ERROR

- **RANDOM ERROR** = chance, random variation = remains unexplained

Random error has many components:

- **UNEXPLAINED VARIATION**: E in 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 ≠ 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.

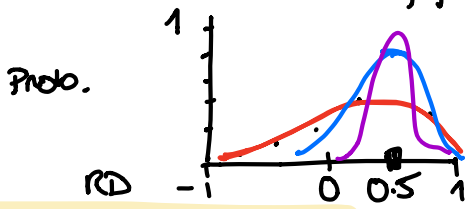
ASSUMPTIONS OF SAMPLING: Samples are

- 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:



True RD = 0.5
n = 5
n = 30
n = 1000

STATISTICAL PRECISION: {inverse of the variance of the measurements. / opposite of random error.

- IMPROVED BY
  - ↑ sample size
  - ↑ statistical efficiency: modifying the design of a study to ↓ variance:
    - = nº of cases and controls
    - = nº of exposed / non-exposed

} if there is no effect or confounding.

- DECREASED BY:

• STRATIFICATION: if there are many factors to stratify on, there can be strata that vary severely between low/high values. = SPARSE DATA

MATCHING on stratification vbles ↑ precision without ↑ sample size

When matching is not feasible → ↑ 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:

we need more than a decision as to if chance alone could produce the association;

we need to estimate { the MAGNITUDE of association
the PRECISION of the estimation method.

Solutions: use { CONFIDENCE INTERVALS: range of values for the association
INTERVAL ESTIMATION!

MISINTERPRETATION OF SIG. TEST:

Ho: Hypothesis of NO ASSOCIATION in the SUPERPOPULATION that was sampled between 2 vbles

• Ho: not stat. significant ($p > 0.05$) refers to the SUPERPOPULATION

it means = cannot reject Ho 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 the different: there could be

- Sources of uncontrolled bias
- chance alone (5%) that is was SS without difference

P-VALUES:

Better interpretation: there's a problem w/ {the test hypothesis / the study} or both

MISINTERPRETATION:

- p-values represent probabilities of test hypotheses
- p-value is the probability of the observed data under the test hypothesis + probability of all other possible data configurations where the test statistic was more extreme than observed.

  wrong bc p-value includes =>

- 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 / too few observations} or both

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/association. unlikely that it's the true parameter due to bias, random error...
- 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 ass. / the random variability of the point estimate}

2-SIDED P-VALUE: provides NOT degree of consistency between the data and a single H

CAUSAL INFERENCE.

CAUSAL TYPES:

PROB. OF DISEASE	E^+	E^-	Prop. in cohort 1 (E^+)	Prop. 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

MEASURE 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: MEASURE OF EFFECT ≠ MEASURE OF ASSOCIATION

CONFOUNDER: factor that explains the difference $\underbrace{p_1+p_3}_{\text{counterfactual}} \neq \underbrace{q_1+q_3}_{\text{observed}}$.

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

OR/RR = 1 and RD = 0 can mean { no effect
= number of susceptible and protecteds.
balance of causal / preventive effects

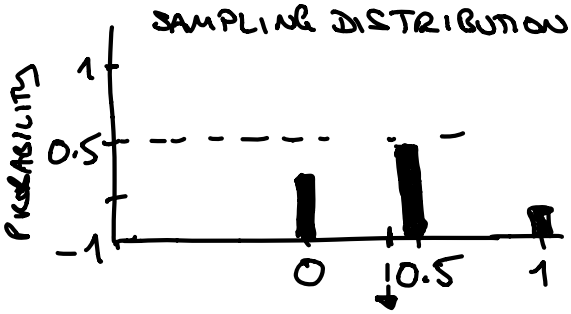
Ex:

counterfactual of $D^+|E^+$? = $D^+|E^-$ = D^+

counterfactual of $SUSCEPT.|E^-$? = $SUSCEPT.|E^+$ = D^+

EXAMPLE.

Sample ⇒ n = 2 (n=2) Options:	$D^+\|E^+$	$D^+\|E^-$	RD	Probability
Susceptible = 2 ⇒ 2/2 = 1	1×1=1	1×0=0	1−0=1	0.4×0.4 = 0.16
Immune = 0 ⇒ 0/2 = 0	0	0		
Susceptible = 0 ⇒ 0/2 = 0	0×1=0	0	0−0=0	0.4×0.6 = 0.24 } 0.48
Immune = 2 ⇒ 2/2 = 1	1×0=0	1×0=0		0.6×0.4 = 0.24 }
Susceptible = 1 ⇒ 1/2 = 0.5	0.5×1=0.5	0.5×0=0	0.5−0=0.5	0.6×0.6 = 0.36
Immune = 1 ⇒ 1/2 = 0.5	0.5×0=0	0.5×0=0		



RD. 0.4 = TRUE RD.

TRUE POPULATION

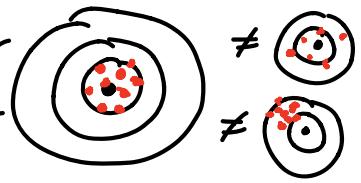
	$D^+\|E^+$	$D^+\|E^-$
Susceptible = 0.4	0.4×1=0.4	0.4×0=0.
Immune = 0.6	0.6×0=0	0.6×0=0
	0.4	0

TRUE RD = 0.4

- SYSTEMATIC ERROR = BIAS (CH. 9 RGL)

VALIDITY = the opposite of bias
estimate w/ little systematic error = close to the true value

ACCURACY = VALID : NO systematic error
+
PRECISE : NO random error



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

EXTERNAL VALIDITY : generalizability to other pop.

- INFORMATION BIAS : bias caused by measurement errors.

- CONTINUOUS VBLES = INF. ERROR.
- CATEGORICAL VBLES = MISSCLASIFICATION

- DIFFERENTIAL : classification error depends on actual values of other vbles
- NON-DIFFERENTIAL: " "does NOT depend on actual values of other vbles
- DEPENDENT : classification error depends on errors {measuring other vbles / classifying
- INDEPENDENT: " "does NOT depend on errors {measuring other vbles / classifying

IN BINARY VBLES:

SENSITIVITY : $P(T^+|E^+)$ — SPECIFICITY : $P(T^-|E^-)$

FALSE ⊖ PROB : $P(T^-|E^+) = 1 - Se$ — FALSE ⊕ P : $P(T^+|E^-) = 1 - Sp$

PREDICTIVE VALUE ⊖ : $P(E^-|T^-)$ — PREDICTIVE VALUE ⊕ : $P(E^+|T^+)$

- DIFFERENTIAL: classification error depends on actual values of other vbles

This bias can be either towards or away from the Ho.

Ex: D = Emphysema E = Smoking. Dx of emphysema ↑ in smokers bc they have ↑ resp. dis and seek ↑ medical care. Smoking → Emphysema, ↑ Detection

Underdiagnosis of D^+ in E^- → Failure to detect true cases ↑ in non-smokers

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

- NON-DIFFERENTIAL: " "does NOT depend on actual values of other vbles

This bias is towards the Ho if the exposure is binary, but can be away from the Ho if { E or D vbles are non-binary (>2 levels) / depends on errors {measuring other vbles / classifying } } => DEPENDENT MISSCLASIFICATION.

- NON-DIFFERENTIAL exposure MISSCLASIFICATION:

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

- Dichotomous exposure + independent misclassification = Bias to Ho

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

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		

→ Ho MISS. OF E^+ : ONE DIRECTION

→ Ho MISS. OF E^+ AND E^-: BIDIRECTIONAL

Nondifferential se and sp = for exposure measurement of {cases / controls

If miss. is extreme can reverse bias in opposite dir. of effect.

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 Ho.

se + sp = 1 ⇒ No effect

se + sp < 1 ⇒ Bias away from Ho in opposite direction of actual effect.

- >2 levels of exposure + independent misclassification = Bias {away from Ho / or / toward Ho

Nondifferential Misclassification with Three Exposure Categories

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

40%

↑ high E^+ in Low E^+ group, compared to E^- → bias away Ho, ↑ OR be high E^+ have higher risk of D^+.

- NON-DIFFERENTIAL DISEASE MISSCLASIFICATION:

Proportion of subjects misclassified on disease does not depend on the status of the subject w/ respect to other vbles (including E+).

In most cases it will produce bias towards the Ho if it is indep 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 classificaton of D+, indep of Exp.

SPECIFICITY 100% = P(D− | T−) No FP
SENSITIVITY 70% = P(D+ | T+) in E+/E−
↳ non-differential and independent of exposure classification errors.

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

■ Sp < 100%
Se = 100%
Non-diff classificaton of D+, indep of Exp.

SENSITIVITY 100% = P(D− | T−) No FN
SPECIFICITY 80% = P(D+ | T+) in E+/E−
↳ non-differential and independent of exposure classification errors.

	D+	D−	
E+	52	100−40 = 60 ×0.8 = 48	100
E−	56	200−20 = 180 ×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 = $\frac{52/100}{56/200} = \frac{52}{28} = 1.9$ → Ho
RD	RD = $\frac{40}{100} - \frac{20}{200} = 0.4 - 0.1 = 0.3$	NO	RD = $\frac{28}{100} - \frac{14}{200} = 0.28 - 0.07 = 0.21$ → Ho		RD = $\frac{52}{100} - \frac{56}{200} = 0.52 - 0.28 = 0.24$ → Ho
			$\frac{0.21}{0.3} = 0.7 = 70\%$ of the true RD = sensitivity 0.7 or as RR remain biased towards Ho		$\frac{0.24}{0.3} = 0.8$ = sensitivity

(10) E → D E'

E− is determined by E+ but not D, so misclasification is non differetial
Any ass. between E− and D is due only to causal effect of E → D (OR_{ED})

(11) E → D → E'

E− is determined by E+ and also D, so misclasification is differetial
Ex: recall bias
Part of the ass. between E− and D is spurious because of the causal effect D → E−

- CONFOUNDING

CONFOUNDER: vble { ass. w/ OUTCOME conditional on the exposure
ass w/ EXPOSURE "↑ (ex: in the exposed group)
NOT on the causal pathway between E → 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 → 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 G

E is statistically associated w/ D: the overall associcion is

either or both

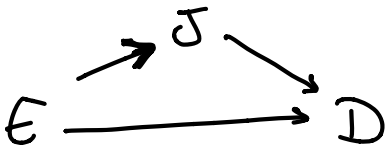
CAUSAL EFFECT → what we want to estimate
E causes → D

SPURIOUS EFFECT → CONFOUNDING → we want to eliminate:
C: common cause
E ↙ ↘ D
- stratify
- adjust
- condition

COMMON EFFECT → COLLIDER

C: common effect
E ↗ ↖ D

THE CAUSAL EFFECT can be:

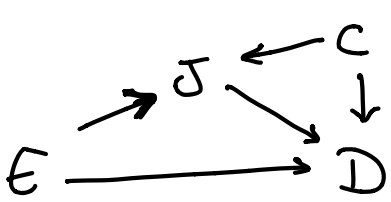


INDIRECT EFFECT : E → J → D

DIRECT EFFECT : E → D

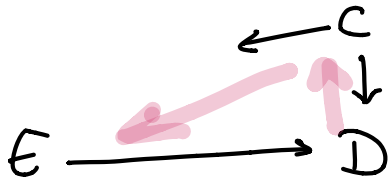
TOTAL EFFECT = INDIRECT + DIRECT

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



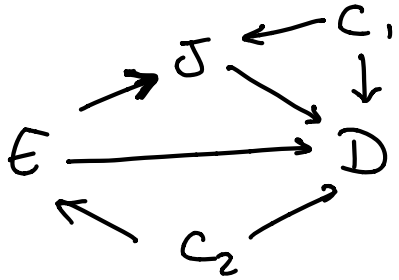
J is a COLLIDER C-E

NOT conditioning on C or J => Total effect.



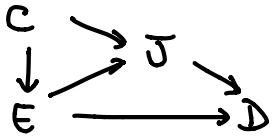
condition only on J => Backdoor p.

condition on { J, C => Direct effect



For total effect: C_2

For direct effect: { C_2, C_1, J



For total effect: C NOT J

For direct effect: J

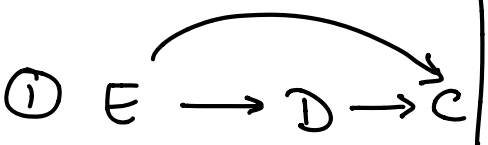
C_1 is not a confounder bc it's not ass. w/ D.

Ex : E : low folate intake ; D : birth defects ; C : low birth weight

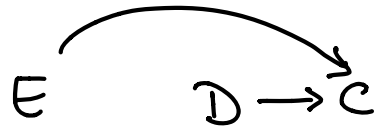
U : unmeasured vde

C Not common cause = not confounder, no adjustment necessary

⇒

	COLLIDER	CONDITION ON.
① 	C	NOT C! It's harmful. Ex:

Assume E NOT CAUSE D



Ex E: diet; D: cancer; C: = 1 > 5kg ↓; C = 0 < 5kg ↓
condition on C: created spurious association E → D
If a person lost weight, if not dieting, cancer likely
so within those who lost weight E and D more ass.

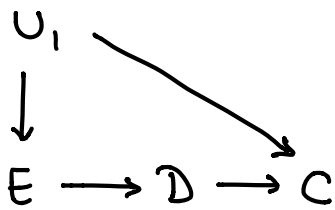
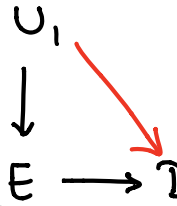
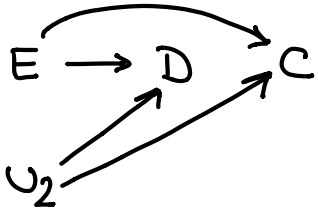
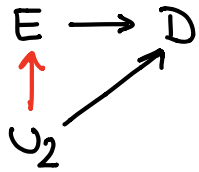
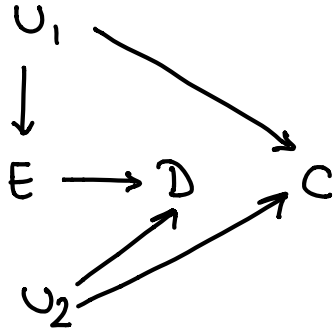
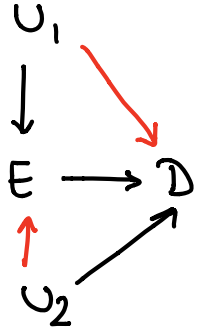
	$D = 1$	$D = 0$
$E = 1$	100	100
$E = 0$	200	200

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

	$C = 1$		$C = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$E = 1$	55	25	45	75
$E = 0$	70	10	130	190

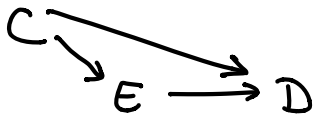
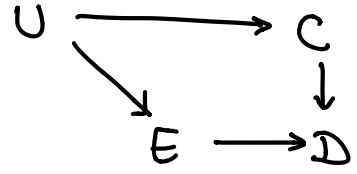
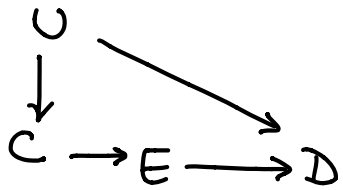
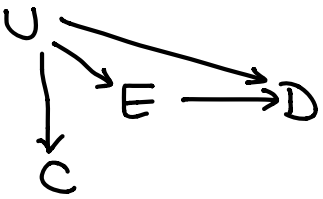
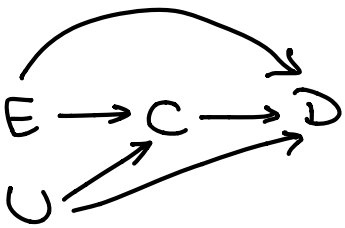
$RR_{ED|C=1}^{*} = 0.79$; $RR_{ED|C=0} = 0.92$ → spurious non-causal ass.

E and D statistically independent

	COLLIDER	CONFOUNDER	CONDITION ON.
② 	C	U_1	C: spurious ass. $U_1 \rightarrow D$ 
③ 	C	U_2	C: spurious ass. $U_1 \rightarrow E$ 
④ 	C	U_1, U_2	C: spurious ass. $U_1 \rightarrow$ {E, D} 

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

Association has a spurious component and the OR_{ED} is a biased estimator of the causal component.

		CONFOUNDER	CONDITION ON
⑤		C	C: OR_{EDIC}. If D is rare $\simeq RR_{EDIC}$ Valid estimator of the caused effect E→D
⑥		U: affects D by changing the value of C	C: OR_{EDIC}. If D is rare $\simeq RR_{EDIC}$ Valid estimator of the caused effect E→D
⑦		U: affects E by changing the value of C	C: OR_{EDIC}. If D is rare $\simeq RR_{EDIC}$ Valid estimator of the caused effect E→D
⑧		U C: surrogate confounder	C: if the actual confounder cannot be measured, adjusting for a surrogate confounder is better than crude OR. If C is strongly correlated w/ U, it will remove most of the confounding: OR_{EDIC}
⑨		U	C: not a confounder → not on causal pathway.

Here we are defining confounding based on background knowledge of the causal structure, contrasting w/ strategies ① and ②: only stat. ass. ③ stat ass + partial a priori knowledge.

These strategies would hold for ①-⑧

Is C a confounder? Depends on nº of restrictions — we can't gauge this just from stat. ass.

⑨ NO → not on causal pathway

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

①-③ NO → affected by exposure or outcome

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

C not in causal pathway. All vbles perfectly measured.

CRUDE.

logit $P(D=1|E) = \beta_0 + \beta_1 E$

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

crude $OR_{ED} = 0.65$

STRATIFIED on C

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

C = 1

	D = 1	D = 0
E = 1	19	8
E = 0	100	46

C = 0

D = 1	D = 0
24	231
94	658

$OR_{EDIC} = 0.80 = e^{\beta_1}$

=> NO heterogeneity of OR
Assume NO interaction

which analysis more appropriate?

STRATEGY

① C is added if the p-value associated w β_2 < 0.10. Here it's p < 0.001 → select C

② OR_{EDIC} is 23% greater than OR_{ED} → select C. → Adjustment for C is appropriate.

③ Check { C is ass w/ E : $OR_{CE|D=0} = 0.5$
C is ass w/ D : $OR_{CD|E=0} = 15.22$
C not on causal pathway. } → select C.

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 VBLES: SUFFICIENT SET

- Delete arrows emerging from E or D
- Identify unblocked backdoor paths D → 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:

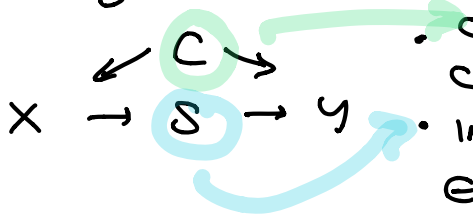
$A \rightarrow B$ A causes B and precedes B temporally.

CAUSAL MARKOV ASSUMPTION: COMPATIBILITY: if 2 vbles are d-separated 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, conditional on its direct parents (S).

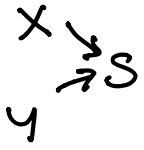
Assuming X precedes Y temporally; only 2 sources of association X-Y:



- CONFOUNDING EVENTS
  common causes (shared ancestors) = confounding
- INTERMEDIATE EVENTS
  Effects of X on Y: directed paths from $X \rightarrow Y$

If we {condition = hold constant} confounding & intermediate events interrupts 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 vbles are independent, they must be d-separated.

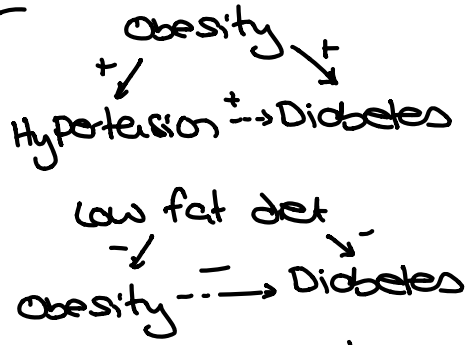
SUFFICIENT: set that controls bias in the $X \rightarrow Y$ association.

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

DIRECTION OF CONFOUNDING

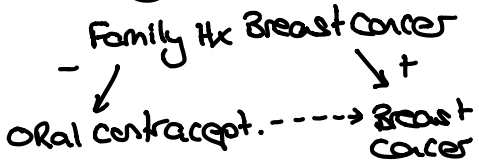
Depends on:

- Direction of ass. w/ D and E
  - Same direction w/ D and E → UPWARD BIAS Away from Ho

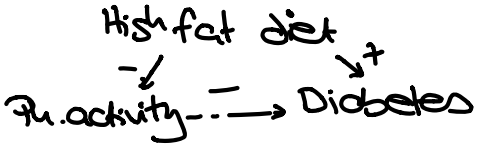


  - ≠ direction w/ D and E: varies

DOWNWARD BIAS : TOWARDS Ho.

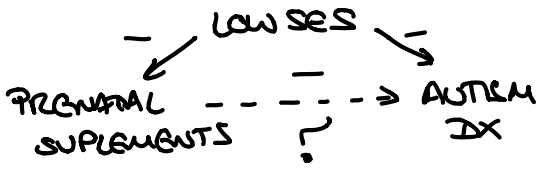


UPWARD BIAS : Away Ho



- Effect on outcome

UPWARD BIAS : TOWARDS Ho



- True ass. of E → outcome
  - ↑ risk : conditions that bias UPWARD, away from Ho are ?
  - ↓ risk : conditions that bias DOWNWARD, toward Ho are ?

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 = levels of the confounder
  Limitations: ↓ eligible ind, can't analyze/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.
  - ONE TO ONE ; - ONE TO MANY ; - FREQUENCY M.: = n° cases/controls for a category : multiple controls to multiple cases.
  - CALIPER M.: ↳ within a range (± 2 yrs) - PROPENSITY SCORE M.: one estimate of likelihood of receiving trt.
  Useful: case-controls, controlling for complex vbles
  Limitations: expensive, can't analyze/evaluate a factor that is matched on, pot. enrollment. sample size limits, you need to collect the vbles before pot. enrollment.

STUDY ANALYSIS

- **STRATIFICATION**: Clarifies degree of homogeneity. $OR_{mh} = \frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i}$ $RR_{mh} = \frac{\sum a_i(c_i+d_i)/n_i}{\sum c_i(a_i+b_i)/n_i}$
  How: 2x2 tables → weighted avg. of RR and OR across strata ⇒ Cochran-Mantel-Haenszel estimate
  Limitations: residual confounding if confounder not evenly distributed in E and D.
  feasible only for few vbles.

- **STANDARIZATION**: Reweight strata-spp rates so that Exp. categories are 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 cofounding vble.
IPTW creates a pseudo-pop without cofounding by re-weighting, yet the true exp. is not altered.

⊕ Can be used in time-dependent cofounding: wh earlier exp influence the cofounding effect of later exp.

Ex: ind. w/ Sx: early exp → behavior change → Alteration of future exp.

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

CRUDE

	D+	D−	TOTAL
E+	230	770	1000
E−	240	960	1200
	470	1730	2200

SMOKING C+

	D+	D−	TOTAL
E+	180	720	900
E−	40	360	400
	220	1080	1800

NON-SMOKING C−

	D+	D−	TOTAL
E+	50	50	100
E−	200	600	800
	250	650	900

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

	N	P(E=e)	P(E=e|C=c)	WEIGHT	PSEUDO-N
E^+C^+	900	$\frac{1000}{2200} = 0.45$	$\frac{900}{1800} = 0.69$	$\frac{0.45}{0.69} = 0.65$	$0.65 \cdot 900 = 530.85 \simeq 591$
E^-C^+	400	$\frac{1200}{2200} = 0.5454$	$\frac{400}{1800} = 0.30$	$\frac{0.54}{0.3} = 1.8$	$1.8 \cdot 400 = 709$
E^+C^-	100	0.45	$\frac{100}{900} = 0.1\hat{}$	$\frac{0.45}{0.1} = 4.09$	$4.09 \cdot 100 = 409$
E^-C^-	800	0.5454	$\frac{800}{900} = 0.89$	$\frac{0.5454}{0.89} = 0.606$	$0.606 \cdot 800 = 490.38 \simeq 491$

NON-CONFOUNDED TABLE

	D+	D−	TOTAL
E+			
E−			
			2200

⟸

SMOKING C+

	D+	D−	TOTAL
E+	0.65× 180=118	0.65× 720=473	591
E−	1.8×40 =72	1.8×360 =648	709
	190		1800

NON-SMOKING C−

	D+	D−	TOTAL
E+	4.09×	4.09×	409
E−	0.6×	0.6×	491
			900

- SELECTION BIAS : non-comparability of E^+ and E^- induced by restriction the analysis a certain level/s of a common effect of E and D or vbles corr w/ E & D.

≠ confounding due to unmeasured common cause

Pop.	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^-	$\frac{c}{C} = f_c$	$\frac{d}{D} = f_d$

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

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

Example :

Pop.	D^+	D^-
E^+	1000	1 600 000
E^-	5000	1 200 000

Study sample	D^+	D^-
E^+	100	320
E^-	500	240
	10% sample	0.02% sample

NO SELECTION BIAS because % of cases and non-c is = for E^+ and E^- (balanced) = 1

$$OR = \frac{1000 \times 1\,200\,000}{5000 \times 1\,600\,000} = 0.15 \qquad OR = \frac{100 \times 240}{500 \times 320} = 0.15 \qquad \frac{f_a \cdot f_d}{f_c \cdot f_b} = \frac{0.1 \cdot 0.02}{0.1 \cdot 0.02} = 1$$

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

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

SELECTION BIAS

Study sample	D^+	D^-	
E^+	100	320	0.02% = f_b of sample
E^-	500	480	0.04% = f_d of sample
	10% sample		

unbalanced sampling fractions :

$$OR = \frac{100 \times 480}{500 \times 320} = 0.3$$

$$\frac{f_a f_d}{f_b f_c} = \frac{0.1 \cdot 0.04}{0.1 \cdot 0.02} = 2 \neq 1$$

INTERACTION

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

2 exposures / risk factors act in:

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

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

RELATIVE RISK = RR

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

B \ 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 vbles. What is the combined effect?

B \ A	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^- Unexp	A^{LOW} Res Exp	A^{HIGH} Smelterworker
B^-: Nonsmokers	1	2.3	8.4
B^+: Smokers	8.3	17.5	26.2

	A^{LOW} Res Exp	A^{HIGH} Smelterworker
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 $2.3 \simeq 2.1$	YES: ANT. $8.4 \gg 3.15$
MULTIPLICATIVE INTERACTION	$8.3 \times 2.3 \simeq 17.5$	$8.3 \times 8.4 \simeq 64 \neq 26^2$

	A^{LOW} Res Exp	A^{HIGH} Smelterworker
B^-: Nonsmokers	$2.3 - 1 = 1.3$	$8.4 - 1 = 7.4$
B^+: Smokers	$17.5 - 8.3 = 9.2$	$26.2 - 8.3 = 17.9$
EFFECT MOD of EXCESS RR?	YES: SYN $1.3 << 9.2$	YES: SYN $7.4 << 17.9$
ADDITIVE INTERACTION	$1.3 + 7.3 = 8.6 \neq 17.5 - 1 = 16.5$	$7.4 + 16.5 = 23.9 \simeq 26.2 - 1 = 25.2$

INDEPENDENTLY ACTING FACTORS

EXAMPLE

ADDITIVE EFFECT:

Background risk (⊖A, ⊖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 ⊖A = 3 + 12 = 15
⊕A = 9 + 12 = 21

		Incidence Rates		Attributable Risks	
		Factor A		Factor A	
		−	+	−	+
Factor B	−	3.0	9.0	0	6
	+	15.0	21.0	12	18

combined risk A⊕ B⊕ = 21
background risk = 3
18

+ Attributable risk A = 6
Attributable risk B = 12
18

MULTIPLICATIVE EFFECT

Background risk (⊖A, ⊖B) = 3

- A⊕ : 9/3 = 3 triples risk

to both { B⊖ 3 × 3 = 9
B⊕ 15 × 3 = 45

- B⊕ : 15/3 = 5 → ×5 risk

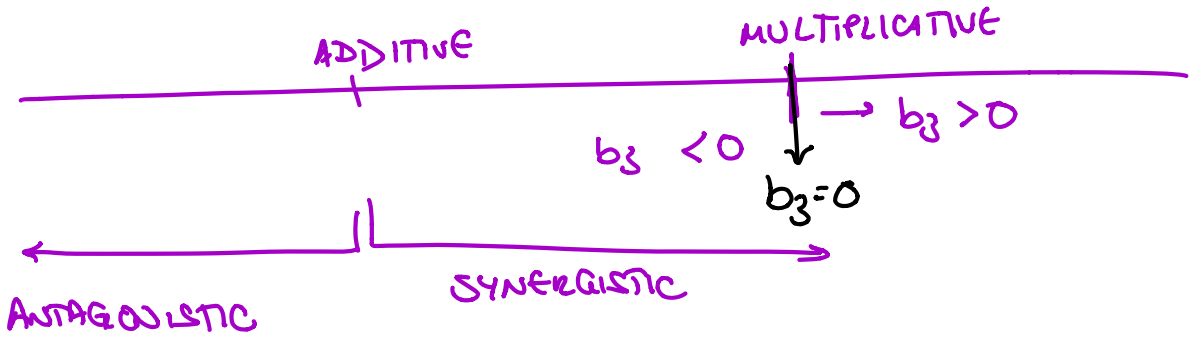
to both { A⊖ 3 × 5 = 15
A⊕ 9 × 5 = 45

		Incidence Rates		Relative Risks	
		Factor A		Factor A	
		−	+	−	+
Factor B	−	3.0	9.0	1	3
	+	15.0	45.0	5	15

45/3 = 15

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

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



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

Ex: Is A an effect modifier of B?
Effect of B^+ vs. B^-

- In the presence of $x_1 = 1$

A { $B^+ x_2 = 1$ — $\log ODDS = \beta_0 + \beta_1 + \beta_2 + \beta_3$
$B^- x_2 = 0$ — $\log ODDS = \beta_0 + \beta_1$

$\beta_2 + \beta_3$

- without $x_1 = 0$

A { $B^+ x_2 = 1$ — $\log ODDS = \beta_0 +$ β_2
$B^- x_2 = 0$ — $\log ODDS = \beta_0$

β_2

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

MARGINAL AND CONDITIONAL ASSOCIATION

	D^+	D^-
E^+		
E^-		

RR = 1 NO association

MARGINALLY INDEPENDENT

C = 1

	D^+	D^-
E^+		
E^-		

C = 0

	D^+	D^-
E^+		
E^-		

RR = 2 RR = 2

CONDITIONALLY ASSOCIATED / DEPENDENT

	D^+	D^-
E^+		
E^-		

RR = 2 Association = unblocked path

MARGINALLY ASSOCIATED

C = 1

	D^+	D^-
E^+		
E^-		

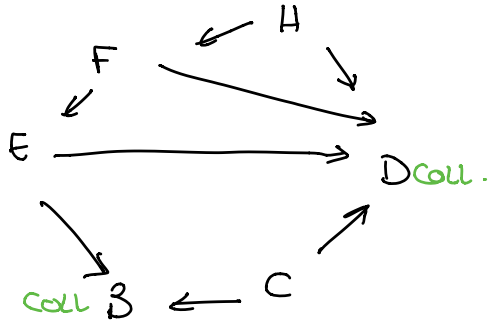
C = 0

	D^+	D^-
E^+		
E^-		

RR = 0.5 RR = 2

EFFECT MEASURE MODIFICATION

Ex:



marginally
H . ass. w
everything except C

C marg. indp. of everything except B and D.

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

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^-
D^+	A	B
D^-	C	D

study pop	D^+	D^-
E^+	a	b
E^-	c	d

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

B: " " of D^+ that were E^-

C: " " D^- E^+

D: " " D^- that were E^-

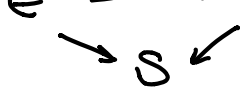
ORDIE using sampling fractions → assess if selection bias is present.

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

$$OR = \frac{\frac{a}{A}}{\frac{c}{C}} = \frac{\frac{b}{B}}{\frac{d}{D}} = \frac{\frac{0.05}{0.01}}{\frac{1}{0.2}} = 1$$ for example

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

SELECTION BIAS DAG



DIRECTION OF CONFOUNDING.