

ERROR

- RANDOM ERROR = chance, random variation = remains unexplained

Random error has many components:

- UNEXPLAINED VARIATION: E is occurrence measures.
- SAMPLING ERROR: subjects of the study are viewed as a sample of possible people who could have been included in the study / or the \neq experiences the study subjects could have had.

The degree to which a sample pop. deviates from the total pop.

It is unpredictable and due to the sampling process.

most studies rely on samples because it is too costly to measure whole pop.

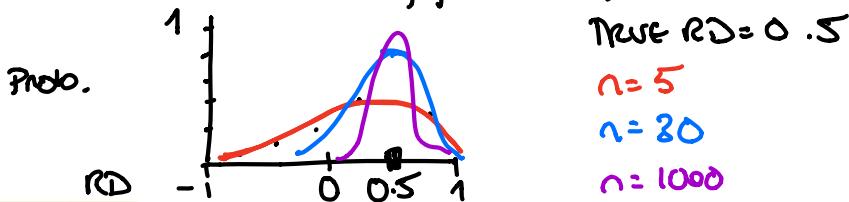
ASSUMPTIONS OF SAMPLING: Samples are

- Random = RANDOMNESS ASSUMPTION
- representative of the population.

SAMPLING DISTRIBUTION:

Different samples will result in different measures of occurrence; the larger the sample the more likely that the measure of occurrence corresponds to the truth in that population; however there is always some variation.

ex:



- STATISTICAL PRECISION: { inverse of the variance of the measurements.
- improved by
• ↑ sample size } opposite of random error.
- ↑ statistical efficiency: modifying the design of a study to ↓ variance:
- = n^e of cases and controls } if there is no effect or confounding.
- = n^e of exposed / non-exposed }

- DECREASED BY:
 - STRATIFICATION: if there are many factors to stratify on, there can be strata that vary scarcely between low/high values. = SPARSE DATA
MATCHING on stratification variables \uparrow precision without \uparrow sample size
when matching is not feasible $\rightarrow \uparrow$ sample size.

■ STATISTICAL INFERENCE ; APPROACHES TO RANDOM ERROR

what conclusion can be drawn about population from the information provided by a sample of that population?

SIGNIFICANCE vs. HYPOTHESIS TESTING: Assume that the model is correct.

Criticism:

CRITICISM :
 we need more than a decision as to if choice above could produce the association;
 we need to estimate {
 the MAGNITUDE of association
 the PRECISION of the estimator method.

Solutions: use
CONFIDENCE INTERVALS, range of values for the association
INTERVAL ESTIMATION?

MISINTERPRETATION OF SIG. TEST :

H_0 : Hypothesis of no association in the superpopulation that was sampled between 2 variables

- H_0 : not stat. significant ($p > 0.05$) refers to the SUPERPOPULATION
it means = cannot reject H_0 that the SUPERPOPULATION groups are the same
NOT: the two observed groups sampled from the SUPERPOPULATION are the same.

- H_0 : stat. significant ($p < 0.05$) :

Does NOT mean the superpopulation groups are different: there could be
- sources of uncontrollable bias
- chance alone (5%) that is was sig without difference

P-VALUES:

Better interpretation: there's a problem w/the test hypothesis for both the study

MISINTERPRETATION:

- p-values represent probabilities of test hypotheses
- p-value is the probability of the observed data under the test hypothesis
wrong bc p-value includes \Rightarrow +
probability of all other possible data configurations where the test statistic was more extreme than observed.
- p-value is the probability that the observed data would show as strong an association or stronger than that observed under the test hypothesis.

HYPOTHESIS TESTS:

Type II errors result when: magnitude of effect, biases, random variability combine to give results that are insufficiently inconsistent with the H_0 to reject it. This failure to reject can occur bc the effect is small for both too few observations

ALTERNATIVE HYPOTHESIS: should be formulated "either the H_0 is false or the model is wrong".

STATISTICAL ESTIMATION:

Epidemiological analysis of data is a measurement, not a decision-making problem. Effect measures are measured on a continuous scale w/ ∞ n° of possible values

POINT ESTIMATE: estimate of the target parameter = magnitude of effect / unlikely that it's the true parameter due to bias, random error... association.

- PRECISION = CONFIDENCE INTERVAL = uncertainty of point estimate:
depends on the amount of random variability and the α -level

INTERVAL ESTIMATION provides an idea of the direction and magnitude of the eff. NOT \rightarrow the random variability of the point estimate

2-SIDED P-VALUE: provides degree of consistency between the data and a single H

■ CAUSAL INFERENCE.

CAUSAL TYPES :

PROB. OF DISEASE	E^+	E^-	Prob. in cohort 1 (E^+)	Prob. in cohort 2 (E^-)
DOomed	1	1	p_1	q_1
SUSCEPTIBLE	1	0	p_2	q_2
PROTECTED	0	1	p_3	q_3
IMMUNE	0	0	p_4	q_4

MEASURING OF EFFECT : effect of E^+ on the prob of being D^+ in the SAME pop.
= counterfactual

MEASURE OF ASSOCIATION : effect of E^+ on the prob of being D^+ between E^+ and E^- pops.

CONFOUNDING : MEASURING OF EFFECT \neq MEASURING OF ASSOCIATION

counterfactual observed

CONFOUNDER : factor that explains the difference $p_1 + p_3 \neq q_1 + q_3$

Ex: MEASURES OF EFFECT ASSOCIATION	COUNTERFACTUAL	REAL LIFE
$RD = P(D^+ E^+) - P(D^+ E^-) =$	$(p_1 + p_2) - (p_1 + p_3)$	$(p_1 + p_2) - (q_1 + q_3)$
$OR: \frac{P(D^+ E^+)/P(D^- E^+)}{P(D^+ E^-)/P(D^- E^-)}$	$\frac{(p_1 + p_2)}{1 - (p_1 + p_2)} / \frac{(p_1 + p_3)}{1 - (p_1 + p_3)}$ $(p_2 + p_4)$	$(p_2 + p_4)$ $\frac{(p_1 + p_2)}{1 - (p_1 + p_2)} / \frac{(q_1 + q_3)}{1 - (q_1 + q_3)}$ $(q_2 + q_4)$
$RR: \frac{P(D^+ E^+)}{P(D^+ E^-)} =$	$\frac{(p_1 + p_2)}{(p_1 + p_3)}$	$\frac{(p_1 + p_2)}{(q_1 + q_3)}$

$OR/RR = 1$ and $RD = 0$ mean $\begin{cases} \text{no effect} \\ = \text{number of susceptible and protected} \\ \text{balance of causal / preventive effects} \end{cases}$

Ex:

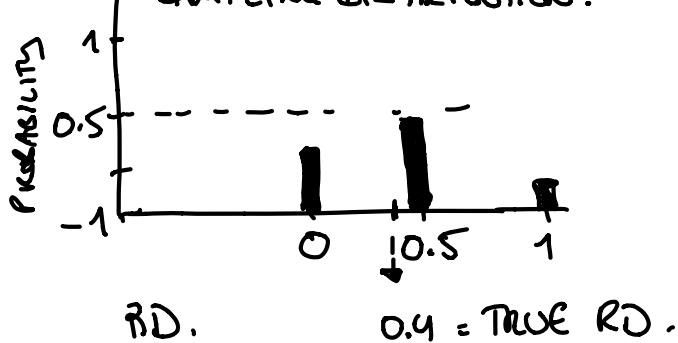
counterfactual of DOSED E^+ ? = DOSED E^- = D^+

counterfactual of SUSCPT. E^- ? = SUSCPT. E^+ = D^+

EXAMPLE.

Sample $\Rightarrow n=2$ Options:	$D^+ E^+$	$n=2$ $D^+ E^-$	RD	Probability
Susceptible = 2 $\Rightarrow \frac{2}{2} = 1$ Immune = 0 $\Rightarrow \frac{0}{2} = 0$	1x1=1 0	1x0=0 0	1-0=1 0	$0.4 \times 0.4 = 0.16$
Susceptible = 0 $\Rightarrow \frac{0}{2} = 0$ Immune = 2 $\Rightarrow \frac{2}{2} = 1$	0x1=0 1x0=0	0 0	0-0=0 0	$0.4 \times 0.6 = 0.24$ $0.6 \times 0.4 = 0.24$
Susceptible = 1 $\Rightarrow \frac{1}{2} = 0.5$ Immune = 1 $\Rightarrow \frac{1}{2} = 0.5$	0.5x1=0.5 0.5x0=0	0.5x0=0 0	0.5-0=0.5 =0.5	$0.6 \times 0.6 = 0.36$

SAMPLING DISTRIBUTION.



TRUE POPULATION

	$D^+ E^+$	$D^+ E^-$
Susceptible = 0.4	$0.4 \times 1 = 0.4$	$0.4 \times 0 = 0$
Immune = 0.6	$0.6 \times 0 = 0$	$0.6 \times 0 = 0$

TRUE RD
= 0.4

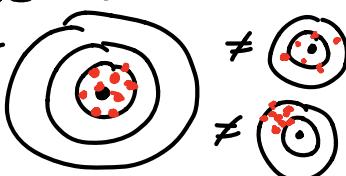
• SYSTEMATIC ERROR = BIAS (CH. 9 RGL)

VALIDITY - the opposite of bias

estimate w/ little systematic error = close to true value

Accuracy = **VALID** : NO systematic error

+
Precise : NO random error



INTERNAL VALIDITY : accurate measure of effects apart from random variation. in some pop

EXTERNAL VALIDITY : generalizability to other pop.

- **INFORMATION BIAS**: bias caused by measurement errors.
- CONTINUOUS VBLFS = INF. ERROR.
- CATEGORICAL VBLFS = MISCLASSIFICATION
- DIFFERENTIAL: classification error depends on actual values of other vbls
- NON-DIFFERENTIAL: " does NOT depend on actual values of other vbls
- DEPENDENT: classification error depends on errors measuring other vbls
classifying
- INDEPENDENT: " does NOT depend on errors measuring other vbls
classifying

IN BINARY VBLFS:

SENSITIVITY: $P(T^+|E^+)$

FALSE NEG PROB: $P(T^-|E^+) = 1 - Se$

PREDICTIVE VALUE \ominus : $P(E^-|T^-)$

SPECIFICITY: $P(T^-|E^-)$

FALSE POS \oplus : $P(T^+|E^-) = 1 - Sp$

PREDICTIVE VALUE \oplus : $P(E^+|T^+)$

- DIFFERENTIAL: classification error depends on actual values of other vbls
This bias can be either towards or away from the Ho.

Ex: D = Emphysema E = Smoking. D x of emphysema ↑ in smokers bc they have ↑ resp. dis and seek T medical care. Smoking → Emphysema ↑ Detection
Underdiagnosis of D⁺ in E⁻ → Failure to detect true cases ↑ in non-smoker

Ex RECALL BIAS: D = congenital malformation. Mothers of D⁺ babies recall T E⁺

- NON-DIFFERENTIAL: " does NOT depend on actual values of other vbls
This bias is towards the Ho if the exposure is binary, but can be away from the Ho if { E or D vbls are non-binary (> 2 levels)
depends on errors measuring other vbls
classifying => DEPENDENT MISCLASSIFICATION.

- NON-DIFFERENTIAL EXPOSURE MISCLASSIFICATION:

Proportion of subjects misclassified on exposure does not depend on the status of the subject w/ respect to other vbls (including D⁺).

② Dichotomous exposure + independent misclassification = Bias to H₀

Effect of Nondifferential Misclassification of Alcohol Consumption on Estimation of the Incidence-Rate Difference and Incidence-Rate Ratio for Laryngeal Cancer (Hypothetical Data)

TOTAL 1,500,000	Incidence Rate ($\times 10^5$ y)	Rate Difference ($\times 10^5$ y)	Rate Ratio
No misclassification			
1,000,000 drinkers	50	40	5.0
500,000 nondrinkers	10		
Half of drinkers classed with nondrinkers			
500,000 drinkers	50	20	1.7
1,000,000 "nondrinkers" (50% are actually drinkers)	30		
Half of drinkers classed with nondrinkers and one-third of nondrinkers classed with drinkers			
666,667 "drinkers" (25% are actually nondrinkers)	40	6	1.2
833,333 "nondrinkers" (60% are actually drinkers)	34		

BIAS

→ H₀ miss. of E⁺ : one direction

→ H₀ miss. of E⁺ AND E⁻: bidirectional

Nondifferential bias se and sp = for exposure measurement of cases (control)

Nondifferential Misclassification with Two Exposure Categories

	Exposed	Unexposed	
Correct data			
Cases	240	200	
Controls	240	600	OR = 3.0
Sensitivity = 0.8			
Specificity = 1.0			
Cases	192	248	
Controls	192	648	OR = 2.6
Sensitivity = 0.8			
Specificity = 0.8			
Cases	232	208	
Controls	312	528	OR = 1.9
Sensitivity = 0.4			
Specificity = 0.6			
Cases	176	264	
Controls	336	504	OR = 1.0
Sensitivity = 0.0			
Specificity = 0.0			
Cases	200	240	
Controls	600	240	OR = 0.33

OR, odds ratio.

} Bias towards H₀.

se + sp = 1 \Rightarrow no effect

se + sp < 1 \Rightarrow Bias away from H₀ in opposite direction of actual effect.

③ > 2 levels of exposure + independent misclassification = Bias away from H₀ or towards H₀

↑ high E⁺ in low E group,

compared to E⁻ \rightarrow bias away H₀, ↑ OR

bc high E⁺ have higher risk of D⁺.

	Unexposed	Low Exposure	High Exposure
Correct data			
Cases	100	200	600
Controls	100	100	100
	OR = 2		OR = 6
40% of high exposure \rightarrow 4 low exposure			
Cases	100	440	360
Controls	100	140	60
	OR = 3.1		OR = 6

← 40%

- NON-DIFFERENTIAL DISEASE MISSCLASSIFICATION:

Proportion of subjects misclassified or disease does not depend on the status of the subject w/ respect to other variables (including E^+).
In most cases it will produce bias towards the H₀ if it is independent from other errors.

Bias depends on { measure of association
SP, SE

Ex: TRUTH.

	D ⁺	D ⁻	
E ⁺	40		100
E ⁻	20		200

■ SP = 100%
SE < 100%
Non-diff classification
of D⁺, indep of Exp.

Specificity 100%: P(D⁻|IT⁻) = 100%
Sensitivity 70% = P(D⁺|IT⁺) = E⁺/E⁺
↳ non-differential ad independent
of exposure classification errors.

	D ⁺	D ⁻	
E ⁺	40 × 0.7 = 28		100
E ⁻	20 × 0.3 = 14		200

■ SP < 100%
SE = 100%
Non-diff classification
of D⁺, indep of Exp.

Sensitivity 100%: P(D⁺|IT⁺) = 100%
Specificity 80% = P(D⁻|IT⁻) = E⁻/E⁻
↳ non-differential ad independent
of exposure classification errors.

	D ⁺	D ⁻	
E ⁺	52	$\frac{100-40}{100} = \frac{60}{100} \times 0.9 = 48$	100
E ⁻	56	$\frac{200-20}{200} = \frac{180}{200} \times 0.8 = 144$	200

BIAS

RR	$RR = \frac{40/100}{20/200} = 4$	NO	$RR = \frac{28/100}{14/200} = \frac{28}{7} = 4$	NO	$RR = \frac{52/100}{56/200} = \frac{52}{56} = 1.9 \rightarrow H_0$
RD	$RD = \frac{40-20}{100-200} = 0.4 - 0.1 = 0.3$	NO	$RD = \frac{28-14}{100-200} = 0.28 - 0.07 = 0.21 \rightarrow H_0$	$RD = \frac{52-56}{100-200} = 0.52 - 0.28 = 0.24 \rightarrow H_0$	

$\frac{0.21}{0.3} = 0.7$: 70% of the true RD = sensitivity 0.7
or RR more biased towards H₀

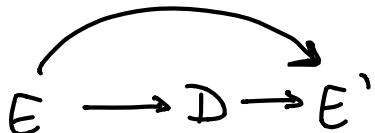
$\frac{0.24}{0.3} = 0.8$ = sensitivity

(10)



E' is determined by E^+ but not D,
so missclassification is non-differential
Any ass. between E' and D is due
only to causal effect of $E \rightarrow D$ (or E')

(11)



E' is determined by E^+ and also D,
so missclassification is differential
Ex: recall bias
Part of the ass. between E' and D
is spurious because of the causal effect
 $D \rightarrow E'$

CONFOUNDING

confounder: role (ass. w/ outcome conditional on the exposure)
 ass w/ exposure "↑" (ex: in the exposed group)
 NOT on the causal pathway between $E \rightarrow D$

IDENTIFYING ConfoundERS (Hernán 2002) STRATEGIES:

① Automatic variable selection procedures: ex: stepwise selection
 Assumption: all important confounders will be selected

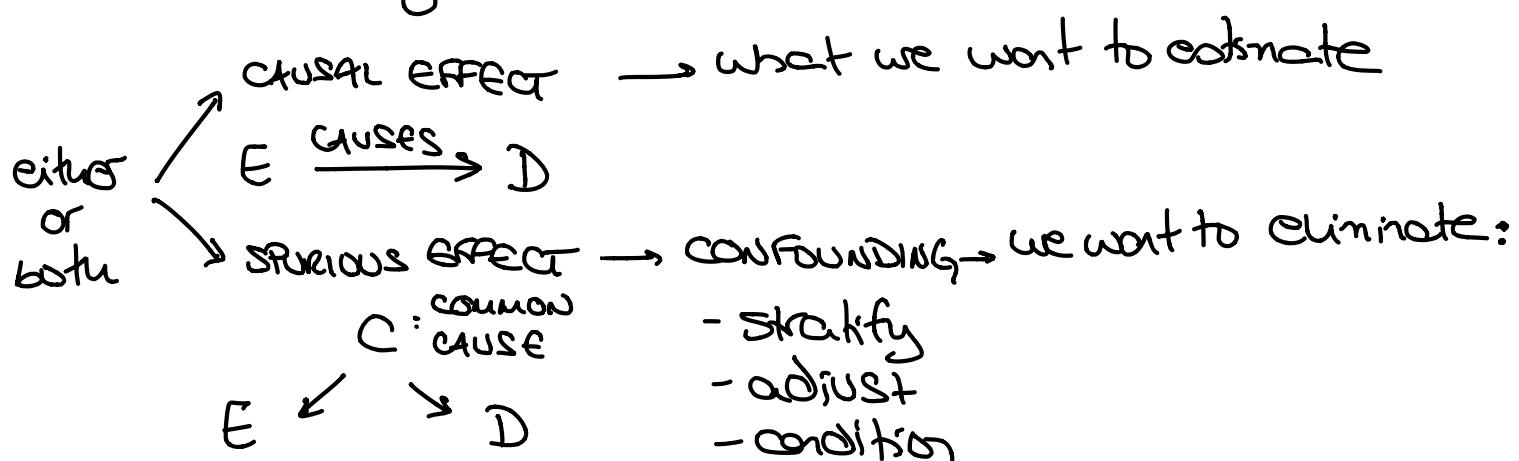
② Change in effect estimate: comparison of adjusted/unadjusted effect estimates; if $> 10\%$ relative change \rightarrow vble is selected

③ Check whether vble meets criteria for a confounder:
 combines statistical associations + background knowledge.

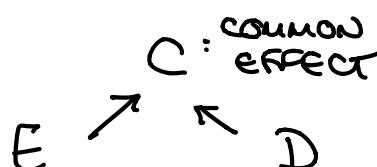
CAUSAL DIAGRAMS = DAGS = Directed Acyclic Graphs

Because cause precedes effect, the cannot be a closed C

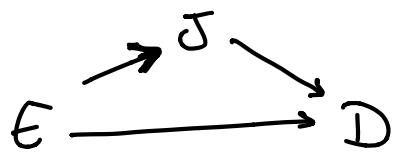
E is statistically associated w/ D : the causal association is



COMMON EFFECT \rightarrow COLLIDER



THE CAUSAL EFFECT can be:

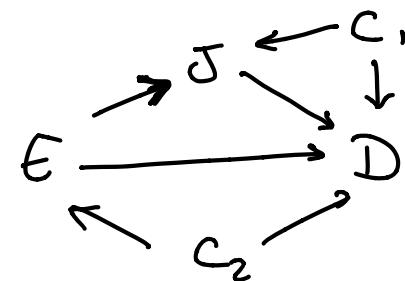
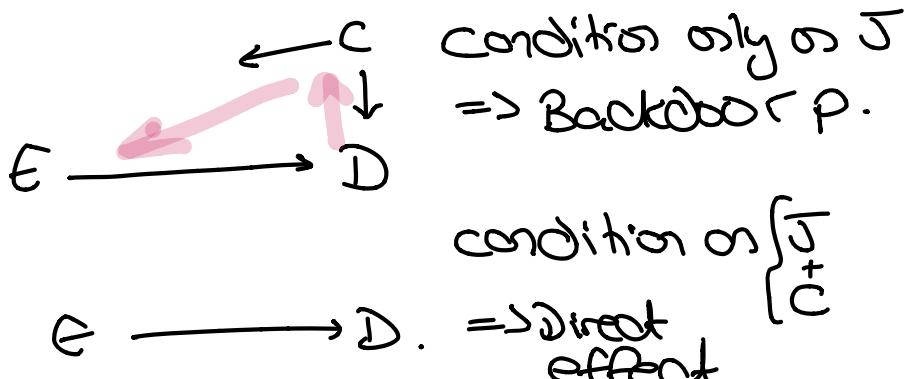
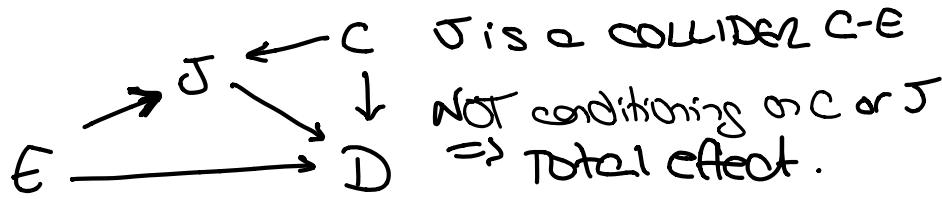


$$\text{INDIRECT EFFECT : } E \xrightarrow{\quad} J \xrightarrow{\quad} D$$

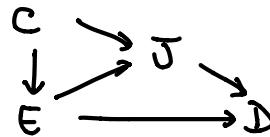
$$\text{DIRECT EFFECT : } E \xrightarrow{\quad} D$$

$$\text{TOTAL EFFECT}$$

To obtain an unbiased estimate of the direct effect, requires measuring and adjusting for all confounders of the $J \rightarrow D$ association,
AND $E \rightarrow D$ association



For total effect: C
For direct effect: $\{C_2, C_1, J\}$



for total effect: C
NOT J



For direct effect: J
 C is not a confounder bc
it's not ass. w/ D .

Ex : E : low folic acid intake ; D : birth defects ; C : low birth weight
 U : unmeasured variable

1 C Not common cause = not confounder, no adjustment necessary

\Rightarrow

	COLLIDER	CONDITION ON.
		C
① E \rightarrow D \rightarrow C	C	NOT C! It's harmful. Ex:

ASSUME E $\not\rightarrow$ D
NOT CAUSE



D = 1 D = 0

E = 1	100	100
E = 0	200	200

$$RR_{ED}^* = 1$$

Ex E: diet; D: concert; C: $1 > 5 \text{ kg} \uparrow; C = 0 < 5 \text{ kg} \downarrow$
condition on C: creates spurious associations E \rightarrow D

If a person lost weight, if not dieting, concert \uparrow likely
so within those who lost weight E and D more ass.

C = 1

C = 0

D = 1	D = 0
E = 1	55
E = 0	70

D = 1	D = 0
C = 1	45
C = 0	130

$$RR_{ED|C=1}^* = 0.79; RR_{ED|C=0} = 0.92$$

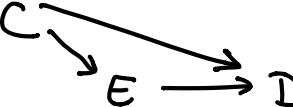
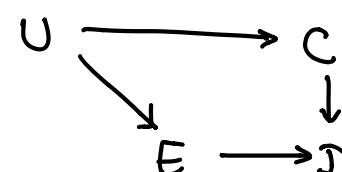
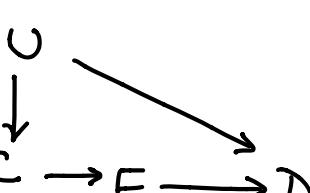
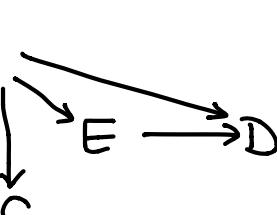
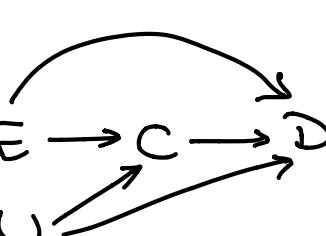
\rightarrow SPURIOUS NON-CAUSAL ASS.

E and D statistically independent

	COLLIDER	CONFOUNDER	CONDITION ON.
②	C	U1	C: spurious ass. U1 \rightarrow D
③	C	U2	C: spurious ass. U1 \rightarrow E
④	C	U1, U2	C: spurious ass. U1 \rightarrow E

• C a common cause for E and D = confounder.

Association has a spurious component and the OREDIC is a biased estimator of the causal component.

	CONFOUNDER	CONDITION ON
⑤		C C: OREDIC . If D is rare $\approx RR_{EDIC}$ valid estimator of the causal effect $E \rightarrow D$
⑥		U: affects D by changing the value of C C: OREDIC . If D is rare $\approx RR_{EDIC}$ valid estimator of the causal effect $E \rightarrow D$
⑦		U: affects E by changing the value of C C: OREDIC . If D is rare $\approx RR_{EDIC}$ valid estimator of the causal effect $E \rightarrow D$
⑧		U C: surrogate confounder C: if the actual confounder cannot be measured, adjusting for a surrogate confounder is better than none or. if C is strongly correlated w/ U, it will remove most of the confounding: OREDIC
⑨		U C: Not a confounder \rightarrow Not on causal pathway.

Here we are defining confounding based on background knowledge of the causal structure, contrasting w/ strategies ① ad ②: only stat. ass.
These Strategies would hold for ①-⑧

Is C a confounder? depends on # of restrictions

③ stat ass + partial a priori knowledge.

⑨ NO \rightarrow not on causal pathway

we can't gauge this just from stat. ass.

③-⑧ NO \rightarrow not a risk factor (ass w. exp)

①-③ NO \rightarrow affected by exposure or outcome

Example : D=1 mums w kids w neural tube defects | E=1: folic acid supp.
 D=0 " " " other defects. | E=0: NO "

C not in causal pathway. All vbls perfectly measured.

CRUDE.

STRATIFIED ON C

$$\text{logit } P(D=1|E) = \beta_0 + \beta_1 E \quad \text{logit } P(D=1|E, C) = \beta_0 + \beta_1 E + \beta_2 C$$

	D=1	D=0
E=1	43	239
E=0	194	704

	C=1		C=0	
	D=1	D=0	D=1	D=0
E=1	19	8	24	231
E=0	100	46	94	658

\Rightarrow no heterogeneity of OR
Assume No interaction

crude OR_{ED} = 0.65

$$OR_{ED|C} = 0.80 = e^{\beta_1}$$

which analysis more appropriate?

STRATEGY

- ① C is added if the p-value associated w $\beta_2 < 0.10$. Here it's $p < 0.001 \rightarrow$ select C
- ② OR_{ED|C} is 23% greater than OR_{ED} \rightarrow select C. \rightarrow Adjustment for C
- ③ Check $\left\{ \begin{array}{l} C \text{ is ass w/ E : } OR_{C|E=D=0} = 0.5 \\ C \text{ is ass w/ D : } OR_{C|D=E=0} = 15.22 \end{array} \right\} \rightarrow$ select C. \nearrow is appropriate.
C not in causal pathway.

If C = stillbirth or abortion. To know if it's a confounder we need to know the underlying causal structure. It's not ⑤-⑦ bc they assume C occurs before E or D.

STRATEGY FOR SELECTION OF CONTROL VBLS : SUFFICIENT SET

- Delete arrows emanating from E or D
- Identify unblocked backdoor paths $D \rightarrow E$
- Select a set S of vbles that block all backdoor paths unless they are colliders: if you adjust for a collider, you may be unblocking a backdoor path.
- Select vbles that block remaining / new b. paths, preferably closer to D/outcome.
If they are colliders, control for their ancestors
- S is now a sufficient set.
- Smallest set of confounders: drop further vbles until the b. path test fails

ASSUMPTIONS DAG :

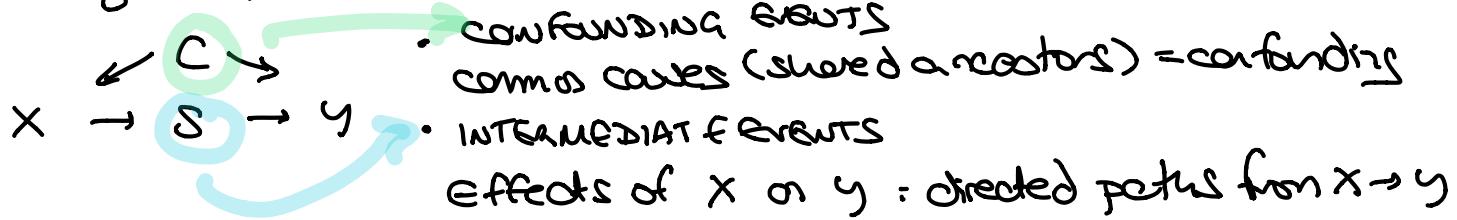
$A \rightarrow B$ A causes B and precedes B temporally.

if 2 r_bs are d-separated
CAUSAL MARKOV ASSUMPTION : COMPATIBILITY: they must be independent.

$X \rightarrow S \rightarrow Y$: X and Y are independent/unassociated given S.

Y is independent of all other variables it does not affect,
conditioned on its direct parents(S).

Assuming X precedes Y temporally; only 2 sources of association X-Y,

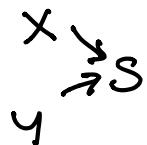


• INTERMEDIATE EVENTS

Effects of X on Y: directed paths from X to Y

condition =
If we hold constant confounding and intermediate events it respects the association $X \rightarrow Y$.

WEAK FAITHFULNESS: X and Y might be associated given S
Presence of open paths alert to the possibility of association.



FAITHFULNESS: If 2 r_bs are independent, they must be d-separated.

SUFFICIENT: set that controls bias in the $X \rightarrow Y$ association.

MINIMALLY SUFFICIENT: set that controls bias in the $X \rightarrow Y$ association,
if no proper subset of S is sufficient (removing any set of r_bs leaves an insufficient set).

DIRECTION OF CONFOUNDING

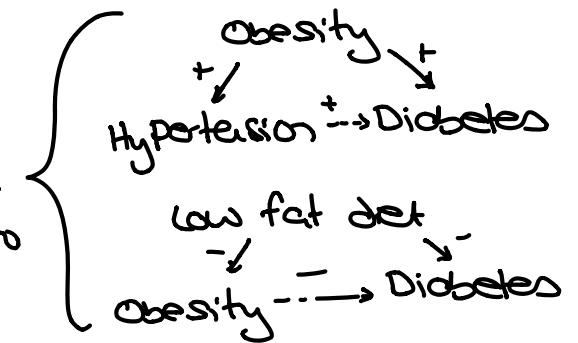
Depends on:

D

■ Direction of ass. w/ E

- Same direction w/ D and E → UPWARD BIAS
Away from H₀

- ≠ direction w/ D and E varies



DOWNTWARD BIAS : Family Hx Breast cancer
TOWARDS H₀, Oral contracept. → Breast cancer

UPWARD BIAS : High fat diet
Away H₀ Physical activity ← Diabetes

UPWARD BIAS : Low SES
TOWARDS H₀ PREGNATAL SUPPLEMENTS → AUTISM DX

■ Effect on outcome

■ True ass. of E → outcome

- ↑ risk : conditions that bias UPWARD, away from H₀ or ?
- ↓ risk : conditions that bias DOWNTWARD, toward H₀ or ?

METHODS TO CONTROL CONFOUNDING.

■ STUDY DESIGN :

- RANDOMIZATION: Limits confounding by unmeasured factors probabilistically + accounts quantitatively for any residual confounding they produce.

NOT identical distribution of unmeasured factors

YES similar distribution"

" not affected by the trt.

Limitations: feasibility, ethics

- RESTRICTION: Restrict study selection to ind. w = locus of the confounder

Limitations: ↓ eligible ind, can't analyse/evaluate a factor that is matched on, possible residual confounding if not restrictive enough.

■ STUDY DESIGN + STUDY ANALYSIS :

- MATCHING: does not itself control for confounding, but increases the efficacy of the study that does; the confounding factor must still be controlled in the analysis.
= no cases/controls for a category

- ONE TO ONE ; - ONE TO MANY ; - FREQUENCY M.: multiple controls to multiple cases.

- CUTOFF M.: within a range (± 2 yrs) - PROGNOSIS SCORE M.: estimate of likelihood of receiving trt.

useful: case-controls, controlling for complex idios

Limitations: expensive, can't analyse/evaluate a factor that is matched on, sample size limits, you need to collect the idios before post. enrollment.

■ STUDY ANALYSIS

- STRATIFICATION: clarifies degree of homogeneity.

$$OR_{\text{strata}} = \frac{\sum a_{ij}/n_i}{\sum b_{ij}/n_i} \quad RR = \frac{\sum a_{1j}/n_1}{\sum b_{1j}/n_1} \quad OR_{\text{overall}} = \frac{\sum a_{1j}/n_1}{\sum b_{1j}/n_1}$$

How: 2x2 tables → weighted avg. of RR and OR across strata \Rightarrow Cochran-Mantel-Haenszel estimate

Limitations: residual confounding if confounder not evenly distributed in C and D. feasible only for few idios.

- STANDARDIZATION: Reweight strata - spp rates sorted exp. categories or comparable

using a reference pop. from the dataset or external source.

Limitations: hard to adjust for multiple confounders.

- MULTIVARIATE ADJUSTMENT: logistic/linear reg., ANOVA etc -

⊕ Handles multiple covariates simultaneously

⊖ Confounding can persist

• INVERSE PROBABILITY OF TREATMENT WEIGHTING (IPTW)

Confounding → prob. of exp. for ind. differs among levels of confounding var.
IPTW creates a pseudo-pop without confounding by re-weighting, yet the true exp. is not altered.

- ⑤ Can be used in time-dependent confounding: who earlier exp. influences the confounding effect of later exp.
Ex: ind. w/ sx: early exp → behavior change → Attacher of future exp.

Ex. E: hypertension D: CHD . C: smoking.

CRUDE			
	D ⁺	D ⁻	TOTAL
E ⁺	250	770	1000
E ⁻	240	960	1200
	470	1730	2200

SMOKING C ⁺			
	D ⁺	D ⁻	TOTAL
E ⁺	180	720	900
E ⁻	40	360	400
	220	1080	1300

NON-SMOKING C ⁻			
	D ⁺	D ⁻	TOTAL
E ⁺	50	50	100
E ⁻	200	600	800
	250	650	900

IPTW calculation: $w_1 = \frac{1}{\frac{P(E=e|C=c)}{P(E=e)}} = \frac{P(E=e)}{P(E=e|C=c)} \Rightarrow$ overall prob. of exp
 \Rightarrow conditional prob. of exposure

	N	P(E=e)	P(E=e C=c)	WEIGHT	PSEUDO-N
E ⁺ C ⁺	900	$\frac{1000}{2200} = 0.45$	$\frac{900}{1300} = 0.69$	$\frac{0.45}{0.69} = 0.65$	$0.65 \cdot 900 = 570.85 \approx 591$
E ⁻ C ⁺	400	$\frac{1200}{2200} = 0.5454$	$\frac{400}{1800} = 0.80$	$\frac{0.5454}{0.3} = 1.8$	$1.8 \cdot 400 = 709$
E ⁺ C ⁻	100	0.45	$\frac{100}{900} = 0.1$	$\frac{0.45}{0.1} = 4.09$	$4.09 \cdot 100 = 409$
E ⁻ C ⁻	800	0.5454	$\frac{800}{900} = 0.89$	$\frac{0.5454}{0.89} = 0.606$	$0.606 \cdot 800 = 490.38 \approx 491$

NON-COCONFUSED TABLE

	D ⁺	D ⁻	TOTAL
E ⁺			
E ⁻			
	2200		

SMOKING C ⁺			
	D ⁺	D ⁻	TOTAL
E ⁺	$0.65 \times 180 = 118$	$0.65 \times 720 = 473$	591
E ⁻	$1.8 \times 40 = 72$	$1.8 \times 360 = 648$	709
	190	1300	

NON-SMOKING C ⁻			
	D ⁺	D ⁻	TOTAL
E ⁺	$4.09 \times 50 = 204.5$	$4.09 \times 50 = 204.5$	409
E ⁻	$0.6 \times 200 = 120$	$0.6 \times 600 = 360$	480
	250	650	900

- SELECTION BIAS**: non-comparability of E^+ and E^- induced by restriction the analysis to certain level/s of a common effect of E and D or \neq confounding due to unmeasured common cause?
 vbls corr w/
 E and D .

Popn.	D^+	D^-
E^+	A	B
E^-	C	D

Study sample	D^+	D^-
E^+	a	b
E^-	c	d

SAMPLING FRACTIONS	D^+	D^-
E^+	$\frac{a}{A} = f_a$	$\frac{b}{B} = f_b$
E^-	$\frac{c}{C} = f_c$	$\frac{d}{D} = f_d$

$$OR : \frac{ad}{bc} : \frac{f_a \cdot A \cdot f_d \cdot D}{f_b \cdot B \cdot f_c \cdot C} = \frac{A \cdot D}{B \cdot C}$$

we want $\frac{f_a f_d}{f_b f_c} = 1$ to preserve the original OR.

Example :

Popn.	D^+	D^-
E^+	1000	1600 000
E^-	5000	1200 000

Study sample	D^+	D^-
E^+	100	320
E^-	500	240

10% sample 0.02% sample

NO SELECTION

BIA^S

because % of cases and non-c
rs = for E^+ and E^-
(balanced) = 1

$$OR = \frac{100 \times 240}{500 \times 320} = 0.15 \quad OR = \frac{100 \times 240}{500 \times 320} = 0.15 \quad \frac{f_a \cdot f_d}{f_c \cdot f_b} = \frac{0.1 \cdot 0.02}{0.1 \cdot 0.02} = 1$$

- CONTROL factors to ↓ BIAS
- factors that were part of the study design
 - factors that correlate w/ participation.

LIMITATIONS: can be hard to adjust for selection bias when there is no info on the non-participants.

Study sample	D^+	D^-
E^+	100	380
E^-	500	480

10% sample

Selection BIAS

$0.02\% = f_b$ unbalanced sampling fractions:

$0.04\% = f_d$

$$OR = \frac{100 \times 480}{500 \times 380} = 0.3$$

$$\frac{f_a f_d}{f_b f_c} = \frac{0.1 \cdot 0.04}{0.1 \cdot 0.02} = 2$$

INTERACTION

occurs when the measure of association between risk factor and disease depends on the level of another factor { effect measure modification heterogeneity of measure}

2 exposures/risk factors act n:

SYNERGISM: the additive measure of joint exposure > MORE THAN

the sum of the measures of association between D and E

ANTAGONISM: the additive measure of joint exposure and disease < LESS THAN the sum of the measures of association for each exp. alone

INTERACTION TABLE:

RISK OF DISEASE

RELATIVE RISK = RR

A

A

	A -	A +
B -	R_{A-B^-}	R_{A+B^-}
B +	R_{A-B^+}	R_{A+B^+}

	A -	A +
B -	1	R_{A+B^-} / R_{A-B^-}
B +	R_{A+B^+} / R_{A-B}	R_{A+B^+} / R_{A-B^+}

EFFECT MEASURE MODIFICATION

One variable of interest, how does the other variable affect it?
How does B change at ≠ levels of A?

	A -	A +
B -	1	R_{A+B^-} / R_{A-B^-}
B +	R_{A+B^+} / R_{A-B}	R_{A+B^+} / R_{A-B^+}

INTERACTION: equally interested in both variables.

What is the combined effect?

	A -	A +
B -	1	R_{A+B^-} / R_{A-B^-}
B +	R_{A+B^+} / R_{A-B}	R_{A+B^+} / R_{A-B^+}

Ex:	A - nonexp	A low Res.Exp	A high Smaller worker
B: NonSmokers	1	2.3	8.4
B: Smokers	8.3	17.5	26.2

	A low Res.Exp	A high Smaller worker
B: NonSmokers	$2.3/1 = 2.3$	$8.4/1 = 8.4$
B: Smokers	$17.5/8.3 = 2.1$	$26.2/8.3 = 3.15$

EFFECT MOD OF RR?	NO	YES: ANT.	EFFECT MOD OF Excess RR?	YES: SYN	YES: SYN
MULTIPLICATIVE INTERACTION	$2.3 \approx 2.1$	$8.4 \gg 3.15$			
	$8.3 \times 2.3 \approx 17.5$	$8.3 \times 8.4 \approx 64 \neq 26.2$	ADDITIONAL INTERACTION	$1.3 \ll 9.2$	$7.4 \ll 17.9$

INDEPENDENTLY ACTING FACTORS

ADDITIVE EFFECT:

Background risk $f(\Theta_A, \Theta_B) = 3$

- Exposure A adds $9 - 3 = 6$ units of risk (attributable)

to both non-exposed to B = $3 + 6 = 9$
or exposed to B = $15 + 6 = 21$

- Exposure B adds $15 - 3 = 12$ (attributable)

to both $\Theta_A = 3 + 12 = 15$

$$\Theta_A = 9 + 12 = 21$$

		Attributable Risks	
		Incidence Rates	
		Factor A	
Factor B	-	-	3.0
	+	+	9.0
	-	-	15.0
	+	+	21.0

$$\begin{aligned}
 & \text{combined risk } A+B = 21 \\
 & \text{background risk } = 3 \\
 & + \text{Attributable risk A} = 6 \\
 & + \text{Attributable risk B} = \frac{12}{18}
 \end{aligned}$$

MULTIPLICATIVE EFFECT

Background risk $f(\Theta_A, \Theta_B) = 3$

- $A\oplus$: $9 = 3$ triples risk

to both $\begin{cases} B\ominus \\ B\oplus \end{cases} 3 \times 3 = 9$
 $15 \times 3 = 45$

- $B\oplus$: $15 = 5 \rightarrow \times 5$ risk

to both $\begin{cases} A\ominus \\ A\oplus \end{cases} 3 \times 5 = 15$
 $9 \times 5 = 45$

		Relative Risks	
		Factor A	
		Factor A	
Factor B	-	-	3.0
	+	+	9.0
	-	-	15.0
	+	+	45.0

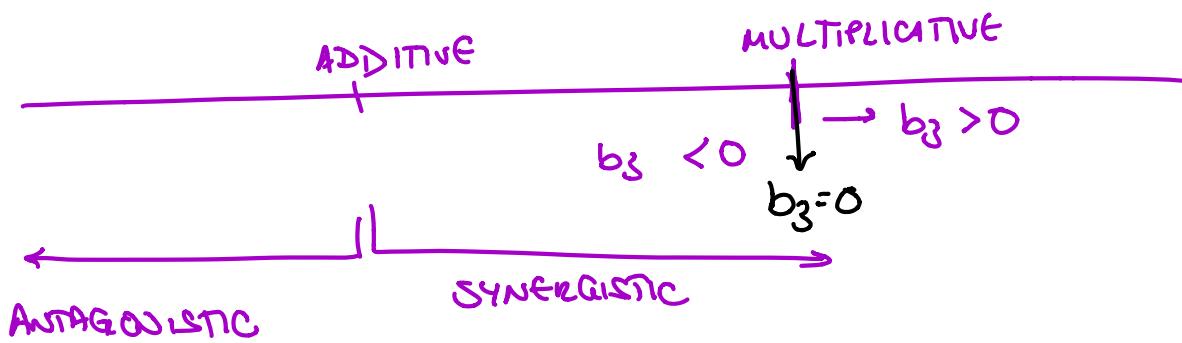
$$\frac{45}{3} = 15$$

CONFOUNDING

INTERACTION

DEFINITION	property of dist. of Exp / D in source pop.	Bidirectional or socio-logic result → Report can be statistically independent of exposure
ACTION	Bias → Control	not necessarily
DISTRIBUTION	different by exp. status	Heterogeneous
PREDICTS OUTCOME	YES	Between stratum - spp M.O.A makes no sense
STRATUM-SPP M.O.A.	Homogeneous across strata	
CRUDE M.O.A	Differs from Homogeneous ↑	
SUMMARY ADJUSTED M.O.A.	Makes sense	

Logistic, Pois, prop. hazards models all assume factors act multiplicatively unless an interaction effect is included. The joint effect lies on a continuous spectrum



$$\log \left(\frac{P}{1-P} \right) = b_0 + b_1 x_1 + b_2 x_2 + b_3 (x_1 x_2) \Rightarrow P = \frac{e^{(b_0+b_1 x_1 + b_2 x_2 + b_3 (x_1 x_2))}}{1+e^{(b_0+b_1 x_1 + b_2 x_2 + b_3 (x_1 x_2))}}$$

Ex: Is A an effect modifier of B?

Effect of B⁺ vs. B⁻

- In the presence of A

$$\begin{cases} B^+ x_2 = 1 & \underline{\log ODDS} = \beta_0 + \beta_1 + \beta_2 + \beta_3 \\ B^- x_2 = 0 & \underline{\log ODDS} = \beta_0 + \beta_1 \end{cases}$$

$$\frac{\beta_2 + \beta_3}{\beta_2 + \beta_3}$$

- without A

$$\begin{cases} B^+ x_2 = 1 & \underline{\log ODDS} = \beta_0 + \beta_2 \\ B^- x_2 = 0 & \underline{\log ODDS} = \beta_0 \end{cases}$$

$$\frac{\beta_2}{\beta_2}$$

effect modification $\Rightarrow \beta_2 \neq \beta_2 + \beta_3$

MARGINAL AND CONDITIONAL ASSOCIATION

		C = 1		C = 0	
		D ⁺	D ⁻	D ⁺	D ⁻
E ⁺				E ⁺	
E ⁻				E ⁻	

RR = 1 No association

MARGINALLY INDEPENDENT

RR = 2

CONDITIONALLY ASSOCIATED / DEPENDENT

RR = 2

		D ⁺	D ⁻
E ⁺			
E ⁻			

		D ⁺	D ⁻
E ⁺			
E ⁻			

		D ⁺	D ⁻
E ⁺			
E ⁻			

RR = 2 Association = unblocked path

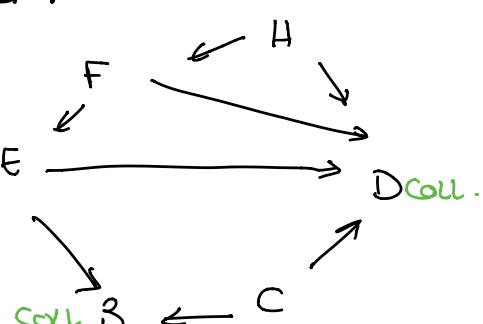
MARGINALLY ASSOCIATED

RR = 0.5

RR = 2

EFFECT MEASURE MODIFICATION

Ex:



H . marginally ass. w everything except C

B ass. w. { E
F
H
D
C } everything

C marginally indep. of everything except B and D.

Conditional ass. between

C - F
C - H
C - E } conditioning on B

RANDOMIZATION ASSUMPTION: the exposure to which the subject is assigned doesn't depend on the risk of disease.
= absence of confounding = comparability of E+ and E-

DIEGO : "For Science!"

source pop.	E^+	E^-	study pop	D^+	D^-
D^+	A	B	E^+	a	b
D^-	C	D	E^-	c	d

A: total n° of diseased people that were exposed - max value of a

that were E^-

B: " " of D^+

E^+

C: " " D^-

that were E^-

D: " " D^-

OR DIE using sampling fractions \rightarrow assess if selection bias is present.

Your exposed is the disease in your study pop. don't represent the exposed in the disease in the source pop.

$$OR = \frac{\frac{a}{A}}{\frac{c}{C}} = \frac{\frac{0.05}{0.01}}{\frac{1}{0.2}} = 1 \quad \left. \begin{array}{l} \\ \\ \end{array} \right\} \text{for example}$$

Selection bias DOES NOT depend only on exposure, it depends on all 4 cells, that is, also on disease.

SELECTION BIAS DAG $E \dashrightarrow D$
 $\downarrow S \swarrow$

DIRECTIONS OF CONFONDING.

