Samenvatting Statistiek powerpoints + aantekeningen

College 1: Met name codes in R en een inleiding

Descreptieve statistiek: Een samenvatting van de data geven. Hiermee bereken je het gemiddelde en de spreiding, zoals het gemiddelde, de mediaan en de modus. Je kan er ook mee visualiseren. Hiermee bepaal je de range, IQR, variantie en de standaardafwijking. Dit geeft je inzicht in de data.

Inferentiële statistiek: Je gaat kijken naar relaties binnen de data. Dit zijn de testen die je verder doet, zoals de t-testen, non-parametrische testen en de X^2-testen enz. Hiermee kun je 2 groepen vergelijken, of 1 groep met een vaste waarde.
Interne consistentie: Vragen die over hetzelfde gaan, meten die ook hetzelfde?

Feitjes voor R:

- NA betekent niet aanwezig/not available. Dit kunnen we fixen met na.rm = FALSE, waarbij alle NA genegeerd worden.
- Voor de komma zijn de rijen en na de komma zijn de kolommen. Vergeet de komma niet aan het einde van de blokhaken!
- Bij variabelen mag je geen spaties invoeren! Op andere plekken negeert ie het gewoon.
- R doet een grafiek op alfabetische volgorde. Dus als je het geslacht gaat bekijken, dan zal hij man eerst doen en daarna vrouw op de x-as.

College 2: Beschrijvende statistiek

Het willekeurig selecteren van de hele groep (populatie) is een steekproef. Een steekproef moet altijd representaties zijn. Dit houdt in dat de juist mensen uit de juiste groep kiest.
Bij statistiek gaat het altijd om variabelen. Dat houdt in dat er altijd variatie in de data zit. Een 'geval'/case is te zien in de rijen en de variabelen in de kolommen.

Er zijn verschillende variabelen:

1. Nominale variabelen: Deze heeft alleen categorieën zoals het geslacht kan man of vrouw zijn. Je kan niet van beide een beetje hebben. Verder mag er geen ordening of getallen in zitten. Een man kan dus niet hoger zijn als een vrouw. Hiervoor geldt dat A niet gelijk is aan B.
2. Ordinale variabelen: Hier zit wel een ordening in, maar het verschil is niet goed weer te geven, zoals vmbo, havo en vwo. We weten dat vwo hoger is dan vmbo, maar we kunnen niet goed aangeven dat het vwo bijvoorbeeld 1 hoger is dan vmbo. Het kunnen ook getallen zijn, zoals de Likert schaal, waarop je aangeeft hoe tevreden je bijvoorbeeld bent op een schaal van 1-5. Hiervoor geldt dus dat A groter is dan B, of dat A kleiner is dan B.
3. Numerieke variabelen. Deze zijn op te delen in:

- Interval schaal: Er is geen echt nulpunt, maar er is wel een verschil tussen de getallen, zoals het geboortejaar of de temperatuur. De temperatuur kan in principe oneindig doorgaan in de min en heeft geen exact nulpunt. Hiervoor geldt: A + B en A - B
- Ratioschaal: Hier is wel een duidelijk nulpunt en de verhouding tussen de getallen is ook duidelijk en heeft een betekenis, zoals meeteenheden en de leeftijd. Hiervoor geldt: A*B en A/B.

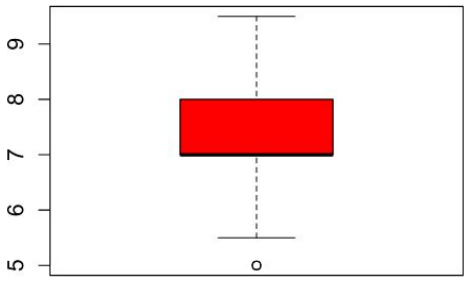
De interval- en ratio schaal worden vaak samengenomen, omdat we hier ook dezelfde statistiek op kunnen toepassen.
Cirkeldiagrammen zijn geen goede weergave, omdat de absolute frequenties niet zichtbaar zijn en we er slecht in zijn om de grootte van een oppervlak te bepalen.

Een distributie geeft alle getallen samen weer in een normaal verdeling. Dit laat de variabiliteit van de variabele zien. Een density plot/dichtheidscurve geeft aan hoeveel er in een bepaald gebied zit. Het totale gebied onder de curve van je plot is 1. Dus 100%, oftewel alle data. De pieken geven aan waar de meeste data

zit. De dichtheidscurve geeft de normaalverdeling aan als deze er is. Je weet alleen niet precies hoeveel data overal zit, maar het geeft wel een idee hoeveel er ongeveer tussen de 2 waarden zit.

Centrummaten:

- Modus: Welk getal het vaakst voorkomt
- Mediaan: Wat de middelste waarde is
- Gemiddelde: Alles optellen en delen door n.



Spreidingsmaten:

- Quartilen:

Q1 = punt tussen groep 1 en groep 2
Q2 = punt tussen groep 2 en groep 3 = mediaan
Q3 = punt tussen groep 3 en groep 4
Dit kun je ook met percentielen doen, waarbij Q1 = 25%, Q2 = 50% en Q3 = 75%.

- Minimum, maximum en de range (verschil tussen het maximum en minimum)
- IQR: verschil tussen Q3 en Q1. Dit is 50% van de data en kan met een boxplot gevisualiseerd worden. De afstand van IQR is groter dan 1,5 van de doos.
- Variantie: verschil tussen het populatiegemiddelde en de waarde. We vergelijken iedere waarde met het gemiddelde en tellen dit op en kwardrateren het. Vervolgens vermenigvuldigen we nog met 1/n.

De formules hiervoor zijn:

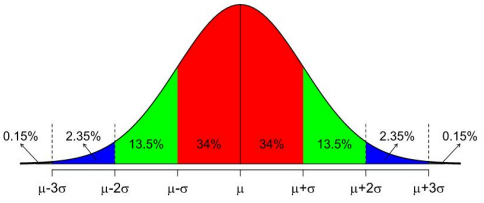
$\sigma^2 = 1/n * \Sigma(xi - \mu)^2 \quad = populatie$

$s^2 = 1/n-1 * \Sigma(xi - \bar{x})^2 \quad = steekproef$

We delen bij de steekproef door n-1, omdat we niet helemaal zeker zijn om het steekproefgemiddelde wel helemaal hetzelfde is als het populatiegemiddelde. Als we dus delen door een iets kleiner getal, dan geeft dat een iets grotere s^2, wat dus weer een betere weerspiegeling geeft en we dus minder snel fouten maken.

- De standaardafwijking is de wortel van de variantie.

De z-score is de afstand tot het gemiddelde in het aantal standaardafwijkingen. De z-scores bepalen is het standaardiseren van een verdeling, waarbij het gemiddelde 0 is en de standaaardafwijking 1. We kunnen de z bepalen volgens: $z = x - \mu / \sigma$.



Vervolgens moeten we nog de SE (standaardfout) bepalen. Dit doen we door: σ / wortel n. De z-toets wordt overigens niet veel geruikt, omdat hij alleen gebruikt kan worden als we de standaardafwijking weten.

College 3: Steekproeven

Het betrouwbaarheidsinterval (BI) is een interval waarin waarschijnlijk is dat hier het populatiegemiddelde ligt. Bij een BI van 95% is de kans 95% dat het populatiegemiddelde in het 95% gebied ligt. Met de SE kunnen we het BI bepalen.

We gebruiken de standaardafwijking als we praten over een individu ten opzichte van het gemiddelde van de populatie of steekproef. De standaardfout gebruiken we als we redeneren over een populatie door middel van een steekproef.

Hypothese-toetsen testen of een hypothese over de populatie waar is. Statistische significantie houdt is dat de fouten/verschillen door willekeurigheid komen en niet door onsystematische variantie. Dus het effect komt niet door toeval, maar door datgene wat je aan het onderzoeken bent. Bij de H0 hypothese gaan we er vanuit dat er niks aan de hand is en dat de dingen die we vergelijken gelijk zijn. Bij de Ha hypothese gaan we ervan uit dat er wel iets aan de hand is. We kunnen zeggen dat de dingen die we vergelijken niet gelijk zijn of een

verschil heeft wat groter of kleiner kan zijn dan de andere variabele. Als we de Ha aannemen, hebben we niet genoeg bewijs om H0 aan te nemen.

De p-waarde is de kans op de steekproef gegeven dat de H0-hypothese waar is. Als de p-waarde lager is dan de grens die we stellen (alpha-niveau) dan is de meting significant en nemen we H0 aan en verwerpen we Ha. Bij tweezijdig testen hebben we een alpha-niveau van 0,05 en bij eenzijdig 0,025.

Het heeft geen nut om extra data te gaan verzamelen, zodat iets significanter wordt, omdat de effectgrootte niet gaat veranderen. We willen een zo laag mogelijke n en een zo laag mogelijke p. Bij veel metingen wordt data snel significant, omdat s kleiner wordt. De effectgrootte verandert niet, omdat er geen n wordt gebruikt in de formule. Bij een grotere steekproef kijken we dus naar de effectgrootte. We kunnen hiermee alleen de grootte van het effect bekijken en niet of iets significant is of niet.

Hoe behandelen we hypothese testen:

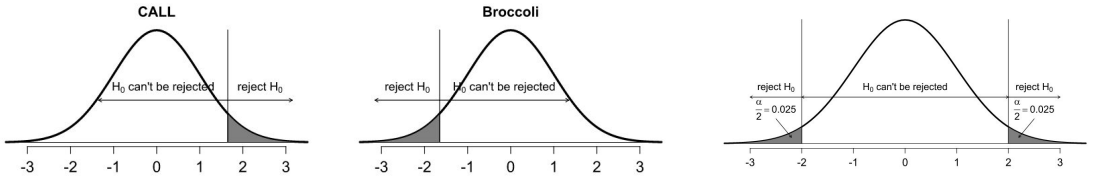
1. Bepaal H0 en Ha
2. Bepaal de test en de onderliggende verdeling
3. Bepaal het alpha-level wanneer je H0 wil verwerpen
4. Bepaal de waarde van de statistiek op basis van je steekproef
5. Bereken de p-waarde en vergelijk deze met het alpha-niveau

Is de p-waarde lager (<) dan het alpha-niveau dan verwerpen we H0 (significant resultaat)
Is de p-waarde hoger of gelijk (>,=) aan het alpha-niveau dan nemen we Ha aan (niet-significant resultaat)

Welke waarde is voldoende om de H0 te verwerpen? Hiervoor heb je een gegeven p-waarde. Deze p-waarde kun je weer omzetten dan een z-score. Met de z-score, kun je de x uitrekenen en heb hiermee de maximale of minimale waarde om H0 te verwerpen. Hoe je van z-score naar p-waarde hoef je niet te kunnen berekenen voor het tentamen.

Bij een z-test berekenen we de z-score met de formule:
$z = (x - \mu) / SE$
Vervolgens krijgen we een gelinkte p-waarde en kunnen we het vergelijken met het alpha-niveau.

Eenzijdige z-test: We kijken maar aan 1 kant van de verdeling. Hierbij is in Ha gezegd dat het gaat om een groter of kleiner dan gemiddelde. We verwerpen H0 als de waarde buiten het BI van 95% valt. Hierbij is het alpha-niveau 0,025.



Tweezijdige z-test: We voorspellen dat Ha anders is en niet hoger of lager. Hiervoor zeggen we bij de Ha hypothese dat de waardes niet gelijk zijn. Het alpha-niveau is hier 0,05.

College 4: Parametrische testen - t-testen

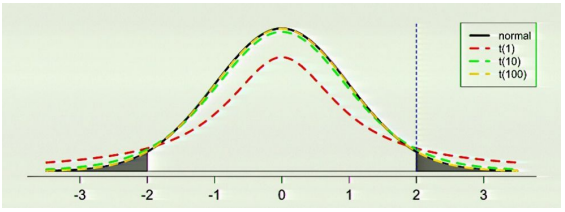
De t-distributie vergeleken met de normaal verdeling

- Het verschil tussen de 2 is groot als we weinig vrijheidsgraden hebben, maar als we meer dan 100 vrijheidsgraden hebben is het verschil niet meer aanwezig. Bij 10 vrijheidsgraden is er meer dan 2,5% aan beide grenzen van de verdeling. Dus bij een t(10) =2 is de meting niet meer significant, terwijl dat bij een normaal verdeling wel significant is.

- De vorm is een klein beetje anders, dus is het punt van de 2,5% ook een klein beetje anders. Oftewel de p-waarde is iets anders.

Hierom moet je altijd bijhouden hoeveel vrijheidsgraden er zijn! Anders weet je niet hoe de verdeling eruit ziet.

Voor alle t-toetsen geldt dat de data normaal verdeeld is als we minder dan 30 metingen hebben. Er zijn 3 soorten t-toetsen



In R gebruiken we voor alle t-testen die genoemd worden dezelfde test, namelijk t.test. Voor alle t-testen gebruiken we ook de Cohen's d voor de effectgrootte. Het stappenplan is ook overal gelijk.

Stappenplan voor t-testen:

1. Kloppen de vereisten?
2. Visualiseren met een boxplot
3. Berekenen van de t-waarde en de p-waarde
4. Is de waarde significant of niet? Vergelijk je dus met de hypotheses.
5. Bereken de effectgrootte.

Cohen's d: small (0,2 - 0,3), medium (0,5), large (<0,8).

Single sample t-test: Wanneer we een gemiddelde vergelijken met een vaste waarde.
Hypotheses zijn:
H0: $\mu = \mu 0$
Ha: $\mu \neq \mu 0$
Een grotere t-waarde geven reden om H0 te verwerpen. De effectgrootte is het verschil tussen het gemiddelde van de steekproef - het gemiddelde van de populatie / steekproefgemiddelde. Bereken je met Cohen's d.

De vereisten voor de test zijn:

- De data moet willekeurig geselecteerd zijn uit de populatie
- De data moet gemeten op interval of ratioschaal
- Observaties zijn onafhankelijk
- De observaties zijn ongeveer normaal verdeeld onder n=30. Dit komt omdat de standaardafwijking anders te hoog wordt. Daarboven maakt het niet uit

Independent sample t-test: Wanneer we 2 onafhankelijke groepen met elkaar vergelijken.
Hypotheses zijn:
H0: $\mu 1 = \mu 2$
Ha: $\mu 1 \neq \mu 2$
Dit kan zowel eenzijdig als tweezijdig. Het aantal vrijheidsgraden is: (n1 -1) + (n2 -1). Wanneer we aannemen dat er een gelijke spreiding in beide groepen is.

De vereisten voor de test zijn:

- De data moet willekeurig geselecteerd zijn uit de populatie
- De data moet gemeten op interval of ratioschaal
- Observaties zijn onafhankelijk, zowel in 1 groep als tussen de 2 groepen. Mogen geen verbanden zijn tussen de groepen.
- De observaties zijn ongeveer normaal verdeeld onder n=30. Dit komt omdat de standaardafwijking anders te hoog wordt. Daarboven maakt het niet uit. Geldt voor beide groepen.
- De varianties moeten gelijk zijn in beide groepen. In beiden groepen moeten de spreidingen dus liggen tussen bijvoorbeeld 2 en 6 en niet in de ene groep tussen 1 en 5 en in de andere groep tussen 2

en 3. Als het wel zo is, dan fixt de t-toets het zelf. Hierdoor wordt het aantal vrijheidsgraden naar beneden gesteld, dus iets minder dan 95% in het middenstuk, dus een hogere t waarde voor significantie.

De effectgrootte is hier te berekenen met d = gemiddelde 1 - gemiddelde 2 / s.

Gepaarde t-test: Wanneer we gepaarde data gebruiken. Als we bijvoorbeeld naar een indiviu kijken, die sochtends en savonds gemeten wordt. We kunnen zien of iemand omhoog of omlaag is gegaan bijvoorbeeld. We kunnen heel veel fouten verwijderen, door paren met elkaar te vergelijken. Bij 2 onafhankelijke groepen zit er veel variatie binnen de groepen, maar dat is bij gepaarde data niet zo.
Hypotheses zijn:
H0: $\mu (x1-y1) = 0$
Ha: $\mu (x1-y1) \neq 0$
Aantal vrijheidsgraden is het aantal paren -1.

De vereisten voor de test zijn:

- De data moet willekeurig geselecteerd zijn uit de populatie
- De data moet gemeten op interval of ratioschaal
- Observaties zijn onafhankelijk tussen de groepen/momenten.
- De observaties zijn ongeveer normaal verdeeld onder n=30. Dit komt omdat de standaardafwijking anders te hoog wordt. Daarboven maakt het niet uit. Geldt binnen beide momenten.
- De schalen moeten gelijk zijn.

Met een onafhankelijke t-test krijgen we een kleinere t-waarde dan bij een gepaarde t-test. Voor de p-waarde geldt dan dat deze groter wordt bij een gepaarde t-test. Met een goede test kunnen de dus beter de effecten bepalen. Als we een onafhankelijke test gebruiken voor een gepaarde test dan is de kans groter dat we H0 aannemen. En vergroten we de kans op een type 2 fout.

We gebruiken t-testen bij numerieke variabelen. In de rapportage moet altijd vermeld worden:

- Hypotheses en waar zijn we in geïnteresseerd. Ook de onderzoeksvragen.
- Hoeveel mensen worden getest, dus de steekproefgrootte, hoe we aan de data komen. Verwijs naar je tabellen of grafieken.
- Welke test en waarom? Grafiek toevoegen als het kan.
- Wat zijn de vereisten van de test en wordt hier aan voldaan?
- Effectgrootte, alpha-level, t-waarde + vrijheidsgraden (ook bij X^2 test), p-waarde, gemiddelde per groep (bij t-test)
- Welke hypothese wordt aangenomen en welke verwerpen we? Interpreteer de resultaten als dat nodig is.

Dit geldt ook voor de non-parametrische testen uit college 5.

Met meerdere toetsen achter elkaar, vinden we sneller een toeval. Om dit toch significant te krijgen, moeten we de significantiegrens delen door het aantal toetsen.

College 5: non-parametrische testen
Niet-parametrische testen hebben geen parameters, zoals vrijheidsgraden. We kijken vooral naar de rangorde. We gebruiken deze testen als:

- Je weet niet of de data normaal verdeeld is
- Wanneer de vereisten van de t-testen niet gelden

Ze kunnen wel gebruikt worden als er wel aan voldaan wordt. Alleen de power (kans op een type 2 fout) is groter bij een niet-parametrische test.

Eigenschappen van niet-parametrische toetsen:

- Kun je vaker gebruiken dan parametrische toetsen
- Voor normaal verdeelde data is minder kans om een effect te vinden dan met parametrische testen.

Mann-Whitney U test: Is de alternatief voor independent sample testen. Kan ook gebruikt worden voor ordinale data, zoals de Likert schaal.
H0: Als we een willekeurige waarde uit de ene steekproef trekken en uit een andere, dan is het wisselend welke groter of kleiner is. De verdelingen zijn ongeveer gelijk.
Ha: De verdelingen zijn niet gelijk. Dus bijvoorbeeld dat de waarde van de ene steekproef bijna altijd groter is dan die dan van de andere steekproef.

Er zijn geen vereisten over het gemiddelde. Als de distributies van de groepen gelijk zijn, dan betekent de nulhypothese ook dat de medianen gelijk zijn. Bij de Ha geldt dit niet. Wordt ook wel Wilcoxon's rank sum test genoemd.

Van de beide steekproeven sorteren we de waardes van laag naar hoog en dan tellen we hoe vaak de items van de ene groep na de items van de andere groep komen. Dan tellen we hoe vaak de een na de ander komt. De minimale waarde is U. Hoe lager deze waarde, hoe groter de kans dat dit significant is.

De wilcox.test werkt hetzelfde als de t.test. In R geeft het een W ipv een U, maar is verder hetzelfde. De effectgrootte berekenen we met de Cliff's delta. Je hoeft de grenzen niet te kennen, maar je moet het wel altijd vermelden.

Ipv een Mann-Whitney U test, zouden we ook een Wilcoxon signed-rank test kunnen gebruiken, waarbij we de waardes converteren naar ranks (rangorde aangeven). De p-waarde is bijna identiek.

Wilcoxon signed-rank test: Is de alternatief voor gepaarde testen en de single sample test. De data is niet-normaal verdeeld. Maar er is wel een eis dat de data symmetrische moet zijn. Kunnen we gebruiken voor ordinaal en geschaalde data.

Hypotheses:
Voor gepaarde data is de H0 dat de mediaan van de verschillen 0 is. Voor de Ha is het dan dat de mediaan van de verschillen niet 0 is.
Voor een single sample test is: H0: de distributie is symmetrische om de waarde x die we gegeven hebben. Doordat we eisen dat de data symmetrisch is, dus is de H0 hetzelfde als voor de t-test. De Ha is dus dat de distributie niet symmetrisch is.

Het idee is dat we de paarsgewijse verschillen berekenen. Met 1 waarde vergelijken we alles met de waarde. Bij gepaarde data vergelijken we de waardes gewoon. De absolute verschillen (min-tekens negeren) sorteren we van laag naar hoog. Verschillen van 0 gooien we eruit. De verschillen voegen we toe aan de rangordes. We berekenen een T waarde. Dit is de som van alle positieve waarden. Hierna voegen we de minnen weer toe.

Als 1 van de groepen niet normaal verdeeld is, dan hebben we niet genoeg informatie. We weten namelijk niet of de verschillen normaal verdeeld zijn. Als er niks gezegd wordt over de symmetrie, hebben nog steeds niet genoeg informatie. Dus kunnen we uit deze informatie niet bepalen welke toets we kunnen gebruiken.

Als H0 waar is, dan ligt de T waarde dichtbij de helft van de totale som van de T waardes zonder alle mintekens. We gebruiken weer de wilcox.test. Bij deze test verliezen we informatie.

De effectgrootte berekenen we hier met r. Dit is de z-waarde delen door wortel n. De z-waarde kan bepaald worden via de p-waarde, die gegeven is in de wilcox-test.

Sign test: Gebruiken we als de data niet symmetrisch is. Ook hier kunnen we de wilcox-test gebruiken, maar voegen we nog toe dat de data gepaard is (paired = TRUE). De effectgrootte wordt uitgerekend met de Cliff's delta. We kijken of het verschil positief of negatief is. Hierdoor verliezen we een heleboel informatie. De data moet niet-normaal en asymmetrisch zijn. Het test of de verschillen willekeurig of structureel is. H0: er is geen verschil tussen de plusjes en minnetjes. Ha juist meer plussen of minnen. Is gebaseerd op de binominale verdeling, met 2 parameters: n (hoeveel) en p (kans op succes). Een voorbeeld is hoe vaak vinden we kop als we 100 keer een muntje opgooien? Het lijkt op een normale verdeling. Het gemiddelde is n x p.

Doordat we alleen maar plussen en minnen overhouden en alle getallen eigenlijk weggegooid hebben, verliezen we enorm veel informatie. De test die we gebruiken is de binom.test (x = aantal keren succes, n = aantal metingen, p = kans, alternative = less/greater/two.sided). p is altijd 0,5 bij een tekentoets. De tekentoets vindt ook minder snel een effect. Het is wel de beste optie voor niet-nummerieke data.

Met de Shapiro-Wilk test kunnen we testen of de data normaal verdeeld. Als de p-waarde van deze test, lager is dan de alpha-waarde, dan is de data niet normaal verdeeld. Dit is lastiger voor een groep met weinig meetpunten. Daarom visualiseren we eerst en als we het dan echt niet weten, gaan we de toets gebruiken.

College 6: X^2-test, correlatie en Cronbach's alpha

Relatie tussen categorische variabelen kun je weergeven met kruistabellen. Dit doe je met prop.table. Dit kun je doen met:

- Rijen en kolommen. Dit geeft de proporties per kolom of per rij, in plaats van beiden totaal.

X^2 test: Relatie tussen 2 nominale variabelen bepalen. Kunnen we zeggen dat de kans op A|B, dus A gegeven dat B waar is, is die anders dan gewoon de kans op A? Dus is de kans dat iemand die tweetalig onderwijs heeft gevolgd en man is, ander dan de kans op tweetalig onderwijs? Dit kunnen we testen door te tellen.

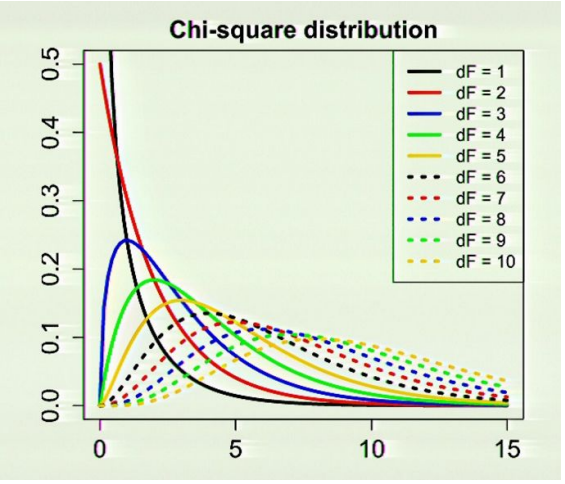
We tellen de echte frequenties en die vergelijken we met de frequenties als er geen relatie zou zijn. Als de verschillen groot zijn, dan is er waarschijnlijk een relatie. Dit geldt niet voor gepaarde data.

De vereisten zijn:

- Data moet willekeurig geselecteerd worden uit de populatie
- Observaties moeten onafhankelijk zijn → Geen gepaarde data dus
- Elke observatie kan geplaatst worden in 1 categorie en niet in beiden
- De verwachte frequentie moet tenminste 5 zijn. Is dit niet zo, dan gebruiken we de Fisher test.

De verwachte frequenties kun je berekenen met de rij proporties * het kolomtotaal. Kun je ook weergeven in een kruistabel. Dit is hoeveel je verwacht als het niet gerelateerd is. Kun je in R doen, door $expected na de chisq.test te zetten.

De formule is de som van: (geobserveerde frequentie - verwachte frequentie)² / verwachte frequenties. Dit is de X^2 waarde. Het aantal vrijheidsgraden is (R-1)*(C-1). Dit is zo, omdat de rijen en kolommen de parameters zijn. Met de ene waarde kunnen we de andere uitrekenen omdat de som altijd 1 is. Hierdoor moet je er 1 van aftrekken.



De verdeling ziet er anders uit dan een normaal verdeling. Hoe meer vrijheidsgraden, hoe zo'n groter gebied na een bepaalde waarde. Is er een lagere of een hogere X^2 waarde nodig voor significantie wanneer het aantal vrijheidsgraden hoger is? Voor een hoger aantal vrijheidsgraden (rode lijn met zwarte lijn

vergelijken) dan kunnen we kijken hoeveel kans er nog onder de waarde van bijvoorbeeld 5 is. Het totale oppervlak onder de lijn is 1. Vanaf 5 is er nog maar een klein beetje oppervlak onder de lijn, maar voor de rode lijn is er meer oppervlak onder de lijn dan bij de zwarte. Dus is er meer kans onder de rode lijn. Voor de rode lijn moet je dus verder naar rechts om dezelfde kans te krijgen als voor de zwarte lijn. Dus we hebben een hogere X^2 waarde nodig voor significantie.

In R kun je dit berekenen met chisq.test waarbij correct op FALSE staat. We doen dan geen continuïteitscorrectie (Yates). Anders worden de p-waardes te hoog, waardoor we een type 2 fout kunnen krijgen.

Fisher test: Als de verwachte frequenties lager is dan 5. Je krijgt in R ook een waarschuwing bij de chisq.test. Deze test doe je met fisher.test.

	Ziekte aanwezig	Ziekte afwezig
Positief	A	B
Negatief	C	D

OR: (A/B) / (C/D). Dit is de **odds ratio**. Berekent de Fisher test voor je. Wordt niet verteld door Martijn, maar staat in het boek volgens hem. De Ha van deze test zegt dat de odds ratio niet gelijk is aan 1. Als de ratio hoger is dan 1, dan houdt dat in dat X groter is dan Y. Als X kleiner is dan Y, dan is de ratio ook lager dan 1.

Bij de Fisher test kijken we verder naar de p-waarde. Als de p-waarde hoger is dan de alpha-waarde, dus we nemen de H0 hypothese aan. Met Cramer's V kunnen we de effectgrootte bepalen. In R doe je dit met assocstats. Een andere optie is de Pearson's test, maar deze wordt minder vaak gebruikt. Dit kun je allemaal doen met assocstats.

De effectgrootte voor de X^2 testen is minder informaties dan bij de t-test, omdat de complexe tabellen moeilijk weer te geven zijn in 1 enkele waarde. Het zegt iets over de hele tabel, maar niet om een bepaalde waarde.

Hypotheses zijn:
H0: De 2 variabelen zijn independent/onafhankelijk
Ha: De 2 variabelen zijn dependent/afhankelijk

Stappenplan:

1. Verwachte frequenties bepalen, zodat je weet welke test je moet hebben
2. Visualiseren met barplot
3. Berekenen van X^2-waarde en p-waarde
4. Effectgrootte bepalen
5. Waar zitten de grootste verschillen? Kun je zien in de proportietabellen.

Met de residueen kun je dit ook zien. Dit doe je door na de chisq.test $stdres te plaatsen. Als de variabelen een waarde hebben >2, dan heeft deze een groot verschil. Dit wordt gedaan met de z-score. Omdat we z-score vinden, weten we dat een waarde hoger dan 2, dus 2 SE van het gemiddelde afzitten en dit niet meer significant is.

Simpson's paradox: Vertel genoeg informatie en houd geen informatie achter, want dit kan invloed hebben op je onderzoek.

De relatie tussen 2 nominale variabelen kun je weergeven met een scatter plot. Doe je in R met plot. Is gewoon een plot met alle punten. Je zet je 2 variabelen, zoals english score en grade tegen elkaar uit. Hoe groot de

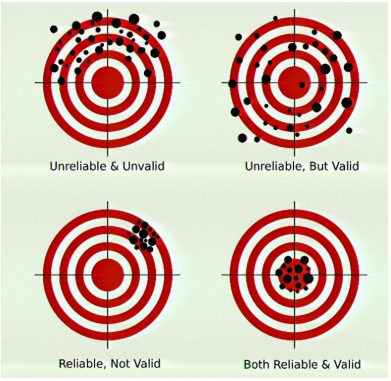
relatie is kunnen we aangeven met de **correlatie test r.** In R kunnen we dit berekenen met cor(dat1, dat2). Als er een NA waarde in de je variabelen zit, werkt het niet en moet je use = pairwise.complete.obs toevoegen als je cor. Hij negeert dan alle NA.

We berekenen de z-scores (standaardiseren). Van de 2 z-scores van de verschillende groepen kunnen we de r berekenen volgens: r = 1/n-1 * som van Zx1 * Zy1. We doen n-1 omdat dat matcht met de standaardafwijking. Een negatieve correlatie geeft aan dat de lijn naar beneden en een positieve naar boven.

Visualisatie is belangrijk. Kun je beter altijd doen. De correlatie is gevoelig voor uitbijters. Hoe meer uitbijters, hoe dichter de correlatie bij 0 komt ipv 1. Door een uitbijter, wordt de z-waarde hoger, waardoor de standaardafwijking ook groter wordt. Dit heeft een groot effect op de correlatie.
Door een hele hoge uitbijter, dan kan de correlatie omklappen van een negatieve lijn, naar een positieve lijn. Hierdoor kunnen we beter wel even visualiseren, want dan ontdekken we dit soort dingen.

Altijd de verborgen factoren wegwerken en onderzoeken, zodat je kan zien of er wel of niet een relatie is. Je moet de test kunnen gebruiken, kunnen visualiseren en aangeven welke richting de correlatie opgaat en de samenhang tussen de variabelen. We doen geen toetsen.

Er zijn ook niet-parametrische alternatieven, zoals Spearman's ρ .



Onderzoekers stellen heel vaak ongeveer dezelfde vragen, zodat je een beter idee krijgt van de onderliggende gedachte. Dus meerdere vragen over hetzelfde. Dit noem je een schaal. Bij een vragenlijst zijn altijd 2 belangrijke elementen:

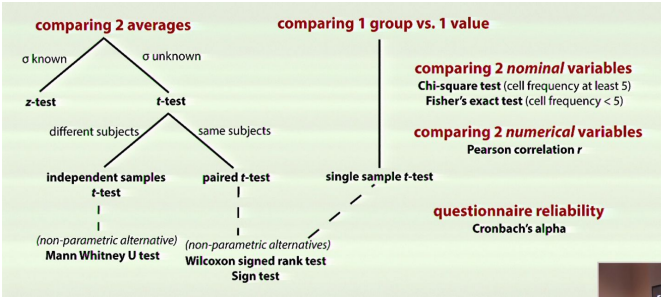
- Validity/valide: Meet je wat je ook daadwerkelijk wil weten?
- Reliability/betrouwbaarheid: Hoe betrouwbaar zijn je vragen voor de maat? Als je vragenlijst nog een keer zou laten maken, zou je dan dezelfde resultaten krijgen? De betrouwbaarheid kunnen we bepalen met de Cronbach's alpha.

We willen eigenlijk altijd een hoge validiteit en een hoge betrouwbaarheid.

Cronbach's alpha: Het idee is dat als we de vragen opsplitsen in 2 groepen, hoe goed zijn die 2 helften van de vragenlijst het dan met elkaar eens? Hoe correleren alle individuele vragen met elkaar? Cronbach's alpha is afhankelijk van de gemiddelde correlatie van alle vragen en het aantal vragen. Een Cronbach's alpha hoger dan 0,7 is acceptabel. Vanaf 0,8 is goed en vanaf 0,9 is heel goed. Cronbach's alpha is afhankelijk van de n en de r.

Vragen kunnen omgekeerd zijn. Deze kunnen we omdraaien, zodat we een goed Cronbach's alpha kunnen bepalen. We kunnen ook vragen weglaten, als ze een te hoge waarde geven bijvoorbeeld.

Voor Cronbach's alpha gebruiken we in R alpha. Hij geeft aan dat sommige waardes omgekeerd zijn in een waarschuwing en dat we ze moeten omdraaien. Met summery kunnen we het resultaat opvragen. We krijgen dan een raw-alpha. Bij een hele lage waarde, zal er waarschijnlijk een omgekeerde waarde inzitten. R weet niet of de waardes positief of negatief moeten zijn. We moeten zelf even kijken welke vragen het zijn.
Bij 'Ik haat statistiek' geeft een hogere waarde aan dat je negatiever bent. Deze waardes wil je inverteren/omkeren. Dit doe je met keys. Hierdoor vergroot je de raw-alpha. Dan is nu de vraag: Kunnen we de betrouwbaarheid verhogen door een item te laten vallen?



Met result$alpha.drop kunnen we zien wat er gebeurt als we een item laten vallen. Als we een item laten vallen, zijn de raw-alpha en de std.alpha gelijk.