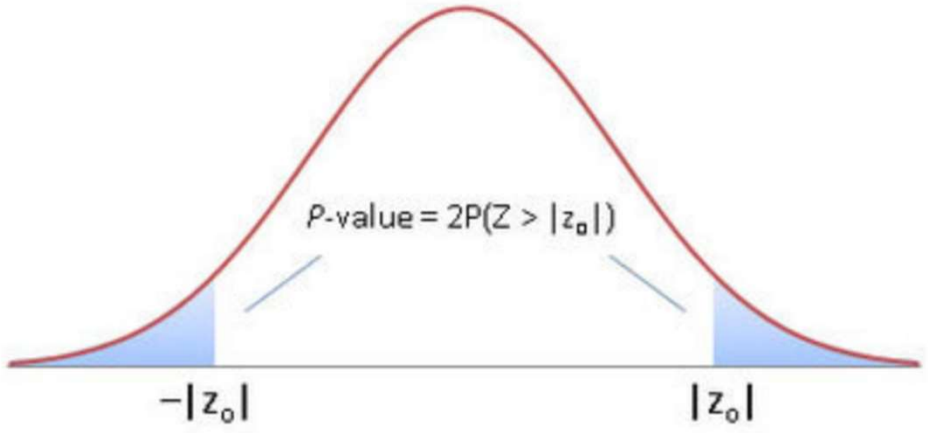
Where:

- X is the measure of disease occurrence or association of interest that is estimated from the data
- $E(X|H_0)$ is the expected value of X under the null hypothesis
- $Var(X|H_0)$ is the variance of X under the null hypothesis
- and Z^2 follows a χ^2 distribution with one degree of freedom

Note: not the expected value or variance in the observed data

Chi-square only has one tail, but it gives two-sided information

The p-value represents the probability that you will correctly reject the null hypothesis when it is truly incorrect.



- True
- False
- Not sure

Total Results: 0

The p-value is the probability of rejecting the null, when the fact the null is true. It can be conceived of as the probability of obtaining the result observed, or one even more extreme when the null is true.

Hypothesis Testing
Limitation: No Information on Direction or Magnitude

- Hypothesis tests and p-values when taken alone give no indication of either the direction or the magnitude of the effect, especially when they are two-sided as they almost always are in epidemiology
  - Note: $\chi 2$ test is always two-sided

- In the above example, although we could reject the null hypothesis in a testing framework and conclude that there was a statistically significant difference, we have not reported the direction of the difference:
  - Is there more or less childhood obesity in poor rural communities?

- or its magnitude:
  - How much more (or less) childhood obesity is there?

Hypothesis Testing
Limitation: No Information on Range of Values

- No information about the range of effects that are consistent with the observed data
  - Hypothesis tests only inform whether or not the data are compatible with the state of nature described by the null hypothesis

- With 95% probability, were these data consistent with a two-fold increase in childhood obesity among the rural poor? A 5% decrease? Both of these?

Hypothesis Testing
Limitation: No Information on Power

- No information about the *power* of the study to detect differences of specified magnitude

- The *power* of a hypothesis test, and the study designed to permit calculation of this test, is the probability that the test will reject the null hypothesis when the null hypothesis is in fact false, i.e. the probability that an effect of the magnitude specified under the alternative hypothesis will be detected given the proposed study design.

- Since most studies are designed to permit calculation of a key hypothesis test with specified power, the power of the study is equated with the power of the test.

Hypothesis Testing
Limitations

Hypothesis tests, when taken alone, yield information only about the consistency of the data with the state of nature described by the null hypothesis; they provide no information on the consistency of the data with alternative, *non-null* states of nature.

Confidence Intervals
Introduction

- Confidence intervals are usually preferred by epidemiologists for statistical inference because they are not beset by the above limitations

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

Confidence Intervals
Estimation

- For "large enough" sample size, the 100(1-α)% confidence interval has the following simple form:

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

Where

- X = the measure of disease occurrence or association of interest estimated in the study
- $\hat{V}ar(X)$= the estimated variance of X
- $Z_{1-\alpha/2}$ = the $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution
  - Most common case: Confidence level α = 5%, the 100(1-α/2) = 97.5^{th} percentile of the standard normal distribution is 1.96

Confidence Intervals
Interpretation

The confidence interval

- has the interpretation that 100(1-α)% of the time, an interval constructed in this way will include the true value of the parameter of interest

- typically gives the range of values of the parameter of interest that would not be rejected at the α significance level

- gives a sense of the magnitude and direction of the parameter of interest that are consistent with the data

The rate ratio of lung cancer from insulation work 30-34 years from first exposure compared to 20-24 years from exposure is 4.9, with a 95% confidence interval of (3.5, 6.7). If the sample size were reduced to 1/10th of the current sample size, the expected value of the point estimate of the rate ratio would be

- Smaller
- Larger
- Same
- Do not know

Total Results: 0

The expected value of the estimate would be the same if the sample size were larger or smaller.
If there were no bias, both estimates would be unbiased estimates of the underlying population parameter of interest.

The rate ratio of lung cancer from insulation work 30-34 years from first exposure compared to 20-24 years from exposure is 4.9, with a 95% confidence interval of (3.5, 6.7). If the sample size were reduced to 1/10th of the current sample size, the expected width of the confidence interval around the rate ratio would be

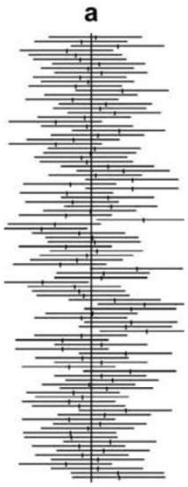
- Smaller
- Larger
- Same
- Do not know

Total Results: 0

The expected value of the estimate would be the same if the sample size were larger or smaller. Both estimates would be unbiased estimates of the underlying population parameter of interest if no bias.

But, sample size is related to the variance of estimator. As studies get larger, they contain more information, the variance of the estimators get smaller, the confidence intervals get narrower and the estimates are more precise.

If the null were true, a 95% confidence interval is expected to include

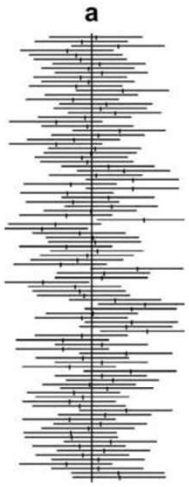


- the null value 95% of the time
- the estimated measure 95% of the time
- the estimated variance 95% of the time
- do not know

Total Results: 0

A valid 95% CI will include the true value of the parameter of interest 95% of the time, and the estimate measure 100% of the time.

Confidence intervals are first calculated on the natural log scale for:



Difference measures of association (IRD; CID; Prev diff)

Ratio measures of association (IRR; CIR; OR; Prev ratio)

All of the above

None of the above

Total Results: 0

HAVE A GOOD WEEK