Inhoudsopgave

1. Grafieken gebruiken om data te beschrijven

Population, een populatie, is de complete set van alle items dat interessant is om te onderzoeken. De populatie grootte, *N* (population size), is over het algemeen groot of zelfs oneindig groot. Een sample, een steekproef, is een geobserveerde subset, of een deel van, een populatie, met een sample grootte die gegeven wordt door *n* (sample size).

Simple random sampling is een procedure die gebruikt wordt om een sample te selecteren van het aantal *n* objecten van een population op zodanige manier dat elke lid van population strict toevallig is gekozen. Dit betekent dat de selectie van een lid geen invloed heeft op de selectie van een ander lid: elk lid van de populatie heeft even veel kans om gekozen te worden en elke mogelijke sample van een gegeven grootte, *n*, heeft dezelfde kan om gekozen te worden. Deze methode van samples is zo vaak voorkomend dat in het begrip simple random sampling het bijvoegelijke naamwoord simple vaak wordt weggelaten en het begrip random sampling vaak wordt gebruikt.

Een parameter is een numerieke maatregel die de specifieke kenmerken van een populatie beschrijft. Een statistic is een numerieke maatregel die de specifieke kenmerken van een sample beschrijft.

Descriptive statistics richten zich op grafische en numerieke procedures die gebruikt worden om data te samenvatten of te verwerken. Inferential statistics richten zich op het gebruik van data om voorspellingen, prognoses en schattingen te maken waardoor betere besluitvorming mogelijk is.

Categorial variables produceren antwoorden die behoren tot een groep of een categorie. Zo zijn ja/nee vragen categorial. Daarbij kunnen we bij categorial variables ook de level of measurement aangeven, deze is dan ordinal of nominal. Ordinal betekent dat het te maken heeft met een ranking, zoals minimum, medium of maximum, en nominal betekent dat het niet te maken heeft met een ranking maar meer zoals man of vrouw.

Numerical variables bestaat uit twee categorieën: discrete en continuous. Een discrete numerical variable kan (maar hoef niet perse) een eindigende hoeveelheid getallen hebben, deze variabele heeft dus te maken met tellen. Een continuous numerical variable kan elke waarde binnen een gegeven bereik van reële getallen zijn en komt meestal voort uit een meetproces (dus niet een telproces).

Data kunnen ook worden beschreven als qualitative (kwalitatief) of quantitative (kwantitatief). Met kwalitatieve data er is geen meetbare betekenis aan de ‘verschillen’ in getallen (bijvoorbeeld rugnummers bij voetbalspelers). Kwantitatieve data heeft wel een meetbare betekenis aan de ‘verschillen’ in getallen (bijvoorbeeld toetsscores bij studenten). Kwalitatieve data omvat nominal en ordinal levels of measurement en kwantitatieve data omvat intervallen en ratio levels of measurement. Nominal data wordt als het zwakste typen van data beschouwd, sinds numerical identificatie strikt gekozen is als meest geschikt en geen rankingen of reacties omvat.

Een frequency distribution is een tabel die wordt gebruikt om data te ordenen. De linker kolom (ook wel groepen of klassen genoemd) omvat alle mogelijke reacties op de bestudeerde variabele. De rechterkolom is een lijst van alle frequenties, of observaties, of getallen, voor elke groep of klas.

Staafgrafieken (bar charts) of cirkelgrafieken (pie charts) worden gebruikt om categorial data te representeren. Als we de intentie hebben om nadruk te leggen op de frequentie

van elke categorie dan gebruiken we voornamelijk een staafgrafiek. Als we de intentie hebben om nadruk te leggen op de proportie van de frequentie in elke categorie, dan gevruiken we voornamelijk de cirkelgrafiek.
Een paretodiagram is een staafgrafiek dat de frequentie van defecte zaken afbeeldt. De staaf aan de linkerkant van de grafiek indiceert de meest frequente zaak en staven rechts ervan indiceren zaken met afnemende frequenties. Een paretodiagram wordt gebruikt om de 'vitale weinige' van de 'triviale vele' te scheiden.

Een lijngrafiek, wordt ook wel een tijd-serie diagram genoemd, is een serie van data gerepresenteerd in een grafiek op verschillende tijdsintervallen. Hierbij wordt tijd op de horizontale x-as geplaatst en de numerical quantity op de verticale y-as geplaatst. Het toevoegen van punten in de grafiek op nabijgelegen tijdsperiode en die verbinden door middel van een lijn laat een time-serie diagram ontstaan.

Om een frequency distribution op te stellen zijn een paar regels vast gelegd. Regel 1: stel *k*, het aantal intervallen (groepen, klassen), vast. Regel 2: intervallen moeten de zelfde breedte *w* (width) zijn, die als volgt wordt bepaald:

$$w = \text{interval witdth} = \frac{\text{largest data value} - \text{smallest data value}}{\text{number of intervals}}$$

Waarbij *k* en *w* naar boven afgerond moeten worden, zo mogelijk tot het volgende hele getal. Regel 3: intervallen moeten inclusief en niet-overlappend zijn.

Een relative frequency distribution wordt berekend door elke frequentie te delen door het aantal observaties en de resulterende proportie te vermenigvuldigen met 100%. Een cumulative frequenct distribution houdt in het totale aantal observaties waarvan de waarden minder zijn dan het limiet voor elke interval. We kunnen een cumulative frequency distribution construeeren door de frequentie van alle frequency distribution intervallen toe te voegen, inclusief het huidige interval. In een relative cumulative frequency distribution kunnen cumulative frequencies worden weergegeven als cumulatieve proporties of percentages.

Een histogram is een grafiek dat bestaat uit verticale staven die gecontrueerd worden op een horizontale lijn die afgebakend is met intervallen voor de variabele die wordt afgebeeld. De intervallen komen overeen met die in een frequency distribution table worden weergegeven. De hoogte van elke staaf is proportioneel aan het aantal observaties in dat interval. Het aantal observaties kan boven de staaf worden afgebeeld.

Een ogive, ook wel een cumulatieve lijngrafiek genoemd, is een lijn die de punten die cumulatieve percentages van observaties onder het limiet van elk interval in een cumulative frequency distribution verbind.

Een stem-and-leaf display is een EDA diagram. Dit diagram wordt soms als alternatief gebruikt voor een histogram. Naar aanleiding van hun toonaangevende cijfers (ook wel de stem genoemd) zijn de data gegroepeerd, waarbij ook rekening wordt gehouden met de eindcijfers (ook wel de leaves genoemd) van elke groep. De 'leaves' worden individueel weergegeven in oplopende volgorde na elke 'stem'.

We kunnen een scatter plot maken door elk punt van elk paar van twee variabelen die een observatie in de dataset representeren te lokaliseren. Op deze manier geeft de scatter plot inzicht in de data waarbij ten eerste het bereik van elke variabele duidelijk is. Ten tweede geeft het een patroon van waarden weer over het bereik, ten derde kan het een mogelijke

relatie tussen twee variabelen suggereren. Ten vierde geeft het een indicatie van uitschieters (extreme punten). Een scatter plot kan je prepareren door de individuele punten op een grafiekpapier te plotten.

Een cross table (kruistabel), ook wel contingency table, somt het aantal observaties voor elke combinatie van waarden voor twee categorial of ordinal variables. De combinatie van alle mogelijke intervallen voor de twee variabelen definieert de cellen in een tabel. Naar een kruistabel met het aantal rijen r en aantal kolommen k wordt gerefereerd als een $r \times k$ kruistabel.

Slecht getekende grafieken kunnen makkelijk leiden tot een vertekening van de werkelijkheid. In proces- en besluitvorming kan dit leiden tot verkeerde beslissingen op basis van deze onjuiste data. Bij het maken van een grafiek moet hiermee dus rekening gehouden worden en nauwkeurig gewerkt worden. In een histogram kan het zijn dat sommige staven breder zijn en andere staven kleiner zijn, maar we weten dat alle intervallen eigenlijk hetzelfde zouden moeten zijn. Dit kan leiden tot verwarring. Ook Time-serie plots kunnen misleidend zijn wanneer er gekozen wordt voor een groot bereik (bijvoorbeeld van 0 tot 1000) terwijl de uiterste punten tussen de 500 en 550 liggen. Hierdoor lijkt een trend relatief stabiel terwijl het deze, wanneer de schaal juist aangepast is, helemaal niet zo stabiel is.

Systematic sampling (het nemen van systematische steekproeven) is een statistische procedure die vaak gebruikt wordt als een alternatief voor random sampling (het nemen van willekeurige steekproeven). Als je aanneemt dat de populatie is ingedeeld op een of andere manier waarbij het niet in verband staat met het onderwerp dat van belang is. Systematic sampling bevat de selectie van elke j^{de} item in de populatie, waarbij j de verhouding is tussen de populatiegrootte N en de gewenste sample grootte n is: $j = \frac{N}{n}$. Wanneer je random een getal tussen 1 en j selecteert, verkrijg je het eerste item dat je systematic sample bevat.

Non-sampling fouten kunnen zijn ten eerste dat de populatie die gesampled wordt niet relevant is, ten tweede kan het onderzoeken van onderwerpen onjuiste of oneerlijke antwoorden geven, dit kan gebeuren omdat vragen op een manier geformuleerd zijn waardoor ze moeilijk te begrijpen zijn of daardoor verkeerd geïnterpreteerd worden. En ten derde omdat er geen antwoord gegeven wordt op enquêtevragen door de geselecteerde personen die antwoord horen te geven op de vragen uit de enquête, waardoor de sample fout ontstaat doordat de onderzochte hoeveelheid personen minder is dan werd aangenomen. Dit betekent eigenlijk dat de populatie niet bereidt is m antwoord te geven en dat de populatie die onderzocht is hoogstwaarschijnlijk geen belang heeft bij het onderwerp.

2. Numerieke maatregelen gebruiken om data te beschrijven

In hoofdstuk 1 werden de begrippen parameter en statistiek geïntroduceerd. Een parameter refereert naar een kenmerk van een populatie en een statistiek verwijst naar de een kenmerk van een specifiek sample. Maatregelen om centrummaten te beschrijven worden meestal gevormd uit sample data, in plaats van population data. Een maatregel van centrummaten is arithmetic mean (rekenkundig gemiddelde) ook wel gewoon alleen de mean (gemiddelde) genoemd. De arithmetic mean van een set van data is de som van alle datawaarden gedeeld door het aantal observaties. Als het gemiddelde niet van de dataset van een sample maar van de gehele populatie wordt berekend dan wordt het het population mean, m, genoemd, die wordt gegeven door $m = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + ... + x_N}{N}$ waarbij N de populatie grootte is, en Σ betekent de 'som van'. Als het gemiddelde van een sample wordt berekend dan wordt de sample mean $\bar{x}$ (een statistic) berekent door middel van $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$ waarbij n de grootte van de sample is en Σ staat voor de 'som van'.

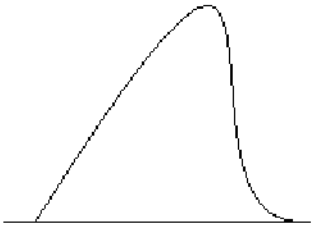
De median (mediaan) is de middelste observatie van een set van observaties die geordend zijn in een oplopende (of aflopende) volgorde. Als de sample grootte, n, een oneven aantal heeft, dan is de middelste observatie de median. Als de sample grootte, n, een even aantal heeft, dan is de median het gemiddelde van de twee middelste observaties. De mediaan zal het getal zijn dat op de $0.5(n+1)th \quad ordered \quad position$ is gelokaliseerd. Let op dat dit dus niet het gemiddelde is maar alleen de plek waarop de mediaan staat.

De mode, als die al bestaat, is de meest frequent voorkomende waarde in een dataset.

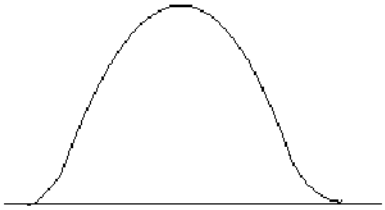
Numerical data worden normaal het beste beschreven door het gemiddelde. Maar, afhankelijk van welk soort type data, een andere soort factor die je in ogenschouw moet nemen is de aanwezigheid van uitschieters (extremen). De median wordt namelijk niet beïnvloedt door uitschieters, maar de mean wel. Daarom moet gecheckt worden of er uitschieters in de dataset zijn en of er een fout in de data zit. Het gemiddelde zal groter uitvallen wanneer er grote uitschieters zijn, en zal minder groot zijn als de dataset kleine uitschieters bevat. De relatie tussen de mean en de median leidt tot het begrijpen van de vorm van de verdeling.

De vorm van de verdeling toont of data gelijkmatig gespreid zijn vanaf het midden (centrum). Soms deelt het centrum de data afgebeeld in een distribution (verdelings) grafiek in tweeën waardoor de twee helften spiegelbeelden van elkaar zijn, de portie aan de ene kant van het midden is (bijna) identiek aan de portie aan de andere kant van het midden. Dit wordt ook wel symmetrie genoemd. Grafieken die dit niet hebben zijn asymmetrisch of scheef. Dus de vorm van een verdeling is symmetrisch als de observaties gebalanceerd zijn, of ongeveer gelijk verdeeld, rond het centrum. En de vorm van een verdeling is assymetrisch (of scheef: skewed) als de observaties niet symmetrisch verdeeld zijn rond het centrum. Een positieve asymmetrische verdeling heeft een top die naar links hangt en een 'staart' die zich uitstrekt in de richting van de positieve waarden. Een negatieve asymmetrische verdeling heeft een top die naar rechts hangt en een 'staart'

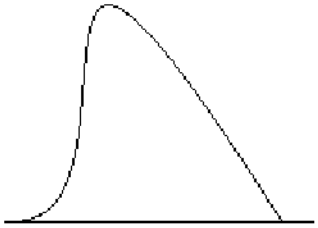
die zich uitstrekt in de richting van de negatieve waarden.



Negatieve assymetrische distribution
(mean < median)



Symmetrische distribution
(mean = median)



Positive asymmetrische distribution
(mean > median)

De mediaan wordt geprefereerd bij de beschrijving van de verdeling van inkomens in een stad, staat of land. De verdeling van inkomens is vaak positief asymmetrisch omdat inkomens een relatief klein deel bestaan uit hoge inkomens. Een groot deel van de bevolking heeft een modaal inkomen. De mean van de verdeling zal hoger liggen dan de median van de verdeling, omdat de weinigen met een hoog inkomen een aanzienlijk bereik hebben waar hun inkomen kan liggen.

De geometric mean (geometrische gemiddelde), $\bar{x}_g$, is de n^{de} wortel van het product van n getallen. $\bar{x}_g = \sqrt[n]{x_1 \times x_2 \times ... \times x_n} = (x_1 \times x_2 \times ... \times x_n)^{\frac{1}{n}}$. De geometric mean rate of return, $\bar{r}_g$, geeft het gemiddelde percentage rendement van een investering over een periode. En wordt berekend door $\bar{r}_g = (x_1 \times x_2 \times ... \times x_n)^{\frac{1}{n}} - 1$. Het berekenen van de geometric mean is belangrijk voor bedrijfsanalisten en economen die belang hebben bij het weten van de groei over een aantal perioden.

De range (bereik) is het verschil tussen de grootste en de kleinste observatie. De first quartile (eerste kwartiel), Q_1 (of het 25ste percentiel), scheidt de kleinste 25% van de data van de rest van de data. Deze wordt berekend door:

$Q_1 = \quad the \quad value \quad in \quad the \quad 0.25(n+1)th \quad ordered \quad position$

Het tweede kwartiel, Q_2 (of het 50ste percentiel), is de median, en wordt berekend door:

$Q_2 = \quad the \quad value \quad in \quad the \quad 0.5(n+1)th \quad ordered \quad position$

Het derde kwartiel, Q_3 (of het 75ste percentiel), scheidt de kleinste 75% van de dat a van de overige grootste 25% van de data en wordt berekend door:

$Q_3 = \quad the \quad value \quad in \quad the \quad 0.75(n+1)th \quad ordered \quad position$

De interquartile range (IQR, interkwartielafstand bereik) meet de spreiding in de middelste 50% van de data; het is het verschil tussen de observatie op Q^3, het derde kwartiel (of 75ste percentiel), en de observatie op Q^1, het eerste kwartiel (of 25ste percentiel). IQR = $Q^3 - Q^1$

De five-number summary (vijf getallen samenvatting) refereert naar de vijf beschrijvende

getallen: het minimum, het eerste kwartiel, de mediaan, het derde kwartiel en het maximum: Minimum < Q^1 < Mediaan < Q^3 < Maximum.

De population variance (populatie variantie), s^2, met respect tot de variantie, is de som van de wortel van de verschillen tussen elke observatie en het populatiegemiddelde gedeeld door de populatiegrootte N: $s^2 = \frac{\sum_{i=1}^{N}(x_i - m)^2}{N}$

De sample variance (steekproef variantie), s^2, is de som van de wortel van de verschillen tussen elke observatie en het steekproefgemiddelde (sample mean) gedeeld door de sample grootte n min één: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Let dus op dat bij een sample gedeeld word door n – 1 en niet door n. Het is door middel van ingewikkelde berekeningen uiteindelijk namelijk aan te tonen dat door n – 1 te gebruiken een nauwkeuriger antwoord zal uitkomen.

De standard deviation (standaard afwijking) van de populatie, s, met respect tot de standard deviation, is de (positieve) wortel van de population variance: $s^2 = \sqrt{\frac{\sum_{i=1}^{N}(x_i - m)^2}{N}}$

De sample deviation is de (positieve) wortel van de sample variance: $s^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

De coefficient of variation (variatiecoëfficiënt), CV, is een maatstaf van de relatieve spreiding dat de standard deviation uitdrukt als een percentage van de mean (mits de mean positief is). De population coefficient of variation is $CV = \frac{s}{m} \times 100\%$ als $m > 0$ en de sample coefficient of variation is $CV = \frac{s}{\bar{x}} \times 100\%$ als $\bar{x} > 0$.

De Chebychev's theorie is dat voor elke populatie met de mean m, de standard deviation s en k > 1, het percentage van observatie dat binnen het interval $[m \pm ks]$ liggen is ten minste $100[1 - \frac{1}{k^2}]\%$ waarbij k het aantal standard deviations is. De Russische wiskundige Chebychev ontwikkelde data intervallen voor welke data set dan ook, onafhankelijk van de vorm van de verdeling.

Voor veel grote populaties (klokvormig) biedt de empirical rule (empirische regel) een schatting van het te benaderen percentage van observaties die binnen een, twee of drie standard deviations van de mean zijn opgenomen: de te benaderen 68% van de observaties liggen in het interval $m \pm 1s$, de te benaderen 95% van de observaties liggen in het interval $m \pm 2s$ en bijna alle observaties liggen in het interval $m \pm 3s$.
De weighted mean (gewogen gemiddelde) van een dataset is $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{n}$ waar w_i is het gewicht (de weight) van de i[de] observatie en $n = \Sigma w_i$.

Als je aanneemt dat data gegroepeerd is in K groepen, met de frequenties f_1, f_2, f_3, …, f_k, dat de middelpunten van deze groepen m_1, m_2, m_3, …, m_k, zijn, dan zijn de sample mean en de sample standard deviation van de gegroepeerde data te benaderen op te volgende manieren. De mean is te benaderen door $\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$ waarbij $n = \sum_{i=1}^{K} f_i$ en is de variance te benaderen door de volgende formule: $s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1}$.

De coveriance (covariantie), Cov, is een manier om de lineaire relatie tussen twee variabelen te meten. Een positieve waarde indiceert een directe of toenemende lineaire relatie en een negatieve waarde indiceert een afnemende lineaire relatie. De covariance van de populatie is te berekenen door $Cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{N} (x_1 - m_x)(y_i - m_y)}{N}$ waar x_i en y_i de geobserveerde waarden zijn, m_x en m_y de populatiegemiddelde en N de populatiegrootte.

De sample covariance is te berekenen door $Cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})(y_i - \bar{y})}{n-1}$ waar x_i en y_i de geobserveerde waarden zijn, $\bar{x}$ en $\bar{y}$ de sample gemiddelde en n de sample grootte. De waarde van de covariantie verschaft geen maatstaf van de sterkte van de relatie tussen twee variabelen.

De correlation coefficient (correlatiecoëfficiënt) kan worden berekend door de covariance te delen door het product van de standard deviations van de twee variabelen. De formule van de populatie correlatiecoëfficiënt, r, is $r = \frac{Cov(x,y)}{s_x s_y}$ waar s_x en s_y de population standard deviations van de twee variabelen zijn en Cov(x,y) de population covariance.
De formule van de sample correlatiecoëfficiënt, r, is $r = \frac{Cov(x,y)}{s_x s_y}$ waar s_x en s_y de sample standard deviations van de twee variabelen zijn en Cov(x,y) de sample covariance.
Een handige regel om te controleren of er wel een relatie bestaat is: $|r| \geq \frac{2}{\sqrt{n}}$. Het bereik van de correlatiecoëfficiënt heeft een bereik van -1 tot 1, des te dichter r ligt bij de +1 des te dichter de datapunten op een stijgende lineaire lijn liggen. We noemen dit een positieve lineaire relatie. Des te dichter r bij de -1 ligt des te dichter de datapunten op een afnemende lineaire lijn liggen. We noemen dit een negatieve lineaire relatie. Wanneer r =

0 betekent dit dat er geen lineaire relatie is tussen x en y, maar let op: dit betekent niet dat er geen relatie is (alleen geen lineaire).

In de appendix van hoofdstuk 2:

In de meeste situaties, zal scheefheid (skewness) met een statistisch programma worden geconstrueerd of door midde van excel. Als skewness nul is of dicht bij de nul ligt, dan is de verdeling symmetrisch of neigt naar symmetrisch. Een negatieve waarde van skewness betekent dat de verdeling neigt naar links. Voor een positieve waarde van skewness neigt de verdeling naar rechts. $skewness = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$. Het belangrijke deel van deze vergelijking is de noemer. De teller geeft de toepassing van standaardisatie, die units van meting onbelangrijk maakt.

3. Kans elementen: waarschijnlijkheidsmethoden (Exclusief 3.5)

Een random experiment is een proces dat leidt tot twee of meer mogelijke uitkomsten, waarbij je niet precies weet welke uitkomst zich voor zal doen. Een goed voorbeeld hiervan is kop of munt, je weet dat er een van de twee zal uitkomen, maar je weet van te voren niet welke.

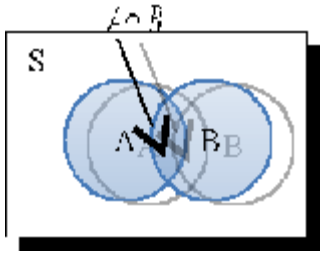
De mogelijke uitkomsten van een random experiment worden de basisuitkomsten genoemd en de set van de basisuitkomsten wordt de sample space genoemd. Om de sample space te noteren wordt de letter S gebruikt.

We moeten de basisuitkomsten op zo'n manier definiëren dat niet twee uitkomsten tegelijkertijd kunnen voorkomen; een random experiment moet dus noodzakelijk leiden tot het voorkomen van één van de basisuitkomsten.

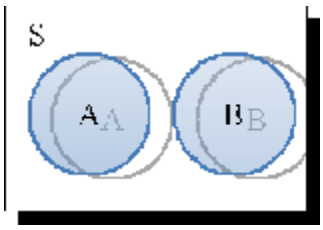
Een event, E, is een subset van basisuitkomsten van de sample space. Een event komt voor als het random experiment resulteert in een van zijn mogelijke basisuitkomsten. Het null event representeert de afwezigheid van een basisuitkomst en het symbool hiervoor is $\varnothing$.

Als A en B twee events zijn in de sample space S, dan wordt hun intersection (snijpunt) genoteerd door het volgende: $A \cap B$. Hun intersection is de set van alle basisuitkomsten in S die behoren tot zowel A al tot B. De intersection van A en B is alleen mogelijk als A en B allebei voorkomen. De term joint probability van A en B gebruiken we om de mogelijkheid tot intersection van A en B te kunnen aanduiden. Als we K gegeven events E_1, E_2, E_3, …, E_k, hebben en ook gegeven is dat $E_1 \cap E_2 \cap E_3 \cap ... \cap E_K$ dan is hun intersectie de set van alle basisuitkomsten die behoren tot elke $E_i (i = 1, 2, 3, ..., K)$.

Het kan voorkomen dat de intersection van twee events een lege set is.
Als de events A en B geen gemeenschappelijke basisuitkomsten hebben, dan worden ze mutually exclusive (beiden exclusief), er is dan ook geen intersection en van $A \cap B$ wordt dus gezegd dat het een lege set is, wat aantoont dat $A \cap B$ niet kan voorkomen.



Links een voorbeeld van intersection en rechts een voorbeeld waar geen intersection mogelijk is.

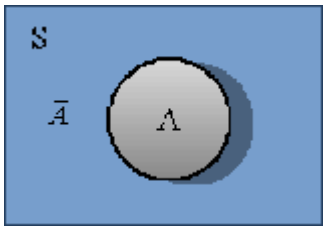


Als A en B twee events zijn in een sample space S. Hun union (unie) wordt genoteerd door $A \cup B$. Het is de set van alle basisuitkomsten in S dat ten minste aan een van deze twee events toebehoort. Let hierbij op dat de union $A \cup B$ alleen kan voorkomen wanneer zowel A als B voorkomen. Als we K gegeven events E_1, E_2, E_3, …, E_k, hebben en ook gegeven is dat $E_1 \cup E_2 \cup E_3 \cup ... \cup E_K$ dan is hun union de set van alle basisuitkomsten die behoren tot ten minste een van deze K events.

Als een union van verschillende events de hele sample space S bedekt, dan wordt ook wel gezegd dat de events collectief exhaustief zijn. Omdat elke basisuitkomst in S is, is elke uitkomst van het random experiment in ten minste een van deze events. Gegeven de

K events E_1, E_2, E_3, …, E_k in de sample space S, als $E_1 \cup E_2 \cup E_3 \cup ... \cup E_K$ = S, dan zijn deze K events collectief exhaustief (collectively exhaustive).

Als A een event is in de sample space S, de set van basisuitkomsten van een random experiment behoort toe aan S maar niet aan A, dan wordt dit de complement van A genoemd. Deze complement van A geven we weer met het volgende symbool $\bar{A}$ (uitspraak: 'A-bar'). Het is dus zo dat de events A en $\bar{A}$ mutually exclusive zijn, geen enkele basisuitkomst behoort tot beiden, en collectively exhaustive, elke basisuitkomst moet of tot A of tot $\bar{A}$ behoren.



Een classical probability is het aantal keer dat een event zal voorkomen, aangenomen dat alle uitkomsten in een sample space gelijk zijn aan de hoeveelheid dat ze zullen voorkomen. Door het aantal uitkomsten in een sample space die voldoen aan het event te delen door het aantal uitkomsten in een sample space weet je de probability van een event. Als je event A hebt dan is de probability van event A: $P(A) = \frac{N_A}{N}$ waarbij N_A het aantal uitkomsten is die voldoen aan de condities van event A en N het aantal uitkomsten in de sample space. Het idee hierachter is dat een probability door een fundamentele reden gevormd kan worden. De classical statement of probability vereist dat we de uitkomsten in de sample space geteld hebben. En dan gebruiken we onze telling om de probability te bepalen.

Het tellingsproces kan worden gegeneraliseerd door het gebruik van de volgende vergelijking om zo het aantal combinaties van n items ingenomen door k op een bepaald moment: $C_k^n = \frac{n!}{k!(n-k)!}$ waarbij 0!=1

Het totale aantal van mogelijke maier van het ordenen van x objecten in volgorde is gegeven door x(x – 1)(x – 2) …(2)(1) = x! Waarbij x! wordt uitgesproken als 'x faculteit'.

Het totale aantal permutations (omzettingen) van x objecten gekozen van n, P_x^n, is het aantal van mogelijke wijzen waarop x objexten geselecteerd zijn van een totaal van n dat op volgorde is gegeven: $P_x^n = n(n-1)(n-2)...(n-x+1)$, dit vermenigvuldigd en gedeeld door (n – x)(n – x – 1)…(2)(1)=(n – x)! geeft $P_x^n = \frac{n(n-1)(n-2)...(n-x+1)(n-x)(n-x-1)...(2)(1)}{(n-x)(n-x-1)...(2)(1)}$ is gelijk aan $P_x^n = \frac{n!}{(n-x)!}$.

Het aantal combinaties C_x^n van x objecten gekozen uit n is het aantal mogelijke selecties die gemaakt kan worden, dit getal is $C_x^n = \frac{P_x^n}{x!}$ oftewel $C_x^n = \frac{n!}{x!(n-x)!}$

De relative frequency probability is het limiet van de verhouding van het aantal keer dat

event A voorkomt in een groot aantal onderzoeken, n, $P(A)=\frac{n_A}{n}$ waarbij n_A het aantal keer dat A voorkomt is en n is het totale aantal van onderzoeken of uitkomsten. De probability is het limiet als n groter wordt (of neigt naar oneindig).

De subjective probability drukt de individuele mate van geloof op een kans dat een event zal voorkomen uit. Deze subjective probability wordt gebruikt in managementbesluitvorming.

Stel dat S de sample space van een random experiment is en O_i de basisuitkomsten en A een event. Voor elk event A in de sample space S gaan we er vanuit dat P(A) gedefinieerd is. Dan hebben we de volgende probability hypotheses: (1) Als A een event in sample space S is dan geldt $0 \le P(A) \le 1$. (2) Als A een event is in S, en O_i duidt de basisuitkomsten aan dan $P(A)=\sum_A P(O_i)$ waarbij de notatie impliceert dat de summatie zich uitstrekt over alle basisuitkomsten in A. (3) En daarbij is P(S)=1.

De eerste hypothese vereist dat de probability tussen de 0 en 1 ligt. De tweede hypothese kan uitgedrukt worden in relatieve frequenties. Neem aan dat een random experiment N keer herhaald wordt, laat Ni het aantal keer dat de basisuitkomst O_i voorkomt zijn, en laat N_i het aantal keer zijn dat event A voorkomt zijn. Omdat de basisuitkomsten elkaar uitsluitend (mutually exclusive) zijn, is N_A gewoon de som van N_i voor alle basisuitkomsten in A. Oftewel: $N_A=\sum_A N_i$ door te delen door het aantal proeven krijgen we $\frac{N_A}{N}=\sum_A \frac{N_i}{N}$.

Maar onder het concept van de relatieve frequentie van de probability is het zo dat N_A/N neigt naar P(A), en elke N_i/N neigt naar $P(O_i)$ als N oneindig groot wordt. Dus de tweede hypothese kan worden gezien als logische vereiste als je probability op deze manier bekijkt. De derde hypothese kunnen we herschrijven: wanneer een random experiment wordt uitgevoerd, moet er iets gebeuren.

Het vervangen van A door de sample space S in de tweede hypothese geeft: $P(S)=\sum_s P(O_i)$ waar de sommatie alle basisuitkomsten in de sample space beslaat. Maar omdat P(S) = 1 in gesteld in de derde hypothese volgt er dus dat $\sum_s P(O_i)=1$. Dit betekent dus dat de som van de probabilities voor alle basisuitkomsten in de sample space is 1.

We kunnen ook een aantal consequenties naar aanleiding van de drie hypotheses noemen. De eerste is dat als de sample space S bestaat uit n die gelijk zijn aan de basisuitkomsten E_1, E_2, E_3, …, E_k dan geldt: $P(O_i)=\frac{1}{n} i=1,2,3,...,n$. Dit is zo omdat de n uitkomsten de sample space dekken en ze zijn gewoonlijk gelijk. Ten tweede als de sample space S, bestaat uit gewoonlijk gelijke basisuitkomsten en event A bestaat uit n_A van deze uitkomsten, dit betekent $P(A)=\frac{n_A}{n}$. Ten derde als A en B mutually exclusive events zijn, dan is de probability van hun union de som van hun individuele probabilities: $P(A \cup B)=P(A)+P(B)$. In het algemeen als E_1, E_2, E_3, …, E_k mutually exclusive events zijn dan is $P(E_1 \cup E_2 \cup E_3 \cup ... \cup E_K)=P(E_1)+P(E_2)+P(E_3)+...+P(E_K)$. Dit komt voort uit de tweede hypothese: de probability van de union van A en B is gelijk aan $P(A \cup B)=\sum_{A \cup B} P(O_i)$ waarbij de sommatie alle basisuitkomsten in $A \cup B$ beslaat.

Maar omdat A en B mutually exclusive zijn, betekent het dat er geen enkele basisuitkomst aan allebei toebehoort: $\sum_{A\cup B} P(O_i) = \sum_{A} P(O_i) + \sum_{B} P(O_i) = P(A) + P(B)$. Ten vierde als E_1, E_2, E_3, …, E_k collectively exhaustive events zijn, dan is de probability van hun union $P(E_1 \cup E_2 \cup E_3 \cup ... \cup E_K) = 1$.

De complement regel is als volgt: $P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1$ wanneer A een event is en $\bar{A}$ de complement van A.

De addition rule of probabilities, als A en B twee events zijn, dan is de probability van hun union $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

De conditionele probability - ervan uitgaand dat A en B twee events zijn - van event A, gegeven dat event B voorkomt, wordt genoteerd als P(A|B) en wordt berekent als volgt: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ waarbij P(B) > 0 (want door nul mag je niet delen). Op dezelfde manier is de conditionele probability van event B, gegeven dat A voorkomt, genoteerd als P(B|A) berekent als volgt: $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$ waarbij P(A) > 0.

Door het gebruik van de multiplication rule of probabilities – ervan uitgaand dat A en B twee events zijn – kan de kans op intersection afgeleid worden van de conditionele kans als volgt: $P(A \cap B) = P(A \mid B)P(B)$ en ook $P(A \cap B) = P(B \mid A)P(A)$.

Stel dat A en B twee events zijn, deze events zijn dan statistically independent (statistisch onafhankelijk) als en alleen als $P(A \cap B) = P(A)P(B)$. Uit de multiplication rule volgt ook dat P P(B|A) = P(B) als P(A) > 0 en (A|B) = P(A) als P(B) > 0.

Meer algemeen als de events E_1, E_2, E_3, …, E_k zijn statistisch onafhankelijk als en alleen als $P(E_1 \cap E_2 \cap E_3 \cap ... \cap E_K) = P(E_1)P(E_2)P(E_3)...P(E_K)$.

Stel we hebben twee sets van events die we als volgt labelen A_1, A_2, A_3 …, A_h en B_1, B_2, B_3 …, B_k. Deze events Ai en Bj zijn mutually exclusive en collectively exhaustive binnen hun eigen set, maar intersections $(A_i \cap B_j)$ kunnen voorkomen tussen alle events van de twee sets. De intersections kunnen gezien worden als basisuitkomsten van een random experiment. Twee sets van deze events, gezamenlijk behandeld op deze manier, worden bivariate genoemd en de kansen worden bivariate probabilities genoemd.

In de context van bivariate kansen worden de intersection probabilities, P $(A_i \cap B_j)$, joint probabilities genoemd. De kansen voor individuele events $P(A_i)$ of $P(B_j)$ worden marginale kansen genoemd. Marginale kansen staan aan de randen in een tabel en kunnen gevormd worden door een corresponderende rij of kolom bij elkaar op te tellen.

Stel A en B zijn een paar van events, elk is opgedeeld in mutually exclusive en collectively exhaustive event categorieën die herkenbaar zijn aan de labellen A_1, A_2, A_3 …, A_h en B_1, B_2, B_3 …, B_k. Als elk event A_i is statistisch onafhankelijk van elk even B_j dan zijn A en B onafhankelijke events (independent events).

De odds (kansen) ten behoeve van een bepaald event is gegeven door de verhouding tussen de kans op het event gedeeld door de kans op de complement van het event. De kansen ten behoeve van A zijn als volgt: $Odds = \frac{P(A)}{1-(PA)} = \frac{P(A)}{P(\bar{A})}$

De overinvolvement ratio is de kans op event A_1, afhankelijk van event B_1, gedeeld door de kans op A_1, afhankelijk van event B_2: $\frac{P(A_1 \mid B_1)}{P(A_1 \mid B_2)}$. Een overinvolment ratio groter dan 1, $\frac{P(A_1 \mid B_1)}{P(A_1 \mid B_2)} > 1.0$ impliceert dat event A_1, stijgt de conditionale odds ratio ten behoeve van B_1: $\frac{P(A_1 \mid B_1)}{P(A_1 \mid B_2)} > \frac{P(B_1)}{P(B_2)}$.

4. Discrete probability verdelingen (Exclusief 4.5, 4.6)

Een random variabele is een variabele die de numerical waarde is die bepaald wordt door de uitkomst van een random experiment.

Een random variabele is een discrete random variabele als het niet meer dan een telbaar aantal waardes op zich kan nemen.

Een random variabele is een continuous (continue) random variabele als het elke waarde binnen een interval op zich kan nemen.

De probability distribution function, P(x), van een discrete random variabele X drukt de kans dat X de waarde x aanneemt uit, als functie van x: P(x) = P(X = x) voor alle waarden van x. Omdat de probability functie alleen waardes aanneemt die geen nul zijn op alleen discrete punten x, wordt het soms ook de probability mass function genoemd. Als eenmaal de kansen berekend zijn, kan een grafiek geplot worden.

Als X een discrete random variabele is met een probability distribution function P(x) dan ligt ten eerste P(x) voor elke waarde van x tussen de 0 en 1: $0 \le P(x) \le 1$ voor elke waarde van x. Ten tweede is de som van de individuele kans gelijk aan 1 voor alle mogelijke waarden van x: $\sum_x P(x) = 1$ (de notatie laat zien dat de sommatie voor alle mogelijke waarde van x geldt). De eerste regel laat zien dat kansen niet negatief kunnen zijn of groter dan 1 en de tweede regel komt voort uit het feit dat de events 'X = x', voor alle mogelijke waarden van x, mutually exclusive en collectively exhaustive zijn. De som van de kansen voor deze events moeten dus, gelijk zijn aan 1 (dus als je alle kansen optelt van deze events zal er 1 uitkomen). Het is een simpele manier om te zeggen dat wanneer een random experiment wordt uitgevoerd er iets moet gebeuren, er een uitkomst uit moet komen.

De cumulatieve probability function $F(x_0)$ voor een random variabele X, drukt de kans dat X niet meer bedraagt dan de waarde x_0 uit, als een functie van x_0: $F(x_0) = P(X \le x_0)$ waarbij de functie is onderzocht voor elke waarde van x_0.

Voor een discrete random variabele wordt de cumulatieve probability functie ook wel de cumulative mass function genoemd. Dit kan worden gezegd omdat als x0 stijgt, de waarden van de cumulatieve probability functie alleen zullen veranderen op de punten van x_0 die kunnen worden aangenomen door de random variabele met positieve kans.

Als probability functie P(x) van de random variabele X is en de cumulatieve probability functie $F(x_0)$ is dan is $F(x_0) = \sum_{x \le x_0} P(x)$ waarbij de notatie betekent dat de sommatie over alle mogelijke waarden van x mogelijk is die kleiner dan of gelijk zijn aan x_0. Deze vergelijking ontstaat omdat het event '$X \le x_0$' de union van de mutually exclusive events 'X = x' is voor alle mogelijke waarden van x kleiner dan of gelijk aan x_0. De kans op de union is dan de som van deze individuele event kansen.

Laat X een discrete random variabele met cumulatieve probability functie $F(x_0)$ zijn dan ligt ten eerste $F(x_0)$ tussen de 0 en de 1 voor elk waarde x_0: $0 \le F(x_0) \le 1$ voor elk waarde x_0. Ten tweede als x_0 en x_1 twee getallen zijn waarbij x_0 kleiner is dan x_1 dan is $F(x_0)$ kleiner dan of gelijk aan $F(x_1)$: $x_0 < x_1$ dan is $F(x_0) \le F(x_1)$. De eerste regel duidt aan dat kansen

niet kleiner kunnen zijn dan 0 en daarmee niet negatief kunnen zijn en dat ze niet groter kunnen zijn dan 1. De tweede regel impliceert dat de kans van een random variabele niet groter kan zijn dan de kans als geheel.

De verwachte waarde, E(X), van een discrete variabele X is gedefinieerd als volgt: $E(X)=m=\sum_x xP(x)$ waar de notatie laat zien dat de sommatie over alle mogelijke waarden van x geldt. De verwachte waarde van een random variabele wordt ook wel het gemiddelde genoemd en daarom ook vaak aangeduid met m.

We laten X een discrete random variabele zijn. De verwachting van de kwadratische verschillen met het gemiddelde, $(X-m)^2$, wordt de variance (variantie) genoemd, die we aanduiden met het symbool s^2 en te berekenen is door $s^2=E[(X-m)^2]=\sum_x (x-m)^2 P(x)$. De variantie van een discrete random variabele X kan ook worden weergegeven met de volgende vergelijking: $s^2=E(X)^2-m^2=\sum_x P(x)\cdot x^2-m^2$.

De standard deviation (standaard deviatie), s_x, is de positieve wortel van de variantie.

We laten X een discrete random variabele met probability functie P(x) en we laten g(X) een soort functie zij van X. De verwachte waarde, E[g(X)], kunnen we als volgt definiëren $E[g(X)]=\sum_x g(x)P(x)$. We definiëren de verwachting van een functie van een random variabele X door middel van de vergelijking $E[g(X)]=\sum_x g(x)P(x)$ omdat de verwachting namelijk kan worden gezien als de gemiddelde waarde dat g(X) zal aannemen over een groot aantal herhaalde onderzoeken. Over het algemeen is $E[g(X)]\neq g(m_x)$. Maar als g(x) een lineaire functie van x is, dan zijn er een aantal simpele resultaten voor het gemiddelde (mean) en de variantie (variance). Deze resultaten zijn handig voor economie en bedrijfseconomie omdat veel applicaties worden benaderd door lineaire functies te gebruiken. Als we de verwachte waarde en variantie voor een lineaire functie van een random variabele kunnen we proberen te bereken door middel van het gebruik van de lineaire functie a + bX, waarbij a en b constanten zijn. We laten X een random variabele zijn dat de waarde x aanneemt met probability P(x) en we gebruiken een nieuwe random variabele Y, die als volgt gedefinieerd wordt: Y = a + Bx. Wanneer de random variabele X een specifieke waar x aanneemt, dan moet Y de waarde a + bx aannemen. Het gemiddelde en de variantie voor zulke variabelen zijn nodig voor de berekeningen.

We laten X een random variabele zijn met variantie s_X^2 en gemiddelde (mean) m_X en a en b zijn constante vaste getallen. Het definiëren van random variabele Y is dus a + bX. Dan kunnen we het gemiddelde van Y berekenen door middel van de vergelijking $m_Y=E(a+bX)=a+bm_X$ en ook kunnen we de variantie van Y berekenen door middel van de vergelijking $s_Y^2=\mathrm{var}(a+bX)=b^2 s_X^2$. Ook kunnen we de standaard deviatie van Y berekenen door de vergelijking $s_y=|b|s_x$.

Er zijn drie speciale voorbeelden van de lineaire vergelijking W = a + bX die van belang zijn. Het eerste voorbeeld is de constante functie W = a voor elke gegeven constante a. Het gemiddelde is gelijk aan constante a: E(a) = a; en de variantie van a is gelijk aan nul: Var(a) = 0. Als een random variabele altijd de waarde a aanneemt is het dus altijd zo dat het gemiddelde gelijk is aan a en de variantie gelijk is aan nul.

In deze situatie is de tweede coëfficiënt b gelijk aan nul (b = 0). In het tweede voorbeeld is a gelijk aan nul (a = 0) waardoor de vergelijking wordt aangepast aan W = bX. Het gemiddelde is dan gelijk aan E(bX) = bm_X; de variantie is dan gelijk aan $Var(bX) = b^2 s_X^2$.

Ten derde kunnen we het gemiddelde en de variantie van de vergelijking $Z = \frac{X - m_X}{s_X}$ ook berekenen. Hierbij is het handig om te stellen dat constante $a = -\frac{m_X}{s_X}$ en dat constante $b = \frac{1}{s_X}$ in de lineaire functie Z = a + bX. Als we uitwerken zien we dat $Z = a + bX = \frac{X - m_X}{s_X} = \frac{X}{s_X} - \frac{m_X}{s_X}$. Het gemiddelde is dan dus gelijk aan $E\left(\frac{X - m_X}{s_X}\right) = -\frac{m_X}{s_X} + \frac{1}{s_X} \cdot m_X = 0$ en de variantie is dan dus gelijk aan $Var\left(\frac{X - m_X}{s_X}\right) = \frac{1}{s_X^2} \cdot s_X^2 = 1$.

We gaan een Bernoulli model ontwikkelen, dit is namelijk de basis van een binomiale verdeling. We gaan uit van een random experiment dat tussen twee verschillende mutually exclusive en collectively exhausive uitkomsten kan liggen. We kunnen deze uitkomsten de labels 'falend' en 'succesvol' geven. We laten P de kans op succesvol zijn. Hieruit kunnen we de kans op falend berekenen want dat is dan (1 – P). Daarna definiëren we de random variabele X zodat X de waarde 1 aanneemt als de uitkomst van het experiment succesvol is en anders de waarde 0. De probability functie van deze random variabele is dus P(1) = P en anders P(0) = (1 – P). Deze verdeling is beter bekend als de Bernoulliverdeling.

Het gemiddelde van een Bernoulliverdeling is $m_X = E(X) = \sum_x x \cdot P(x) = (0)(1-P) + 1(P) = P$ en de variantie van een Bernoulliverdeling is $s_X^2 = E[(X - m_X)^2] = \sum_X (x - m_X)^2 \cdot P(x) = (0 - P)^2 \cdot (1 - P) + (1 - P)^2 \cdot P = P \cdot (1 - P)$.

Een belangrijke generalisatie van de Bernoulliverdeling betreft het geval dat een random experiment met twee mogelijke uitkomsten verscheidende keren herhaald wordt en dat de herhaling onafhankelijk van elkaar zijn. We kunnen de kansen door middel van de binomial probability distribution bepalen.

Het aantal sequenties (reeksen) met x successen in n onafhankelijke onderzoeken is $C_x^n = \frac{n!}{x!(n-x)!}$ waarin $n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 1$ en 0! = 1. Deze C_x^n sequenties zijn mutually exclusive, omdat niet twee van ze op hetzelfde moment kunnen voorkomen. Het event 'x successen als gevolg van n onderzoeken' kan in C_x^n op elkaar uitsluitende manieren voorkomen, elk met de kans $P^x(1-P)^{n-x}$. Daarom door de addition rule of probabilities uit hoofdstuk 3 is de vereiste kans de som van deze C_x^n individuele kansen.

Als we van een random experiment uitgaan dat kan resulteren in twee mogelijke elkaar uitsluitende (mutually exclusive) en collectief exhaustieve (collectively exhaustive) uitkomsten, namelijk 'falend' en 'succesvol', en de P is de kans op succes in slechts een proef. Als n onafhankelijke onderzoeken gedaan zijn, dan is de verdeling van het aantal resulterende successen, x, de binomial distribution, oftewel de binomiale verdeling. De probability verdelingsfunctie voor de binomiale random variabele X = x is als volgt: P(x successen in n onafhankelijke onderzoeken) = $P(x) = \frac{n!}{x!(n-x)!} \cdot P^x (1-P)^{(n-x)}$ voor x = 0, 1, 2, 3, …, n.

Het gemiddelde van deze binomiale verdeling, waarin X het aantal successen in n onafhankelijke onderzoeken is, elk met kans op succes P, is: $m = E(X) = nP$ en de variantie van deze binomiale verdeling is dan $s_X^2 = E[(X - m_X)^2] = nP(1-P)$.

De binomiale verdeling wordt in bedrijfseconomie en economische applicaties over de hele wereld gebruikt. Echter voordat de binomiale verdeling kan worden opgesteld, moet de specifieke situatie goed geanalyseerd zijn om te kijken of het wel voldoet aan een aantal regels. De eerste regel is dat een applicatie uit verscheidende onderzoeken moet bestaan, die telkens maar uit twee uitkomsten kan bestaan zoals 'aan of uit', 'ja of nee' en 'succesvol of falend'. De tweede regel is dat de kans van de uitkomst hetzelfde is voor elk onderzoek. De derde regel is dat de kans van de uitkomst op een van de onderzoeken geen invloed mag hebben op de andere onderzoeken.

De joint probability van de discrete random variabele X en Y drukt de kans uit dat tegelijkertijd X de specifieke waarde x aanneemt en Y de specifieke waarde y aanneemt, als functie van x en y. De notatie die gebruikt wordt is P(x,y) en de joint probability wordt als vergelijking geschreven als volgt: $P(x,y) = P(X = x \cap Y = y)$.

De marginale probability functie is - ervan uitgaand dat X en Y een paar jointly distributed random variabelen zijn - in deze context de probability functie van de random variabele X en deze marginale probability functie wordt verkregen door opsomming van de joint probabilities over alle mogelijke waarden. Voor variabele X is dit $P(x) = \sum_y P(x,y)$. Voor de variabele Y is dit $P(y) = \sum_x P(x,y)$.

Als Y en X discrete random variabelen met joint probability functie P(x,y) dan zijn er twee regels. De eerste is dat 0 < P(x,y) < 1 voor elk paar waarden van x en y en ten tweede is de som van de joint probabilities P(x,y) voor alle mogelijke paren van combinaties van waarden van x en y gelijk aan 1.

De conditionele probability functie van een random variabele, gegeven de specifieke waarden van de andere random variabele, is de collectie van de conditionele kansen.

Als we Y en X een stel jointly distributed discrete random variabelen laten zijn. Dan is de conditionele probability functie van de random variabele Y, gegeven dat de random variabele X de waarde x aanneemt, laat de kans zien dat Y de waarde y aanneemt, als functie van y, wanneer de waarde x een vaste waarde is voor X. Dit wordt genoteerd als P(y|x) en de vergelijking is dus: $P(y \mid x) = \frac{P(x,y)}{P(x)}$. De conditionele probability functie van X, gegeven dat Y = y is dan dus $P(x \mid y) = \frac{P(x,y)}{P(y)}$.

De jointly distributed random variabelen X en Y zijn afhankelijk als en alleen als hun joint probability functie het product is van hun marginale probability functies: P(x,y) = P(x)P(y) voor alle mogelijke waarden van x en y. En k random variabelen zijn afhankelijk als en alleen als geldt $P(X_1, X_2, X_3, \ldots, X_k) = P(X_1)P(X_2)P(X_3)\ldots P(X_k)$.

Het conditionele gemiddelde (conditional mean) kan worden gecreëerd als volgt $m_{Y|X} = E[Y \mid X] = \sum_Y (y \mid x)P(y \mid x)$.

De conditionele variantie (conditional variance) kan worden gecreëerd op eenzelfde manier $s^2_{Y|X} = E[(Y - m_{Y|X})^2] = \sum_Y ((y - m_{Y|X})^2 \mid x) P(y \mid x)$.

Eerder hebben we al de verwachting van een functie van een single random variabele gedefinieerd. Nu kunnen we dit ook doen voor verschillende random variabelen. We laten X en Y een paar discrete random variabelen zijn met joint probability functie P(x,y). De verwachting van elke functie g(X,Y) van deze random variabelen is $E[g(X,Y)] = \sum_x \sum_y g(x,y) P(x,y)$.

Voor de algemene vergelijking W = aX + bY, een lineaire combinatie van random variabelen, kunnen we soortgelijks berekenen. W is de totale opbrengst random variabele (total revenue random variable) en ontstaat door de verkoop van elk product (producten X en Y) met de verkoopprijzen a en b. Het gemiddelde is te berekenen met de volgende formule $m_W = E[W] = am_X + bm_Y$. De variantie is te berekenen met de volgende formule $s^2_w = a^2 s^2_X + b^2 s^2_Y + 2abCov(X,Y)$.

De vergelijking kunnen we uitbreiden tot een lineaire combinatie van verschillende random variabelen $W = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_kX_k = \sum a_i X_i$ waarbij het gemiddelde te berekenen is door $m_W = E[W] = \sum_{i=1}^{k} a_i m_i$ en de variantie te berekenen is door $s^2_w = \sum_{i=1}^{k} a_i^2 s^2_i + 2\sum_{i=1}^{k-1} \sum_{j>1}^{k} a_i a_j Cov(X_i Y_j)$. De term Cov(X,Y) is de covariantie tussen de twee random variabelen die in het vervolg van de samenvatting zal worden uitgelegd.

De covariantie (covariance) is de sterkte van de joint variability voor twee random variabelen. De covariantie kan worden gebruikt bij het opstellen van de variantie van lineaire combinaties van random variabelen, zoals de variantie voor de totale waarde voor de combinaties van obligaties in een portefeuille.
Stel dat m_X het gemiddelde van de random variabele X is en m_Y het gemiddelde van de random variabele Y. De verwachte waarde van $(X - m_X)(Y - m_Y)$ wordt de covariantie tussen X en Y genoemd. Dit noteren we dus als Cov(X,Y). Voor discrete random variabelen is de covariantie op deze manier te berekenen:
$Cov(X,Y) = E[XY] - m_X m_Y = \sum_x \sum_y xyP(x,y) - m_X m_Y$.

De covariantie geeft een indicatie van de richting van de relatie tussen random variabelen (in dit geval waren dit X en Y), maar het geeft niet de sterkte van de relatie tussen de random variabelen aan. De correlatie geeft de sterkte van het verband wel aan. De correlatie tussen de jointly distributed random variabelen X en Y is als volgt:
$r = Corr(X,Y) = \frac{Cov(X,Y)}{s_X s_Y}$. Je ziet aan de vergelijking dat de correlatie van X en Y de covariantie gedeeld door de standaard deviatie van de twee variabelen is. De correlatie kan tussen de -1 en de 1 variëren. Een correlatie van 0 betekent dat er geen lineaire relatie is tussen de twee variabelen; als de twee variabelen onafhankelijk zijn is de correlatie gelijk aan nul. Een positieve correlatie betekent dat als een random variabele groot (klein) is, dat dan de andere random variabele een grotere kans heeft om groot (klein) te zijn, we zeggen dan dat de variabelen positief afhankelijk zijn. Perfecte positieve lineaire afhankelijkheid heeft een correlatie van +1.0. Een negatieve correlatie betekent dat als een random variabele groot (klein) is, dat dan de andere variabele een grotere kans heeft om klein (groot) te zijn, we zeggen dan dat de variabelen negatief afhankelijk

zijn. Perfecte negatieve lineaire afhankelijkheid heeft een correlatie van -1.0. Des te dichter het getal bij de +1.0 ligt des te sterker de positieve lineaire afhankelijkheid. Des te dichter het getal bij de -1.0 ligt des te sterker de negatieve lineaire afhankelijkheid. Des te dichter het getal bij de 0 ligt des te zwakker de lineaire relatie tussen de twee variabelen.

Let op: Als twee variabelen statistisch onafhankelijk zijn, dan is de covariantie tussen de twee variabelen gelijk aan nul. Maar dit betekent niet dat er bij voorbaat geen relatie is, maar er is geen lineaire relatie.

Stel dat m_X het gemiddelde van de random variabele X is en m_Y het gemiddelde van de random variabele Y. Hun varianties zijn gegeven door s_X^2 en s_Y^2. Dan gelden de volgende regels. Ten eerste is de verwachte waarden van hun som de som van hun verwachte waarden: $E(X+Y)=m_X+m_Y$. Ten tweede is de verwachte waarde van hun verschil het verschil tussen hun verwachte waarden: $E(X-Y)=m_X-m_Y$. Ten derde als de covariantie tussen X en Y gelijk is aan nul, dan is de variantie van hun som gelijk aan de som van hun variantie: $Var(X+Y)=s_X^2+s_Y^2$, maar als de covariantie tussen X en Y niet gelijk is aan nul, dan geldt: $Var(X+Y)=s_X^2+s_Y^2+2\mathrm{cov}(X,Y)$.

Ten vierde als de covariantie tussen X en Y gelijk is aan nul dan is de variantie van hun verschil de som van hun varianties: $Var(X-Y)=s_X^2+s_Y^2$, maar als de covariantie tussen X en Y niet gelijk is aan nul dan geldt: $Var(X+Y)=s_X^2+s_Y^2-2\mathrm{cov}(X,Y)$.
Als we X_1, X_2, X_3, …, X_K de random variabelen laten zijn met de gemiddelden $m_1,m_2,m_3,...,m_K$ en de varianties $s_1^2,s_2^2,s_3^2,...,s_K^2$ dan gelden de volgende regels. Ten eerste is de verwachte waarde van hun som als volgt: $E(X_1+X_2+X_3+...X_K)=m_1+m_2+m_3+...+m_K$. Ten tweede als de covariantie tussen elk paar van deze random variabelen gelijk aan nul is dan is de variantie van hun som als volgt: $Var(X_1+X_2+X_3+...X_K)=s_1^2+s_2^2+s_3^2+...+s_K^2$.

Nu gaan we het gemiddelde en de variantie voor de marktwaarde van waardepapieren opstellen. We stellen de random variabele X de prijs voor aandeel A, en de random variabele Y is de prijs voor aandeel B. De portfolio market value (de marktwaarde van het waardepapier), W, is gegeven door de lineaire functie W = aX + bY. Waar a is het aantal aandelen in voorraad A, en b is het aantal aandelen in voorraad B. De gemiddelde waarde voor W wordt als volgt bepaald $m_W=E[W]=am_X+bm_Y$, de variantie voor W is gelijk aan $s_w^2=a^2s_X^2+b^2s_Y^2+2abCov(X,Y)$, of als we de correlatie gebruiken om de variantie van W te berekenen $s_w^2=a^2s_X^2+b^2s_Y^2+2abCorr(X,Y)s_Xs_Y$.

5. Continue kansverdelingen (Exclusief 5.5)

De cumulatieve distributiefunctie (cumulative distribution function), F(x), voor een continue random variabele X drukt de kans dat X niet meer bedraagt dan de waarde van x, als functie van x: $F(x) = P(X \leq x)$.

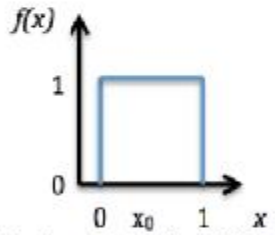
Laat X een continue random variabele met een cumulatieve distributiefunctie F(x), met a en b twee mogelijke waarden van X, waarbij a < b. De kans dat X tussen a en b ligt is P (a < X < b) = F(b) – F(a). Voor continue variabele maakt het niet uit of je schrijft 'kleiner dan' (<) of 'kleiner dan of gelijk aan' (≤), omdat de kans dat X precies gelijk is aan b gelijk is aan nul.

We laten X een continue random variabele zijn. We laten x elk getal in het bereik van deze random variabele zijn. De probability density function, f(x), van de random variabele is een functie die voldoet aan de volgende eisen. Ten eerste is f(x) > 0 voor alle waarden van x. Ten tweede is de oppervlakte onder de probability density functie, f(x), voor alle waarden van de random variabele, X, gelijk aan 1.0. Ten derde als we een geplotte density functie hebben waar a en b twee mogelijke waarden van de random variabele X, met a < b. Dan kan de kans dat X tussen a en b ligt de oppervlakte onder de density functie tussen de punten a en b: $P(a \leq X \leq b) = \int_a^b f(x)\,dx$. Ten vierde is de cumulatieve distributiefunctie, $F(x_0)$, de oppervlakte onder de probability density functie, f(x) tussen de minimum waarde van de random variabele X, x_m, tot de waarde x_0: $F(x_0) = \int_{x_m}^{x_0} f(x)\,dx$.

We laten X een continue random variabele zijn, met probability density functie f(x) en cumulatieve distributiefunctie F(x) dan is dus ten eerste de totale oppervlakte onder de probability density functie, f(x), gelijk aan 1. Ten tweede is de oppervlakte onder de curve f(x) links van x_0 gelijk aan $F(x_0)$ waarbij x_0 elke waarde is die de random variabele kan aannemen.

Nu overwegen we een probability density functie die een probability distributie, met een bereik van 0 tot 1, representeert. Een uniforme probability density functie is in een plaatje hieronder aangegeven. Voor elke uniforme random variabele gedefinieerd over het bereik van a tot b, ziet de probability density functie er als volgt uit:

$$f(x) = \begin{cases} \frac{1}{b-a} \rightarrow a \leq x \leq b \\ 0 \rightarrow else \end{cases}$$



Uniforme probability density functie

Deze functie kan worden gebruikt om de kans te vinden dat de random variabelen vallen binnen een bepaald bereik. We hadden al gezien dat de kans dat een random variabele tussen binnen bepaalden waarden ligt het oppervlakte onder de probability density functie tussen deze twee waarden is. De oppervlakte onder de gehele density functie is gelijk aan 1 en de cumulatieve probability $F(x_0)$ is de oppervlakte onder de density functie links van

het punt x_0.

De verwachte waarde van de random variabele is het gemiddelde van de genomen waarden, als het aantal replicaties oneindig groot wordt. Hiervoor gaan we ervanuit dat eren random experiment tot een uitkomst leidt dat door een continue variabele kan worden gerepresenteerd. En we gaan er vanuit dat er N onafhankelijke replicaties van dit experiment uitgevoerd zijn. De verwachte waarde van een random variabele noteren we door middel van E(X).

Op dezelfde manier als g(X) een functie van de random variabele X is, dan is de verwachte waarde van deze functie de gemiddelde waarden die door de functie wordt aangenomen over een herhaalde onafhankelijk aantal onderzoeken, als het aantal onderzoeken oneindig groot is. Deze verwachting noteren we door middel van E[g(X)]. Door het gebruik van de integraal kunnen we de verwachte waarden voor continue random variabelen berekenen op eenzelfde manier die wordt gebruikt voor discrete random variabelen: $E[g(x)] = \int_x g(x) f(x) dx$.

We kunnen ook het gemiddelde, de variantie en de standaard deviatie berekenen voor continue random variabelen. Hiervoor laten we X een continue random variabele zijn. Er zijn twee belangrijke verwachte waarden die over het algemeen worden gebruikt om continue probability distributions te definiëren. Het gemiddelde van X, als symbool m_X, is gedefinieerd als de verwachte waarde van X: $m_X = E(X)$. De variantie van X, als symbool s_X^2, wordt gedefinieerd als de verwachting van de kwadratische deviatie, $(X - m_X)^2$, het verschil tussen de random variabele en het gemiddelde: $s_X^2 = E[(X - m_X)^2]$, een alternatieve manier om dit te berekenen is $s_X^2 = E(X^2) - m_X^2$. De standaard deviatie van X, met symbool s_X is de wortel van de variantie: $s_X = \sqrt{s_X^2} = \sqrt{E[(X - m_X)^2]} = \sqrt{E(X^2) - m_X^2}$.

Voor een uniforme distributie $f(x) = \frac{1}{b-a}$ met een bereik van a tot b $a \le X \le b$ zijn ook gemiddelde, variantie en standaard deviatie te berekenen door middel van de volgende formules te berekenen. Het gemiddelde is te berekenen door $m_X = E(X) = \frac{a+b}{2}$, de variantie is te berekenen door $s_X^2 = E[(X - m_X)^2] = \frac{(b-a)^2}{12}$. De standaard deviatie is de wortel van de variantie.

Voor de lineaire functie van random variabelen kunnen we ook het gemiddelde, de variantie en de standaard deviatie berekenen. We hebben de functie W = a + bX. Het gemiddelde is te berekenen door $m_W = E(a + bX) = a + bm_X$. De variantie van deze functie is te berekenen door $s_W^2 = Var(a + bX) = b^2 s_X^2$. De standaard deviatie is te berekenen door $s_W = |b| s_X$.

Een speciale vorm is de gestandaardiseerde random variabele $Z = \frac{X - m_X}{s_X}$ die altijd een gemiddelde van 0 en een variantie van 1 heeft.

De normal probability distribution (de normale kansverdeling) is de continue random variabele kansverdeling die vaak in economische en bedrijfseconomische vraagstukken wordt gebruikt. Er zijn veel redenen om deze brede applicatie te gebruiken. De eerste is dat de normale verdeling de kansverdeling met een groot bereik van random variabelen dichtbij benadert. Ten tweede benaderen verdelingen van samples de normale distributie, als een 'grote' samplegrootte gegeven is. Ten derde is de berekening van kansen direct.

En ten vierde, en de meest belangrijke reden, heeft de normale kansverdeling tot goede besluitvorming voor een aantal applicaties geleid.

Een formele definitie van de normal probability density functie voor de random variabele X is $f(x)=\frac{e^{\frac{-(x-m)^2}{2s^2}}}{\sqrt{2ps^2}}$ voor $-\infty<x<\infty$, hierbij kunnen m en s^2 elk getal zijn zodat $-\infty<m<\infty$ en $-\infty<s^2<\infty$ geldt, waarbij e en p fysieke constanten zijn: e = 2.71828 en p = 3.14159…

Stel dat de random variabele X een normale distributie met parameters m en s^2. Dan moeten we de volgende eigenschappen van deze normale distributie in overweging nemen. Ten eerste is het gemiddelde van de random variabele gelijk aan m= E(X). Ten tweede is de variantie van de random variabele gelijk aan s^2= Var(X) = $E[(X-m)^2]$. Ten derde is de vorm van de probability density functie een symmetrische klokvormige curve met een top op het gemiddelde m. Ten vierde als we het gemiddelde en de variantie weten, kunnen we een definitie van de normale distributie geven door middel van de volgende notatie: $X\sim N(m,s^2)$.

Stel dat X een normale random variabele is met gemiddelde m, variantie s^2 en normale distributie $X\sim N(m,s^2)$. Dan is de cumulatieve distributiefunctie gelijk aan $F(x_0)=P(X\le x_0)$: dit is de oppervlakte onder de normale probability density functie links van x_0. Zoals voor elke goede density functie geldt dat de gehele oppervlakte onder de curve gelijk is aan 1: $F(\infty)=1$. We hebben geen simpele algebraïsche vergelijking om de cumulatieve distributiefunctie voor een normale gedistribueerde random variabele.

De algemene vorm van een cumulatieve distributiefunctie heeft een uitgerekte s vorm. We laten X een normale random variabele met cumulatieve distributiefunctie F(x) zijn en we laten a en b twee mogelijke waarden van X zijn waarbij a < b. Dan geldt dus $P(a<X<b)=F(b)-F(a)$. De probability is de oppervlakte onder de corresponderende probability density functie tussen de waarden a en b.

Als Z een normale random variabele met gemiddelde 0 en variantie 1 is, dan geldt dus $Z\sim N(0,1)$, hierdoor kunnen we zeggen dat Z de standaard normale distributie volgt. We noteren de cumulatieve distributiefunctie F(z) en a en b als twee mogelijke waarden van Z met a < b als volgt $P(a<Z<b)=F(b)-F(a)$.

We kunnen probabilities voor elke normale gedistribueerde random variabele vinden door eerst de random variabele naar de standaard normale gedistribueerde random variabele, Z, om te schrijven. Er is altijd een directe relatie tussen elke normale gedistribueerde random variabele en Z, deze relatie gebruikt de transformatie $Z=\frac{X-m_X}{s_X}$ waarbij X een normale gedistribueerde random variabele is: $X\sim N(m,s^2)$.

Een belangrijk antwoord geeft ons de mogelijkheid om de standaard normale tabel te gebruiken om de probabilities te berekenen die gelinkt zijn aan elke normale gedistribueerde random variabele. Een probability kan berekend worden voor de standaard normaal Z door de cumulatieve distributiefunctie van de standaard normale distributie in tabel 1 in de appendix te gebruiken. Deze tabel vertelt ons dat $F(z)=P(Z\le z)$ voor positieve (let op: dus voor niet-negatieve) waarden van z. Een voorbeeld hiervan is dat F(1.79) = 0.9633.

Het vinden van probabilities voor normale gedistribueerde random variabelen gaat als volgt. Hiervoor laten we X een normale gedistribueerde random variabele zijn met gemiddelde m en variantie s^2. Dan is de random variabele $Z = \frac{X-m}{s}$ en heeft deze random variabele Z een normale distributie $Z \sim N(0,1)$. Als we a en b mogelijke waarden van X laten zijn met a < b dan betekent dit dat $P(a<X<b) = P\left(\frac{a-m}{s} < Z < \frac{b-m}{s}\right) = F\left(\frac{a-m}{s}\right) - F\left(\frac{b-m}{s}\right)$ waarbij Z de standaard normale random variabele is en F de cumulatieve distributiefunctie.

De normale distributie kan worden gebruikt om de discrete binomial te benaderen ook de proportie random variabele die uitgebreid in de economie worden gebruikt kunnen hiermee worden benaderd. We nemen een probleem met n onafhankelijke onderzoeken in overweging. Elk van deze onderzoeken heeft een probability van succes P. We hadden in het voorgaande stuk al gezien dat de binomial random variabele X geschreven kion worden als de som van n onafhankelijke Bernoulli random variabelen: $X = X_1 + X_2 + X_3 + \ldots + X_n$ waarbij de random variabele X_i de waarde 1 aan neemt als de uitkomst van de i^{de} onderzoek succesvol is als dit niet succesvol (dus falend) is dan neemt dit de waarde 0 aan, met de respectievelijke probabilities 1 en (1-P). Het aantal X dat succesvol heeft als uitkomst heeft de binomiale distributie een gemiddelde $E(X) = m = nP$ en een variantie $Var(X) = s^2 = nP(1-P)$. Een regel is dat de normale distributie een goede benadering is om te gebruiken voor de binomiale distributie wanneer $nP(1-P) > 5$. Wanneer deze waarde kleiner is dan 5 dan gebruiken we de binomiale distributie om de probabilities te bepalen.

Door het gemiddelde en de variantie van de binomiale distributie te gebruiken en we hebben gevonden dat het aantal onderzoeken n groot is – zodat $nP(1-P) > 5$ – dan is de distributie van de random variabele $Z = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X - nP}{\sqrt{nP(1-P)}}$ de te benaderende standaard normale distributie.

Dit resultaat is belangrijk omdat we op deze manier, voor een grote n, de kans kunnen vinden dat het aantal successen in het gegeven bereik liggen. Als we de kans willen berekenen dat het aantal successen tussen a en b ligt dan doen we dit door middel van de volgende formule:

$$P(a \le X \le b) = P\left(\frac{a-nP}{\sqrt{nP(1-P)}} \le \frac{X-nP}{\sqrt{nP(1-P)}} \le \frac{b-nP}{\sqrt{nP(1-P)}}\right)$$

$$= P\left(\frac{a-nP}{\sqrt{nP(1-P)}} \le Z \le \frac{b-nP}{\sqrt{nP(1-P)}}\right)$$

In een aantal toegepaste problemen moeten we de kans berekenen voor een deel of een percentage intervallen. We kunnen dit doen door rechtstreeks te berekenen uit de normale distributiebenadering voor de binomiale distributie.

Een proportie random variabelen, P, kunnen berekend worden door het aantal successen, X, te delen door de sample grootte n: $P = \frac{X}{n}$. Daarna berekenen we het gemiddelde en de variantie van P door gebruik te maken van de lineaire transformatie van random variabelen. Het gemiddelde bereken we door middel van de formule $m = P$ en de variantie berekenen we door middel van de formule $s^2 = \frac{P(1-P)}{n}$. We kunnen het gemiddelde en de

variantie gebruiken met de normale distributie om de gewenste probability uit te rekenen.

Nu gaan we de joint cumulatieve distributiefunctie berekenen hiervoor laten we X_1, X_2, X_3, …, X_k een continue random variabele voor zijn. De joint cumulatieve distributiefunctie die hierbij hoort is $F(X_1, X_2, X_3, \ldots, X_k)$. Deze definieert de kans dat tegelijkertijd X_1 kleiner is dan x_1 en dat X_2 kleiner is dan x_2 en dat X_3 kleiner is dan x_3 enzovoort zodat $F(X_1, X_2, X_3, \ldots, X_k) = P(X_1 < x_1 \cap X_2 < x_2 \cap X_3 < x_3 \cap \ldots \cap X_k < x_k)$. De cumulatieve distributiefuncties – $F(x_1)$, $F(x_2)$, $F(x_3)$, …, $F(x_k)$ – van de individuele random variabelen worden de hun marginale distributiefuncties genoemd (marginal distribution functions). Voor elke i, is $F(x_i)$ de kans dat de random variabele X_i niet groter is dan de specifieke waarde x_i. De random variabelen zijn onafhankelijk als en alleen als $F(x_1, x_2, x_3, \ldots, x_k) = F(x_1), F(x_2), F(x_3), \ldots, F(x_k)$. Het is dus zo dat de onafhankelijkheid voor de continue random variabelen precies hetzelfde is als in de bij de discrete random variabelen. De onafhankelijkheid van een set van random variabelen impliceert dat de probability distributie van elk van hen niet wordt beïnvloed door de andere waarden.

Om de covariantie te berekenen laten we X en Y een paar continue random variabelen zijn met respectievelijk de gemiddelden m_X en m_Y. De verwachte waarde van $(X - m_X)(Y - m_Y)$ wordt de covariantie (Cov) tussen X en Y genoemd. En is dus te berekenen door $Cov(X,Y) = E[(X - m_X)(Y - m_Y)]$. Een alternatief om de covariantie te berekenen is $Cov(X,Y) = E(XY) - m_X m_Y$. Als de variabelen X en Y van elkaar onafhankelijk zijn, dan is de covariantie tussen de variabelen X en Y gelijk aan 0. Het omgekeerde hoeft niet per definitie waar te zijn (dat als de covariantie gelijk is aan nul dat X en Y dan van elkaar onafhankelijk zijn).

Om de correlatie (Corr) te berekenen laten we X en Y een stel jointly gedistribueerde random variabelen zijn. De correlatie tussen X en Y is dan te berekenen door $r = Corr(X,Y) = \frac{Cov(X,Y)}{s_X s_Y}$.

We laten X_1, X_2, X_3, …, X_K random variabelen zijn met de gemiddelden $m_1, m_2, m_3, \ldots, m_K$ en de varianties $s_1^2, s_2^2, s_3^2, \ldots, s_K^2$, dan moeten we de volgende regels in acht nemen. Ten eerste is het gemiddelde van de som van hun gemiddelden gelijk aan $E(X_1, X_2, X_3, \ldots, X_K) = m_1 + m_2 + m_3, + \ldots + m_K$. Ten tweede als de covariantie tussen elk paar van deze random variabelen gelijk is aan 0 dan is de variantie van hun som de som van hun varianties: $Var(X_1, X_2, X_3, \ldots, X_K) = s_1^2 + s_2^2 + s_3^2 + \ldots + s_K^2$. Maar als de covariantie tussen elk paar van deze variabelen niet gelijk is aan 0 dan is de variantie van hun som als volgt:

$$Var(X_1, X_2, X_3, \ldots, X_K) = s_1^2 + s_2^2 + s_3^2 + \ldots + s_K^2 + 2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} Cov(X_i, X_j)$$

We X en Y een paar continue random variabelen zijn met respectievelijk de gemiddelden m_X en m_Y en varianties s_X^2 en s_Y^2, dan moeten we de volgende regels in acht nemen. Ten eerste is het gemiddelde van hun verschil het verschil tussen hun gemiddelden: $E(X - Y) = m_X - m_Y$. Ten tweede als de covariantie tussen X en Y gelijk is aan 0 dan is de variantie van hun verschil als volgt: $Var(X - Y) = s_X^2 + s_Y^2$. Ten derde als de covariantie tussen X en Y niet gelijk is aan 0 dan is de variantie van hun verschil gelijk aan: $Var(X - Y) = s_X^2 + s_Y^2 - 2Cov(X - Y)$.

De lineaire combinatie van twee random variabelen X en Y is W = aX + bY waarbij a en b constante getallen zijn. De gemiddelde waarde voor W is gelijk aan $m_W = E[W] = E[aX + bY] = am_X + bm_Y$ en de variantie voor W is gelijk aan

$s_W^2 = as_X^2 + bs_Y^2 + 2abCov(X,Y)$ of als je voor de variantie te bereken de correlatie wilt gebruiken $s_W^2 = as_X^2 + bs_Y^2 + 2abCorr(X,Y)s_X s_Y$.

Als de lineaire combinatie in der vergelijking W = aX + bY een verschil is dan is dat W = aX – bY. Dan is het gemiddelde van W gelijk aan $m_W = E[W] = E[aX - bY] = am_X - bm_Y$ en is de variantie van W gelijk aan $s_W^2 = as_X^2 + bs_Y^2 + 2abCov(X,Y)$ of als je de correlatie wilt gebruiken om de variantie te berekenen is dit gelijk aan $s_W^2 = as_X^2 + bs_Y^2 + 2abCorr(X,Y)s_X s_Y$. De vergelijking W = aX – bY ontstaat doordat de coëfficiënt b in de vergelijking een negatieve waarde heeft. Als X en Y beiden joint normale gedistribueerde random variabelen zijn, dan is de resulterende random variabele W ook een normale gedistribueerde random variabele met een gemiddelde en variantie als hiervoor beschreven. Dit resultaat zorgt ervoor dat we de kans dat de lineaire combinatie, W, binnen een specifiek interval ligt kunnen berekenen.

6. De verdeling van een sample statics

Bij statistiek worden vaak samples gebruikt, omdat het moeilijk te meten is hoe het item in de populatie is en het is duur. Samples komen dicht in de buurt van de echte kenmerken van het item.

Simple random sample: bij een simple random sample hebben alle items gelijke kansen en zijn ze onafhankelijk gekozen. De kansen van elk item blijven steeds aan elkaar gelijk, ongeacht of ze al eerder getrokken zijn (bv. een dobbelsteen)

Sample informatie gebruik je om gevolgen te kunnen voorzien voor de hoofd populatie. Echter zijn er verschillen tussen de kenmerken van de sample en de hoofd populatie. Zo zal de *mean* van de sample een kleine afwijking hebben van de mean van de hoofd populatie.

Sampling distributions: voor een willekeurige sample wordt de mean weergegeven met $\bar{x}$. Elke sample heeft een verschilende waarde voor $\bar{x}$. Door alle verkregen waardes van de samples bij elkaar op te tellen en te delen door de hoeveelheid gebruikte samples is het gemiddelde, μ.

Door alle mogelijke sample mean $\bar{x}$ in een tabel te zetten. valt af te lezen wat de mogelijke kans op sample mean $\bar{x}$ in totaal is. Ook is het mogelijk om deze *probability* in een figuur te zetten met de x op de x-as en de probability op de y-as. Wat opvalt aan zo'n grafiek is dat de probability bij de $\bar{x}$ van de samples hoger ligt bij de $\bar{x}$ van de totale populatie.

Door samples te gebruiken, vergroot je de zekerheid van de gevonden uitkomst voor de populatie. Hoe meer samples er gebruikt worden, hoe waarheidsgetrouwer de uitkomst is. Samples worden op de volgende manier gebruikt: een n aantal samples worden gebruikt voor het onderzoek, waarbij μ het gemiddelde van de populatie is en σ^2 de variantie. X_1.., X_n zijn de hoeveelheid gebruikte samples.

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi \rightarrow E(\bar{X}) = E(\frac{1}{n}(X1 + \cdots Xn)) = \mu$, waarbij E de *expected value* is

Hieruit valt te concluderen dat de *sample mean* de *mean* van de populatie is. Hoe groter de populatie is, hoe dichterbij de sample mean bij de daadwerkelijke mean komt.

Vervolgens kijk je naar de variantie, die te bereken is met de formule:

$\text{Var}(\bar{X}) = \sum_{i-1}^{n}(\frac{1}{n})^2 * \sigma_i^2$, waarbij σ_x de *standard deviation* is. Hierbij zal bij en grote populatie de dichtheid van de samples toenemen.

Finite population correction factor: om de populatie variantie te bereken, wordt de correctie factor gebruikt. De formule ziet er als volgt uit: $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} * \frac{N-n}{N-1}$

De sample distribution valt weer onder te verdelen in verschillende soorten. De standaard vorm is de *normal distribution***.** De sample heeft een mean en standard deviation die gelijk zijn aan de populatie mean en deviation.

Standard normal distribution for the sample means: de formule $Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ maakt het mogelijk om de mean voor de samples te vinden. De normale verdeling is belvormig en hoog in het midden, bij de mean. De probability ligt overal op de grafiek tussen de 0 en de 1.

Central limit theorem: De *central limit theorem* laat zien dat veel probability distributions normaal verveeld zijn. De mean wordt namelijk gedeeld door de hoeveelheid samples, wat het uiteindelijke gemiddelde oplevert van de totale populatie.

Bij een normale verdeling is dit ook het geval, de mean ligt in het midden en heeft een *standard deviation*. Ook bij de central limit theorem geldt $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$, met μ als de mean, σ^2 als variance en $\bar{X}$ als de mean van de variabelen.
Law of large numbers: hoe groter de populatie, hoe nauwkeuriger je bij de oorspronkelijke mean komt en de variantie kleiner wordt.

Acceptance interval: een interval waarin een sample mean een hoge probability heeft, gegeven dat we de populatie mean en variantie weten.

Als de sample binnen dit interval valt, mag je aannemen dat de sample uit de populatie komt en daadwerkelijk gebruikt mag worden. Vallen de waarden buiten het interval, dan zal de productie niet doorgaan. De acceptance interval is te bereken aan de hand van: $\mu \pm z_{\alpha/2} * \sigma_{\bar{X}}$, waarbij α/2 de grootste mogelijkheid is. Door het acceptance interval te plotten, is duidelijk te zien of μ ook de juiste mean is van de populatie. Is dit niet het geval, dan zullen ingenieurs een andere mogelijkheid zoeken om een zo klein mogelijke devaition te behouden én om dichterbij de mean te zitten van de populatie. Wanneer de mean wel binnen het acceptance interval valt, zal er geen verdere actie worden ondernemen en kan het proces door gaan.

Sample proportion: X is de kans op succes in een *binomial sample* met n observaties en parameter P, waarbij P leden van de populatie met de te onderzoeken kenmerken heeft. De formule voor sample proportion ziet er als volgt uit: $\hat{p} = \frac{X}{n}$, waarbij $\hat{p}$ de *mean* van variabelen is.

De kans op succes in de binominale verdelingen komen dicht in de buurt van de kans op succes in de normale verdelingen, waardoor de mean en variance van $\hat{p}$ makkelijk te berekenen zijn. De waarde van mean moet wel aan nP(1-P) > 5 voldoen, anders is er geen sprake van een normale verdeling.

$E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n} E(X) = P$, waarbij P de mean is:

$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$, waarbij σ de standard deviation is;

Als de sample een grote waarde is, dan gebruik je $Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}}$.

Ook hier geldt weer dat hoe groter de samples zijn, hoe nauwkeuriger de sample mean bij de daadwerkelijke mean komt.

Naast sampling distributions van de sample means en proportions, bestaat er ook nog de sampling distribution van sample variances. Bij productie is het van belang dat de sample variance zo laag mogelijk ligt, zodat alle producten zo optimaal mogelijk gebruikt kunnen worden. Producten van een zeer hoge kwaliteit hebben over het algemeen een lagere variance dan de producten van lage kwaliteit. Om de variance van een sample te bereken gebruik je de formule: $\sigma^2 = E[(X-\mu)^2]$, waarbij μ de onbekende mean is, die ook weergegeven kan worden met $(x_i - \bar{x})^2$

De formule geherformuleerd, levert het volgende op: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(xi - \bar{x})^2$. S is de *sample standard deviation*.
Als de populatie normaal verdeeld is, dan zijn de sample variance en populatie variance verbonden aan elkaar door een probability distribution, ook wel de *chi-square distribution* genoemd.

Chi-square distribution: $= \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{\sigma^2}$. De distribution is χ^2 (dit is chi-square) en n-1 is de vrijheidsgraad. Het bepalen van de kansen voor sample variances is belangrijker om aan te kunnen tonen dat het een normale verdeling is dan het bepalen van de kansen van de sample means.

Wanneer je een density functie plot, zal deze positief zijn, ook bij een chi-square functie. Het verloop van een chi-square functie zal bepaald worden door de vrijheidsgraad *v*. In formule vorm ziet het er als volgt uit χ^2_v,
De functie van de chi-square mean is $E(\chi^2_v) = v$. Als de functie verder wordt uitgewerkt, krijg je:
$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1) \rightarrow E(s^2) = \sigma^2$
De functie van de variance van de chi-square is $Var(\chi^2_v) = 2v$. Als je deze functie verder uitwerkt, krijg je:
$Var\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$ **à** $Var(s^2) = \frac{2\sigma^4}{(n-1)}$

Om de sample variance te berekenen heb je dus een hoeveelheid samples nodig, *n*,de standard devation, *σ*, een chi-square waarde, K = v = n-1 of χ^2_{n-1} bij het limiet. De waarden vul je in in de formule: $P(s^2 </> K) = P\left[\frac{(n-1)s^2}{\sigma^2}\right] </> \chi^2_{n-1}$ = limiet. Daarna herschrijf je de formule naar $K = \frac{\sigma^2\chi^2_{n-1}}{n-1}$. Als de K-waarde overeenkomt met σ^2 dan is er sprake van het optimale punt, wanneer dit niet het geval is, zal er actie ondernomen moeten worden.

7. The confidence Interval: een enkele populatie

Estimator: een willekeurige variabele dat afhankelijk is van de sample informatie; De waarde bepaalt de benaderingen van de onbekende parameter.
Estimate: een specifieke waarde van de willekeurige variabele

De estimator is het proces en de estimate een resultaat van dit proces. De sample variance, s^2, kan ook een estimator zijn.

Om een onbekende parameter te bereken, zijn er verschillende soorten *estimates*. De eerste is het de *point estimate* en de tweede de *confidence interval*. Bij de point estimate is één nummer, terwijl bij de confidence interval de schatting tussen twee waardes ligt.

Point estimator: een functie van de sample informatie dat een enkel nummer geeft, bijvoorbeeld $\bar{X}$ is de sample mean van de populatie mean, μ
Point estimate: het enkele nummer van de *point estimator*, bijvoorbeeld de waarde van $\bar{X}$, die wordt weergegeven met $\bar{x}$.

Estimators zijn weer onder te verdelen in *unbiasedness* en *efficiency*.
Unbiased estimator: de point estimator $\bar{\theta}$ is gelijk aan een onbekende populatie parameter θ, dus $E(\bar{\theta}) = \theta$.

Het kan voorkomen dat er een klein verschil tussen θ en $\bar{\theta}$ zit, door meerdere samples te gebruiken zullen de waarden uiteindelijk wel gelijk aan elkaar zijn. De sample mean, sample variance en sample proportion zijn ook allemaal unbiased estimators, omdat deze gelijk zijn aan de populatie mean, variance en proportion.
Bias estimator: een niet unbiased estimator. De *bias* is het verschil tussen de *mean* en θ: $\text{Bias}(\bar{\theta}) = E(\bar{\theta}) - \theta$ en bij een unbiased estimator is de bias 0.

Efficient estimator: de estimator die de kleinste variance heeft is de *most efficient estimator* of de *minimum variance unbaised estimator.* Dit komt alleen voor bij verschillende *unbaised estimators.* Het verband dat tussen de verschillende estimators bestaat heet de **relative efficiency** en ziet er in formule vorm als volgt uit:
Relative efficiency $= \frac{Var(\bar{\theta}2)}{Var(\bar{\theta}1)}$. Hoe hoger deze waarde ligt, hoe minder efficiënt $\bar{\theta}2$ is, dus is $\bar{\theta}1$ betrouwbaarder om mee te werken. De relative efficiency ratio geldt alleen tussen de mean μ en median, tussen de mean μ en de $\bar{x}$, tussen de proportion P en $\bar{p}$ en de variance σ^2 en s^2.

Confidence interval estimator: het bepalen van het interval waarin de parameter van een populatie valt, gebaseerd op de sample informatie.
Confidence interval estimate: de confidence interval estimator waarden, a en b

Als er steeds opnieuw samples uit dezelfde populatie worden gehaald, zal er een hoog percentage steeds hetzelfde interval hebben waar de waarde van de onbekende parameter op moet liggen. De *confidence interval estimator* wordt een *confidence interval estimate*. Hierbij is de *variance* bekend.

Confidence interval: bij parameter θ geldt $P(A < \theta < B) = 1 - \alpha$, waarbij α tussen de 0 en 1 ligt. Vermenigvuldig 1- α met 100% en je hebt je confidence level gevonden. Vaak is er een kleine afwijking van de daadwerkelijke waarde van de parameter, dit valt te corrigeren door het *best point estimate± error factor* te doen.

Bij een normale verdeling is de formule $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ van toepassing en $z_{\alpha/2}$ is de waarde van de *standard normal distribution* en α/2 geeft de vorm weer. Om de *confidence level* te bereken, ziet de formule er als volgt uit:

$1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = P\left(\frac{-z_{\alpha/2}\sigma}{\sqrt{n}} < \bar{x} - \mu < \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = P\left(\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)$.

Margin of error/sampling error: de correctie die wordt toegepast op de gevonden waarde om op de echte waarde van de parameter te komen. Deze correctie is $\frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ = ME.
Aan zowel de linker- als de rechterkant moet een correctie worden toegepast. Aan de linkerkant heet deze correctie de *lower confidence limit (LCL)* en wordt weergegeven door $\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ en aan de rechterkant heet deze de *upper confidence limit (UPL)* en wordt weergegeven door $\bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$. Om de totale correctie te vinden, gebruik je de width (w): w = 2(ME). De *margin of error* wordt bepaald door de *standard deviation*, *sample size n* en de *confidence level*.
Door de *standard deviation* kleiner te maken, en de overige parameters constant te houden, zal de ME ook kleiner worden. Ook de *sample size* vergroten of het laten afnemen van de *confidence level* (dus α kleiner maken), zal een kleinere ME opleveren.

Het komt vaak genoeg voor dat de *variance* onbekend is. Gosset heeft echter een manier gevonden om toch met de gegevens te kunnen werken. De verdelingen die Gosset heeft gecreeërd, worden de *Student's t distributions* genoemd. De *standard normal distribution* functie $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ heeft Gosset omgezet in de functie: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. De *standard deviation* van de populatie, σ, heeft hij vervangen door de *standard sample deviation*, s.

Onder de *student's distribution* familie valt ook de vrijheidsgraad, v, functie, t_v. Deze functie lijkt veel op de standaard normale verdeling, het heeft een *mean* van 0, maar heeft een grotere *standard deviation*. Wordt v een steeds groter getal, dan zal de grafiek meer op de standaard normale grafiek lijken. Om de kansen van de vrijheidsfunctie te bereken, is de formule: $P(t_v > t_{v,\alpha/2}) = \alpha/2$ nodig.

Student's distribution functies hebben ook een *confidence interval*. In de formule veranderen alleen de z en de σ door de t_v en de s. De nieuwe formule ziet er als volgt uit: $\bar{x} - \frac{t_{n-1,\alpha/2}s}{\sqrt{n}} < \mu < \bar{x} + \frac{t_{n-1,\alpha/2}s}{\sqrt{n}}$. en de *margin of error* wordt: $ME = \frac{t_{n-1,\alpha/2}s}{\sqrt{n}}$

Bij zeer grote samples wordt de *sample propotion* steeds nauwkeuriger. De formule $Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ wordt gebruikt om de *confidence interval* voor de popuatie *proportion* te bereken. Hierbij is p̈ de *proportion of successes* en P de *probability of success*.
Om de *proportion* te bereken, gebruik je de zelfde formule als die van de *mean*. De nieuwe formule is:

$\hat{p} \quad z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Ook de *margin of error* verandert: ME = $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

De *variance* bij een *confidence interval* zal als volgt veranderd:

Voorheen was er bij een *chi-square distribution* de formule $(\chi^2_v) = \frac{(n-1)s^2}{\sigma^2}$, waarbij v = n – 1. Om de kansen te bereken, voeg je α nog toe in de formule. De nieuwe formule ziet er als volgt uit:

$P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$

Bij de *confidence interval* zal de formule niet α zijn, maar α/2. De formule wordt als volgt:

$P(\chi^2_{n-1,1-\alpha/2} < \chi^2_{n-1,\alpha/2} < \chi^2_{n-1,\alpha/2}) = 1 - \alpha$

Met deze formule kan nu de *population variance* gevonden worden:

$1 - \alpha = P(\chi^2_{n-1,1-\alpha/2} < \chi^2_{n-1,\alpha/2} < \chi^2_{n-1,\alpha/2}) = P(\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}) = P(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$

Wanneer de *sample size n* meer dan 5% van de totale populatie grootte bevat, wordt er gerekend met de *confidence interval*. De *confidence interval* wordt aan de hand van de volgende stappen berekend:

a. Eerst bereken je de *mean*: $\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
b. Vervolgens bereken je de *variance*: $\sigma^2_{\bar{x}} = \frac{s^2}{n}\frac{(N-n)}{(N-1)}$
c. Als laatst vul je de waarden in in de formule: $\bar{x} - t_{n-1,\alpha/2}\widehat{\sigma_{\bar{x}}} < \mu < \bar{x} + t_{n-1,\alpha/2}\widehat{\sigma_{\bar{x}}}$

Bedrijven zijn niet geïnteresseerd in het de gemiddelde populatie, maar in de totaal populatie. Alle formules vermenigvuldig je met N en zien er als volgt uit:

mean: $N\bar{x}$

variance: $N\sigma^2_{\bar{x}} = \frac{Ns}{\sqrt{n}}\frac{(N-n)}{(N-1)}$

confidence interval: $N\bar{x} - t_{n-1,\alpha/2}N\widehat{\sigma_{\bar{x}}} < \mu < N\bar{x} + t_{n-1,\alpha/2}N\widehat{\sigma_{\bar{x}}}$