Belangrijkste begrippen Toegepaste Statistiek

BASISBEGRIPPEN

Centrummaten:

- modus (de meest voorkomende uitkomst, voor variabelen gemeten op ieder niveau)
- mediaan (de middelste uitkomst, of gemiddelde van de middelste twee bij even n, minimaal ordinaal meetniveau)
- gemiddelde (alle waarden opgeteld gedeeld door n waarnemingen, minimaal intervalniveau)

Spreidingsmaten:

- (Standaard)deviatie en variantie
  - De deviatie d^2 van één uitkomst x is het kwadraat van het verschil tussen die specifieke uitkomst (x) en het gemiddelde van alle uitkomsten ($\bar{x}$): $d^2 = (x - \bar{x})^2$. Je centreert je variabele.
  - De variantie s^2 van een serie uitkomsten is het gemiddelde van de deviaties d^2 van die uitkomsten (zie formularium)
  - De standaarddeviatie s is de wortel van de variantie (dus $\sqrt{s^2}$)
- Range (verschil tussen grootste en kleinste uitkomst: maximum - minimum)
- Interkwartielafstand (verschil tussen Q3 en Q1 - hoogte van de doos in een boxplot)
- Variantiecoëfficient (of relatieve standaardafwijking): de standaarddeviatie s gedeeld door de absolute waarde van het gemiddelde $|\bar{x}|$ (ofwel σ gedeeld door μ). Soms CV * 100 of in %.

Frequentietabel en staafdiagram: voor nominale en ordinale variabelen, registreert scores, frequenties, percentages en cumulatieve percentages.
Histogram: bij interval/ratio variabelen, aantal klassen is $\sqrt{n}$, klassenbreedte is range / aantal klas.
Frequentiedichtheid = frequentie / klassenbreedte (gebruik bij ongelijke klassebreedte)
Tak-blad-grafiek: metingen altijd met 1 decimaal, blad = het decimaalgetal = 1 meting.
Boxplot: wordt bepaald door 5 getallen; Q1 en Q3 bepalen je doos, Q2 is de mediaan daarbinnen, min en max zijn streepjes daarboven en daaronder en de outliers/ uitbijters zijn punten daarbuiten met een uitzonderlijk hoge of lage score. Elk kwartiel bevat 25% van de uitkomsten.

KANSBEREKENING EN STATISTISCHE TOETSEN

Standaard normaalverdeling/Gaussverdeling (symmetrisch, modus=mediaan=gemiddelde): μ (gemiddelde van de distributie in een populatie) is 0 en sigma σ (de standaarddeviatie in een populatie) is 1: ±68% waardes liggen tussen μ-σ en μ+σ, en ±95% waardes tussen μ-2σ en μ+2σ.

Centrale limietstelling: steekproefgemiddelden zijn normaal verdeeld (bij groot genoege n) want ze hebben minder variabiliteit dan de oorspronkelijke populatie. μ in de steekproef is hetzelfde als μ in de populatie MAAR variantie en standaarddeviatie zijn dat niet: in de populatie is de variantie σ^2_p en de standaarddeviatie σ_p, maar in de steekproef is de variantie σ^2/n en de standaarddeviatie $\sigma/\sqrt{n}$.

De toetsingsgrootheid z-score is een transformatie van de standaarddeviatie (bij voldoende grote n is de z-score altijd normaal verdeeld met gemiddelde 0 en variantie 1). De absolute waarde van z ligt bijna altijd tussen 0 en 2. Als hypothese H_0 geldt en de centrale limietstelling geldt dan geldt: $z = \bar{x} - \mu_0 / \sigma$ [waarbij σ is $\sigma/\sqrt{n}$ want je gebruikt de standaarddeviatie van de steekproef]. Vervolgens toets je: is z te groot, en dat doe je door de P-waarde te berekenen. De kans P dat een waarde gegeven μ,σ voorkomt is te berekenen als een oppervlakte onder de curve van de normaalverdeling. Beslisregel: alpha α is significantieniveau. Kans van de eerste soort = H_0 verwerpen terwijl H_0 waar is, die wil je zo klein mogelijk maken. Vaak α = 0.05 (5%). Als P<α: verwerp H_0. H_0 simpel gezegd: er is niets aan de hand/ er is geen verschil (bij vergelijken van populaties)/ er is geen correlatie. H_A: er is iets aan de hand/ er is wel een verschil/ er is wel een correlatie. Aannames voor de Z-test: populatieverdeling waarnemingen normaal verdeeld OF n >25 EN waarnemingen onafhankelijk EN **σ in de populatie is bekend**.

Als μ en σ **onbekend** zijn dan gebruik je een andere toetsingsgrootheid: de t-toets. Je gebruikt dan niet standaarddeviatie σ uit de populatie maar standaarddeviatie S uit de steekproef.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

n-1 = aantal vrijheidsgraden, degrees of freedom, df). Zie formularium voor t-toets.
Aannames voor de t-toets: populatieverdeling van de waarnemingen is normaal OF steekproef n>25 EN waarnemingen zijn onafhankelijk. NB: hoe groter de steekproef hoe meer vrijheidsgraden, hoe meer S gelijk is aan σ en hoe meer de t-score op een z-score lijkt.

Een P-waarde is een **kans**. 3 stappen:
- veronderstel H_0 geldt
- bereken |t|
- de P-waarde is de kans om iets extremers dan t te observeren

Betrouwbaarheidsinterval: de kans dat μ hierin ligt is nul of een, maar met dit interval weet je dat in 95% van de gevallen de waarde μ in je interval zal liggen (bij α=5%). Ook hier geldt de aanname: populatieverdeling waarnemingen normaal OF n >25 EN waarnemingen zijn onafhankelijk. Dit interval geeft je dus meer informatie dan een puntschatting want je krijgt een heel interval van betrouwbare waarden: het zegt iets over hoe groot de range van een uitkomst of een verschil is. NB: de T-waarde met je specifieke α en n zoek je op in een tabel (zie Toledo).

Bij het vergelijken van populaties gebruik je twee steekproeven, daarvan zijn 2 soorten:
A onafhankelijke of niet-gepaarde steekproef (bv vergelijken experimentele groep en controlegroep die niets met elkaar te maken hebben)
B afhankelijke of gepaarde steekproef (bv bij longitudinaal onderzoek met meting voor en na, de metingen zijn hier gecorreleerd en dus niet onafhankelijk)

Bij B bereken je het verschil d tussen de meting voor en na en dat gebruik je voor een gepaarde t-toets (je brengt het probleem terug tot één steekproef en gaat uit van H_0: $\mu_d = 0$ en een onbekende σ_d). Zie formularium voor formule gepaarde t-toets, je gebruikt de gemiddelde d van alle verschillen. Voorwaarde: d moet normaal verdeeld zijn.

Bij A heb je twee populaties (met ieder een eigen μ en σ) en twee onafhankelijke steekproeven, en ga je uit van H_0: $\mu_1=\mu_2$, ofwel: $\mu_1-\mu_2 = 0$. Ook daarvoor ga je het verschil berekenen: $\Delta = \bar{x}_1 - \bar{x}_2$. Omdat $\bar{x}_1$ en $\bar{x}_2$ normaal verdeeld zijn, ga je ervanuit dat Δ ook normaal verdeeld is. Als je σ voor de twee populaties niet kent gebruik je de S uit de steekproeven. Vervolgens kijk je of σ danwel S gelijk is in de twee steekproeven.
- Zo ja dan gebruik je een Two Sample t-toets voor pooled/equal variance S_p. De t-scores zijn niet normaal verdeeld, maar bij benadering $\gamma = n_1 + n_2$ - twee vrijheidsgraden.
- Zo nee dan gebruik je een Two Sample t-toets voor unequal variance.

Hoe weet je of σ hetzelfde is in de steekproeven? Je gebruikt Hartley's of Levene's test, daar komt een F-waarde uit en een Pr(>F) die bepaalt of je equal of unequal variance moet gebruiken (en die je afleest zoals een gewone P-waarde, dus P>0.05 betekent: geen evidentie dat de varianties niet gelijk zijn, je gebruikt dan de tabel voor Equal variance. Hoe groot het verschil is tussen de populaties haal je uit het betrouwbaarheidsinterval (het verschil ligt 95% zeker ergens tussen die twee waarden).

Wat doe je als de kansverdeling van de uitkomstvariabele **niet normaal** is en/of de **steekproefgrootte te klein**? Dan moet je een verdelingsvrije/ niet parametrische toets gebruiken (want er zijn geen parameters zoals μ of σ die de kansverdeling karakteriseren). De toetsen kunnen altijd worden toegepast (ook als verdeling wel normaal is).
- Bij onafhankelijke groepen: Wilcoxon Rangsom (Mann-Whitney) toets (rank sum). Equivalent t-toets. Aanname: scores zijn onafhankelijk en meetniveau tenminste ordinaal.
- Bij afhankelijke/gecorreleerde steekproeven: Wilcoxon Rangteken toets (signed rank). Equivalent gepaarde t-toets.

SAMENHANG VAN VARIABELEN
Drie maten voor de sterkte van de samenhang bij continue variabelen:
- Covariantie: Gemeenschappelijke variantie van 2 variabelen x en y. Gevoelig voor verschillen in schaal. Zie formularium voor de formule.
- Pearson correlatiecoëfficient r: Een standaardisatie van de covariantie en maat voor lineaire samenhang voor twee continue variabelen. Geeft richting en sterkte aan van het verband, bij kleiner dan 0 een negatief lineaire correlatie, bij groter dan 0 een positief lineaire correlatie. Altijd tussen -1 en +1, bij 0 geen lineaire samenhang. r is een schatting van de correlatiecoëfficient ρ

in de populatie. Toetsen van Pearson (en dus van de hypothese over de correlatie in de populatie) doe je ook met een t-test en een bijbehorende P-waarde, uitgaande van H_0: $\rho = 0$. Ook in dit geval geldt: als $P<0.05$ dan H_0 verwerpen, er is waarschijnlijk een correlatie aanwezig, ρ in de populatie is niet 0.
- Odds ratio, RR (zie Epidemiologie)

Som van de kwadraten (zie formularium): $SS_{Total} = SS_{Regression} + SS_{Error}$
SS_{Total}: Totale variatie in de waarnemingen/responsvariabele
SS_{Error}: Het niet door het regressiemodel verklaarde deel van de totale variatie i/d waarnemingen
$SS_{Regression}$: Het wel door het model verklaarde deel van de totale variatie
NB: Als het model perfect is dan zou dit alle variatie verklaren, dan is SS_{Error} nul.

De determinatiecoefficient r^2 (getal tussen 0 en 1, zie formularium) is een maat voor het deel van de variatie in de waarnemingen dat door het lineaire regressiemodel wordt verklaard. Hoe groter hoe beter. r^2 is numeriek gelijk aan het kwadraat van Pearson's r.

Lineaire regressie: lijn $Y = \beta_0 + \beta_1 x + \varepsilon$. Regressiemodel kiest een lijn die voor elke punt op de x-as het gemiddelde van y kiest. Daarbij komt het residu ε: de afwijking t/o het gemiddelde. De som van die afwijkingen moet dus zo klein mogelijk zijn om de lijn zo goed mogelijk bij de punten te laten passen. Parameters β_0 en β_1 ga je schatten met b_0 en b_1 uit de steekproef.
β_0 is de intercept: de geschatte waarde van y als x nul is (dus hoe hoog de lijn op de y-as begint bij x=0). β_1 is de hellingshoek van de lijn (dus hoe schuin de lijn gaat). Als de standaarddeviaties van x en y hetzelfde zijn ($s_x = s_y$) dan is de hellingshoek b_1 gelijk aan de correlatiecoëfficient $r_{x,y}$, en uiteraard is de hoek positief als de correlatie positief is (en andersom).
Beste lijn: $\hat{y} = b_0 + b_1 x$. Zie formularium, je hebt 5 elementen nodig om de lijn te schatten.
Let op: $\hat{y}$ geeft alleen maar een voorspelling voor de punten x binnen het bereik van je gegevens.
Als je lineaire regressie uitrekent krijg je standaard 2 tabellen: ANOVA en Coefficients: twee manieren om de relatie tussen x en y te toetsen.

ANOVA
- Total Sum Sq (de totale som van kwadraten, onderste regel) = SS_{Total}
- Regression Sum Sq (de variatie in y die verklaard wordt door je model) = $SS_{Regression}$
- Residuals Sum Sq (de variatie in y die het model niet kan verklaren) = SS_{Error}

[Regression Sum Sq] gedeeld door [Total Sum Sq] is r^2: het % in de respons dat de variatie in je model kan verklaren. De laatste 3 kolommen van deze tabel hoef je niet te kennen.

Coefficients
- Intercept (bovenste regel) * Estimate is de schatter b_0 die je gebruikt voor β_0
- **onderste regel** * Estimate is de schatter voor b_1 die je gebruikt voor β_1

In deze tabel zit **rechtsonder** de P-waarde (kans) die je gebruikt om je hypothese H_0: $\beta_1 = 0$ te toetsen. Veronderstel dat H_0 geldt ($Y = \beta_0 + \varepsilon$, want β_1 valt weg), bereken $|t|$; de P-waarde is de kans iets extremers te observeren. Als $P<0.05$: H_0 verwerpen. NB: β_1 is 0 betekent correlatie is 0 betekent geen lineaire relatie, dit hangt altijd samen. Dus als je H_0 verwerpt dan is β_1 **niet** 0 en is er dus **wel** een correlatie in de populatie.

KEUZE VOOR STATISTISCHE METHODE
Als je steekproef hetzelfde is als je populatie, of je bent alleen geïnteresseerd in je steekproef, dan heb je voldoende aan beschrijvende statistiek (bv gemiddelde, mediaan, odds ratio, correlatie, histogram, boxplots etc). Als je steekproef (veel) kleiner is dan je populatie en als je wilt veralgemenen, dan heb je inferentiële statistiek nodig om om hypothesen over je populatie te toetsen o.b.v. je steekproef.
Verder van belang:
- Observationeel onderzoek (analysetechnieken vallen vaak buiten deze cursus want complex) versus experimenteel onderzoek (vanwege randomisatie is eenvoudige statistiek vaak voldoende, zoals t-toets, Wilcoxon rangsom, ANOVA).
- Longitudinaal onderzoek (ieder individu op meer tijdstippen gemeten, ingewikkelder statistiek buiten deze cursus, bv multilevel analyse, GEE) versus transversaal onderzoek (ieder individu op één tijdstip één keer gemeten, gebruik t-toets, Wilcoxon rangsom, lineaire regressie).

Belangrijkste begrippen Epidemiologie

Epidemiologie bestudeert de distributie, de determinanten en de frequentie van een ziekte bij (groepen) mensen. Einddoel:
- Etiologie (oorzakenleer)
- Preventie

Hoe? Onderzoek tussen determinanten (alle mogelijke factoren die de frequentie van een ziekte kunnen beïnvloeden, heet ook blootstelling/exposure of variabelen) en de ziekte (outcome), binnen een duidelijk gedefinieerde populatie (groep mensen met gelijke kenmerken). Bij een vaste (fixed) populatie zijn de leden permanent gedefinieerd door een gebeurtenis, bij een dynamische populatie kunnen leden tijdelijk in de groep aanwezig zijn. Sequentie van epidemiologisch onderzoek: Vermoeden dat een blootstelling een ziekte veroorzaakt -> Hypothese over exposure-disease association -> Opzetten/uitvoeren epidemiologisch study design -> Is de associatie significant en is er causaliteit? -> Evalueer behandeling of preventie. Exposure oriented epidemiology start bij de blootstelling, Outcome oriented epidemiology start bij de ziekte.

BASISBEGRIPPEN

Frequentiematen/ verhoudingen
- Ratio (x/y: de verhouding tussen twee kwantitatieve variabelen, bv aantal vrouwen t/o aantal mannen = 3:1)
- Proportie (x/x+y, verhouding waar teller voortkomt uit de noemer, bv proportie vrouwen in een rusthuis met 300 vrouwen en 100 mannen = 300/400 ofwel 75%.
- Rate (x/x+y in t1-t2: meet de kans dat een nieuwe gebeurtenis zich **in de loop van de tijd** zal voordoen. Teller is **aantal gebeurtenissen in een periode**, noemer is de populatie die aan de gebeurtenis bloot staat). Bijvoorbeeld Mortality Rate: de totale of bruto mortaliteit = alle overlijdens in een bepaalde periode in een bepaalde populatie. De totale specifieke mortaliteit is de verdeling van de mortaliteit naar bv leeftijd, geslacht, SES, regio.

Hou bij het meten van frequentie rekening met de volgende 3 factoren:
1. de grootte van de oorspronkelijke populatie
2. aantal individuen die de ziekte hebben
3. de tijdsduur dat de populatie gevolgd werd

- Prevalentie (n aantal personen ziek gedeeld door totaal aantal personen in de (risico)populatie **op ogenblik t.** Het is de verhouding van het aantal gevallen/zieken in een populatie op een bepaald tijdstip (bv in een jaar) -> de totale hoeveelheid water in de badkuip).
- Incidentie, ookwel incidentie rate / IR (n aantal personen met **nieuwe** ziekte gedeeld door totaal aantal personen in de (risico)populatie **in periode t1-t2**. Het is het aantal nieuwe gevallen/ zieken binnen een bepaald tijdsinterval -> **in**cidentie is het aantal nieuwe druppels dat **in** de badkuip stroomt).

NB: het 'water in de badkuip', dus het totaal aantal zieken, wordt kleiner als ofwel mensen sterven ofwel mensen genezen.
- Cumulatieve incidentie / CI (n aantal nieuwe gevallen binnen een (risico)populatie over een bepaalde tijd. De tijd maakt geen deel uit van de deling maar er wordt wel verwezen naar het tijdsinterval in de omschrijving van de studie. Het is het mogelijke risico dat een persoon de ziekte zal ontwikkelen over een welbepaalde tijd, bv lifetime risk). Vaak gemeten in een vaste populatie.
- Person-Time incidentie rate (n aantal nieuwe gevallen binnen een (risico)populatie gedeeld door de aantallen in **persoonsjaren**. NB: tijd is al in de noemer verwerkt. Heet ook incidence density. Uitgedrukt in jaren, maanden of dagen).

Vergelijken van relatie tussen blootstelling en ziekte door verschillende groepen te vergelijken. Minstens twee, bijvoorbeeld: blootgestelde groep en niet-blootgestelde (controle)groep

2x2 Tabel

	Ziekte JA	Ziekte NEE	TOTAAL
Exposure JA	a	b	a+b
Exposure NEE	c	d	c+d
TOTAAL	a+c	b+d	a+b+c+d

- Relatief Risico RR (a/a+b gedeeld door c/c+d). De verhouding van de kans op het krijgen van de ziekte in de blootgestelde groep t.o.v. de kans op het krijgen van de ziekte in de niet blootgestelde groep. Uitkomst bv: de blootgestelde groep loopt x keer meer risico op de ziekte dan de niet-blootgestelde groep.
- Odds van de ziekte (a/(a+b) gedeeld door b/(a+b) = a/b. De kans dat een blootgestelde persoon ziek werd t.o.v. de kans dat hij niet ziek werd). Je kunt dit ofwel uitrekenen binnen de blootgestelde populatie ofwel binnen de niet-blootgestelde populatie. In het laatste geval: Odds = c/d.
- Odds Ratio (a/b gedeeld door c/d = a*d/b*c. Dit is het verhoogde risico op de ziekte wanneer je blootgesteld werd t.o.v. wanneer je niet blootgesteld werd. NB: als het aantal zieken klein is t.o.v. de hele populatie dan is de Odds Ratio bijna gelijk aan het Relatieve Risico RR (en geeft het daar dus een schatting van).
- Odds van de exposure (a/(a+c) gedeeld door c/(a+c) = a/c. De kans dat een zieke persoon/ patiënt blootgesteld werd t.o.v. de kans dat hij niet blootgesteld werd. Of: b/(b+d) gedeeld door d/(b+d) = b/d. De kans dat een persoon uit de controlegroep blootgesteld werd t.o.v. de kans dat hij niet blootgesteld werd.

OORZAAK & OORZAKELIJK VERBAND

Bij een direct oorzakelijk verband is er geen intermediaire factor, binnen een indirect oorzakelijk verband is die er wel. In de biologie zijn die er eigenlijk **altijd**: tussen oorzaak en gevolg, op de causal pathway, ligt dus altijd een intermediaire factor/ is er causal interference. Vraag je dus altijd af: wat is oorzaak, wat is gevolg, en zijn er nog andere factoren die de associatie kunnen beïnvloeden (confounders, zie verderop). Associatie betekent nog geen oorzakelijk verband, en een risicofactor is niet hetzelfde als 'de oorzaak'.

Bradford Hill's criteria (basis van evidence based medicine):

1. Sterkte van de associatie (sterke associaties maken meer kans 'juist' te zijn dan zwakke).
2. Consistentie / reproductiviteit (uitkomst opnieuw gevonden na reproductie/replicatie van de uitkomst in verschillende studies).
3. Specificiteit van de associatie (één blootstelling leidt tot één ziekte. Is meestal niet zo).
4. Temporele relatie (de blootstelling moet aan de ziekte voorafgaan. Belangrijkste criterium voor causaliteit! Hou wel rekening met incubatieperiode en latente periode).
5. Dosis-respons relatie / biologische gradiënt (het risico wordt groter naar mate de blootstelling frequenter of groter wordt)
6. Biologische verklaarbaarheid / plausibiliteit (is het consistent met de huidige medische en moleculair biologische kennis?)
7. Coherentie (er mag geen conflict ontstaan met wat algemeen gekend is omtrent het verloop en de biologie van de ziekte)
8. Experiment - Interventie (zodra de blootstelling wegvalt of vermindert, valt ook de outcome weg of vermindert. Sterke evidentie voor causaliteit, maar niet altijd ethisch verantwoord!)
9. Analogie (wanneer een gelijkenis aangetoond kan worden met andere blootstellingen en ziekten)

Gebruik deze criteria als richtlijnen om verschil te kunnen maken tussen een associatie en een oorzakelijk verband in een epidemiologische studie, om deze studies kritisch te beoordelen, om zelf optimaal een study design op te zetten en je eigen resultaten ook kritisch te beoordelen.

STUDY DESIGNS

- Cross-sectioneel study design. Momentopname, ook wel prevalentie-onderzoek, dwarsdoorsnede onderzoek of transversaal onderzoek genoemd. Voorbeeld: een nationale screening. Doel: prevalentie van aandoening en risicofactoren in kaart brengen, zoeken naar associaties tussen blootstelling en ziekte, distributie van ziekte in kaart brengen, opsporen niet gekende ziektegevallen, bepalen nood aan bepaalde voorzieningen vanuit de volksgezondheid. Nadeel: **zegt niks over oorzaak en gevolg** en er is een groter risico op bias omdat het slechts een snapshot is: je mist bv snel mensen die niet op dát moment ziek zijn. Voordeel: relatief goedkoop, goede representatie van het probleem mits steekproef goed gekozen is.
- Cohort studies. Ook wel incidentie-onderzoek of longitudinaal onderzoek genoemd. Je volgt een groep mensen over een zekere tijd, kan zowel retrospectief (terugkijkend) als prospectief (vooruitkijkend) zijn vanaf het eerste meetmoment. Doel: incidentie, etiologie en natuurlijk verloop van ziekte in kaart brengen. Bij een prospectieve cohortstudie start je met een duidelijk

bepaalde populatie die vrij is van ziekte maar wel een zeker risico heeft de ziekte te krijgen. Vervolgens verzamel je veel data, kan zowel directe metingen/tests zijn (beter maar duur) of indirecte gegevens (self reported vragenlijst, goedkoper maar risico op bias en misclassificatie). Nadeel: tijdrovend, complex, follow-up fase kost veel geld en je verliest mensen die verhuizen of geen zin meer hebben. Voordeel: biedt zeer veel informatie, er is een 'forward direction' dus **je kunt een oorzaak-gevolg relatie bepalen** en vaak ook dosis-respons, en je kunt het relatieve risico berekenen. Als de n groot genoeg is kun je ook subgroepen bestuderen. Voorbeeld: Framingham Heart Study, waar het originele cohort is uitgebreid met offspring cohort (de kinderen) en third generation cohort (de kleinkinderen). Verhuizingen waren hier geen probleem want je meet alleen díe regio; voorbeeld van fixed population/ gesloten cohort. Voorbeeld 2: Nurses Health Study, net als Framington een population based cohort studie.

- Case Controle Studie. Je onderzoekt een groep mensen met een ziekte (cases) en die vergelijk je met een groep mensen die de ziekte niet hebben (controls). Het is een outcome oriented study (je vertrekt vanuit de ziekte) en **altijd retrospectief**. Doel: zoeken naar etiologie & evalueren van interventies. Voorwaarden: controls mogen niet lijden aan de ziekte, en je moet cases en controls in de twee groepen zo goed mogelijk matchen (bv op leeftijd, geslacht, SES), en opnemen van cases/controls moet onafhankelijk zijn van de blootstelling die onderzocht wordt (vermijd selectie bias). **Nadeel:** j**e kunt moeilijk causaliteit bewijzen**, het is niet dubbel blind (want jij weet wie er ziek zijn), er is bias (want mensen weten zelf dat ze ziek zijn), en je kunt incidentie niet meten (want je stelt zelf je populatie samen). **Voordeel:** i**deaal om zeldzame ziektes te onderzoeken** (je hoeft niet te wachten tot ze zich voordoen), het is relatief goedkoop, en meerdere exposures kunnen bestudeerd worden (multifactorieel). Je gebruikte dit design dus als de ziekte zeldzaam is, als het lang duurt voor ze zich openbaart, als gegevens omtrent blootstelling moeilijk te verkrijgen of te duur zijn, als er weinig gekend is omtrent de ziekte, en als de populatie heel dynamisch is.

BIAS & CONFOUNDERS

Bias. Een systematische fout, resulterend in niet correct inschatten van de associatie tussen blootstelling en ziekte. Kan in elk studietype voorkomen, kan zowel door onderzoeker als door participant geïntroduceerd worden, en zowel tijdens het design als tijdens het verloop zelf. In de analytische fase kun je de bias wel evalueren maar niet oplossen/ corrigeren. Twee soorten:

- Selectie bias. Gevolg van verkeerd selecteren studiepopulatie waardoor deze verschilt van de doelgroep. Vaak bij case-control en retrospectieve cohort studies omdat hier de blootstelling en de ziekte zich reeds voordeden vóór de start van de studie. Kan echter ook voorkomen bij prospectieve cohorts en experimentele studies bij selectief verlies van deelnemers of selectief niet deelnemen aan een studie.
- Observatie/ informatie bias. Bias door systematisch fout bij ingeven van informatie. Gebeurt na start van de studie, nadat participanten geselecteerd werden. **Types**:
  - Recall bias (zieke mensen herinneren zich meer over het verleden dan niet-zieke mensen)
  - Interviewer bias (als studie niet blind is)
  - Classificatie bias (blootstelling of ziekte verkeerd geclassificeerd)

NB: details omtrent observatie bias (differential/ non differential) zijn niet te kennen!

Confounders. Een confounder is een derde variabele die de effecten tussen blootstelling en ziekte kan verstoren. Zorgt net als bias voor een vertekend beeld tussen blootstelling en ziekte. Simpel gezegd het mixen van effecten tussen exposure en outcome. Kan vaak met logistische regressie-analyse en via stratificatie achterhaald worden. Let op: een intermediaire factor die **op de causal pathway ligt** is **geen confounder** maar is een variabele die tussen de blootstelling en het resultaat ligt. In dat geval geldt: blootstelling -> leidt tot intermediaire factor -> leidt tot ziekte. Voorbeeld: rookgedrag moeder leidt tot laag geboortegewicht leidt tot perinatale sterfte. Een confounder ligt niet op de causal pathway, maar is **zowel geassocieerd met de exposure als met de outcome**. Anders gezegd: de confounder is een onafhankelijke oorzaak of voorspeller van de ziekte, ofwel: de associatie tussen confounder en outcome is onafhankelijk van de exposure. NB: als er **per stratum** een verschil is in associatie en het effect dus **anders is** in de relatie tussen exposure en outcome dan spreek je van effect modificatie in plaats van confounding. Dit is een biologisch fenomeen, en is het belangrijk om de analyse vervolgens uit te voeren volgens stratum (bv verdeling populatie over verschillende leeftijdsgroepen als blijkt dat leeftijd een effect modifier is). Je gaat hier dus **niet controleren** voor deze variabele, wat bij confounding wel gebeurt. Effect modificatie gaat over een derde variabele die wordt geëvalueerd d.m.v. stratificatie. Deze derde variabele ligt niet op de causal pathway en is dus geen intermediaire factor.

SAMENVATTENDE SCREENSHOTS UIT DE CURSUSSEN

Population versus Sample (Statistiek)

Characteristic	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Proportion	π	p
Standard Deviation	σ	s
Variance	σ^2	s^2
Correlation Coefficient	ρ	r
Slope of a line	β	b

Type variabele (Statistiek)

- Dichotoom oftewel Binair: Geslacht, wel/niet roken, wel/niet ziek...
- Polytoom nominaal: Regio, burgerlijke staat...
- Polytoom ordinaal: SES, Sport prestatie, Opleiding...
- Continu interval: Temperatuur, IQ...
- Continu ratio: Lengte, gewicht, bloeddruk...

Type toets (Statistiek)

		Y: Continu	Y: Binair
X	1 groep	gepaarde t-toets z-toets, tekentoets Wilcoxon rangteken toets...	z-toets, tekentoets binomiale toets (exact)...
X	Dichotoom (2 groepen)	t-toets Mann-Whitney toets...	χ^2 toets voor kruistabel...
X	Polytoom (K groepen)	1-weg ANOVA Kruskal-Wallis toets...	χ^2 toets voor kruistabel...
X	Continu	(rang) correlatie lineaire regressie...	logistische regressie...
X	Meerdere X	lineaire regressie ANOVA...	logistische regressie...

Overzicht van Epidemiologische studies en de tijdslijn