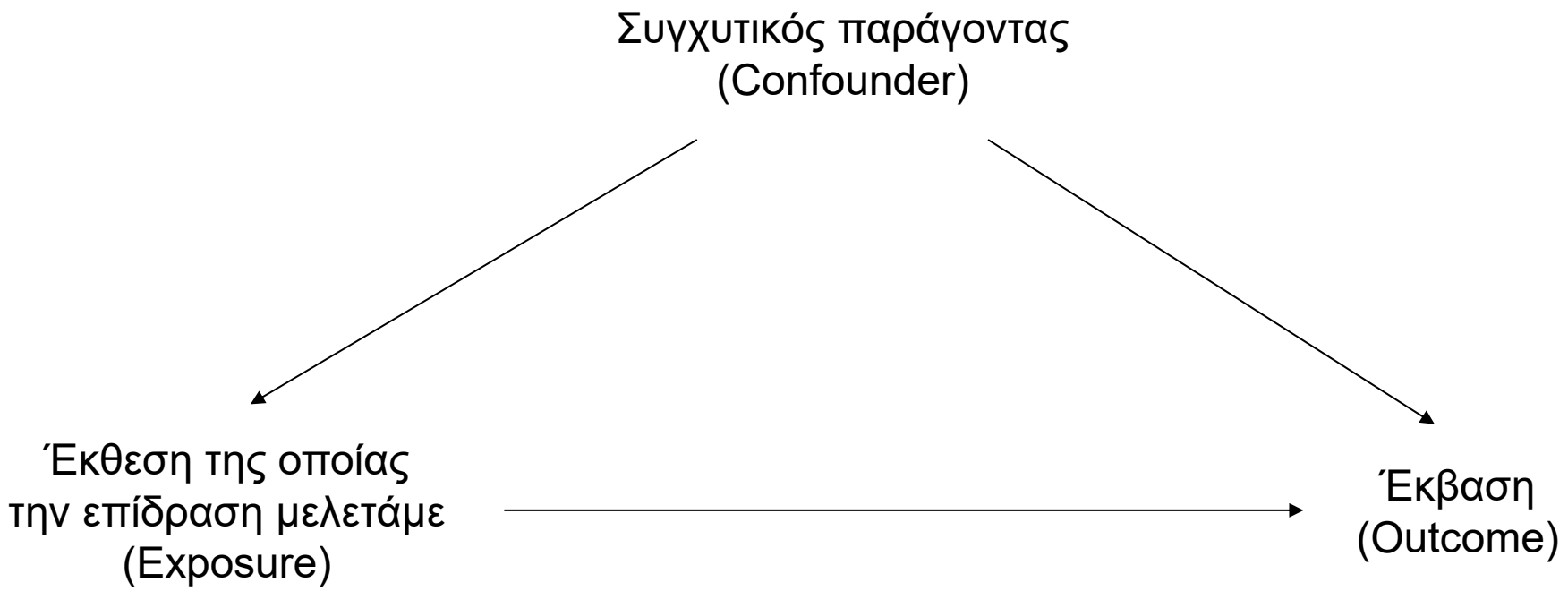
Στατιστικές Μέθοδοι στην Επιδημιολογία

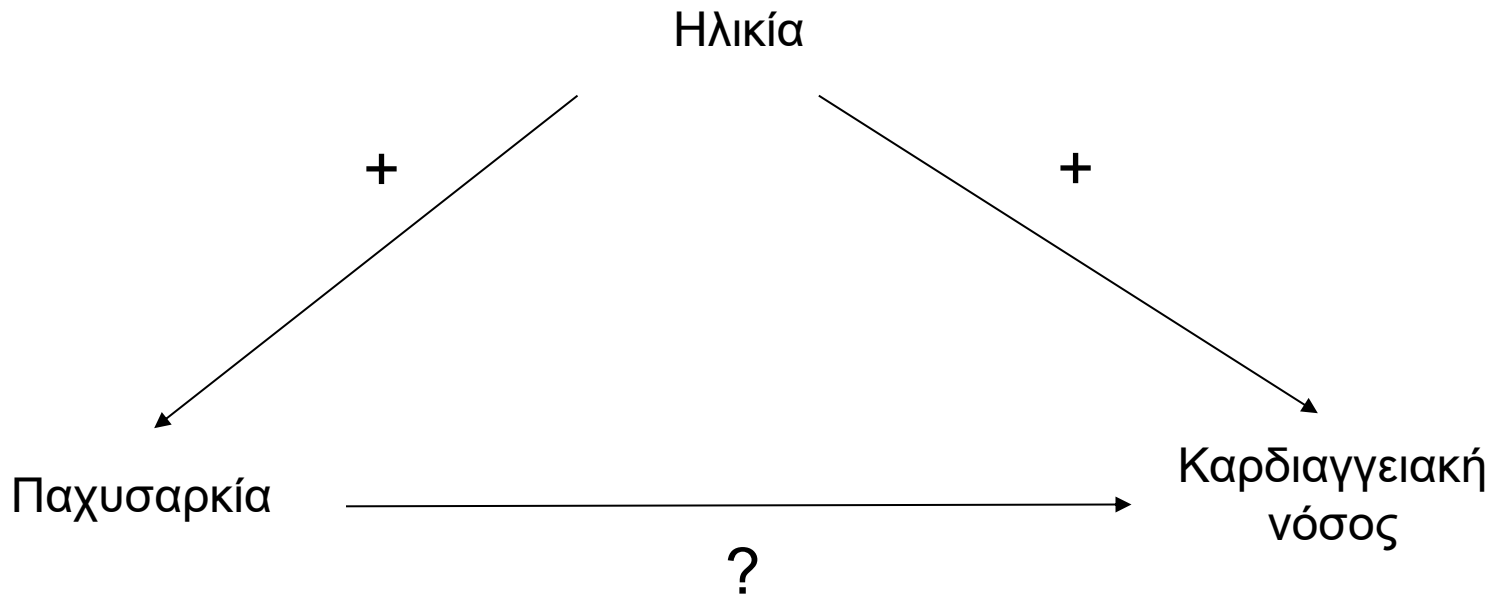
Επαναληπτικό μάθημα

20/01/2026

Chapter I
Epidemiologic Research

- Επιδημιολογία: Η μελέτη της κατανομής των νοσημάτων ή των συμβάντων και των παραγόντων κινδύνου για αυτά, σε καθορισμένους πληθυσμούς και η εφαρμογή της μελέτης αυτής για τον έλεγχο των προβλημάτων υγείας
- Μεροληψία: Συστηματικό Σφάλμα (π.χ selection bias)
- Συγχυτικοί παράγοντες: παράγοντες που επηρεάζουν την έκβαση και σχετίζονται και με την έκθεση στον παράγοντα του οποίου την επίδραση μελετάμε.
- Μοναδική μέθοδος εξάλειψης της επίδρασης Συγχυτικών παραγόντων: Τυχαιοποίηση
- Αντιμετώπιση στο στάδιο της ανάλυσης: Μόνο δεδομένης της γνώσης όλων των Συγχυτικών παραγόντων και της καταγραφής των αντίστοιχων δεδομένων


Συγχυτικός παράγοντας
(Confounder)
Έκθεση της οποίας
την επίδραση μελετάμε
(Exposure)
Έκβαση
(Outcome)



All individuals

Obese	46	254	300
Not Obese	60	640	700
Total	106	894	1000

Risk ratio=(46/300)/(60/700)=1.79

	<50		
	CVD	No CVD	Total
Obese	10	90	100
Not Obese	35	465	500
Total	45	555	600

	>50		
	CVD	No CVD	Total
Obese	36	164	200
Not Obese	25	175	200
Total	61	339	400

<50 years old:
Risk ratio=(10/100)/(35/500)=1.43

>50 years old:
Risk ratio=(36/200)/(25/200)=1.44

Η ηλικία σχετίζεται θετικά με την παχυσαρκία (άτομα >50 ετών είναι πιθανότερο να είναι παχύσαρκα συγκριτικά με άτομα <50 ετών).

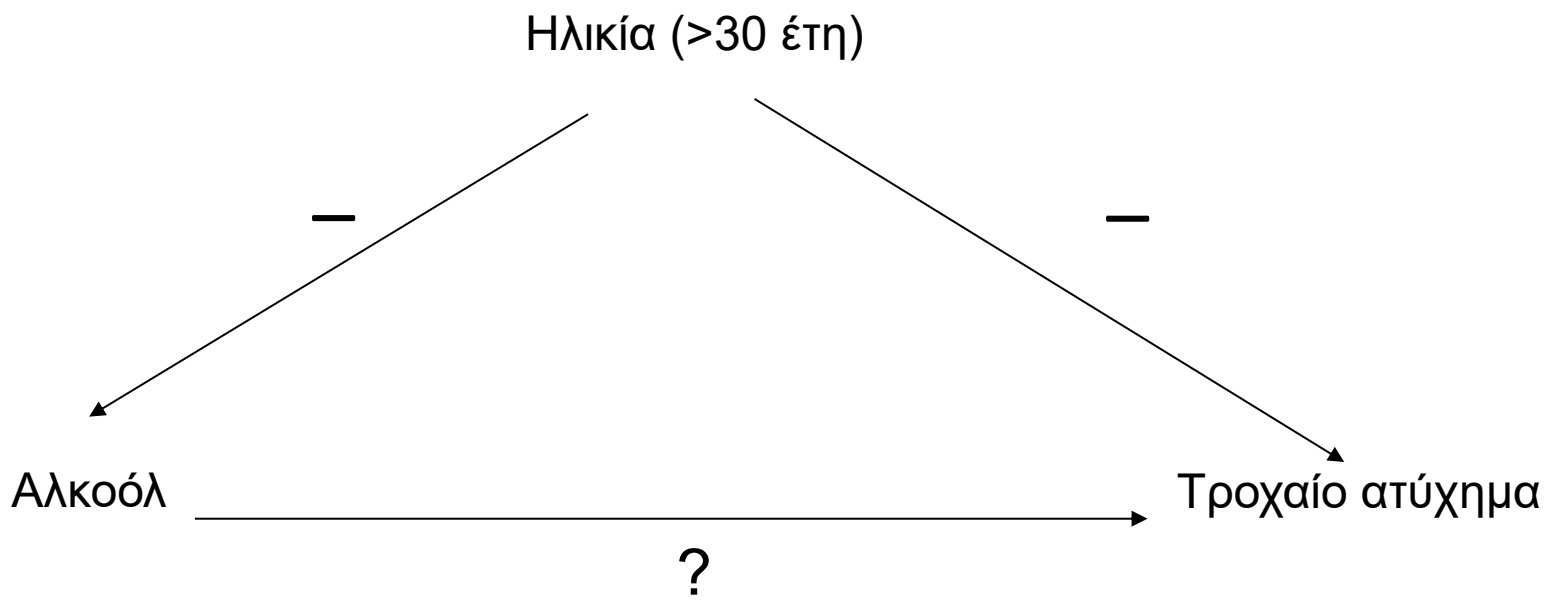
Άρα στην ομάδα των παχύσαρκων έχουμε περισσότερα άτομα >50 ετών.

Επομένως, ακόμα και στην περίπτωση που η παχυσαρκία δεν έχει καμία επίδραση στον κίνδυνο καρδιαγγειακής νόσου, αν αγνοήσουμε την ηλικία στην ανάλυσή μας, αυτό που τελικά θα δούμε είναι σχέση προς την ίδια κατεύθυνση με αυτή της ηλικίας,

Δηλαδή αύξηση του κινδύνου για καρδιαγγειακή νόσου.

Εφόσον και η παχυσαρκία αυξάνει τον κίνδυνο, προσθέτοντας και την επίδραση της ηλικίας θα έχουμε υπερεκτίμηση της επίδρασης (απομάκρυνση από την μηδενική).

Διερευνούμε τη σχέση αλκοόλ και τροχαίου ατυχήματος



Άτομα ηλικίας >30 ετών είναι λιγότερο πιθανό να καταναλώνουν αλκοόλ, επομένως στην ομάδα αυτών που καταναλώνουν αλκοόλ έχουμε λιγότερους >30 ετών.

Αγνοώντας την ηλικία στην ανάλυση, η εκτίμηση της επίδρασης του αλκοόλ θα κινηθεί σε κατεύθυνση αντίθετη της επίδρασης της ηλικίας.

Άρα, ακόμα και αν στην πραγματικότητα το αλκοόλ δεν έχει καμία επίδραση στον κίνδυνο του ατυχήματος, η ανάλυση θα μας δείξει επιβαρυντική επίδραση

Εφόσον το αλκοόλ έχει επιβαρυντική επίδραση, αγνοώντας την ηλικία απομακρυνόμαστε από την μηδενική και υπερεκτιμάμε την επιβαρυντική του επίδραση.

Relative measures of Exposure Effect

	Exposure	No exposure	
Disease	a	b	a+b
No Disease	c	d	c+d
	a+c py1	b+d py2	N

- Risk Ratio
- Rate Ratio
- Odds Ratio

Έλεγχοι Υποθέσεων

- Test ομοιογένειας για τα rates - έλεγχος X^2

- H_0: όλα τα rates ίσα

  H_1: κάποιο/α διαφορετικά

- Ελέγχουμε για γραμμική τάση, ουσιαστικά ελέγχουμε αν το μέσο RR διαφέρει σημαντικά από τη μονάδα

- Αν υπάρχει trend (αν δηλαδή απορριφθεί η μηδενική υπόθεση) → μπορούμε να υποθέσουμε όμοια επίδραση για τη μεταβολή από οποιαδήποτε κατηγορία στην επόμενη

Confounding *vs.* Interaction

Estimate the effect of exposure of interest within levels of suspected confounder

Test for heterogeneity/effect modifier

- H0: True Rate Ratio $RR_i=RR_j$
- H1: True Rate Ratio $RR_i \neq RR_j$ για τουλάχιστον ένα i/j

- If effect of exposure same across levels of suspected confounder → confounder
  - *Present overall OR (e.g., Mantel Haenszel , adjusted for the confounder)*

- If effect of exposure different among levels of suspected confounder → effect modifier (Interaction)
  - *Present separate OR for each stratum*

- *Not adjusting for a confounding variable can result either on a more intense effect or a less intense effect of the exposure.*

Chapter VIII
Matched Case-Control Studies

- To control for confounders
- Analyze using Conditional Logistic Regression
- β_0 is meaningless
- Main effects of matching not estimated, BUT their interaction with other variables CAN be studied
  - If matching variable confounder → gain efficiency
  - If matching variable associate with exposure only → Overmatching loss of efficiency

Chapter X
Choice of Controls in Case-Control Studies

- What is your scientific question?
- What type of disease you study?
- What is the exposure of interest?

- What relative measure of disease you need to estimate?
- Which group of controls will be suitable to provide you with least biased estimates?

- **<u>Exclusive sampling</u>**: Controls are selected from those people in the population who are disease-free in the end of the study period. We have seen that this design can provide an estimate of the **odds ratio of disease**.

➢ *Παίρνουμε μάρτυρες (controls) μόνο από όσους **δεν** είχαν αναπτύξει τη νόσο μέχρι το τέλος της περιόδου.*

- **<u>Inclusive sampling</u>**: Controls selected among those who were initially (in the beginning of the study) at risk, no matter if they subsequently developed the disease. The odds ratio from the study can be shown to estimate the **risk of disease**.

➢ *Παίρνουμε μάρτυρες (controls) από τον πληθυσμό που ήταν στην αρχή at risk (δηλαδή όλη η κοόρτη), χωρίς να αποκλείσω αυτούς που αργότερα έγιναν cases.*

- **<u>Concurrent sampling</u>**: In this design the controls are selected simultaneously with the cases when (at the same “time of diagnosis”) the latter occur, from the group of people who are still at risk of the disease at the time of diagnosis of the case. Provides estimate of the **rate ratio**.

- Για κάθε case στον χρόνο t, επιλέγουμε μάρτυρες (controls) από το “risk set” στον ίδιο χρόνο t, δηλαδή όσοι δεν έχουν νοσήσει ακόμη και παραμένουν υπό παρακολούθηση. Ένα άτομο μπορεί να γίνει control σε νωρίτερο χρόνο και αργότερα να γίνει case.

Multiplicative interactions

	Exposure 1: E_1^-	Exposure 1: E_1^+
Exposure 2: E_2^-	e^{a}	$e^{a}e^{\beta_1}$
Exposure 2: E_2^+	$e^{a}e^{\beta_2}$	$e^{a}e^{\beta_1}e^{\beta_2}e^{\beta_3}$

Fit interaction term between E_1&E_2

$$\textbf{Odds (disease)} = e^{(a)}e^{(\beta_1 X_1)}e^{(\beta_2 X_2)}e^{(\beta_3 X_1 X_2)}$$

Testing: the "effect" of **E_1^+&E_2^+** is NOT multiplicative, because of term e^{β_3}

e^{β_3} = Multiplicativity index

Can be null (e^{β_3}=1 i.e, β_3=0) and thus the "effect" of **E_1^+& E_2^+** is multiplicative

Can be sub-multiplicative (e^{β_3}<1 i.e, β_3<0)

Can be super-multiplicative (e^{β_3}>1 i.e, β_3>0)

Evaluating additive interaction through relative risks

		Exposure A	
		No(A=0)	Yes (A=1)
Exposure B	No(B=0)	$RR_{00}=p_{00}/p_{00}=1$	$RR_{10}=p_{10}/p_{00}=1.875$
	Yes (B=1)	$RR_{01}=p_{01}/p_{00}=2.00$	$RR_{11}=p_{11}/p_{00}=3.125$

• But usually we do not have estimates of incidence or prevalence rates but relative risks such as rate ratios or odds ratios.

Previous equation can be re-expressed as :

$RR_{11}- RR_{10}- RR_{01}+ RR_{00}= RR_{11}- RR_{10}- RR_{01}+ 1$

•Relative Excess Risk due to Interaction (RERI)

RERI = (RR11-1)-[(RR10-1)+ (RR01-1)]=3.125 - 2 - 1.875 + 1 = 0.25 > 0

•Synergy index S:

$$S = (RR_{11}-1)/(RR_{10}+ RR_{01}-2) = 2.125/1.875 =1.13 > 1$$

Indices for additive interaction

RERI & S

RERI>0 ⟺ S>1 (super-additivity – additive interaction)

RERI=0 ⟺ S=1 (additive interaction)

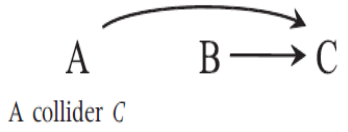
RERI<0 ⟺ S<1 (sub-additive interaction)

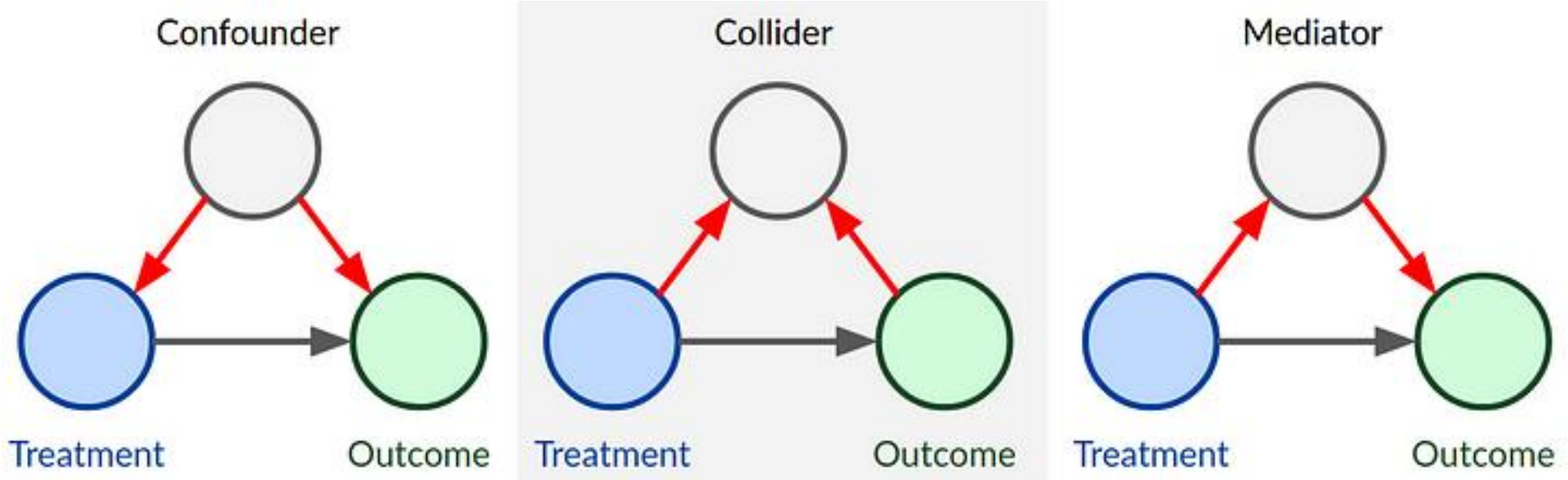
INTERPRETATION

RERI: The actual amount (difference) of excess relative risk for disease D due to the combined presence of factors A and B that is added to or substracted from the sum of excess relative risk for disease D due to factor A only and factor B only (acting separately)

S: the excess relative risk for disease D caused by the combined presence of factors A and B, relative to the sum of excess relative risk for disease D due to factor A only and factor B only (acting separately)

Collider = Common effect





The common effect C is referred to as a collider because two arrowheads collide into it.

Example: A: genetic factor, B:diabetes, C: heart disease.

Propensity score: Definition

- **The propensity score is the probability of treatment assignment (A) conditional on observed baseline characteristics (X).**

- It allows to design and analyze an observational (nonrandomized) study so that it **mimics** some of the particular characteristics of **a randomized controlled trial**.

Why use a propensity score method

- To reduce or eliminate the effects of confounding when using observational data to estimate treatment effects.

- It allows the estimation of the *marginal* rather than the *conditional* treatment effect

So, how does it works?

- It is a balancing score
  - This means that, conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated subjects.

Step 1: Estimation of the Propensity Score

- The propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics.

- The estimated propensity score is the predicted probability of treatment derived from the fitted regression model.

Step 2: Balancing

- The reason we estimated the propensity score, was to use it in a way to provide balance of baseline characteristics between treated and untreated individuals

- This can be done in more than one ways

Propensity score methods

- 1. Matching on the propensity score
- 2. Stratification on the propensity score
- 3. Inverse probability of treatment weighting using the propensity score
- 4. Covariate adjustment using the propensity score.

1. Matching

- Propensity score matching entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score (Rosenbaum & Rubin, 1983a, 1985).

- Once a matched sample has been formed, the treatment effect can be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample.
  - Reporting of treatment effects can be done in same metrics as are commonly used in RCTs.

2. Stratification

- The process of dividing the study population into subgroups based on the estimated propensity scores.

- The idea is to divide the study population into strata, or subgroups, with similar estimated propensity scores within each stratum. By doing so, the treated and control groups within each stratum are more likely to be balanced in terms of the baseline covariates.

- A common approach is to divide subjects into five equal-size groups using the quintiles of the estimated propensity score.

3. Inverse Probability Weighting

- Inverse probability of treatment weighting (IPTW) using the propensity score uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment.

- A subject's weight is equal to the inverse of the probability of receiving the treatment that the subject actually received.

4. Covariate Adjustment

- Using this approach, the outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score.

- this method assumes that the nature of the relationship between the propensity score and the outcome has been correctly modeled.

Step 3: Checking Balancing

Quantitative balance checks:

Standardized mean differences (SMD) and variance ratios (VR) can be examined to see how much imbalance is in each covariate.

Πληθυσμιακές μελέτες

- Μελέτες που αφορούν μεγάλες πληθυσμιακές ομάδες
- Συγχρονικές μελέτες – τυχαίο δείγμα συνήθως του γενικού πληθυσμού
- Στόχοι:
  - Προσδιορισμός επιπολασμού νοσημάτων και παραγόντων κινδύνου
  - Καταγραφή της συχνότητας χρήσης υπηρεσιών υγείας

Σχεδιασμός πληθυσμιακών μελετών

- Ορισμός πληθυσμού αναφοράς
- Ορισμός πληθυσμού από τον οποίο θα προέλθει το δείγμα
- Μέθοδος δειγματοληψίας
- Μέγεθος δείγματος

Είδη δειγματοληπτικών μεθόδων

- Δειγματοληψία μη βασιζόμενη σε πιθανότητα
  - Δείγμα ευκολίας (convenience sampling)
  - Δειγματοληψία οδηγούμενη από τους συμμετέχοντες (respondent driven sampling)

- Μεροληπτικές μέθοδοι – πολλές φορές δεν υπάρχει ωστόσο εναλλακτική

- Υπάρχουν ειδικές μέθοδοι ανάλυσης για convenience sampling και RDS ώστε να οδηγούν σε αμερόληπτα αποτελέσματα

Δειγματοληψία βάσει πιθανότητας

- Κάθε δειγματοληπτική μονάδα έχει συγκεκριμένη μη μηδενική πιθανότητα να επιλεγεί

- Είδη:
  - Απλή τυχαία δειγματοληψία
  - Συστηματική τυχαία δειγματοληψία
  - Στρωματοποιημένη τυχαία δειγματοληψία
  - Κατά συστάδες δειγματοληψία
  - Πολυσταδιακή τυχαία δειγματοληψία
  - Δειγματοληψία σε πολλαπλές φάσεις

Δειγματοληπτικά βάρη

- Πολλές φορές οι μελέτες σχεδιάζονται έτσι ώστε να παίρνουν μεγαλύτερα δείγματα από συγκεκριμένες υποομάδες που έχουν ενδιαφέρον από πλευράς δημόσιας υγείας (oversampling)

- Τα δειγματοληπτικά βάρη αποδίδονται σε κάθε συμμετέχοντα → αντιστοιχούν στον αριθμό των ατόμων που αντιπροσωπεύει και αντανακλούν την άνιση πιθανότητα επιλογής.

→Αμερόληπτες και αντιπροσωπευτικές εκτιμήσεις του γενικού πληθυσμού

Post-stratification weights

- Ορισμένες φορές η κατανομή του δείγματος κατά ηλικία και φύλο διαφέρει από την αντίστοιχη του πληθυσμού

- Κατασκευή των post-stratification weights με την χρήση πληροφορίας από δεδομένα απογραφής ώστε η κατανομή του δείγματος να είναι όμοια του πληθυσμού

Ανάλυση STATA

- svyset psu [pweight=weight], strata(strata)