Week 4 – Tuesday Session

Midterm Review

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

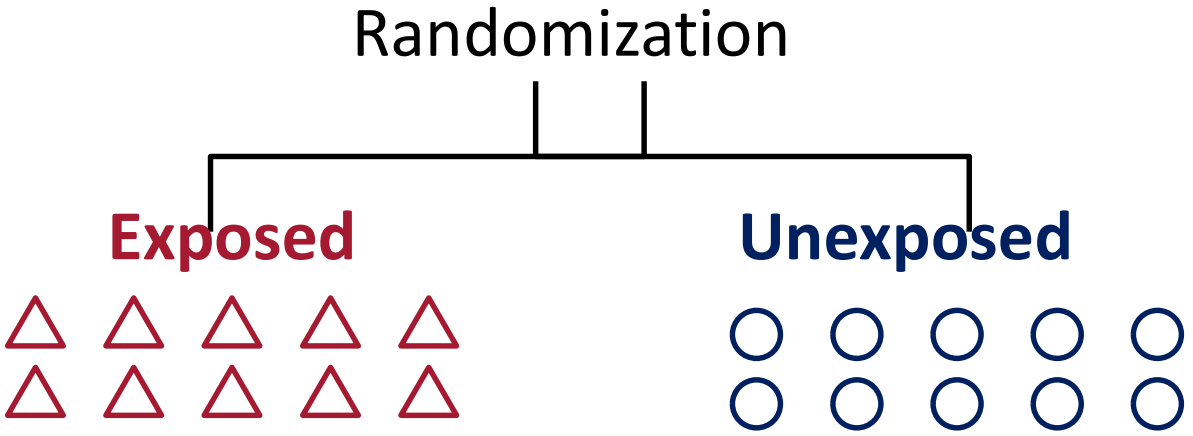
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

EPI202 review session

- **Midterm exam**
  - Thursday, November 17, 2022
  - Exam will be online through Canvas
  - Format will be similar to the quizzes:
    - ✓ multiple choice
    - ✓ fill in the blank
    - ✓ calculations and interpretations
    - ✓ true false

- **Resources**
  - Bring your computer and charger and make sure your computer is fully charged
  - You can also use calculator/software/EPI202Calculator (preferred)
  - We will provide copies of the Roadmap
  - You may not use any other resource (e.g., internet, another classmate answer).
  - Avoid safari browser.

Exchangeability



- The distribution of the outcomes would be identical between the two groups if both were exposed, or if both were unexposed.
- This happens to be true because randomization assures that all other causes of outcome are balanced between the groups.
- The distribution of the outcome in the unexposed serves as a proxy for the unobserved counterfactual (”*time machine experiment*”)

Prevalence (P)

- Proportion of population with **existing** disease/characteristic at a given point in time

- $\Pr[Y=1] = \frac{\text{\# existing cases}}{\text{\# individuals in study population}}$ at a point in time
- Range: 0 to 1
- Dimensionless proportion
- Must provide time referent, e.g.
  - prevalence of 0.80 at age 50
  - prevalence of 0.35 in 1995
  - prevalence of 0.27 at study entry

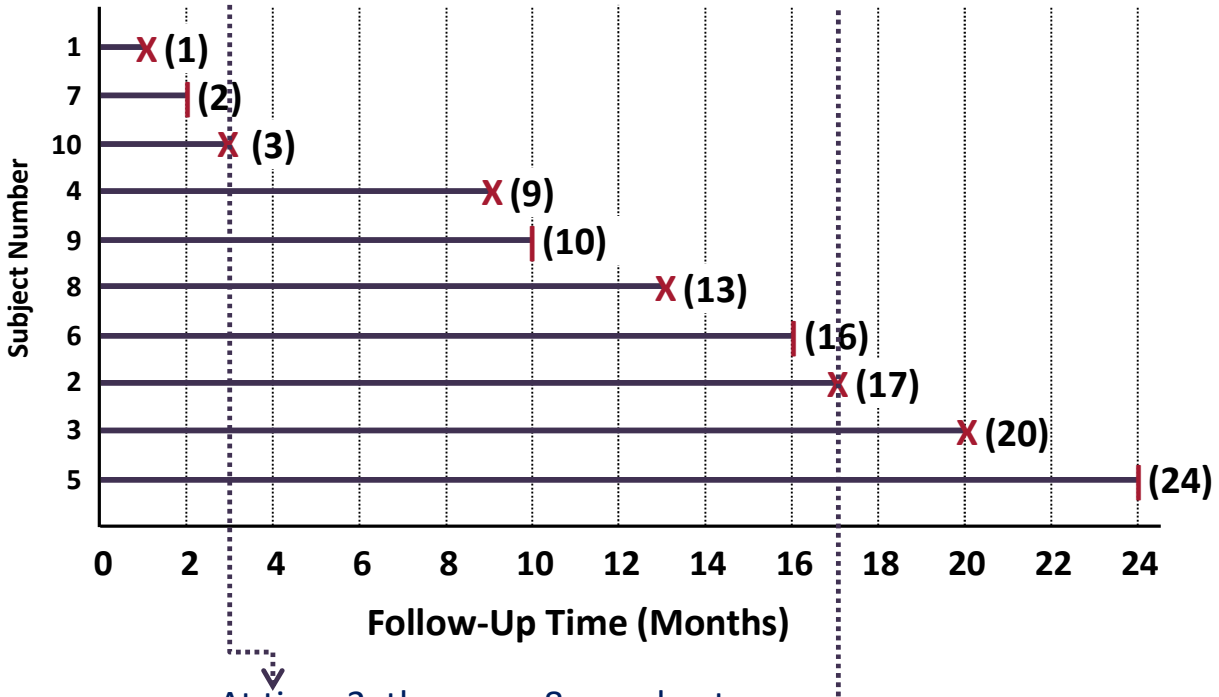
Cumulative Incidence (CI)

- Proportion of people with **incident** (new) events between time t_0 and t_1

$$CI = Pr[Y = 1] = R = \frac{A}{N} = \frac{\text{\# incident cases during } t_0 \text{ to } t_1}{\text{\# individuals at risk at } t_0}$$

- Range: 0 to 1
- Dimensionless proportion, but must state time period
- Anyone included in the denominator must be eligible to move into the numerator
  - No disease at start of follow-up and at risk of developing disease
  - Everyone is followed for fixed amount of time
  - As cases of disease accumulate over time, the CI increases
- Synonyms: incidence proportion, risk, attack “rate”

Hypothetical Cohort
Time on Study, Sorted Shortest to Longest



X =death

|=censor

()=follow-up in months

At time 3, there are 8 people at risk (#10,4,9,8,6,2,3,5)

At time 17, there are 3 people at risk (#2,3,5)

Adapted from *Epidemiology: Beyond the Basics*, 3rd Edition, Szklo and Nieto , Chapter 2

Kaplan-Meier (Product-Limit) Formula

Index	Time (Months)	# Events	# Censored	# At Risk	Conditional Cumulative Incidence	Conditional Survival Proportion	Cumulative Survival Proportion	
K	t_k	A_k		N_k	CI_k	S_k		
1	1	1	0	=10	1/10=0.100	9/10=0.900	(9/10)	=0.900
2	3	1	1	10-1-1=8	1/8=0.125	7/8=0.875	(9/10)(7/8)	=0.788
3	9	1	0	8-1-0=7	1/7=0.143	6/7=0.857	(9/10)(7/8)(6/7)	=0.675
4	13	1	1	7-1-1=5	1/5=0.200	4/5=0.800	(9/10)(7/8)(6/7)(4/5)	=0.540
5	17	1	1	5-1-1=3	1/3=0.333	2/3=0.667	(9/10)(7/8)(6/7)(4/5)(2/3)	=0.360
6	20	1	0	3-1-0=2	1/2=0.500	1/2=0.500	(9/10)(7/8)(6/7)(4/5)(2/3)(1/2)	=0.180

- If incorrectly ignore variable follow-up, S=4/10=0.40 and CI=1-S=6/10=0.60
- The cumulative survival proportion over the whole time interval is the product of the conditional survival proportions for every subinterval t_{k-1} to t_k
- $S = \prod_{k=1}^{6} \frac{N_k - A_k}{N_k}$ =(9/10)(7/8)(6/7)(4/5)(2/3)(1/2)=0.18 and CI=1-S=41/50=0.82

Adapted from *Epidemiology: Beyond the Basics*, 3rd Edition, Szklo and Nieto , Chapter 2

Incidence Rate (IR)

- $IR = \frac{\#\text{ incident cases during } t_0 \text{ to } t_1}{\Sigma \text{ person-time at risk accumulated during } t_0 \text{ to } t_1}$
- Synonyms: incidence density, hazard rate
- Cannot be interpreted without time unit
  - An incidence rate of 100 cases per 1 person-year might be expressed as
    - 100 cases / person-year = 100 cases x person-years^{-1}
    - 8.33 cases / person-month
    - 1.92 cases / person-week
    - 0.27 cases / person-day
  - Units usually chosen to ensure that minimum rate has at least one digit to the left of the decimal place

Relationship Between Prevalence & Incidence Rate

- Prevalence is a function of both incidence rate and duration
- Under steady state,

$$\text{prevalence odds} = \frac{P}{1-P} = \text{incidence rate} \times \text{duration}$$

- Combines information on occurrence and duration
  - If incidence is low but duration is long → prevalence is high
  - If incidence is high but duration is short → prevalence is low
- Prevalence can decrease if
  - Fewer cases added to population; occurrence decreases
  - People remain in population for shorter duration
    - More people are cured
    - People die faster

Interpreting measures of association (I)

- Your interpretation MUST specify all 7 components to clearly communicate findings

Component	Example
Measure	Cumulative incidence
Outcome	Myocardial infarction
Exposed	Hypertension
Value of the measure	7.45
Relative Change	Times (or "fold" or as we will see soon, "% higher")
Comparator/referent	No hypertension
Time frame	10 years (NOT "the study period"!)

The cumulative incidence of MI among women with hypertension was 7.45 times the cumulative incidence of MI among women without hypertension over 10 years

Interpreting measures of association (II)

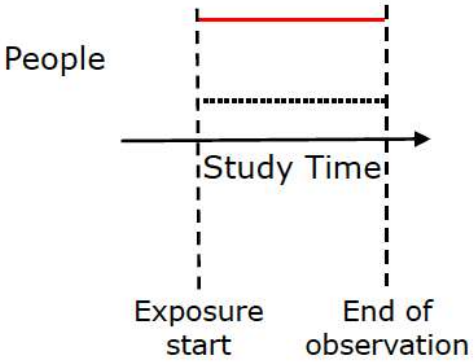
- Your interpretation MUST specify all 7 components to clearly communicate findings

Component	Example
Measure	Cumulative incidence
Outcome	Myocardial infarction
Exposed	Hypertension
Value of the measure	75 cases per 10,000 women
Excess/decrement	Excess, more, decrement, fewer
Comparator/referent	No hypertension
Time frame	10 years (NOT "the study period"!)

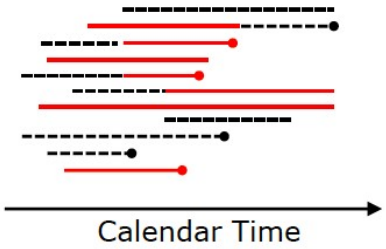
Women with hypertension had an excess cumulative incidence of 75 MIs per 10,000 women over 10 years compared to women without hypertension

Cohort studies

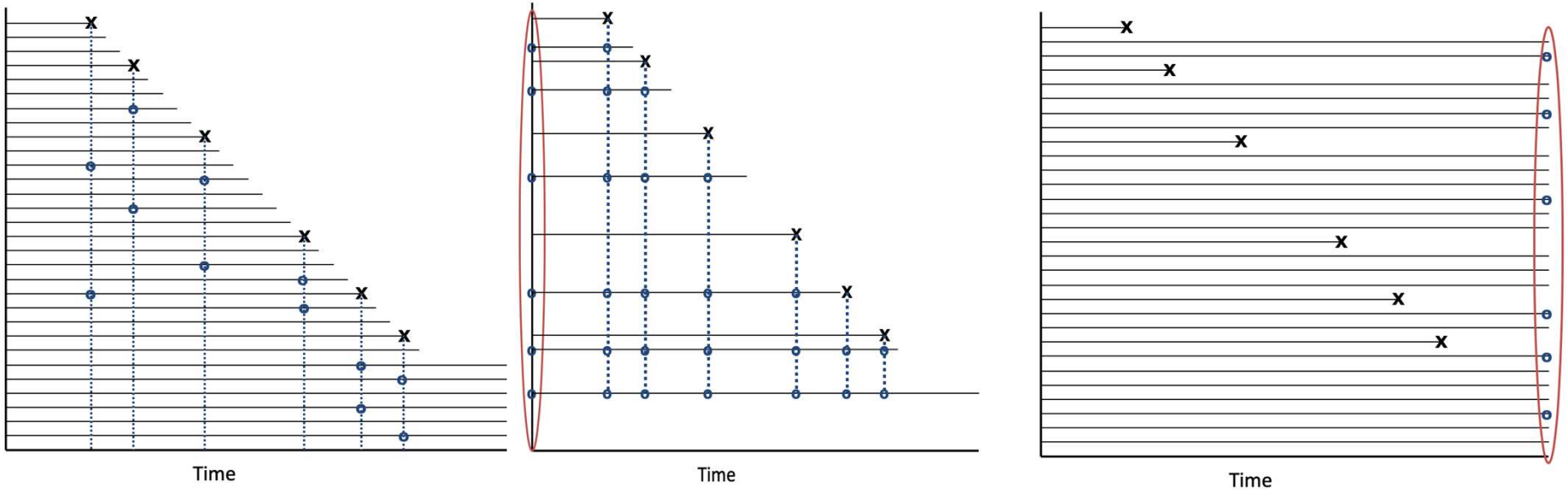
- Closed cohorts:
  - ✓ Membership defining **event**
  - ✓ All measures of association can be directly calculated
    - Unless losses to follow-up or competing risks
  - ✓ The individual is the unit of observation
  - ✓ "*Two-by-two table*" with rows of cases and total number of individuals



- Open cohorts:
  - ✓ Membership defining **state**
  - ✓ Only rates can be directly calculated
  - ✓ Person-time is the unit of observation
  - ✓ "*Two-by-two table*" with rows of cases and total of person-time



Case-control studies (I)

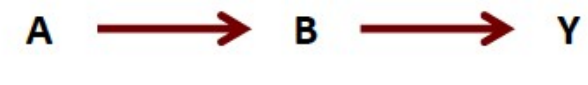


Case-control studies (II)

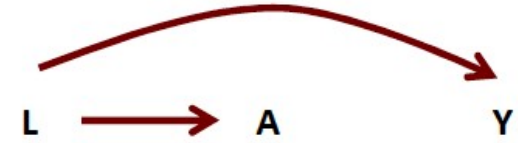
	Risk set	Case-cohort	Case-crossover	Cumulative incidence sampling
Sampling from	Person-time	Person-time	Person-time	Closed cohort (individuals)
OR approximates	IRR	IRR	IRR	OR
Same control twice?	Yes (rare)	Yes	No	No
RDA*	No	No	No	Yes, for the OR ~ CIR
When sampling?	Each time a case occurs	Start (sub-cohort) Each time a case occurs		End (survivors)
Can a control become a case?	Yes	Yes	All	No
Sampling prop. to person-time	Yes	Yes		No

* rare disease assumption

Directed acyclic graphs

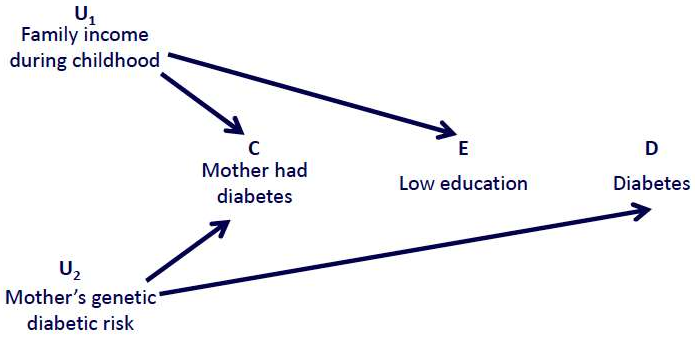


Not to condition on mediators



To recognize confounding

To recognize selection bias

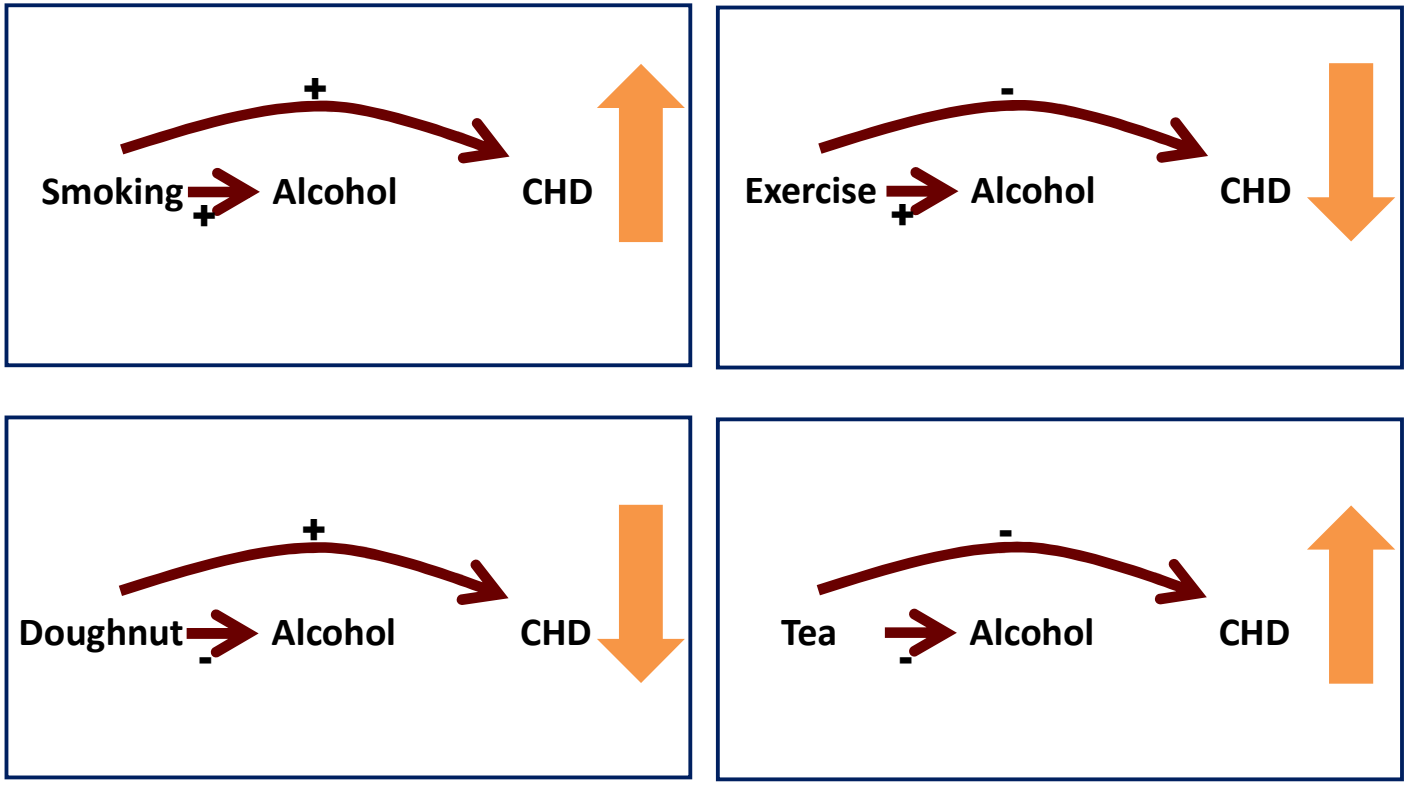


To understand when a variable can meet the properties but not be a confounder

Impact of Confounding on Validity

- Introduces bias into point estimates
- Reduces the validity of statistical inferential procedures and results such as p-values and confidence intervals
- May lead to p-values that are incorrectly too small or too large
- May lead to confidence intervals that are
  - incorrectly centered-due to bias in the point estimate
  - of incorrect width-due to bias in the estimate of the variance of the point estimate

Direction of Confounding



Assume exposures and outcomes are binary and there are no other sources of confounding or bias

Target trial emulation

- Approach to study design and data analysis of observational data
- Explicitly emulating the (hypothetical) target trial (i.e., the target of inference)
- *Seven components*: eligibility criteria, treatment strategies, treatment assignment, outcome, time zero and follow-up, causal contrasts and data analysis
- Control for confounding is necessary to appropriately emulate treatment assignment (assigned at random in the target trial)

Effect Measure Modification

- In the presence of effect measure modification, the magnitude of the association between exposure and disease varies according to the value of (across strata of) a third factor, which is called an effect modifier.
- Effect measure modification is an intrinsic phenomenon and cannot be eliminated from a study through clever design
- Effect measure modification is a finding to be reported rather than a bias to be avoided
- Synonyms: interaction, synergy, antagonism, heterogeneity of treatment effect
- Issue of external validity, therefore, limits generalizability (transportability)
- EMM is scale dependent
- EMM is reciprocal – If X modifies effect of Z on Y then Z modifies the effect of X on Y.

Confounding vs. Effect measure modification

Confounding	EMM
• Confounder associated with exposure <u>in study base</u> • Confounder associated with outcome <u>even in the absence of exposure</u> • Confounder not a downstream consequence of exposure on outcome	In the presence of effect measure modification, the magnitude of the association between exposure and outcome varies according to the value of (across strata of) a *third* variable, called the effect modifier.
Bias in point estimates, p-values, confidence intervals	No bias
Not scale-dependent	Scale dependent (additive, multiplicative). If present in the additive scale, causal interaction.
Stratification, restriction, matching …	Report stratum-specific estimates Standardization
Does not affect generalizability	Affects generalizability

Causal inference in an ideal randomized experiments

ID	$Y^{a=1}$	$Y^{a=0}$
1	0	1
2	1	0
3	1	1
4	0	0

ID	A	Y
1	0	1
2	1	1
3	1	1
4	0	0

causal risk ratio: $Pr[Y^{a=1}=1] / Pr[Y^{a=0}=1]$

associational risk ratio $Pr[Y=1|A=1] / Pr[Y=1|A=0]$

$$\frac{Pr[Y^{a=1}=1]}{Pr[Y^{a=0}=1]} = \frac{Pr[Y^{a=1}=1|A=1]}{Pr[Y^{a=0}=1|A=0]} = \frac{Pr[Y=1|A=1]}{Pr[Y=1|A=0]}$$

- I can do this only if $Y^a \perp\!\!\!\perp A$ for all values of "a"
- This step is called "Exchangeability": i.e., counterfactual outcome is independent of exposure

- Remove counterfactuals
- This step is called "Consistency"

Plus *positivity* and *well-defined interventions*

Causal inference in observational studies

	Marginal RCT	Conditional RCT	Observational study
Exchangeability at baseline			
Causal effect identifiable from data alone?			
Confounding?			

Crucial Difference: Association is NOT Causation

Association	Causation
▪ Different risk in two <u>disjoint subsets of the population</u> determined by the subjects' actual exposure value	▪ Different risk in the <u>same population</u> under two exposure values
▪ Pr[Y=1|A=a] is the risk in subjects of the population that meet the condition "having actually received exposure level a"	▪ $Pr[Y^a=1]$ is the risk in all subjects of the population had they received the counterfactual exposure level a

Formal Definition of Exchangeability

$Y^a \amalg A$ for all a

- Exchangeability implies lack of confounding
- Exchangeability is another causal concept that cannot be represented by associational language

Exchangeability vs. Independence

- Exchangeability $Y^a \perp\!\!\!\perp A$

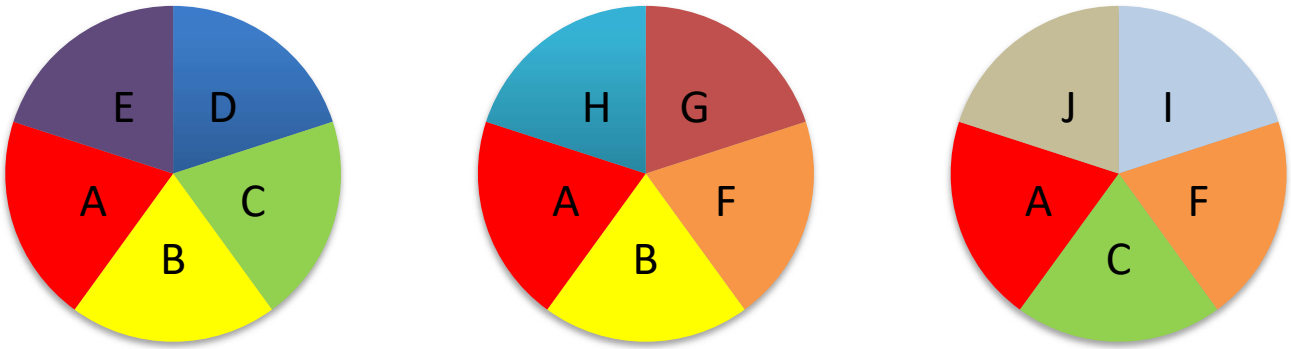
  is different from

- Independence $Y \perp\!\!\!\perp A$

Causal Inference From Observational Data Is Possible

- Under the assumptions of
  - Exchangeability
  - Consistency
  - Positivity
  - Well-defined interventions and outcomes
- Can be computed using methods that account for stratification factors
- Other assumptions are required for both observational and randomized data

Causal inference based on the sufficient component cause theory



- Sufficient vs. Component Vs. Necessary
- If two causes coexist on the same pie, they share the same pathway
  - We will see interaction or effect measure modification in the additive scale
  - Agnostic about multiplicative EMM
- Strength of a given cause is determined by the prevalence of component causes.

Hill criteria

- A list of 9 criteria supposed to bring evidence of causality:
  - Strength
  - Consistency
  - Specificity
  - Temporality
  - Biologic gradient
  - Plausibility
  - Coherence
  - Experimental evidence
  - Analogy

- Only temporality is required

EPI 202

Statistical inference: hypothesis testing

1. State the null and alternative hypotheses:

    a) Explicitly include the measure of association and the study design
        i. H_0: The incidence rate in the exposed is the same as the incidence rate in the unexposed, or IRR=1 or IRD=0
        ii. Ha: The incidence rate in the exposed is not the same as the incidence rate in the unexposed, or IRR=1 or IRD=0

2. Construct a test statistic:

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

Difference between **statistical significance** and **clinical relevance**

Statistical inference: hypothesis testing

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

3. Use the X^2 distribution with 1 df to find out the p-value associated with our estimated Z^2 result

4. Interpretation of the p-value
   - ✓ The probability of observing a result as extreme or more extreme under the assumption that the **null hypothesis is true** due to random variation.

5. Use of the p-value in the context of hypothesis testing:
   - ✓ We fail to reject the null hypothesis at the 0.05 level and hence we conclude that the data do not support the hypothesis that the incidence rate in the exposed is different than the incidence rate in the unexposed.

Statistical inference: drawbacks of hypothesis testing

- The P-values summarize the consistency of the observed data with the state of nature described by the null.

- Limitations of P- values: no information on
  - Direction
  - Magnitude
  - Range of values compatible with the data
  - Power

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

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

Statistical inference: hypothesis testing

- Assumptions necessary for valid interpretation of inferential procedures:
    - No confounding
    - No selection bias
    - No measurement error

- Definition of bias
    - If we repeated the experiment in random samples from the same population and the expected value of my estimate is the same as the true value, we say that my estimate is unbiased
    - e.g., RR_{MH}

- Relationship between statistical efficiency / information / variance / power

Analytic approaches for stratified analysis

- H_0: There is no association between exposure and outcome **after stratifying by X.**
  - H_0: $IRR_{MH} = 1 \leftrightarrow IRD_{Summ} = 0$
- H_A: There is an association between exposure and outcome **after stratifying by X.**
  - H_A: $IRR_{MH} \neq 1 \leftrightarrow IRD_{Summ} \neq 0$

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i\left(X_i|H_0\right)\right]^2}{\sum_{i=1}^{I} Var_i\left(X_i|H_0\right)} \sim \chi_1^2$$

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

$$ln\,\hat{IRR}_{MH} \pm 1.96\sqrt{\hat{Var}\left(ln\,\hat{IRR}_{MH}\right)}$$

Link to positivity violations

Analytic approaches for effect measure modification

- H_0: The rate ratio is the same across all I levels of the stratification variable(s)
  - ↔ There is no effect modification of the IRR by the stratification variable(s)
  - ↔ The rate ratio is *homogeneous* across the strata
  - ↔ $IRR_1 = IRR_2 = \ldots = IRR_I$
  - ↔ $IRR_i = IRR_j$ for all i,j

- H_A: The rate ratio is not the same across all I levels of the stratification variable(s)
  - ↔ There is effect modification of the IRR by one (or more) of the stratification variable(s)
  - ↔ The rate ratio is *heterogeneous* across the strata
  - ↔ **At least one** of the IRRs does not equal at least one of the others

$$H = \sum_{i=1}^{I} \frac{[ln(\hat{IRR}_i) - ln\,\hat{IRR}_{MH}]^2}{\hat{Var}_i[ln(\hat{IRR}_i)]} \sim \chi^2_{I-1}$$

Relative Excess Risk due to Interaction (RERI)

Consider two exposures, A1 and A2:

	$A_1=0$	$A_1=1$
$A_2=0$	CI_{00}	CI_{10}
$A_2=1$	CI_{01}	CI_{11}

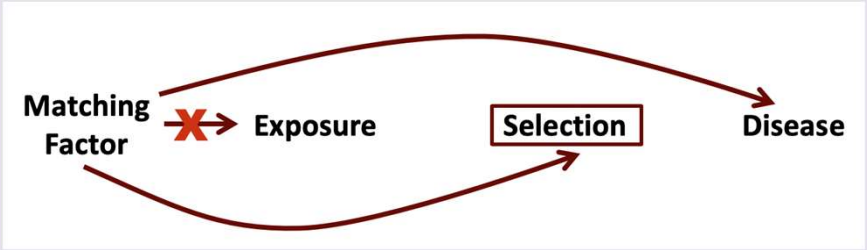
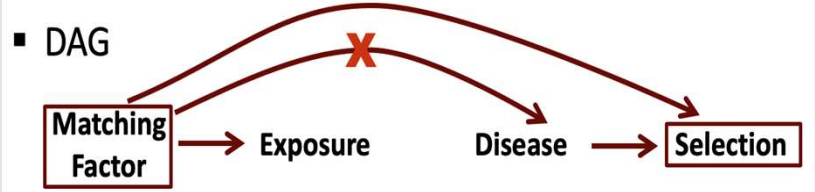
If there is no EMM on the additive scale then
$CI_{10} - CI_{00} = CI_{11} - CI_{01}$ and therefore, $CI_{11} - CI_{01} - CI_{10} + CI_{00} = 0$

Divide each term by CI_{00} and no EMM on the additive scale implies
$CIR_{11} - CIR_{01} - CIR_{10} + 1 = 0$.

This expression is called the Relative Excess Risk due to Interaction (RERI). The RERI allows the use of multiplicative parameters to determine whether there is EMM on the additive scale.

RERI = 0 if there is no EMM on the additive scale
RERI > 0 if the absolute effect of A1 is greater in the presence of A2 (and vice versa)
RERI < 0 if the absolute effect of A1 is weaker in the presence of A2 (and vice versa)

Matching in cohort and case-control studies

Cohort	Case-control
	▪ DAG 
We can evaluate if the matching factor is a risk factor for the outcome	We *cannot* evaluate if the matching factor is a risk factor for the outcome
We can study EMM by the matching factor	We can study EMM by the matching factor
Erasing the arrow from MF to exposure	Erasing the arrow from MF to disease, but adding the arrow from MF to selection
We get an unbiased estimate in the matched population	We need to use stratification-based methods after matching (e.g., M-H or McNemar)

Analysis of matched case-control studies

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

Shorthand for OR_{MH} in matched case-control studies.

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

$$\hat{OR}_{MH} = \frac{f_{10}}{f_{01}}$$

$$\mathrm{Var}[ln(\hat{OR}_{MH})] = \frac{1}{f_{10}} + \frac{1}{f_{01}}$$

Only discordant pairs contribute information

GOOD LUCK ON THURSDAY!