Boeksamenvatting bij Statistics for Business and Economics van Newbold et al. – 8e druk

Inhoudsopgave

Grafieken om data te omschrijven – Chapter 1

1.1 Beslissingen maken in een onzekere omstandigheid

De *populatie* is de complete set van items waar de onderzoeker in geïnteresseerd is. Het onderzoeken van de complete set van items kan onmogelijk zijn vanwege de kosten of tijd gebrek. Voorbeeld van populatie is alle mogelijke kopers van een nieuw product.

Een *sample* is een deel van de populatie. De sample moet representatief zijn voor de gehele populatie.

Simple random sampling: een methode om een sample te bepalen van de populatie op een manier zodat elk lid van de populatie door toeval wordt geselecteerd. De keuze voor een bepaald lid beïnvloed dus niet de selectie van een ander lid.

Systematic sampling: Selectie van ieder j-ste item in de populatie. J de ratio is van de populatie grootte N tot de gewenste sample grootte n. (J = N/n) . Het startpunt is een getal tussen 1 en j, deze wordt random gekozen.

Een *parameter* is een beschrijvende maat die een specifiek kenmerk van een populatie beschrijft (bijv. gemiddelde, mediaan en standaarddeviatie). Een *statistiek* beschrijft een specifiek kenmerk van een sample.

Nonsampling errors:

- De populatie waarvan de steekproef is genomen is niet de relevante populatie.
- Deelnemers kunnen onnauwkeurige of geen eerlijke antwoorden geven. Dit kan gebeuren doordat vragen worden geformuleerd op een wijze die moeilijk te begrijpen is of op een manier die richting een bepaald antwoord stuurt. Vragen kunnen ook te gevoelig zijn.
- Deelnemers geven gaan antwoord op vragen.

Probleem definitie:

Welke informatie is nodig? Wat is de relevante populatie? Hoe moeten steekproef deelnemers worden geselecteerd? Hoe moet informatie worden verkregen van de steekproef deelnemers?

Descriptive statistics: gericht op grafische en numerieke procedures die gebruikt worden voor het samenvatten en presenteren van data op een informatieve manier.

Inferential statistics: gericht op het gebruiken van de gegevens uit een steekproef om voorspellingen en schattingen te maken om betere beslissingen te nemen.

1.2 Classificatie van variabelen

Een *variabele* is een specifieke eigenschap van een individu of object (bijv. leeftijd of gewicht), het gegeven wat interessant is voor de onderzoeker.

Categorische variabelen produceren antwoorden die behoren tot bepaalde categorieën. Bijvoorbeeld: ja/nee, geslacht, ranking (zeer mee oneens, zeer mee eens).

Numerieke variabelen:

Discrete variabele kent een vast aantal waarden, het antwoord is vaak afkomstig van een telproces. Bijvoorbeeld het aantal studenten in een klas.

Continue variabele kan elke waarde binnen een bepaald bereik van reële getallen aannemen en ontstaat meestal door een meetproces. Bijvoorbeeld lengte, gewicht, tijd, afstand, temperatuur.

Bij *kwalitatieve gegevens* is er geen meetbare betekenis tussen het verschil in aantallen. Bijvoorbeeld een voetbalspeler heeft nummer 20 op zijn shirt staan, deze is niet twee keer zo goed als een speler met nummer 10 op zijn shirt.

Meetniveaus:

Nominaal, de waarden zijn woorden die de categorieën van reacties beschrijven. Bijvoorbeeld 1=ja, 2=nee of 1=man, 2=vrouw.

Ordinaal, de waarden zijn woorden die de categorieën beschrijven met een bepaalde ranking. Bijvoorbeeld productkwaliteit: 1=slecht, 2=gemiddeld, 3=goed.

Bij *kwantitatieve gegevens* is er een meetbare betekenis tussen het verschil in aantallen. Bijvoorbeeld de score op een tentamen.

Meetniveaus:

Interval, geeft rang en afstand tot een willekeurig nulpunt gemeten in intervallen. Bijvoorbeeld temperatuur: 80 graden Celsius is niet vier keer zo warm als 20 graden Celsius.

Ratio, geeft rang en afstand tot een natuurlijk nulpunt. Verschil tussen twee metingen heeft betekenis (leeftijd, gewicht, lengte). 200 kilo is twee keer zoveel als 100 kilo.

1.3 Grafieken om categorische variabelen te beschrijven

Frequentieverdeling is een tabel om gegevens te ordenen. De linker kolom bevat alle mogelijke antwoorden over de onderzochte variabele. De rechterkant is het aantal frequenties voor elke klasse. De *relatieve frequentieverdeling* wordt verkregen door elke frequentie te delen door het totaal aantal waarnemingen en deze te vermenigvuldigen met 100%.

Tabellen en grafieken nominaal meetniveau:

- *Staafdiagram* geeft de frequentie per categorie aan met hoogte.
- *Kruis tabel* geeft het aantal waarnemingen voor elke combinatie van waarden voor twee categorische of ordinale variabelen. Bijvoorbeeld een staaf man en een staaf vrouw. In de staven wordt de frequentie van onderzoeksonderwerp weer gegeven apart voor de man en vrouw.
- *Cirkeldiagram*, als we de relatieve frequenties willen zien. De grootte van elk punt staat voor het percentage van deze categorie van het totaal.
- *Pareto diagram* is een staafdiagram waarbij de staven van links naar rechts staan van hoogste frequentie naar laagste frequentie.

1.5 Grafieken om numerieke variabelen te beschrijven

Een *frequentieverdeling* voor numerieke gegevens is een tabel waarin de gegevens samengevat orden door ze te verdelen in bepaalde klassen. Het maken van een frequentieverdeling:

1. Bepaal k, het aantal klassen
2. De klassen moeten allemaal dezelfde breedte hebben, deze wordt berekend door: (Hoogste observatie – laagste observatie) / aantal klassen. De klasse breedte moet altijd omhoog afgerond worden.
3. De klassen moeten inclusief en niet-overlappend zijn. Elke waarneming moet behoren tot één en slechts één klasse. Overlappend wanneer een klasse 30-40 en die daarna 40-50.

Cumulatieve frequentieverdeling is het optellen van de frequentie van een klasse met die van de klassen daarboven.

Relatieve cumulatieve frequentieverdeling zijn de cumulatieve frequenties als percentage van het totaal.

Histogram, bestaat uit staven die aan elkaar vast zitten. De hoogte van de staaf geeft de hoogte van de frequentie aan. Op de horizontale as worden de intervallen weergeven en op de verticale as de frequentie.

Ogive (cumulatieve lijndiagram) geeft de cumulatieve percentages weer in een grafiek. De punten worden verbonden en zo ontstaat er een lijn. Vorm van een verdeling (histogram). Visueel vaststellen of de gegevens gelijkmatig van het midden (centrum) zijn verspreid. De vorm van de verdeling is *symmetrisch* wanneer de observaties ongeveer gelijkmatig verdeeld zijn over het midden. Het midden van de data verdeelt een grafiek in twee "spiegelbeelden". Een scheve verdeling kan naar links zijn of naar rechts. Een scheve (*asymmetrisch*) rechtsverdeling heeft een staart naar rechts. Een scheve linker verdeling een staart naar links. Verdeling van inkomen is vaak scheef naar rechts. Cijfers voor examens zijn vaak scheef naar links. *Stamdiagram* (stem-and-leaf display). De gegevens worden gegroepeerd op basis van hun eerste cijfer(s) (de stam genoemd) en hun laatste cijfers (de bladeren). De lengte van de regel is de frequentie van die klasse-interval. *Scatterdiagram* wordt gebruikt om mogelijke relaties tussen twee numerieke variabelen te onderzoeken. We kunnen een scatterdiagram bereiden door het lokaliseren van één punt voor elk paar van twee variabelen.

1.6 Fouten bij presenteren van gegevens

Tabellen en grafieken moeten overtuigend, duidelijk en waarheidsgetrouw zijn. Kijken naar de vorm van een lijn alleen is ontoereikend voor het verkrijgen van een duidelijk beeld van de data i.v.m. eventuele misleidende tijdreeksen.

Numerieke maten gebruiken om gegevens te beschrijven – Chapter 2

2.1 Maten van centrale tendens en locatie

Numerieke maten (het gemiddelde, de mediaan en de modus) als antwoord op vragen over de locatie van het centrum van een set gegevens. Deze numerieke maten verstrekken informatie over een "typische" observatie in de gegevens en worden aangeduid als de maten van de centrale tendens.

Rekenkundig gemiddelde (arithmetic mean = μ) is de som van alle waarden, gedeeld door het totaal aantal waarnemingen.

μ = $\alpha = \frac{x1+x2+x3\ldots xn}{n}$

Geometrisch gemiddelde, $\acute{X}_g$ is de nde wortel van het product van n nummers:

$\acute{X}_g = \sqrt[n]{X1+x2+\ldots.xn}$

Meetkundig gemiddelde rendement geeft het gemiddelde percentage rendement van een investering door de tijd heen. $r'_g = (x1+x2+\ldots xn)^{n-1}-1$

Mediaan is de middelste waarneming van een reeks gerangschikte waarnemingen. Als de steekproefomvang een even getal is, dan is de mediaan het gemiddelde van de twee middelste waarnemingen.

Modus (indien aanwezig) is de waarde die het meeste voorkomt. Unimodaal is er één modus, bimodaal twee modi en met meer dan twee modi is de verdeling multimodaal. De modus wordt meestal gebruikt bij categorische data.

Categorische gegevens worden het best beschreven door de mediaan of de modus. De modus kan niet de echte centrum van numerieke data vertegenwoordigen.

Numerieke gegevens worden meestal het beste beschreven door de gemiddelde. Het gemiddelde wordt wel beïnvloed door uitschieters. De mediaan is minder gevoelig voor extreme uitschieters.

Vorm van verdeling beschrijven door het berekenen van een maat voor de scheefheid. *Scheefheid* is positief als de verdeling scheef naar rechts is, negatief als de verdeling scheef naar links is en nul als de verdeling symmetrisch is (klokvormige distributie). In een symmetrische verdeling zijn het gemiddelde en de mediaan gelijk.

Bij continue numerieke unimodale data is het gemiddelde meestal minder dan de mediaan in een scheve-linksverdeling en het gemiddelde is hoger dan de mediaan in een scheve-rechtsverdeling.

Percentielen en kwartielen zijn maten die de locatie van een waarde ten opzichte van de gehele set van gegevens laat zien. Om percentielen te vinden moeten de gegevens gerangschikt worden van de kleinste tot de grootste waarde. Pth percentiel = de waarde van de (P/100)(n+1)de van de geordende posities.

Kwartielen scheiden grote datasets in vier kwarten.

Q1 = de waarde van de 0,25(n+1)ste positie

Q2 = de waarde van de 0,50(n+1)ste positie

Q3 = de waarde van de 0,75(n+1)ste positie

De vijf-nummer samenvatting verwijst naar de vijf beschrijvende maten: minimum, eerste kwartiel, mediaan, derde kwartiel, en maximum.

2.2 Maten van variabiliteit

Binnen twee datasets met hetzelfde gemiddelde kunnen de waarnemingen in een set meer variëren dan in de andere. De *range* is het verschil tussen de grootste en kleinste observatie. De range is echter geen goede maat voor de variabiliteit omdat deze beïnvloedt wordt door uitschieters. Om dit te voorkomen kunnen we de laagste 25% en de hoogste 25% van de gegevens elimineren. Zo kunnen we de spreiding van de middelste 50% van de gegevens meten. Hiervoor gebruiken we de *interkwartiel afstand,* deze berekenen we door Q3 – Q1.

Een *box-and-whisker plot* (boxplot) is een grafiek die de vorm van de verdeling in termen van de vijf-nummer samenvatting. Het middelste blok laat de gegevens zien tussen het eerste kwartiel en het derde kwartiel. De box wordt in tweeën gesplitst op de mediaan door een lijn. Er zijn twee "whiskers" (lijnen): eentje van het minimum tot het eerste kwartiel en de andere van het derde kwartiel naar het maximum.

Variantie berekenen:

De *populatievariantie*, σ^2 , is de som van de gekwadrateerde verschillen tussen elke waarneming en populatiegemiddelde gedeeld door de populatiegrootte, N:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

De *steekproef variantie*, s^2 , is de som van de gekwadrateerde verschillen tussen de waarneming en de steekproef gemiddelde gedeeld door de steekproefomvang, n, min 1:

$$s^2 = \frac{\sum_{i=1}^{N} (\chi_i - \bar{\chi})^2}{n-1}$$

Standaardafwijking:

De populatie standaardafwijking, σ, is de wortel van de populatievariantie.

De steekproef standaardafwijking, s, is de wortel van de steekproef variantie.

De variatiecoëfficiënt is de standaardafwijking als percentage van het gemiddelde.

Populatie variatiecoëfficiënt: $CV = \frac{\sigma}{X} \times 100\%$ als $\mu > 0$

Steekproef variatiecoëfficiënt: $CV = \frac{s}{\bar{\chi}} \times 100\%$ als $\bar{\chi} > 0$

Chebyshey's theorem: stelt dat het aantal van de observaties in een steekproef die binnen k standaardafwijkingen van het gemiddelde afliggen op zijn minst $100(1 - \frac{1}{k^2})\%$ als k > 1.

K staat voor hoeveel standaardafwijkingen van het gemiddelde verwijderd.

Wanneer een histogram normaal verdeeld is kan met behulp van de *Empirical Rule* aannames maken hoeveel procent van de observaties binnen één, twee of drie standaardafwijkingen van het gemiddelde zijn.

Ongeveer 68% van de waarnemingen zijn binnen 1 standaardafwijking, ongeveer 95% binnen 2 standaardafwijkingen en bijna alle waarnemingen binnen 3 afwijkingen. Wanneer een waarneming meer verschilt van het gemiddelde dan 3 standaardafwijkingen dan is deze een uitschieter.

Z-score is een waarde die het aantal standaard afwijkingen aangeeft die een waarde van het gemiddelde af zit. Z-score groter dan nul geeft aan dat de waarde groter is dan het gemiddelde; z-score kleiner dan nul geeft aan dat de waarde minder is dan het gemiddelde; z-score is nul geeft aan dat de waarde gelijk is aan het gemiddelde.

Z-score wordt vaak gebruikt bij toelatingstoetsen voor hogescholen en universiteiten.

$$z=\frac{\chi_i-\mu}{\sigma}$$

2.4 Metingen over relaties tussen variabelen

Covariantie is een maat voor de richting van een lineaire relatie tussen twee variabelen. Een positieve waarde geeft een directe of toenemende lineaire relatie aan en een negatieve waarde geeft een afnemende lineaire relatie aan.

Populatie covariantie: $Cov(x,y)=\sigma_{xy}=\frac{\sum_{i=1}^{N}(\chi_i-\mu_x)(y_i-\mu_y)}{N}$

Steekproef covariantie: $Cov(x,y)=s_{xy}=\frac{\sum_{i=1}^{N}(\chi_i-\bar{\chi})(y_i-\bar{y})}{n-1}$

Covariantie geeft geen maat voor de sterkte van het verband tussen twee variabelen, de *correlatiecoëfficiënt* doet dit wel. Deze wordt berekend door de covariantie te delen door het product van de standaard afwijkingen van de twee variabelen.

Populatie correlatiecoëfficiënt: $\rho=\frac{cov(x,y)}{\sigma_x\sigma_y}$

Steekproef correlatiecoëfficiënt: $r=\frac{cov(x,y)}{s_x s_y}$

De correlatiecoëfficiënt varieert van -1 tot +1.

Wanneer r = 0, dan is er geen lineaire relatie tussen x en y.

Belangrijk is om te begrijpen dat de correlatie geen oorzaken impliceert. Als twee variabelen sterk gecorreleerd zijn hoeft dat niet te betekenen dat de ene variabele de andere variabele veroorzaakt.

Toeval elementen: Waarschijnlijkheidsmethoden – Chapter 3

3.1 Random experiment, uitkomsten, en gebeurtenissen

Een random experiment is een proces dat leidt tot twee of meer mogelijke uitkomsten, niet wetende welke uitkomst zal plaatsvinden. Bijvoorbeeld het opgooien van een muntje, de uitkomst is kop of munt.

De mogelijke uitkomsten van een random experiment worden de basis uitkomsten genoemd. Alle basis uitkomsten bij elkaar noemen we de steekproefruimte, deze geven we aan met S. Een basis uitkomst wordt aangegeven met een O.

Vaak zijn we geïnteresseerd in een deel van de basis uitkomsten en niet de individuele uitkomsten. We noemen dit deel van uitkomsten een event (E), dit is bepaalde gebeurtenis. Het nul event staat voor de afwezigheid van basis uitkomsten en wordt genoteerd als een O met een schuine streep erdoor.

Intersectie van events: Is het totaal van alle basis uitkomsten in S die horen bij event A en ook bij event B. Bij K events $E_1, E_2, \ldots, E_k$, hun intersectie, $E_1 \cap E_2 \cap \ldots \cap E_k$, is het totaal van alle basis uitkomsten die horen bij elke $E_i (i = 1,2,\ldots, K)$

Mutually exclusive events (elkaar uitsluitende gebeurtenissen): wanneer event A en event B geen gemeenschappelijke basis uitkomsten hebben. Hun intersectie, $A \cap B$, is dan leeg.

Union of events (unie van gebeurtenissen): als event A of event B plaatsvindt of als beide plaatsvinden. De unie is dan $A \cup B$

De events zijn *collectively exhaustive* (collectief uitputtend) wanneer de unie van een aantal events de hele steekproefomvang omvat. $E_1 \cup E_2 \cup \ldots \cup E_k = S$

De set van basis uitkomsten van een random experiment behorend tot S maar niet bij A wordt de *complement* van A genoemd en schrijven we als $\bar{A}$. Events A en $\bar{A}$ zijn mutually exclusive.

Voorbeeld. Een honkbalwedstrijd, twee gebeurtenissen (events) waarin we geïnteresseerd zijn: "De slagman bereikt het honk" $(Event\ A[O_1, O_2, O_6])$ en "De slagman raakt de bal" $(Event\ B[O_1, O_4, O_5, O_6])$:

1. Complementen van de events: $\bar{A} = [O_3, O_4, O_5]$ en $\bar{B} = [O_2, O_3]$
2. Intersectie van de events: $A \cap B = [O_1, O_6]$
3. Unie van de events: $A \cup B = [O_1, O_2, O_4, O_5, O_6]$
4. Mutually exclusive: events A en $\bar{A}$, events B en $\bar{B}$

3.2 Waarschijnlijkheid (drie definities)

Klassieke waarschijnlijkheid: hoe vaak een event zal plaatsvinden, er vanuit gaande dat alle uitkomsten in een steekproef gelijke kansen hebben om te gebeuren.

De kans op gebeurtenis A is: $P(A) = \frac{N_A}{N}$

Het tellen van alle uitkomsten is erg tijdrovend als we eerst alle mogelijke uitkomsten moeten vinden. Daarom is er een formule om te bepalen hoeveel combinaties mogelijk zijn.

Permutaties en combinaties:

1. *Aantal mogelijke ordeningen van x objecten*: $x(x-1)(x-2)...(2)(1)=x!$ We hebben x objecten, bij het plaatsen van objecten denken we aan een rij dozen. In de eerste doos kunnen x verschillende objecten. We hebben dus een object in de eerste doos gedaan, dus er kunnen nog maar (x-1) objecten in de tweede doos. Bij de laatste doos is er nog maar 1 object over.
2. *Permutaties*: Bij permutaties kijken we naar x objecten gekozen uit n. We hebben dus meer objecten dan dozen (n > x). De eerste box kan op n manieren gevuld worden, de tweede box op (n-1) manieren, de laatste box op (n-x+1) manieren. Er blijven (n-x) objecten over.
   $$P_x^n=\frac{n!}{(n-x)!}$$
3. *Combinaties*: Het aantal mogelijke combinaties x objecten gekozen uit n. $C_x^n=\frac{n!}{x!(n-x)!}$

*Relatieve Frequentie***:** Hoe vaak een event voorkomt vergeleken met andere events. Een event met een waarschijnlijkheid van 0.40 komt vaker voor dan een event met waarschijnlijkheid 0.30.

Subjectieve waarschijnlijkheid: in hoeverre een individu gelooft in de kans dat een event zal gebeuren.

Waarschijnlijkheidsvereisten:

1. $0\leqslant P(A)\leqslant 1$ 2. $P(A)=\sum_A P(O_i)$ 3. $P(S)=1$

1. De waarschijnlijkheid moet tussen 0 en 1 liggen.
2. In termen van relatieve frequenties; de som van kansen op alle basis uitkomsten van een event is gelijk aan de kans op dat event.
3. De som van de kansen voor de basis uitkomsten in de steekproef is gelijk aan 1.

3.3 Kansregels:

> Complement regel: $P(\bar{A})=1-P(A)$

> Toevoegingsregel: Kans op union: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

> Voorwaardelijke kans: Kans op event A, met het gegeven dat event B al is gebeurd:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}\ provided\ that\ P(B)>0$$

> Vermenigvuldigingsregel: De kans op de intersectie van event A en event B kan worden afgeleid van de voorwaardelijke kans: $P(A\cap B)=P(A|B)P(B)$

> Statistische *onafhankelijkheid* als: $P(A\cap B)=P(A)P(B)$ of $P(A|B)=P(A)(if\ P(B)>0)$

3.4 Bivariate kans

Kruistabel waarbij verschillende metingen worden gedaan over een zelfde steekproefgroep. Bijvoorbeeld de hoeveelheid televisie kijken en het inkomen van mensen. Vaak wordt hiervoor een boom diagram gebruikt.

Gezamenlijke kans: de intersectie kans: $P(A_i\cap B_v)$

Marginale kans: Kansen van de individuele events: $P(A_i)$ of $P(B_v)$

Voorwaardelijke kans: $P(A_1|B_1)=\frac{P(A_1\cap B_1)}{P(B_1)}$

Odds (kansen): De ratio van de kans op een event gedeeld door de kans op z'n complement. De kansen in het voordeel van A zijn: $Odds = \frac{P(A)}{1-P(A)} = \frac{P(A)}{P(\bar{A})}$

Overbetrokkenheid ratio: Wanneer een event invloed heeft op de uitkomst voorwaarde, wanneer de ratio van de voorwaardelijke kans niet gelijk is aan 1. Voorbeeld:

$B_1 : Kopers$ $A_1 : De\ reclame\ gezien$

$B_2 : niet-kopers$ $A_2 : De\ reclame\ niet\ gezien$

De kans dat de koper de reclame heeft gezien : $\frac{P(B_1|A_1)}{P(B_2|A_1)}$

Adverteren is effectief wanneer: $\frac{P(B_1|A_1)}{P(B_2|A_1)} > \frac{P(B_1)}{P(B_2)}$

3.5 Theorema van Bayes

Herziening conditionele kansen door het gebruik van beschikbare informatie en een werkwijze om te bepalen hoe kansen worden aangepast aan aanvullende informatie.

Herhaling vermenigvuldigingsregel: $P(A_1 \cap B_1) = P(A_1|B_1)P(B_1)$

Stappen Bayes' Theorema:

1. Bepaal de verschillende events van het probleem
2. Bepaal de kansen en voorwaardelijke kansen voor de events
3. Bereken de complementen van de kansen
4. Pas Bayes' theorema toe om de oplossing van de kans te berekenen.

De voorwaardelijke kans van E_i , gegeven A:

$$P(E_i|A_1) = \frac{P(A_1|E_i)P(E_i)}{P(A_1|E_1)P(E_1) + P(A_1|E_2)P(E_2) + \ldots + P(A_1|E_K)P(E_K)}$$

Published on *WorldSupporter* (www.worldsupporter.org)

Discrete kansverdelingen – Chapter 4

4.1 Random variabelen

Een *random variabele* is een variabele die de numerieke waarden aanneemt uit een steekproefruimte gegenereerd door een random experiment. Een *discrete random variabele* kan niet meer dan telbare waarden aannemen (eindig aantal waarden). Bijvoorbeeld het aantal defecte items in een steekproef van een grote levering, aantal mensen die uitchecken binnen een uur.

Een *continue random variabele* kan elke waarde binnen een interval aannemen (ontelbaar). Bijvoorbeeld temperatuur, tijd, inkomen.

4.2 Kansverdeling voor discrete random variabele

De *kansverdeling functie*, P(x), van een discrete random variabele X staat voor de kans dat X de waarde x aanneemt, als een functie van x. $P(x)=P(X=x), voor\ alle\ waarden\ van\ x$

Kansverdeling grafiek: verticaal P(x), horizontaal x.

Vereiste eigenschappen van kansverdeling discrete random variabelen:

1. $0 \leqslant P(x) \leqslant 1\ for\ any\ value\ x$
2. De som van de individuele kansen is 1: $\sum_x P(x)=1$

Cumulatieve kansverdeling, $F(x_0)$, is het optellen van kansen. Als de kans $P(x_0)=0.15$ en de kans $P(x_1)=0.30$ dan is de cumulatieve kans $F(x_1)=P(x_0)+P(x_1)=0.45$

Dus: $F(x_0)=P(X \leqslant x_0)$

Vereiste eigenschappen van cumulatieve kansverdeling discrete random variabelen:

1. $0 \leqslant F(x_0) \leqslant 1\ voor\ elk\ getal\ x_0$
2. $als\ x_0\ en\ x_1\ twee\ getallen\ zijn\ met\ x_0 < x_1, dan\ F(x_0) \leqslant F(x_1)$

4.3 Eigenschappen van discrete random variabelen

Verwachte waarde (expected value) $E[X]$ van een discrete random variabele X wordt ook wel het gemiddelde μ genoemd en wordt gedefinieerd als: $E[X]=\mu=\sum_x xP(x)$

De *variantie* van een discrete random variabele X kan worden uitgedrukt als:
$\sigma^2=E[(X-\mu)^2]=\sum_x (x-\mu)^2 P(x)$

De *standaardafwijking*, σ, is de wortel van de variantie.

Verwachte waarde van functies van random variabelen: $E[g(X)]=\sum_x g(x)P(x)$

Eigenschappen voor lineaire functies van een random variabele: De random variabele Y is a+bX.

Het gemiddelde: $\mu_Y=E[a+bX]=a+b\mu_X$

De variantie: $\sigma_Y^2=Var(a+bX)=b^2\sigma_X^2$

Standaardafwijking: $\sigma_Y=|b|\sigma_X$

4.4 Binomiale verdeling

Bernoulli verdeling: Een random experiment waarbij twee mutually exclusive uitkomsten mogelijk zijn. We labelen deze uitkomsten "succes" en "falen". De kans op succes is P en de kans op falen is (1-P). X is 1 wanneer de uitkomst van het experiment succes is, anders is X 0. De kansverdeling van de random variabele is dan: $P(0)=(1-P)\,en\,P(1)=P$

Gemiddelde: $\mu_X=E[X]=\sum_x xP(x)$

Variantie: $\sigma^2=E[(X-\mu)^2]=\sum_x (x-\mu)^2 P(x)$

Aantal rijen met x successen in n onafhankelijke proeven:

$C_x^n=\frac{n!}{x!(n-x)!}$ waar $n!=nx(n-1)\times(n-2)\times\ldots\times 1$ en $0!=1$

De binomiale verdeling: P(x successen in n onafhankelijke proeven) =

$P(x)=\frac{n!}{x!(n-x)!}P^x(1-P)^{(n-x)}$ voor x = 0,1,2,....,n.

Gemiddelde: $\mu=E[X]=nP$

Variantie: $\sigma_X^2=E[(X-\mu_X)^2]=nP(1-P)$

Binominale kansverdeling:

1. Meerdere steekproeven, elke steekproef heeft maar twee uitkomsten: succes of falen.
2. De kans op de uitkomst is hetzelfde voor elke proef
3. De kans op de uitkomst in een proef heeft geen invloed op de kans in andere proeven.

Voorbeeld:

Fleur is makelaar, ze heeft 5 contacten en ze gelooft dat voor elk contact de kans op verkoop 0.40 is. Wat is de kans dat ze op z'n meest aan 1 verkoopt?

$P(X\leqslant 1)=P(X=0)+P(X=1)$

$P(0\,verkopen)=P(0)=\frac{5!}{0!\,5!}(0.4)^0(0.6)^5=0.078$

$P(1\,verkoop)=P(1)=\frac{5!}{1!\,4!}(0.4)^1(0.6)^4=0.259$ dus $P(X\leqslant 1)=0.078+0.259=0.337$

4.7 Gezamenlijk verdeelde discrete random variabelen

Het is belangrijk dat kansmodellen het gezamenlijke effect van variabelen op kansen reflecteren. Bijvoorbeeld: het kopen van luxe kookspullen is verschillend voor verschillende leeftijdgroepen. In paragraaf 3.4 hadden we het over de kans op intersectie van bivariate events, $P(A_i\cap B_j)$. Hier gebruiken we random variabelen: $P(x,y)=P(X=x\cap Y=y)$

Marginale kansverdeling: $P(x)=\sum_y P(x,y)$ en $P(y)=\sum_x P(x,y)$

Eigenschappen gezamenlijke kansen:

- $0\leqslant P(x,y)\leqslant 1\,voor\,elk\;paar\,van\,waarden\,x\,en\,y$
- som van de gezamenlijke kansen P(x,y) van alle mogelijke paren van waarden moet 1 zijn.

Voorwaardelijke kansverdeling: $p(y|x)=\frac{p(x,y)}{p(x)}$ en $p(x|y)=\frac{p(x,y)}{p(y)}$

De gezamenlijk verdeelde random variabelen X en Y zijn onafhankelijk wanneer de gezamenlijke kansverdeling het product is van hun marginale kansverdeling: $P(x,y)=P(x)P(y)$

Verwachtte waarden van functies van gezamenlijk verdeelde random variabelen:

$$E[g(X,Y)]=\sum_x \sum_y g(x,y)P(x,y)$$

Covariantie: geeft de lineaire samenhang weer tussen twee random variabelen, de richting van de relatie. De verwachtte waarde van $(X-\mu_x)(Y-\mu_Y)$ wordt de covariantie tussen X en Y genoemd, deze wordt geschreven als $Cov(X,Y)$

$$Cov(X,Y)=E[(X-\mu_x)(Y-\mu_Y)]=\sum_x \sum_y (X-\mu_x)(Y-\mu_Y)P(x,y)$$

Correlatie: een maat voor de sterkte van de lineaire relatie tussen twee random variabelen, de maatregel beperkt tot het bereik van -1 tot +1.

$$p=Corr(X,Y)=\frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Wanneer twee random variabelen statistisch onafhankelijk zijn dan is de covariantie tussen hun 0.

Portfolio analyse: De lineaire combinatie van de gemiddelde waarden van de effecten in de portfolio. Voorbeeld: Het portfolio bestaat uit a aandelen van voorraad A en b aandelen uit voorraad B. We willen het gemiddelde en de variantie gebruiken voor de marktwaard, W, van een portfolio, waar W de lineaire functie $W=aX+bY$ is.

Gemiddelde van W: $\mu_w=E[W]=E[aX+bY]=a\mu_x+b\mu_y$

Variantie van W:

$\sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2abCov(X,Y)$ of $\sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2abCorr(X,Y)\sigma_X\sigma_Y$

Continue kans – Chapter 5

5.1 Continue random variabelen

Cumulatieve verdelingsfunctie, (F(x), voor een continue random variabele X drukt de kans uit dat X niet hoger is dan de waarde van x, als een functie van x: $P(X \leqslant x) = F(x)$

Grafiek: verticale as f(x), horizontale as x.

Kans dat een continue random variabele in een bepaalde reeks valt: $P(a < X < b) = F(b) - F(a)$

Voorbeeld: de cumulatieve verdelingsfunctie in de reeks is F(x) = 0.001x. De kans dat verkopen tussen 250 en 750 liter is: $P(250 < X < 750) = (0.001)(750) - (0.001)(250) = 0.75 - 0.25 = 0.50$

Kansdichtheid functie eigenschappen:

1. f(x) > 0 voor alle waarden van x
2. Het gebied onder de kansdichtheidsfunctie, f(x), voor alle waarden van de random variabele, X binnen zijn reeks, is gelijk aan 1.
3. Grafiek dichtheidsfunctie: a en b zijn de twee mogelijke waarden van random variabele X, met a < b. De kans dat X ligt tussen a en b is het gebied onder de kansdichtheidsfunctie tussen deze punten: $P(a \leqslant X \leqslant b) = \int_a^b f(x)\,dx$
4. De cumulatieve verdelingsfunctie, $F(x_0)$, is het gebied onder de kansdichtheidsfunctie, f(x), tot x_0 : $F(x_0) = \int_{x_m}^{x_0} f(x)\,dx$

5.2 Verwachtingen voor continue random variabelen

Een *uniforme verdeling* over de reeks van a tot b: $f(x) = \frac{1}{b-a} \qquad a \leqslant X \leqslant b$

Gemiddelde: $\mu_X = E[X] = \frac{a+b}{2}$

Variantie: $\sigma_X^2 = E[(X - \mu_X)^2] = \frac{(b-a)^2}{12}$ → standaardafwijking: $\sigma_X = \sqrt{\sigma_X^2}$

Lineaire functies van random variabelen: $W = a + bX$

$\mu_W = E[a+bX] = a + b\mu_X \qquad \sigma_W^2 = Var[a+bX] = b^2\sigma_X^2 \qquad \sigma_W = |b|\sigma_X$

5.3 De normale verdeling

Eigenschappen *normale verdeling*:

1. Gemiddelde van de random variabele is $E[X] = \mu$
2. De variantie van de random variabele is $Var(X) = \sigma^2$
3. De vorm van de kansdichtheidsfunctie is een symmetrische klok-vormige kromme met het gemiddelde als centrum.
4. Notatie normale verdeling: $X \sim N(\mu, \sigma^2)$

Cumulatieve verdelingsfunctie van de normale verdeling: $F(x_0) = P(X \leqslant x_0)$ Dit is het gebeid onder de normale kansdichtheidsfunctie links van x_0 . Totale gebeid onder kromme is 1.

Kans dat X tussen a en b ligt: $P(a<X<b)=F(b)-F(a)$

Standaard normale verdeling: Als Z een normaal random variabele is met gemiddelde 0 en variantie 1: $Z\sim N(0,1)$. Z volgt de standaard normale verdeling.

Relatie tussen een normaal verdeelde random variabele en Z: $Z=\frac{X-\mu}{\sigma}$

Voorbeeld: Appa heeft een investeringsportfolio met gemiddelde waarde van €1.000.000 met een standaardafwijking van €20.000. Hij wil weten wat de kans is dat de waarde van zijn portfolio tussen €970.000 en €1.060.000 is.

1. Bereken de Z waarden:

$$Z_{970.000}=\frac{970.000-1.000.000}{30.000}=-1.0 \qquad Z_{1.060.000}=\frac{1.060.000-1.000.000}{30.000}=+2.0$$

2. De kans dat de portfolio waarde, X, tussen €970.000 en €1.060.000 ligt, is gelijk aan de kans dat Z tussen -1 en +2 ligt: $P(970.000\leqslant X\leqslant 1.060.000)=P(-1\leqslant Z\leqslant +2)$
3. $1-P(Z\leqslant -1)-P(Z\geqslant +2)=1-0{,}1587-0{,}0228=0{,}8185$

Andere manier: $P(a<X<b)=P(\frac{a-\mu}{sigma}<Z<\frac{b-\mu}{sigma})$

5.4 Normaal verdeelde benadering voor binominale verdeling

$X=X_1+X_2+...+X_n$ met kans op succes P en kans op falen 1-P

Gemiddelde: $E[X]=\mu=nP$

Variantie: $Var(X)=\sigma^2=nP(1-P)$

Wanneer het aantal proeven n groot is, zoals $nP(1-P)>5$, dan is de verdeling van de random variabele benaderd als een standaard normale verdeling. $Z=\frac{X-E[X]}{\sqrt{Var(X)}}=\frac{X-nP}{\sqrt{nP(1-P)}}$

Kans aantal keer succes tussen a en b: $P(a\leqslant X\leqslant b)=P(\frac{a-nP}{\sqrt{nP(1-P)}}\leqslant Z\leqslant \frac{b-nP}{\sqrt{nP(1-P)}})$

Kansen uit percentage intervallen. Een proportie random variabele, P, kan worden berekend door het aantal successen, X, te delen door de steekproefgrootte, n.

$\mu=P$ en $\sigma^2=P(1-P)/n$

5.6 Gezamenlijk verdeelde continue random variabelen

De *gezamenlijke cumulatieve verdeling* van continue random variabelen, $F(x_1,x_2,...,x_K)$ staat voor de kans dat tegelijkertijd X_1 is lager dan x_1 , X_2 is lager dan x_2 , enzovoort.

Dus: $F(x_1,x_2,...,x_K)=P(X_1<x_1\cap X_2<x_2\cap...\cap X_K<x_K)$

Covariantie: $Cov(X,Y)=E[XY]-\mu_X\mu_Y$

Correlatie: $p=Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$

Som van gemiddelde: $E[(X_1+X_2+...+X_K)]=\mu_1+\mu_2+...+\mu_K$

Som van variantie wanneer covariantie 0: $Var(X_1+X_2+...+X_K)=\sigma_1^2+\sigma_2^2+...+\sigma_K^2$

wanneer covariantie niet 0: $Var(X_1+X_2+...+X_K)=\sigma_1^2+\sigma_2^2+...+\sigma_K^2+2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}Cov(X_i,X_j)$

Verschillen tussen paren random variabelen:

1. Gemiddelde: $E[X-Y]=\mu_X-\mu_Y$
2. Variantie wanneer covariantie 0: $Var(X-Y)=\sigma_X^2+\sigma_Y^2$
3. Variantie wanneer covariantie niet 0: $Var(X-Y)=\sigma_X^2+\sigma_Y^2-2\,Cov(X,Y)$

Lineaire combinaties van random variabelen:

$W=aX+bY \qquad \mu_W=E[W]=E[aX+bY]=a\mu_X+b\mu_Y \qquad \sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2+2abCov(X,Y)$

$W=aX-bY \qquad \mu_W=E[W]=E[aX-bY]=a\mu_X-b\mu_Y \qquad \sigma_W^2=a^2\sigma_X^2+b^2\sigma_Y^2-2abCov(X,Y)$

Als X en Y gezamenlijk normaal verdeelde random variabelen zijn, dan is W ook normaal verdeeld.

Verdelingen van steekproef statistieken – Chapter 6

6.1 Steekproeven van een populatie

Bij een *aselecte steekproef* worden n objecten uit een populatie zodanig gekozen dat elk lid van de populatie dezelfde kans heeft om geselecteerd te worden. De keuze van het ene lid is onafhankelijk van de keuze van een ander lid en elke mogelijke uitkomst heeft dezelfde kans. Het is belangrijk dat een steekproef de hele populatie vertegenwoordigt. Het is erg moeilijk om elk item in een populatie te meten en dat is bovendien erg duur.

Steekproef verdeling voorbeeld:

Florine is een supervisor over zes werknemers met verschillende jaren aan ervaring: 2,4,6,6,7,8.

Het gemiddelde: $\mu=(2+4+6+6+7+8)/6=5.5$

Twee van de werknemers worden random gekozen voor een bepaalde werkgroep. In dit voorbeeld gaan we uit van een steekproef zonder terugleggen, van een kleine populatie. Als er sprake was van een grote populatie van bijvoorbeeld 1000 werknemers dan hadden we geen rekening hoeven houden met terug leggen, dit is verwaarloosbaar. Er zijn 15 mogelijke uitkomsten (combinaties), elke uitkomst heeft dus een kans van 1/15 om geselecteerd te worden. Bij uitkomst 2&8 is het gemiddelde 5, dit gemiddelde is hetzelfde bij 2 andere combinaties. De kans op een uitkomst met gemiddelde 5 is dus 3/15.

6.2 Steekproef verdeling van steekproef gemiddelden

Steekproef gemiddelde van random variabelen $X_1, X2, \ldots, X_n$ is: $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$

De steekproefverdeling van de steekproef gemiddelden is het populatiegemiddelde:

$$E[\bar{X}]=E[\frac{1}{n}(X_1+X_2+\ldots+X_n)]=\frac{n\mu}{n}=\mu \qquad Var(\bar{X})=\frac{\sigma^2}{n} \qquad \sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$$

Wanneer de steekproefgrootte, n, niet klein is vergeleken met de populatiegrootte, N, dan is de standaard error van $\bar{X}$ als volgt: $\sigma_x=\frac{\sigma}{\sqrt{n}}\times\sqrt{\frac{N-n}{N-1}}$

Standaard normale verdeling van de steekproef gemiddelden: $Z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}$

Central Limit Theorem: $X_1, X_2, \ldots, X_n$ is een set van n onafhankelijke random variabelen met gelijke verdelingen met gemiddelde μ, variantie σ^2 en $\bar{X}$ als het gemiddelde van deze random variabelen. Als n groot is, de central limit theorem zegt dat de verdeling van $Z=\frac{\bar{X}-\mu_X}{\sigma_{\bar{X}}}$ nadert de standaard normale verdeling.

Acceptatie interval: een interval waarbinnen een steekproef gemiddelde een hoge kans van waarschijnlijkheid kent. Wanneer het steekproef gemiddelde binnen de interval ligt,, dan kunnen we de conclusie accepteren dat de random steekproef gekomen is uit de populatie met bekende populatie gemiddelde en variantie. Symmetrische acceptatie interval: $\mu\pm z_{\alpha/2}\sigma_{\bar{x}}$.

6.3 Steekproef verdelingen van steekproef proporties

Steekproef proportie: $\hat{p}$ is de proportie van de populatieleden die een karakteristiek bezitten waarin interesse is. De steekproef proportie is: $\hat{p}=X/n$.

Steekproefverdeling: $\hat{p}$ is de steekproef proportie van successen in een random steekproef van een populatie met proportie van successen P.

1. De steekproefverdeling van $\hat{p}$ heeft het gemiddelde P: $E[\hat{p}]=P$
2. De steekproefverdeling van $\hat{p}$ heeft een standaardafwijking: $\sigma_{\hat{p}}=\sqrt{\frac{P(1-P)}{n}}$
3. Wanneer de steekproefgrootte groot is, de random variabele: $Z=\frac{\hat{p}-P}{\sigma_{\hat{p}}}$ is verdeeld als standaard normaal. De benadering is goed wanneer: $nP(1-P)>5$

6.4 Steekproefverdeling van steekproef variaties

Steekproef variantie: $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ *Steekproef standaardafwijking*: $s=\sqrt{s^2}$

Steekproefverdeling: s^2 is de steekproef variantie voor een random steekproef van n observaties uit een populatie met een variantie σ^2 :

1. De steekproefverdeling van s^2 heeft het gemiddelde σ^2 : $E[s^2]=\sigma^2$
2. De variantie van een steekproefverdeling van s^2 hangt van van de onderliggende populatieverdeling. Als die verdeling normaal is, dan: $Var(s^2)=\frac{2\sigma^4}{n-1}$

Als de populatieverdeling normaal is, dan: $X^2_{n-1}=\frac{(n-1)s^2}{\sigma^2}$ is verdeeld als de chi-squared verdeling met n-1 vrijheidsgraden, $(X^2_{(n-1)})$.

Betrouwbaarheidsinterval schatting: één populatie – Chapter 7

7.1 Eigenschappen van punt schatters

Schatter van een populatie parameter: random variabele die afhangt van de steekproef informatie. Een specifieke waarde van die random variabele wordt een *schatting* genoemd. Bijvoorbeeld: Het steekproefgemiddelde $\bar{X}$ is een punt schatter van het populatiegemiddelde, μ , en de waarde die $\bar{X}$ aanneemt voor een gegeven dataset wordt de punt schatting genoemd.

Een punt schatter $\hat{\theta}$ is een *onpartijdige schatter* (unbiased estimator) van een populatie parameter θ als de verwachtte waarde gelijk is aan de parameter: $E(\hat{\theta})=\theta$

De *bias* in $\hat{\theta}$: $E(\hat{\theta})-\theta$ De bias van een onpartijdige schatter is 0.

Bij meerdere onpartijdige schatters van een parameter, is de onpartijdige schatter met de kleinste variantie de meest *efficiënte schatter*. $\hat{\theta}_1 \, meer \, efficiënt \, dan \, \hat{\theta}_2 \, wanneer : Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$

De *relatieve efficiëntie*: $\dfrac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$

7.2 Betrouwbaarheidsinterval schatting voor het gemiddelde van een normale verdeling: populatie variantie bekend.

Betrouwbaarheidsinterval schatting voor een populatie parameter is een regel om een interval te bepalen waarbij de kans groot is dat de parameter daarin zit.

De interval van a tot b is een $100(1-\alpha)\%$ betrouwbaarheidsinterval van θ . De hoeveelheid % wordt het betrouwbaarheidslevel van de interval genoemd. De betrouwbaarheidsinterval wordt geschreven als: $a<\theta<b \, , met \, 100(1-\alpha)\%$ betrouwbaarheid.

ME, de margin of error, is de error factor: $\hat{\theta} \pm ME$

Een random steekproef van n observaties van een normale verdeling met gemiddelde en variantie σ^2 . Als het steekproefgemiddelde $\bar{x}$ is, dan wordt een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor het populatie gemiddelde met bekende variantie gegeven door:

$$\bar{x} \pm ME \rightarrow ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad width \, , w = 2(ME) \qquad UCL = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad LCL = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Waarbij: ME; margin of error. UCL; Upper confidence limit. LCL; Lower confidence limit.

Verminderen margin of error door de standaardafwijking naar beneden te brengen, de steekproefgrootte omhoog te brengen of een lager betrouwbaarheidslevel.

7.3 Betrouwbaarheidsinterval schatting voor het gemiddelde van een normale verdeling: populatie variantie onbekend.

Student's t verdeling met (n-1) vrijheidsgraden: $t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$

Random variabele met de student's t verdeling met v vrijheidsgraden wordt aangegeven als t_v . Dan is $t_{v,\alpha/2}$ de betrouwbaarheidsfactor, gedefinieerd als het aantal waarvoor: $P(t_v > t_{v,\alpha/2}) = \alpha/2$.

Een random steekproef van n observaties van een normale verdeling met gemiddelde μ en variantie onbekend. Als het steekproefgemiddelde en de standaardafwijking $\bar{x}\,en\,s$ zijn, dan zijn de vrijheidsgraden v=n-1, en de $100(1-\alpha)\%$ *betrouwbaarheidsinterval voor het populatiegemiddelde* zonder de variantie te weten, gegeven door: $\bar{x}\pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ of $\bar{x}\pm ME$

7.4 Betrouwbaarheidsinterval schatting voor populatie proportie (grote steekproef)

Als $\hat{p}$ het waargenomen aandeel van successen in een aselecte steekproef van n waarnemingen uit een populatie met een deel van de successen P is. Dan, als $nP(1-P)>5$, een $100(1-\alpha)\%$ *betrouwbaarheidsinterval voor het populatie aandeel* is gegeven door: $\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ of $\hat{p}\pm ME$.

7.5 Betrouwbaarheidsinterval schatting voor de variantie van een normale verdeling

Een random steekproef met n waarnemingen uit een normaal verdeelde populatie met variantie σ^2 . Als de waargenomen steekproef variantie s^2 is, dan zijn de beneden en boven betrouwbaarheidslimieten van $100(1-\alpha)\%$ betrouwbaarheidsinterval voor de populatie variantie gegeven door: $LCL=\frac{(n-1)s^2}{x^2_{n-1,\alpha/2}}$ $UCL=\frac{(n-1)s^2}{x^2_{n-1,1-\alpha/2}}$

waar $x^2_{n-1,\alpha/2}$ is het getal waarvoor $P(x^2_{n-1}>x^2_{n-1,\alpha/2})=\frac{\alpha}{2}$

en $x^2_{n-1,1-\alpha/2}$ is het getal waarvoor $P(x^2_{n-1}<x^2_{n-1,1-\alpha/2})=\frac{\alpha}{2}$

De random variabele x^2_{n-1} volgt een chi-kwadraat verdeling met (n-1) vrijheidsgraden.

7.7 Steekproefgrootte bepalen: grote populaties

Steekproef grootte voor het gemiddelde van een normaal verdeelde populatie met bekende variantie. Een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor het populatie gemiddelde strekt zich over een afstand ME (sampling error) aan elke kant van het steekproefgemiddelde als de steekproefgrootte, n, als volgt is: $n=\frac{z^2_{\alpha/2}\sigma^2}{ME^2}$ = hoe groot de steekproefgrootte moet zijn om een bepaalde interval te bereiken.

Steekproefgrootte voor een deelpopulatie: $n=\frac{0{,}25(z_{\alpha/2})^2}{(ME)^2}$

7.8 Steekproefgrootte bepalen: Eindige populatie

$n=\frac{n_0 N}{n_0+(N-1)}$ met $n_0=\frac{z^2_{\alpha/2}\sigma^2}{ME^2}$ met $Var(\bar{X})=\sigma\frac{2}{X}=\frac{\sigma^2}{2}(\frac{N-n}{N-1})$

Steekproefgrootte voor een deelpopulatie: $n=\frac{NP(1-P)}{(N-1)\sigma^2 p+P(1-P)}$

Published on *WorldSupporter* (www.worldsupporter.org)

Betrouwbaarheidsinterval schatting – Chapter 8

8.1 Betrouwbaarheidsinterval schatting van het verschil tussen twee normale populatiegemiddelde: afhankelijke samples

Samples zijn afhankelijk als de waarden in het ene sample beïnvloedt worden door de waarden in een andere sample.

We hebben een random sample van n matched pairs van waarnemingen uit normale verdelingen met gemiddelden μ_x en μ_y . Dan zijn $x_1, x_2, \ldots, x_n$ de waarden van de waarnemingen uit de populatie met gemiddelde μ_x en $y_1, y_2, \ldots, y_n$ die van de populatie met gemiddelde μ_y

$\bar{d}$ staat voor het geobserveerde sample gemiddelde en s_d voor de standaardafwijking.

Het verschil tussen twee waarnemingen: $d_i = x_i - y_i$.

Als de populatie verdeling van de verschillen als normaal wordt beschouwd, dan is een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor het verschil tussen twee gemiddelden, afhankelijke samples $(\mu_d = \mu_x - \mu_y)$ wordt gegeven door $\bar{d} \pm t_{n-1,\,\alpha/2} \frac{s_d}{\sqrt{n}}$ ook wel $\bar{d} \pm ME$

De random variabele t_{n-1} heeft een Student's t verdeling met (n-1) vrijheidsgraden.

$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$ Waarbij $\bar{d}$ dus het gemiddelde van de verschillen is.

Voorbeeld:

Pair	Drug X	Drug Y	Difference $d_i = x_i - y_i$
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	30		
6	32	30	2
7	29	26	3
8	31	33	-2
9	30	36	-6

99% betrouwbaarheid level

$\bar{d} = 1.625 \quad s_d = 3.777 \quad t_{n-1,\alpha/2} = t_{7,0.005} = 3.499 \quad n = 8$

Bereken het verschil tussen de effectiviteit van drug X en drug Y:

$\bar{d} \pm t_{n-1,\,\alpha/2} \frac{s_d}{\sqrt{n}}$ → $1.625 \pm 3.499 \frac{3.777}{\sqrt{8}}$ → LCL = -3.05 en UCL = 6.30

Omdat het betrouwbaarheid level de nul bevat is het niet te bepalen of de ene drug effectiever is dan de andere.

Mogelijkheden:

$\mu_x - \mu_y$ kan positief zijn, dit betekend dat drug X effectiever is.

$\mu_x - \mu_y$ kan negatief zijn, dit betekend dat drug Y effectiever is.

$\mu_x - \mu_y$ kan nul zijn, dat zijn drug X en drug Y even effectief.

8.2 Betrouwbaarheidsinterval schatting van het verschil tussen twee normale populatiegemiddelden: onafhankelijke samples.

Drie situaties:

Twee gemiddelden, onafhankelijke samples, bekende populatie variaties.

Stel dat er twee onafhankelijke random samples met n_x en n_y waarnemingen uit normaal verdeelde populaties met gemiddelde μ_x en μ_y en variaties σ_x^2 en σ_y^2 Als de geobserveerde sample gemiddelden $\bar{x}$ en $\bar{y}$ zijn, dan is een $100(1-\alpha)\%$ betrouwbaarheid level voor het verschil tussen twee gemiddelden, onafhankelijke samples, en bekende variaties gegeven door:

$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$ ook wel $(\bar{x} - \bar{y}) \pm ME$

Twee gemiddelden, onafhankelijke samples, onbekende populatie variaties beschouwd als gelijk.

Stel dat er twee onafhankelijke random samples met n_x en n_y waarnemingen uit normaal verdeelde populaties met gemiddelde μ_x en μ_y en de variaties zijn onbekend, maar wel gelijk. Als de geobserveerde sample gemiddelden $\bar{x}$ en $\bar{y}$ zijn, en de geobserveerde sample variaties zijn s_x^2 en s_y^2 dan is een $100(1-\alpha)\%$ betrouwbaarheid level voor het verschil tussen twee gemiddelden, onafhankelijke samples, en onbekende variaties die als gelijk worden beschouwd gegeven door:

$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$ ook wel $(\bar{x} - \bar{y}) \pm ME$

waarbij de pooled sample variantie s_p^2 gegeven is door: $s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$

Twee gemiddelden, onafhankelijke samples, onbekende populatie variaties niet beschouwd als gelijk.

Stel dat er twee onafhankelijke random samples met n_x en n_y waarnemingen uit normaal verdeelde populaties met gemiddelde μ_x en μ_y en de variaties zijn onbekend, worden beschouwd als ongelijk. Als de geobserveerde sample gemiddelden $\bar{x}$ en $\bar{y}$ zijn, en de geobserveerde sample variaties zijn s_x^2 en s_y^2 dan is een $100(1-\alpha)\%$ betrouwbaarheid level voor het verschil tussen twee gemiddelden, onafhankelijke samples, en onbekende variaties die als ongelijk worden beschouwd gegeven door:

$(\bar{x}-\bar{y})\pm t_{v,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$ ook wel $(\bar{x}-\bar{y})\pm ME$

waarbij de vrijheidsgraden, v, gegeven is door: $v=\frac{[(\frac{s_x^2}{n_x})+(\frac{s_y^2}{n_y})]^2}{(\frac{s_x^2}{n_x})^2/(n_x-1)+(\frac{s_y^2}{s_y^2})^2/(n_y-1)}$

Als de sample groottes gelijk zijn dan dalen de vrijheidsgraden tot: $v=(1+\frac{2}{\frac{s_x^2}{s_y^2}+\frac{s_y^2}{s_x^2}})\times(n-1)$

8.3 Betrouwbaarheidsinterval schatting van het verschil tussen twee populatie proporties (grote samples)

P_x is het waargenomen deel successen in een random sample van n_x waarnemingen uit een populatie met proportie P_x successen. $\hat{p}_y$ Is het deel successen waargenomen in een onafhankelijke random sample van n_y waarnemingen uit een populatie met proportie P_y successen. Als de sample grootte groot is dan is een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor het verschil tussen populatie proporties (grote samples), (P_x-P_y) , gegeven door:

$(\hat{p}_x-\hat{p}_y)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$ ook wel $(\hat{p}_x-\hat{p}_y)\pm ME$

Hypothese testen van een enkele populatie – Chapter 9

9.1 Concepten van hypothese testen

Nul-hypothese, H0: een hypothese die wordt beschouwd als waarheid, tenzij voldoende bewijs wordt verkregen om het tegendeel te bewijzen.

Alternatieve hypothese, H1: een hypothese waartegen de nul-hypothese wordt getest en die zal waar zijn als de nul-hypothese verworpen wordt.

Eenvoudige hypothese: een hypothese die één enkele waarde voor een parameter van de populatie.

Samengestelde hypothese: een hypothese die een reeks waarden specificeert voor een populatieparameter.

Eenzijdige alternatief: een alternatieve hypothese waarbij de rejection region zich in één van de staarten van de normale verdeling bevindt. $H_1: \mu > 27$

Tweezijdige alternatief: de rejection region bevindt zich aan beide staarten van de normale verdeling. $H_1: \mu \neq 27$

Twee mogelijke fouten bij het toetsen van hypotheses:

Type I error: Het verwerpen van een kloppende nul-hypothese, α , significantie niveau.

Type II error: Een foute nul-hypothese wordt niet verworpen, β , power.

9.2 Testen van het gemiddelde van een normale verdeling: populatie variantie bekend.

Rejection region: gebied waarbinnen de nul-hypothese verworpen wordt.

Eenzijdige alternatieve hypothese:

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

Tweezijdige alternatieve hypothese:

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0 \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leftarrow z_{\alpha/2} \text{ of } \textit{Verwerp } H_0 \textit{ als } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

P-waarde: $\textit{Verwerp } H_0 \textit{ als} : p - \textit{waarde} < \alpha \qquad p - \textit{waarde} = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geqslant z_p\right)$

P-waarde < 0.01, genoeg bewijs om H0 te verwerpen. Sprake van een hoog significantie niveau.

P-waarde tussen 0.01 en 0.05, sterk bewijs om H0 te verwerpen, sprake van significantie.

P-waarde >0.05, zwak bewijs om H0 te verwerpen, niet statistisch significant.

9.3 Testen van het gemiddelde van een normale verdeling: populatie variantie onbekend.

$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \rightarrow \quad Verwerp\, H_0\, als: t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}$

$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \rightarrow \quad Verwerp\, H_0\, als: t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha}$

9.4 Testen van de populatie proportie (grote samples)

$H_0: P = P_0 \quad H_1: P > P_0 \rightarrow \quad Verwerp\, H_0\, als: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > z_\alpha$

$H_0: P = P_0 \quad H_1: P < P_0 \rightarrow \quad Verwerp\, H_0\, als: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -z_\alpha$

$H_0: P = P_0 \quad H_1: P \neq P_0 \rightarrow$

$Verwerp\, H_0\, als: \dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -z_{\alpha/2}$ of $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > z_{\alpha/2}$

9.5 Beoordeling van het vermogen van een toets

De kans op een Type II fout bepalen:

$\beta = P(\bar{x} < \bar{x}_c | \mu = \mu^*) = P\left(z < \dfrac{\bar{x}_c - \mu^*}{\sigma/\sqrt{n}}\right)$ De sterkte = $1 - \beta$

De waarde van beta en de sterkte zal verschillen voor elke μ^*

Stappen:

1. Vorm de toets beslissingsregel, vind de range van waarden van de sample proportie die leiden tot het falen in het verwerpen van de nul-hypothese.
2. Gebruik de waarde P_1 voor de populatie proportie, vind de kans dat de sample proportie in de niet-verwerpen regio valt.

Voorbeeld:

$H_0: P = P_0 = 0.50 \quad H_1: P \neq 0.50 \quad n = 600 \quad \alpha = 0.05$

De beslissingsregel om H0 te verwerpen is: $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} < -1.96$ of $\dfrac{\hat{p} - P_0}{\sqrt{P_0(1-P_0)/n}} > 1.96$

$\hat{p} > 0.50 + 1.96\sqrt{0.50(1-0.50)/600} = 0.50 + 0.04$ of $\hat{p} < 0.50 - 0.04 = 0.46$

We willen de kans op een type II fout bepalen wanneer deze beslissingsregel wordt gebruikt.

Stel dat de echte populatie proportie $P_1 = 0.55$ is.

We willen de kans bepalen dat de sample proportie tussen 0.46 en 0.54 is als de populatie proportie 0.55 is. Dan is de kans op een type II fout als volgt:

$$P(0.46 \leqslant \hat{p} \leqslant 0.54) = P\left[\frac{0.46 - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}} \leqslant Z \leqslant \frac{0.54 - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right] \quad = \quad P(-4.43 \leqslant Z \leqslant -0.49) = 0.3121$$

De kans op een type II fout is $\beta = 0.3121$ De sterkte van de toets: $Power = 1 - \beta = 0.6879$

9.6 Toetsen van de variantie van een normale verdeling

$$H_0: \sigma^2 = \sigma_0^2 \qquad H_1: \sigma^2 > \sigma_0^2 \quad \rightarrow \quad Verwerp\, H_0\, als: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha}$$

$$H_0: \sigma^2 = \sigma_0^2 \qquad H_1: \sigma^2 < \sigma_0^2 \quad \rightarrow \quad Verwerp\, H_0\, als: \frac{(n-1)s^2}{\sigma_0^2} < X^2_{n-1,1-\alpha}$$

$$H_0: \sigma^2 = \sigma_0^2 \qquad H_1: \sigma^2 \neq \sigma_0^2 \quad \rightarrow \quad Verwerp\, H_0\, als: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha/2} \text{ of } \frac{(n-1)s^2}{\sigma_0^2} < X^2_{n-1,\alpha/2}$$

Voorbeeld:

Bepalen of de variantie van onzuiverheden van verschepen van stoffen binnen de vastgestelde norm is. Met n=100, de variantie mag niet hoger zijn dan 4.

Een random sample van 20 zakken wordt genomen. De sample variantie is berekend op 6.62.

$$H_0: \sigma^2 \leqslant \sigma_0^2 = 4 \qquad H_1: \sigma^2 > 4 \quad \rightarrow \quad Verwerp\, H_0\, als: \frac{(n-1)s^2}{\sigma_0^2} > X^2_{n-1,\alpha}$$

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)(6.62)}{4} = 31.445 \quad > \quad X^2_{n-1,\alpha} = 30.144$$

Dus we verwerpen H0 en concluderen dat de variantie van onzuiverheden de norm overschrijdt. We doen de aanbeveling dat het productieproces beter bestudeerd moet worden en er verbeteringen moeten plaatsvinden om de variantie van product componenten te verlagen.

Twee populatie hypothese test – Chapter 10

10.1 Testen van de verschillen tussen twee normale populatiegemiddelden: afhankelijke samples.

Een random sample van n matched paren van waarnemingen uit verdelingen met gemiddelden μ_x en μ_y . Dan is $\bar{d}$ het sample gemiddelde en s_d de standaardafwijking voor de n verschillen $(x_i - y_i)$. Als de populatieverdeling van het verschil normaal verdeeld is, dan hebben de volgende testen significantie level α :

De nul-hypothese testen, groter dan 0: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y > 0$

H0 verwerpen wanneer: $\frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1,\alpha}$

De nul-hypothese testen, kleiner dan 0: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y < 0$

H0 verwerpen wanneer: $\frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1,\alpha}$

De nul-hypothese testen, niet gelijk aan 0: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$

H0 verwerpen wanneer: $\frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1,\alpha/2}$ of $\frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1,\alpha/2}$

10.2 Testen van het verschil tussen twee normale populatiegemiddelden: onafhankelijke samples.

Drie situaties:

Onafhankelijke samples, variantie bekend.

De nul-hypothese testen: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$

Verwerp H0 als: $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} < -z_{\alpha/2}$ of $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} > z_{\alpha/2}$

Bij een eenzijdige toets → < $-z_\alpha$ of > z_α

Onafhankelijke samples, variaties onbekend, variaties beschouwd als gelijk.

De nul-hypothese testen: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$

Verwerp 0 als: : $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha/2}$ of $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha/2}$

Bij een eenzijdige toets → < $-t_{n_x+n_y-2,\alpha}$ of > $t_{n_x+n_y-2,\alpha}$

Onafhankelijke samples, variaties onbekend, variaties beschouwd als ongelijk.

De nul-hypothese testen: $H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$

Verwerp 0 als: : $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} < -t_{v,\alpha/2}$ of $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}} > t_{v,\alpha/2}$

Bij een eenzijdige toets → < $-t_{v,\alpha}$ of > $t_{v,\alpha}$

10.3 Testen van verschillen tussen twee populatie proporties (grote samples)

Onafhankelijke random samples met grootte n_x en n_y met deel successen $\hat{p}_x$ en $\hat{p}_y$

Als we aannemen dat de populatie proporties gelijk zijn, dan kunnen we een schatting maken van de gewone proportie: $\hat{p}_0 = \frac{n_x \hat{p}_x + n_y \hat{p}_y}{n_x + n_y}$ Voor grote sample groottes hebben de volgende testen significantie level α :

De nul-hypothese testen: $H_0 : P_x - P_y = 0$ $H_1 : P_x - P_y \neq 0$

Verwerp H0 als: $\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_{\alpha/2}$ of $\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{\alpha/2}$

Bij eenzijdige toets → < $-z_\alpha$ of > z_α

$\hat{p}_0 = \frac{n_x \hat{P}_x + n_y \hat{P}_y}{n_x + n_y}$

10.4 Testen van de gelijkheid van de variaties tussen twee normaal verdeelde populaties.

De F verdeling $F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ met numerator vrijheidsgraden $(n_x - 1)$ en de nominator vrijheidsgraden $(n_y - 1)$

Bij twee onafhankelijke random samples met n_x en n_y waarnemingen uit twee normale populaties met variaties σ_x^2 en σ_y^2 , sample variaties s_x^2 en s_y^2

Testen van gelijkheid van variaties van twee normale populaties:

$H_0 : \sigma_x^2 = \sigma_y^2$ $H_1 : \sigma_x^2 \neq \sigma_y^2$

Verwerp H0 als: $\frac{s_x^2}{s_y^2} > F_{n_x-1, n_y-1, \alpha/2}$

Bij eenzijdige toets → $F_{n_x-1, n_y-1, \alpha}$

10.5 Opmerkingen bij hypothese testen

Het definiëren van de nul en alternatieve hypothese vereist een zorgvuldige afweging van de doelstellingen van de analyse.

De testen in dit hoofdstuk zijn gebaseerd op de aanname dat de onderliggende verdeling normaal is of dat de central limit theorie van toepassing is.

Twee variabele regressie analyse – Chapter 11

11.1 Overzicht van lineaire modellen

Least squares regressielijn: $\hat{y}: b_0+b_1 x$ → *Helling* $b_1=\frac{Cov(x,y)}{s_x^2}=r\frac{s_y}{s_x}$ *y-as* $b_0=\bar{y}-b_1\bar{x}$

11.2 Lineaire regressie model

Aannames:

1. De Y's zijn lineaire functies van X plus een random foutterm: $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$
2. De x waarden zijn vaste getallen, $x_i(i=1,\ldots,n)$
3. De fouttermen zijn random variabelen; $E[\varepsilon_i]=0$ $E[\varepsilon_i^2]=\sigma^2$ voor $(i=1,\ldots,n)$
4. De random fouttermen, ε_i , zijn niet gecorreleerd met elkaar $E[\varepsilon_i\varepsilon_j]=0$ voor allen.

Lineaire regressie populatie model → $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

11.3 Least squares procedure

Schattingen van de lineaire vergelijkingscoëfficiënten b_0 en b_1 → $\hat{y}_i=b_0+b_1 x_i$

11.4 De verklarende kracht van een lineaire regressie vergelijking

Analyse van variantie; SST = SSR(uitgelegd door regressie) + SSE(onverklaarde fout)

Sum of squares total: $SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$

Sum of squares error: $SSE=\sum_{i=1}^{n}(y_i-(b_0+b_1 x_i))^2=\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}e_i^2$

Sum of squares regression: $SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$

Determinatiecoëfficiënt, $R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$ een hogere waarde staat voor een betere regressie.

Correlatie → $R^2=r^2$

Schatting van model error variantie : $\hat{\sigma}^2=s_e^2=\frac{\sum_{i=1}^{n}e_i^2}{n-2}=\frac{SSE}{n-2}$

11.5 Statistische gevolgtrekking: Hypothese toetsen en betrouwbaarheidsinterval

Steekproefverdeling van de Least squares schatter coëfficiënt:

Als de standaard least squares aannames gelden, dan is b_1 een zuivere schatter voor β_1 en heeft een populatievariantie: $\sigma^2_{b_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}$

En de unbiased sample variantie schatter: $\sigma^2_{b_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}$

De variantie van de helling coëfficiënt is afhankelijk van twee belangrijke hoeveelheden:

De afstand van de punten op de regressielijn gemeten door s_e^2 . Hogere waarden betekenen grotere variantie voor b_1 .

De totale afwijking van de X waarden van het gemiddelde, wat gemeten wordt door $(n-1)s_x^2$. Grotere afwijkingen in de X waarden en grotere sample groottes veroorzaken een kleinere variantie voor de helling coëfficiënt.

Basis voor gevolgtrekking over de populatie regressiehelling:

Als β_1 de populatie regressiehelling is en b_1 de least squares schatting gebaseerd op n paren van sample observaties. Dan, als de standaard regressie aannames gelden en we ervan uit kunnen gaan dat de errors, ε_i , normaal verdeeld zijn, is de random variabele $t = \frac{b_1 - \beta_1}{s_{b_1}}$ verdeeld als een Student's t met (n-2) vrijheidsgraden.

Het toetsen van de populatie regressiehelling:

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 > \beta_1^* \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{b_1 - \beta_1^*}{s_{b_1}} \geqslant t_{n-2,\alpha}$

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 < \beta_1^* \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{b_1 - \beta_1^*}{s_{b_1}} \leqslant -t_{n-2,\alpha}$

$H_0: \beta_1 = \beta_1^* \quad H_1: \beta_1 \neq \beta_1^* \rightarrow \textit{Verwerp } H_0 \textit{ als } \frac{b_1 - \beta_1^*}{s_{b_1}} \geqslant t_{n-2,\alpha/2}$ of $\frac{b_1 - \beta_1^*}{s_{b_1}} \leqslant -t_{n-2,\alpha/2}$

Betrouwbaarheidsinterval voor de populatie regressiehelling β_1 :

Als de regressie errors, ε_i , normaal verdeeld zijn of de verdeling van b_1 is bij benadering normaal verdeeld en de standaard regressie aannames gelden, een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor de populatie regressiehelling β_1 wordt gegeven door:

$b_1 - t_{(n-2,\alpha/2)} s_{b_1} < \beta_1 < b_1 + t_{(n-2,\alpha/2)} s_{b_1}$

waarbij $t_{n-2,\alpha/2}$ het getal is waarvoor: $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ en de random variabele t_{n-2} volgt een Student's t verdeling met (n-2) vrijheidsgraden.

Hypothesen toetsen voor de populatie helling coëfficiënt gebruik makend van de F verdeling:

F toets voor simple regressie coëfficiënt:

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

De F statistiek: $F = \frac{MSR}{MSE} = \frac{SSR}{s_e^2}$

De beslissingsregel is dan als volgt: $Verwerp\, H_0\, als\, F \geqslant F_{1, n-2, \alpha}$

11.6 Voorspelling

Regressiemodellen kunnen worden gebruikt bij het voorspellen van de afhankelijke variabele, als de toekomstige waarde van de onafhankelijke variabele gegeven is.

Toekomstvoorspelling intervallen en betrouwbaarheidsintervallen voor voorspellingen:

Stel dat het populatie regressie model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (i = 1, …, n), de standaard regressie aannames gelden en de ε_i zijn normaal verdeeld. Laat b_0 en b_1 de least squares schattingen van β_0 en β_1 zijn, gebaseerd op $(x_1, y_1), \ldots, (x_n, y_n)$. Dan kan worden aangetoond dat de volgende $100(1-\alpha)\%$ intervallen zijn.

Voor de prognose van een enkele uitkomst waarde resulterend voor Y_{n+1} , de voorspelling interval is als volgt: $\hat{y_{n+1}} \pm t_{n-2, \alpha/2} \sqrt{[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}]} s_e$

Voor de prognose van het gemiddelde of voorwaardelijke verwachting $E(Y_{n+1} | X_{n+1})$, de voorspelling interval is als volgt: $\hat{y_{n+1}} \pm t_{n-2, \alpha/2} \sqrt{[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}]} s_e$

waarbij: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ en $\hat{y_{n+1}} = B_0 + b_1 x_{n+1}$

De kans is $1-\alpha$ dat de correcte voorspelling binnen de interval ligt.

Hoe groter de interval, hoe groter de onzekerheid rondom de puntvoorspelling.

Hoe groter de steekproefomvang n, hoe smaller de voorspelling interval en de betrouwbaarheidsinterval.

Hoe groter s_e^2 , hoe groter de voorspelling interval en de betrouwbaarheidsinterval.

Een grote dispersion impliceert dat we informatie hebben voor een wijde range van deze variabele, dus een preciezere schatting van de populatie regressielijn en de corresponderende smallere betrouwbaarheidsinterval en voorspellingsinterval.

Grotere waarden van de hoeveelheid $(x_{n+1} - \bar{x})^2$ resulteren in bredere betrouwbaarheidsintervallen en bredere voorspellingsintervallen.

11.7 Correlatie analyse

Hypothese toetsen voor correlatie:

r is de sample correlatiecoëfficiënt, berekend uit een random sample van n paren observaties uit een gezamenlijke normale verdeling. De volgende toetsen volgen de nul populatie correlatie:

$$H_0: \rho = 0 \quad H_1: \rho > 0 \quad \rightarrow \quad verwerp\ H_0\ als \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$$

$$H_0: \rho = 0 \quad H_1: \rho < 0 \quad \rightarrow \quad verwerp\ H_0\ als \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha}$$

$$H_0: \rho = 0 \quad H_1: \rho \neq 0 \quad \rightarrow \quad verwerp\ H_0\ als \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha/2} \text{ of } \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$$

11.8 Beta meting van financieel risico

Metingen en analyse procedures om investeerders te helpen met het meten en controleren van hun financieel risico bij het ontwikkelen van een investeringsportfolio.

Diversifiable risico is het risico verbonden aan specifieke bedrijven en industrieën en omvat: werk conflicten, nieuwe concurrentie, consumentenmarkt veranderingen en vele andere factoren. Dit risico kan gecontroleerd worden door een groter portfolio en door het opnemen van de aandelen waarvan het rendement een negatieve correlatie kent.

Nondiversifiable risico is het risico verbonden aan de gehele economie. Het effect hiervan wordt gemeten door het gemiddelde rendement op aandelen. Het effect op individuele bedrijven wordt gemeten met de beta coëfficiënt.

De beta coëfficiënt voor een specifiek bedrijf is de hellingcoëfficiënt, deze geeft aan hoe afhankelijk het rendement van een bepaald bedrijf is van het gehele markt rendement.

Als het rendement van het bedrijf de markt precies volgt, dan is de beta coëfficiënt 1.

Als het rendement van het bedrijf heftiger reageert op de markt dan is de beta coëfficiënt meer dan 1.

Minder reagerend op de markt, dan is de beta minder dan 1.

De required return op een investering =
$(Risk - free\ rate) + [(beta\ for\ investment) \times ((Market\ return) - (risk - free\ rata))]$

Hoe hoger de beta waarde is, hoe hoger de required return. Dat komt doordat de aandelen heftiger reageert op het nondiversifiable markt risico.

11.9 Grafische analyse

Grafische analyse wordt gebruikt om effecten te laten zien op regressie analyses van punten die extreme X waarden hebben en punten die Y waarden hebben die verschillen van de least squares regression equation.

Extreme punten zijn punten die X waarden hebben die substantieel afwijken van de X waarden van andere punten.

De leverage van een punt is gedefinieerd als: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Outlier punten zijn punten die substantieel afwijken in de Y richting van de voorspelde waarde. Deze punten worden geïdentificeerd door het berekenen van de standardized residual als volgt:

$$e_{is} = \frac{e_i}{s_e \sqrt{1 - h_i}}$$

Inleiding in niet-parametrische statistieken – Chapter 14

14.1 Goodness-of-fit toetsen: gespecificeerde waarschijnlijkheid

Goodness-of-fit toetsen worden gebruikt om de populatie van nominale data te beschrijven.

Chi-kwadraat random variabele

Een random sample van n waarnemingen, elke waarneming komt uit één van de K categorieën. Stel dat de geobserveerde nummers in elke categorie $O_1, O_2, \ldots, O_K$ zijn. Een nul-hypothese specificeert kansen $P_1, P_2, \ldots, P_K$ dat een waarneming in elke categorie valt, de verwachtte aantallen in de categorie, onder H0, zou als volgt zijn: $E_i = nP_i$ voor $i = 1,2,\ldots,K$

Als de nul-hypothese waar is en de sample grootte is groot genoeg zodat de verwachtte waarden op z'n minst 5 zijn, dan is de random variabele in verband met $X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$ bekend als een chi-kwadraat random variabele, en heeft een chi-kwadraat verdeling met (K-1) vrijheidsgraden.

Een goodness-of-fit test met gespecificeerde waarschijnlijkheden, significantielevel α , van H0 tegen het alternatief dat de gespecificeerde waarschijnlijkheid niet correct is gebaseerd op de beslissingsregel: $verwerp\ H_0\ als \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > X^2_{K-1,\alpha}$ De random variabele X^2_{K-1} volgt een chi-kwadraat verdeling met K-1 vrijheidsgraden.

14.2 Goodness-of-fit toetsen: populatie parameters onbekend

Wanneer populatie parameters geschat zijn, schatting van m onbekende populatie parameters. De vrijheidsgraden veranderen voor de chi-kwadraat random variabele in: (K – m – 1), waarbij K het aantal categorieën en m het aantal onbekende populatie parameters.

Een test voor de normale verdeling:

Jarque-Bera test voor normaliteit: We hebben een random sample $x_1, x_2, \ldots, x_n$ van n waarnemingen van een populatie. De test statistiek voor de Jarque-Bera test voor normaliteit is:

$$JB = n\left[\frac{(scheefheid)^2}{6} + \frac{(kurtosis - 3)^2}{24}\right]$$

met $scheefheid = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$ en $kurtosis = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4}$

Als het aantal sample waarnemingen groot wordt, dan heeft deze statistiek, onder de nul-hypothese met normaal verdeelde populatie, een chi-kwadraat verdeling met 2 vrijheidsgraden.

Helaas, de chi-kwadraat benadering van de verdeling van de Jarque-Bera toets statistiek, JB, is alleen dichtbij voor grote sample groottes.

14.3 Contingency tabellen

Stel dat een sample wordt genomen uit een populatie en de leden kunnen uniek kruis geclassificeerd zijn overeenkomstig een paar kenmerken, A en B. Een vliegmaatschappij wil weten of er enige relatie bestaat tussen het geslacht van een klant en de gebruikte methode om een vliegticket te boeken. We kunnen hierbij een kruis-classificatie maken met n waarnemingen in een r x c Contingency Tabel.

Chi-kwadraat random variabele voor Contingency tabel

Onder de nul-hypothese, de random variabele verbonden met $x^2=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij}-E_{ij})^2}{E_{ij}}$ heeft een chi-kwadraat verdeling met (r-1)(c-1) vrijheidsgraden. De benadering werkt goed wanneer meer dan 20% van de geschatte verwachtte getallen E_{ij} minder is dan 5.

H0: er bestaat geen verband tussen twee karakteristieken in de populatie. Dan is $E_{ij}=\frac{R_i C_j}{n}$ waarbij Ri en Cj de corresponderende rij en kolom totalen zijn.

Beslissingsregel: $verwerp\ H_0\ als: \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(O_{ij}-E_{ij})^2}{E_{ij}} > x^2_{(r-1)(c-1),\alpha}$

14.4 Niet-parametrische toetsen voor gepaarde of gematcht samples.

De simpelste niet-parametrische toets is de *sign toets*. De sign toets wordt gebruikt in marktonderzoek om te bepalen of er klantvoorkeur voor één van de twee producten bestaat. Sign van verschil is of het verschil negatief of positief is.

Berekenen van het verschil voor elk paar waarnemingen en de sign van dit verschil vastleggen.

$H_0: P=0.5$

De p-waarde van een sign toets wordt gevonden door gebruik te maken van de binominale verdeling met n = aantal of niet-nul verschillen, S= aantal positieve verschillen en P =0.5.

a. Upper-tail toets: $H_1: P>0.5 \qquad p-waarde=P(x\geqslant S)$

b. Lower-tail toets: $H_1: P<0.5 \qquad p-waarde=P(x\leqslant S)$

c. Two-tail toets: $H_1: P\neq 0.5 \qquad p-waarde=2\,P(x\geqslant S)$

Een nadeel van de sign toets is dat het maar een beperkt deel van de informatie verwerkt, alleen het teken van het verschil.

De *Wilcoxon Signed rank toet*s geeft een methode om informatie over de grootte van het verschil te betrekken in de toets.

Alle verschillen worden op grootte gerangschikt.

Verwerp H0 wanneer: $T\leqslant T\ Appendix\ tabel\ 10$ met $T=min(T_+, T_-)$

Waarbij: n = aantal niet-nul verschillen

T_+ = totaal van de positieve ranks T_- = totaal van negatieve ranks

De *sign toets: normale benadering* (grote samples)

Volgens de central limit theorem kan de normale verdeling gebruikt worden om de binominale verdeling te benaderen als de sample grootte groot is.

Gemiddelde: $\mu=np=0.5n$

Standaardafwijking: $\sigma=\sqrt{np(1-p)}=\sqrt{0.25n}=0.5\sqrt{n}$

De toets statistiek: $Z=\frac{S*-\mu}{\sigma}=\frac{S*-0.5n}{0.5\sqrt{n}}$

a. Two-tail toets: $S*=S+0.5 \, als \, S<\mu$ of $S*=S-0.5 \, als \, S>\mu$

b. Upper-tail toets: $S*=S-0.5$

c. Lower-tail toets: $S*=S+0.5$

De *Wilcoxon Signed Rank toets: normale benadering* (grote samples)

Gemiddelde: $E(T)=\mu_T=\frac{n(n+1)}{4}$

Variantie: $Var(T)=\sigma_T^2=\frac{n(n+1)(2n+1)}{24}$

De toets statistiek: $Z=\frac{T-\mu_T}{\sigma_T}$

Als het aantal, n, van niet-nul verschillen groot is en T is de geobserveerde waarde van de Wilcoxon statistiek, dan hebben de volgende toetsen significantie level α .

De alternatieve hypothese is eenzijdig, verwerp H0 wanneer: $\frac{T-\mu_T}{\sigma_T} \leftarrow z_\alpha$

De alternatieve hypothese is tweezijdig, verwerp H0 wanneer: $\frac{T-\mu_T}{\sigma_T} \leftarrow z_{\alpha/2}$

De sign toets kan ook gebruikt worden bij het testen van hypothesen over de centrale locatie of een populatie verdeling.

14.5 Niet-parametrische toetsen voor onafhankelijke random samples

Twee toetsen die de centrale locaties van twee populatieverdelingen toetsen wanneer onafhankelijke random samples worden getrokken van deze twee populaties.

Mann-Whitney U toets

Deze verdeling benadert de normale verdeling als het aantal sample observaties stijgt. De benadering is alleen adequaat wanneer elk sample op z'n minst 10 observaties bevat.

$$U=n_1 n_2+\frac{n_1(N_1+1)}{2}-R_1$$

Waarbij R_1 het totaal van de ranks van de waarnemingen uit de eerste populatie.

Gemiddelde: $E(U)=\mu_U=\frac{n_1 n_2}{2}$

Variantie: $Var(U)=\sigma_U^2=\frac{n_1 n_2(n_1+n_2+1)}{12}$

De verdeling van de random variabele $Z=\frac{U-\mu_U}{\sigma_U}$ wordt benaderd door de normale verdeling.

Voorbeeld:

Aantal uren per week student aan het studeren voor Statistiek en Gedrag in Organisaties:

Statistiek	10	6	8	10	12	13	11	9	5	11		
Gedrag	13	17	14	12	10	9	15	16	11	8	9	7

Mann-Whitney U test Ranks voor uren studeren per week:

n1 = 10, n2 = 12, R1= 93.5

$$U=(10)(12)+\frac{(10)(11)}{2}-93.5=81.5$$

$$E(U)=\frac{(10)(12)}{2}=60$$

$$Var(U)=\sigma_U^2=\frac{(10)(12)(23)}{12}=230$$

$Z=\frac{81.5-60}{\sqrt{230}}=1.42$ en p-waarde = 0.1556

Statistiek	(Rank)	Gedrag	(Rank)
10	(10.0)	13	(17.5)
6	(2.0)	17	(22.0)
8	(4.5)	14	(19.0)
10	(10.0)	12	(15.5)
12	(15.5)	10	(10.0)
13	(17.5)	9	(7.0)
11	(13.0)	15	(20.0)
9	(7.0)	16	(21.0)
5	(1.0)	11	(13.0)
11	(13.0)	8	(4.5)
		9	(7.0)
		7	(3.0)
	Rank totaal = 93.5		Rank totaal = 159.5

Met de standaard 0.05 significantie level, het toets resultaat is niet voldoende om te concluderen dat studenten meer tijd spenderen aan studeren voor één van de twee vakken dan voor de ander.

Wilcoxon rank totaal statistiek T

Gemiddelde: $E(T)=\mu_T=\frac{n_1(n_1+n_2+1)}{2}$

Variantie: $\sigma_T^2=\frac{n_1 n_2(n_1+n_2+1)}{12}$ dan is $Z=\frac{T-\mu_T}{\sigma_T}$

14.6 Spearman rank correlatie

Stel dat een random sample $(x_1, y_1), \ldots, (x_n, y_n)$ van n paren observaties wordt genomen. Als x_i en y_i elk zijn gerangschikt in oplopende volgorde en de sample correlatie van deze ranken wordt berekend, de resulterende coëfficiënt wordt de Spearman rank correlatiecoëfficiënt genoemd.

Als er geen gebonden X of Y ranken zijn, een equivalente formule voor het berekenen van deze coëfficiënt is $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ waar d_i de verschillen zijn tussen de gerangschikte paren.

Toetsen tegen het alternatief van positieve associatie: $verwerp\ H_0\ als \quad r_s > r_{s,\alpha}$

Toetsen tegen het alternatief van negatieve associatie: $verwerp\ H_0\ als \quad r_s < -r_{s,\alpha}$

Toetsen tegen het tweezijdige alternatief: $verwerp\ H_0\ als \quad r_s < -r_{s,\alpha}$ of $r_s > r_{s,\alpha}$

14.7 Een non-parametrische toets voor randomness

Runs test: Kleine sample grootte.

R is het aantal runs in een opeenvolging van n waarnemingen met $n \leqslant 20$. De nul-hypothese is dat de serie een reeks van random variabelen is.

Appendix tabel 14 geeft het kleinste significantielevel waarbij de nul-hypothese kan worden verworpen tegen de alternatieve van positieve associatie tussen naburige waarnemingen, als functie van R en n.

Als het alternatief tweezijdig is met non-randomness, dan moet het significantielevel verdubbeld worden als die minder dan 0.5 is.

Als het significantielevel groter dan 0.5 is, dan is het juiste significantielevel voor de tweezijdige toets $2(1-\alpha)$.

Runs test: Grote sample grootte

Gegeven dat we een tijd serie hebben met n observaties en n > 20, definieer het aantal runs, R, als het aantal opeenvolgingen boven of onder de mediaan.

$H_0: the\ series\ is\ random$

De verdeling van het aantal runs onder de nul-hypothese kan benaderd worden met de normale verdeling. Onder de nul-hypothese kent $Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{2(n-1)}}}$ een normale verdeling.

Als de alternatieve hypothese van positieve associatie: $verwerp\ H_0\ als\ Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} < -z_\alpha$

Tweezijdige hypothese van non-randomness:

$$verwerp\ H_0\ als\ Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} < -z_{\alpha/2} \quad \text{of} \quad Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2-2n}{4(n-1)}}} > z_{\alpha/2}$$