C. Aspects of inference from studies

1. Causality:
   a. Probabilistic and not deterministic
   i. Ex., smoking
   b. Not just association:
   i. Ex., cholesterol level and cholesterol medication

2. Test of size $\alpha$ is
   a. Choose among 2 hypotheses
      i. Null Hypothesis $H_0$: $\theta = \theta_0$
      ii. Alt. Hypothesis $H_A$: $\theta \in$ range $\neq \theta_0$
   b. Two kinds of $H_A$
      i. One sided
      ii. Two sided

3. Observed significance level or $p$-value is
   a. Probability of getting statistic at least as extreme as what we see
   b. $P[T \geq T_{obs}]$

4. Choice of statistic
   a. Lots of times one-sided test statistics are intuitive
      i. Ex., difference in sample means
   b. Two-sided test statistic balances information from both sides
      i. Doubling minimum one–sided $p$-value works.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

2. Observational Studies

2.1 Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

3. Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

3. Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

3. Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

3. Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$

3. Case Control study

a. Identify separately people with and without disease
b. Determine whether they were exposed or not
c. Often match on traits that are expected to impact disease

d. Study is retrospective, both in time and in expected causation.

B & D 1: 1.3

e. Efficiency good when
   i. cases are rare
      • Sufficient number of cases would require an enormous cohort.
   ii. Exposure overall is rare but exposure among cases is not, and collecting a cohort of exposed is hard
      B & D 1: 1.4

f. Disadvantages
   i. Gives only comparison, and not absolute levels.
   ii. Has all of the above disadvantages of the historical cohort study.
   iii. Differential choice of cases, controls $\Rightarrow$ bias
   iv. 3 requirements for using normal approx. for distn of $\theta$
   v. Need shape approx. normal
   vi. Need expectation right (we just checked this)
   vii. Covariance is $\approx 0$
2. Confidence Intervals for SMR
   a. Hence 95% CI for $\varsigma$ is $\text{SMR} \pm 1.96 \times \sqrt{\frac{\text{SMR}}{e}}$
     i. Trouble if $1.96 \times \sqrt{\frac{\text{SMR}}{e}} > \text{SMR} \Leftrightarrow 1.96^2/e > \text{SMR}$
     ii. Do on log scale
     iii. $\log(\text{SMR}) - \log(\varsigma) \approx (\text{SMR} - \varsigma) \frac{d}{\pi} \log(\varsigma) = (\text{SMR} - \varsigma)/\varsigma$
     iv. $\text{Var} [\log(\text{SMR})] \approx \text{Var} [\text{SMR}] \frac{1}{\varsigma^2} \approx \frac{1}{e \text{SMR}} = \frac{1}{d}$.
    v. CI for $\log(\varsigma)$ is $\log(\text{SMR}) \pm 1.96/\sqrt{d}$
    vi. 95% CI for $\varsigma$ is $\text{SMR} \times \exp(\pm 1.96/\sqrt{d})$
    vii. Similar expression for CI for $\log(\lambda_{1i})$ is
        $\log(\lambda_{1i}) \pm 1.96\sqrt{\frac{Q_{1i}}{d_{1i}}}/Q_{1i} \times (Q_{1i}/d_{1i}) = \log(\lambda_{1i}) \pm 1.96/\sqrt{d_{1i}}$
    viii. Other approximations, some exact methods also available

3. Formulate $H_0$ using SMR
   a. Calculate total number $d_k$ for $k = 0, 1$
      i. $k = \begin{cases} 0 & \text{for unexposed} \\ 1 & \text{for exposed} \end{cases}$
   b. Calculate expected number $e_k$
   c. If relative risk of group 1 to group 0 is $q$ for all time groups, then expect $qe_1/e_0$ times as many individuals in group 1 as in group 0
   d. Both groups are Poisson

4. Hence build $T$ around the difference between $d_0$ and $d_1$
   a. $T = d_1 - d_0$? OK
   b. $T = d_1/(d_1 + d_0)$? Better