Hoofdstuk 5: Steekproevendistributies

Inleiding
Statistische gevolgtrekkingen worden gebruikt om conclusies over een populatie of proces te trekken op basis van data. Deze data wordt samengevat door middel van *toetsen,* zoals gemiddelden, proporties en hellingscoëfficiënten van regressielijnen. Wanneer data het gevolg is van een random wijze van steekproeven trekken, dan is een toets een random variabele die met kansberekeningen begrepen kan worden.

- Een toets van een random steekproef of een gerandomiseerd experiment is een random variabele.
- De kansdistributie van een toets wordt de *steekproevendistributie (sampling distribution)* genoemd. Zo een distributie laat zien hoe een toets (zoals een gemiddelde) zal variëren wanneer herhaaldelijk een steekproef getrokken zou worden.
- De *populatiedistributie* van een variabele is een distributie die alle waarden bevat die een variabele aanneemt bij leden van de populatie. De populatiedistributie is ook de kansdistributie van een variabele wanneer we random een individu uit de populatie trekken.

5.1 De steekproevendistributie van een steekproefgemiddelde

Categorische en kwantitatieve data
Tellingen en proporties zijn discrete random variabelen en beschrijven categorische data. De toetsen om kwantitatieve variabelen mee te beschrijven zijn echter continuerend van aard. Voorbeelden zijn het steekproefgemiddelde, percentielen en de standaarddeviatie. Steekproefgemiddelden worden vaak gebruikt om een algemeen beeld te geven van een steekproef.

Steekproefgemiddelden ($\bar{x}$)
Er zijn twee belangrijke feiten als het aankomt op steekproefgemiddelden:

- Steekproefgemiddelden zijn minder variabel dan individuele observaties.
- Steekproefgemiddelden zijn normaler verdeeld dan individuele observaties.

Het gemiddelde en de standaarddeviatie van ($\bar{x}$)
Het steekproefgemiddelde ($\bar{x}$) is een schatting van het gemiddelde μ van de populatie, net zoals $\hat{p}$ een schatting is van de populatieproportie p. De steekproevendistributie van $\bar{x}$ wordt bepaald door (1) het design dat gebruikt wordt om data te verzamelen, (2) de steekproefgrootte n en (3) de populatiedistributie.

- Het steekproefgemiddelde van een SRS van grootte n is: $\bar{x}$ = 1/n ($X_1 + X_2 + X_3 + \ldots + X_n$). Het n aantal metingen zijn waarden van n random variabelen $X_1, X_2, \ldots, X_n$. Een enkele X_i is een meting van een enkel individu dat random uit de populatie getrokken is en deze meting heeft daarom de distributie van de populatie. Als de populatie in vergelijking tot de steekproef groot is, dan kunnen we $X_1, X_2, \ldots, X_n$ zien als onafhankelijke random variabelen die allemaal dezelfde distributie hebben. De conclusie is dus dat het gemiddelde van $\bar{x}$ hetzelfde is als het gemiddelde van de populatie. Om deze reden is $\bar{x}$ een foutloze schatter van de het onbekende populatiegemiddelde μ.

- De standaarddeviatie van het steekproefgemiddelde is: $\bar{x}$ = σ/$\sqrt{n}$. Zoals bij de steekproefproportie, vermindert ook de spreiding van de steekproevendistributie van een steekproefgemiddelde als de steekproefgrootte stijgt.

Kortom: het steekproefgemiddelde is hetzelfde als het populatiegemiddelde, omdat $\bar{x}$ als een foutloze voorspeller van μ wordt gezien. De standaarddeviatie van de steekproef is de standaarddeviatie van de populatie gedeeld door de wortel van het aantal deelnemers.

Accuraatheid
Hoe precies voorspelt $\bar{x}$ het populatiegemiddelde μ? Omdat de waarden van $\bar{x}$ per steekproef variëren moeten we deze vraag beantwoorden aan de hand van de steekproevendistributie. We weten dat $\bar{x}$ een foutloze voorspeller van μ is en daarom zullen de waarden van $\bar{x}$ bij herhaalde steekproeven niet systematisch te hoog of te laag zijn in relatie tot μ. Veel steekproeven zullen een $\bar{x}$ geven die dichtbij μ ligt als de steekproevendistributie rond de waarde van μ zal liggen. De precisie van de schatting van μ hangt af van de spreiding van de steekproevendistributie.

De centrale limiettheorie
We hebben tot nu toe het middenpunt en de spreiding van de kansdistributie van $\bar{x}$ besproken, maar de vorm van deze distributie is nog niet aan de orde geweest. De vorm van de distributie van $\bar{x}$ hangt af van de vorm van de populatiedistributie. Als de populatiedistributie normaal verdeeld is, dan is de distributie van het steekproefgemiddelde dat ook.

- Als een populatie een N(μ,σ) distributie heeft, dan heeft $\bar{x}$ van n aantal observaties een N(μ.,σ/$\sqrt{n}$) distributie. In de praktijk zijn veel populaties echter niet normaal verdeeld. Toch is het bij grote steekproeven zo dat de distributie van $\bar{x}$ dan toch bij benadering normaal verdeeld is. Het maakt in dat geval dus niet uit welke vorm de populatiedistributie heeft, als de populatie maar een duidelijke standaarddeviatie (σ) heeft. Dit feit wordt ook wel de *centrale limiettheorie* genoemd.

Kortom: als je een grote SRS van n uit welke populatie dan ook trekt (met gemiddelde μ en standaarddeviatie σ), dan zal de steekproevendistributie van het steekproefgemiddelde ($\bar{x}$) bij benadering normaalverdeeld zijn: $\bar{x}$ = N(μ,σ/$\sqrt{n}$).

Andere feiten die gepaard gaan met de centrale limiettheorie
Er zijn drie andere feiten die te maken hebben met de centrale limiet theorie:

- Ten eerste is de normaalbenadering voor steekproefproporties en tellingen een voorbeeld van de centrale limiettheorie. Dit is waar, omdat een steekproefproportie als een steekproefgemiddelde gezien kan worden.
- Daarnaast heeft een lineaire combinatie van normaal verdeelde random variabelen ook een normaal verdeelde distributie. Dus: als X en Y onafhankelijke normaal verdeelde random variabelen zijn en a en b vaste getallen zijn, dan is aX+bY ook normaal verdeeld.
- Tot slot is het zo dat algemene versies van de centrale limiettheorie stellen dat de distributie van een optelling of een gemiddelde van veel kleine random hoeveelheden bijna normaal verdeeld is. Dit is zelfs waar wanneer de hoeveelheden niet onafhankelijk van elkaar zijn. Ze mogen echter niet te sterk met elkaar correleren.

5.2 Steekproevendistributie voor tellingen en proporties

Tellingen

Een random variabele X is een *telling (count)* als we tellen hoe vaak een bepaalde uitkomst voorkomt. Je kunt bijvoorbeeld tellen hoe vaak mensen 'ja' antwoorden op de vraag of prostitutie legaal moet zijn.

- Als het aantal observaties *n* is, dan is de steekproefproportie: $\hat{p}$ = *X/n*. X staat voor de telling, bijvoorbeeld het aantal mensen dat achter de legalisering van prostitutie staat.

De binomiale distributie

Bij een binomiale distributie hoort een aantal kenmerken:

- Er zijn *n* aantal observaties.
- Deze *n* observaties zijn allemaal onafhankelijk.
- Elke observatie valt in één van twee categorieën. Deze categorieën noemen we voor het gemak 'succes' en 'falen'.
- De kans op een succes (we noemen dit *'p'*) is voor elke observatie hetzelfde.

Voorbeeld:
Het n aantal keer werpen met een munt. Elke keer heb je 0.5 kans op kop of munt. De uitkomsten zijn onafhankelijk van elkaar: als je een keer munt hebt gegooid vergroot dat niet de kans dat je de volgende keer ook munt zult gooien. Als we kop 'succes' noemen, dan is *p* de kans op kop en deze kans blijft hetzelfde als we de volgende keer weer een munt werpen.

- De distributie van X (de telling van het aantal successen in een binomiale setting) wordt helemaal bepaald door het aantal observaties (n) en de kans op succes (p).
- De mogelijke waarden van X zijn hele cijfers tussen 0 en n.
- We korten de binomiale distributie af met *B(n,p)*.
- De binomiale distributie is van belang wanneer we conclusies over de populatie willen trekken over de proportie 'successen'. Het kiezen van een SRS uit een populatie is echter niet echt een binomiale situatie.
- Een populatie bevat proportie p van successen. Als de populatie veel groter dan de steekproef is, dan heeft telling X (aantal successen in een SRS van grootte n) *bijna* de binomiale distributie B(n,p).
- Deze benadering van de binomiale distributie wordt groter als de populatie steeds groter wordt in relatie tot de steekproef. De vuistregel is dat we de binomiale steekproevendistributie gebruiken voor tellingen wanneer de populatie minstens 20 keer zo groot is als de steekproef.

Binomiale kansen vinden

Vaak kunnen binomiale kansen berekend worden door middel van software. Het is ook mogelijk om tabel C achterin het boek te raadplegen. Om deze tabel te gebruiken moet de kans op individuele uitkomsten voor de binomiale random variabele X geweten worden.

Het binomiale gemiddelde en de standaarddeviatie

Wat zijn het gemiddelde (μ_x) en de standaarddeviatie (σ_x) van binomiale kansen? Het gemiddelde kunnen we raden. Als Piet 75% van de keren succes heeft, dan is het gemiddelde bij 12 gebeurtenissen 75% van 12, dus 9. Dat is μ_x wanneer X dus B(12,0.75) is.

- Dit betekent dat we *np* kunnen berekenen om het gemiddelde te vinden. We noemen een succes vaak p of 1 en geen succes is een 0 (of 1-p). Kortom: μ_x=np.
- De standaarddeviatie (σ_x) berekenen we als volgt. Eerst berekenen we de np(1-p). Vervolgens trekken we hier de wortel uit.

Steekproefproporties

Hoeveel procent van de volwassenen is voor abortus? Bij steekproevendistributies willen we vaak schatten wat de *proportie* p van 'successen' in een populatie is. Onze schatter van de steekproefproportie van successen is:

- $\hat{p}$ = X/n. Het is hierbij van belang om te weten dat $\hat{p}$ niet hetzelfde is als telling X. De telling X neemt een heel getal aan tussen de 0 en n, maar een proportie is altijd een getal tussen de 0 en 1. In een binomiale situatie heeft telling X een binomiale distributie. $\hat{p}$ heeft juist geen binomiale distributie. We kunnen echter wel kansberekeningen voor $\hat{p}$ uitvoeren door deze op te schrijven in de vorm van telling X. Vervolgens kunnen we gebruik maken van binomiale berekeningen. De eerste stap is het vinden van het gemiddelde en de standaarddeviatie van een steekproefproportie.
- Laat p de steekproefproportie van successen in een SRS van grootte n zijn. Deze SRS is getrokken uit een grote populatie met een populatieproportie p van successen. Het gemiddelde van $\hat{p}$ is p. De standaarddeviatie van $\hat{p}$ is $\sqrt{p(1-p)/n}$. Deze formule voor de standaarddeviatie gebruiken we als de populatie 20 keer zo groot is als de steekproef.
- Het feit dat het gemiddelde van p $\hat{p}$ is, zegt dus dat de steekproefproportie een *foutloze schatter (unbiased estimator)* is van de populatieproportie p. Wanneer een steekproef uit een nieuwe populatie wordt getrokken (met een andere waarde voor de populatieproportie p), dan verandert de steekproevendistributie van $\hat{p}$ richting de nieuwe waarde van p. De spreiding van $\hat{p}$ wordt minder als de steekproefgrootte stijgt. De variantie of standaarddeviatie worden dan dus kleiner Dit betekent dat de steekproefproportie van een grote steekproef dicht zal liggen bij de populatiepropotie p.

Normaalbenadering voor tellingen en proporties

De steekproevendistributie van een steekproefproportie $\hat{p}$ is bijna normaal verdeeld. Nu weten we dat de distributie van $\hat{p}$ een binomiale telling is van steekproefgrootte n. In de praktijk is het zo dat de steekproefproportie $\hat{p}$, maar ook telling X, bijna normaal verdeeld zijn in grote steekproeven.

- Trek een SRS van grootte n uit een grote populatie met populatieproportie p van successen. Als X dan de telling van het aantal successen in de steekproef is, dan geldt: $\hat{p}$=X/n. Als n groot is, dan zijn de steekproevendistributies van zowel $\hat{p}$ als X bijna normaalverdeeld:
- X is bijna N(np, $\sqrt{np(1-p)}$.
- $\hat{p}$ is bijna N(p, $\sqrt{p(1-p)/n}$).
- De vuistregel is dat we deze benaderingen alleen gebruiken voor waarden van p waarbij geldt: np ≥ 10 en n(1-p) is ook ≥ 10. De accuraatheid van deze normaalbenaderingen wordt beter naarmate de steekproefgrootte n stijgt. De benaderingen zijn het beste voor een vaste n wanneer p gelijk is aan 0.5. De benaderingen zijn het minst accuraat wanneer p zich rond de 0 of de 1 bevindt.

De continuïteitscorrectie
Als we een binomiale kansberekening willen maken voor bijvoorbeeld X ≥ 10, dan moet er rekening mee gehouden worden dat het in de praktijk gaat om alle waarden tussen de 9.5 en 10.5. De kans die bij X ≥ 10 hoort, is eigenlijk hetzelfde als de kans die bij X ≥ 10.5 hoort. Er wordt voor de kansberekening daarom uitgegaan van 10.5 in plaats van 10. Dit wordt de *continuïteitscorrectie* voor de normaalbenadering genoemd.

Binomiale coëfficiënt
Om het aantal manieren waarop k aantal successen bij n aantal observaties geschikt kunnen worden uit te rekenen, wordt de binomiaal coefficient gebruikt: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

De formule voor binomiale coefficienten gebruikt de *factoriele notatie.* De factorieel n! voor elk positief getal n is:

- n! = nx(n-1)x(n-2)x…x3x2x1. Ook geldt van 0! = 1.
- De notatie n boven k heeft niets te maken van n/k.
- Als X de binomiale distributie B(n,p) met n observaties en kans p van succes voor elke observatie heeft, dan zijn de mogelijke waarden van X 0,1,2,3…n.

De poissonverdelingen
Een telling X heeft een binomiale verdeling wanneer deze geproduceerd wordt in een binomiale setting. Als één of meerdere facetten van deze setting niet kloppen, zal de telling X een andere verdeling hebben. We komen vaak tellingen tegen die *open* zijn, dat wil zeggen dat ze niet gebaseerd zijn op een vast aantal n observaties. In deze situaties kan de *poissonverdeling* gebruit worden. Deze telling representeert het aantal gebeurtenissen (noem deze 'successen') die voorkomen in een vastgestelde meetunit, bijvoorbeeld binnen een bepaalde tijd, regio of ruimte. Deze verdeling kan gebruikt worden onder de volgende condities:

1. Het aantal successen dat voorkomt in twee niet overlappende meetunits is *onafhankelijk*
2. De kans dat een succes voorkomt in een meetunit is hetzelfde voor alle units van gelijke grootte en is proportioneel aan de grootte van de unit.
3. De kans dat meer dan één gebeurtenis voorkomt in een meetunit is te verwaarlozen voor hele kleine units. Ofwel: de gebeurtenissen komen één voor één voor.

Voor de poissonverdelingen is μ van het aantal successen per meetunit de enige belangrijke kwantiteit. De standaarddeviatie van de verdeling is √μ.

Wanneer het gemiddelde van de poissonverdeling groot is, kan het moeilijk zijn om poissonkansen te berekenen met een rekenmachine of software. Gelukkig is het zo dat wanneer μ groot is, de kansen benaderd kunnen worden door de normaalverdeling met gemiddelde μ en standaarddeviatie √μ te gebruiken.