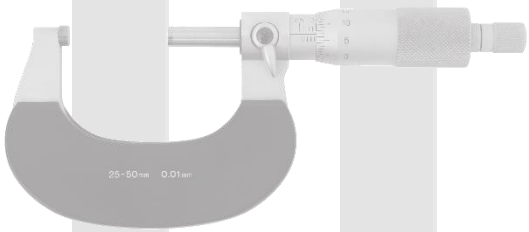
STATISTIK MED TILFÆL-DIGT GENEREREDE TAL

UNDERVISNINGSELEMENT

A6

—

UNDERVISNING
I MÅLETEKNIK

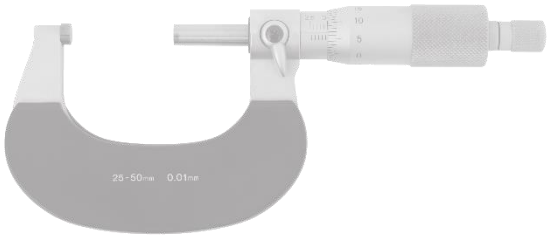


metrologi.dk

STATISTIK MED TILFÆLDIGT GENEREREDE TAL

Mathias Geisler & David Balslev-Harder, DFM A/S

1. udgave – August 2019, redigeret oktober 2019



Metrologi.dk er finansieret af Styrelsen for Forskning og Innovation i perioden 2016 – 2020. Materiale er udarbejdet i et samarbejde mellem GTS-institutterne DFM A/S og FORCE Technology.
Læs mere om projektet på www.metrologi.dk.

Grafisk design af: Henriette Schäfer Høyrup og David Balslev-Harder.

Indholdsfortegnelse

Indledning

Dette kompendium fungerer som baggrundsviden for de to akkompagnerende Excel-ark, *introduktionTilStatisik.xlsx* og *normalfordelinger.xlsx*.

Statistik undervises normalt fra et teoretisk matematisk grundlag eksemplificeret med datasæt fra den virkelige verden. I dette læremiddel tager vi udgangspunkt i, at man typisk skal se virkeligt mange datasæt for ordentligt at få en fornemmelse for, hvordan statistik fungerer. For et af de vigtigste elementer i at kunne se statistikken udfolde sig er, at man skal have mange datapunkter – rigtigt mange. Heldigvis sidder du formentlig og læser dette kompendium på det perfekte værktøj til at generere disse datapunkter: din computer. Computeren kan, ved hjælp af forskellige algoritmer, generere tilfældige tal i store mængder på kort tid, hvilket sparer os turen i laboratoriet, før vi kan analysere vores data. Selvom datasættet er kunstigt genereret, kan vi stadig bruge det som grundlag til at foretage forskellige statistiske test. Fordelen herved er, at vi kontrollerer de statistiske parametre (f.eks. middelværdi og spredning), så vi kan se, hvordan resultaterne af de forskellige test ændrer sig, når vores data gør det – helt uden at flytte os fra stolen.

1 Tilfældigt genererede tal

Før vi begynder, er det vigtigt at forstå, hvad det præcist betyder, når vi bruger udtrykket ”tilfældigt genererede tal”. Mange af dem er nemlig ikke helt så tilfældige, som navnet ellers kunne antyde. Der findes forskellige metoder, der kan generere tallene efter en algoritme, altså en forudbestemt opskrift. På den måde får man det, der kaldes pseudotilfældige tal. Fordi tallene er genereret efter en sådan forskrift, er det muligt at forudsige, hvad det næste tal i rækken er, og de er altså ikke helt så tilfældige alligevel.

Om dette er et problem afhænger af anvendelsen. Til f.eks. kryptering er det helt essentielt, at de brugte tal er 100 % tilfældige. Det giver ikke synderligt god sikkerhed, hvis en spion blot kunne sidde og regne sig frem til den hemmelige nøgle brugt i krypteringen og derefter nemt afkode de fortrolige meddelelser.

Boks 1: Helt tilfældige tal

Rigtigt tilfældige tal er essentielle for mange anvendelser, f.eks. til online kasinoer og kryptering til sikker kommunikation. Der findes forskellige måder at lave dem på. En mulighed er atmosfærisk støj målt med radioer (den skratten man hører, når man ikke er tunet ind på en kanal), som bliver brugt på hjemmesiden random.org.

En anden blev brugt i et storstilet forskningsprojekt, der tilsigtede at vise nogle af de fundamentale egenskaber ved kvantemekanikken. I The Big Bell Test engagerede forskerne almindelige mennesker med forskellige minispil til at generere computerbits (0’ere og 1’ere), som efterfølgende blev anvendt i deres eksperiment – med stor succes!

Til vores formål er det dog ikke helt så vigtigt, tallene kun er pseudotilfældige. Så længe disse tal er genereret ”tilfældigt nok”, kan vi stadig bruge dem, fordi fordelingen af tallene er tilfældig nok (vi får f.eks. ikke otte 2-taller i træk).

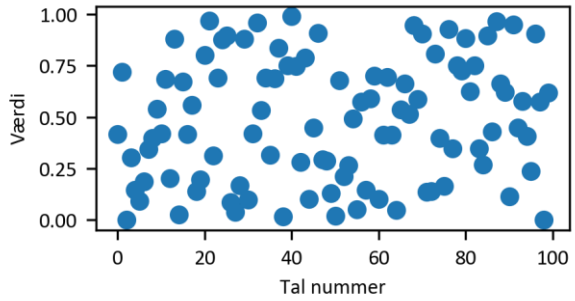
Den mest brugte algoritme til at generere pseudotilfældige tal kaldes The Mersenne Twister, som også anvendes i Excel fra Office 2010 og frem. Dybest set er det en virkelig lang liste af tal (den går igennem $2^{19937} - 1$, før den starter forfra), som bliver udregnet, efterhånden som vi har behov for det. Begyndelsespunktet i listen kan i de fleste implementeringer vælges ved at bruge et såkaldt seed, hvilket gør det muligt at reproducere resultater. I Excel vælges dette seed dog internt i programmet og kan derfor ikke sættes eksplicit af brugeren.

The Mersenne Twister består en række forskellige statistiske test af talrækkens tilfældighed. Dette gør den tilfældig nok til, at den bruges af forskere verden over til deres simuleringer i forskellige sammenhænge – og den er derfor også rigeligt god til vores formål.

Selvom den korrekte benævnelse altså er pseudotilfældige tal, vil vi for nemheds skyld droppe præfikset ”pseudo” i resten af kompendiet og blot betegne de genererede tal som tilfældige.

2 Implementering i Excel

Den helt grundlæggende funktion til generering af tilfældige tal i Excel er RAND(), der kan give et tal i intervallet [0,1[med uniform sandsynlighed. Som notationen her angiver, er 0 med i intervallet, mens 1 ikke er det. Uniform sandsynlighed betyder, at et hvilket som helst tal i intervallet har lige stor sandsynlighed for at fremkomme. Ved at generere et datasæt bestående af et tilstrækkeligt antal elementer baseret på RAND(), kan vi inspicere, om funktionen virker, som vi forventer.

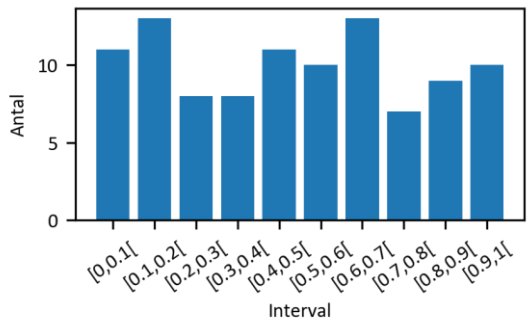


Figur 1: Værdien af 100 tal tilfældigt genereret med RAND().

I Figur 1 ses et plot af et datasæt bestående af 100 genererede talværdier. Som det ses har vi tal i hele intervallet, og der lader ikke til at være klynger af punkter, der kunne indikere en skævvridning. En mere illustrativ måde at inspirere den slags data på er dog med et histogram, som kan ses i Figur 2.

Opgave:

Fra en uniform sandsynlighedsfordeling i intervallet [0,1[ville vi umiddelbart forvente et fladt histogram med 10 tal i hver søjle. Hvorfor er der ikke lige mange tal i hvert interval? Hvad kan vi gøre for at jævne de relative forskelle ud?



Figur 2: Fordelingen af de 100 tal i Figur 1.

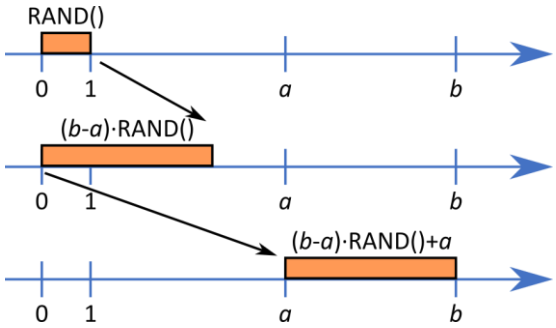
I de kommende underafsnit vil vi undersøge, hvordan vi kan kombinere tilfældige tal for at simulere forskellige situationer og på den måde illustrere nogle af styrkerne ved tilfældighedsgeneratorer.

2.1 Tilfældige tal i et ønsket interval

Ved at manipulere funktionen RAND() kan vi opnå andre sandsynlighedsfordelinger end blot en uniform i intervallet [0,1[. Vi kan f.eks. frembringe tal i et vilkårligt interval [a,b[ved at bruge udtrykket

$$(b-a)\cdot \mathrm{RAND}() + a.$$

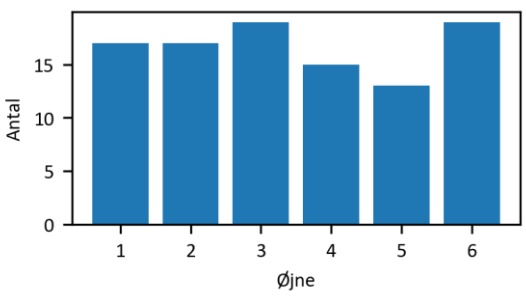
Denne proces er skematisk illustreret i Figur 3, hvor det ses, hvordan intervallet først strækkes med faktoren (b-a), og derefter flyttes med a.



Figur 3: Ændring af intervallet for RAND() fra [0,1[til [a,b[.

2.2 Simulering af sekssidet terning

Ved at bygge videre på ovenstående metode kan vi nemt simulere noget, som de fleste har hverdagserfaring med, nemlig kast med en sekssidet terning. Som bekendt er der seks mulige udfald af et sådant kast. Dette kan vi simulere ved at generere tilfældige tal i intervallet [0,6[og altid runde resultatet op, altså ved



Figur 4: Simulering af 100 kast med en sekssidet terning.

at vælge $a = 0$ og $b = 6$. Den uniforme fordeling sikrer, vi har lige stor sandsynlighed for at lande i alle delintervallerne [0,1[, [1,2[, osv., mens afrundingen udelukkende giver os heltallige udfald som på en terning.

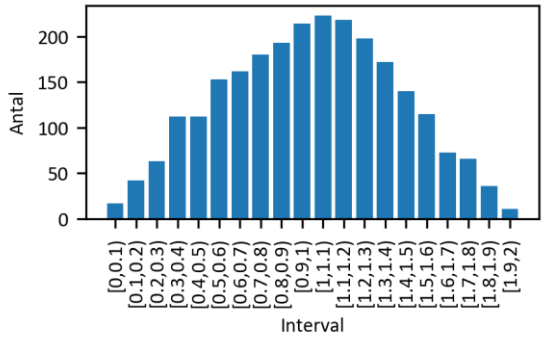
Resultatet af en simulering med disse parametre kan ses i Figur 4 for 100 kast. Som i Figur 2 kan vi se, at forekomsten af de forskellige slag ikke er ligeligt fordelt på de seks udfald, hvilket vi heller ikke ville forvente grundet den statistiske usikkerhed forbundet med simuleringen.

2.3 Summen af to tilfældige tal

Hvad sker der, hvis vi i stedet for fordelingen af et enkelt tilfældigt tal kigger på summen af to tal?

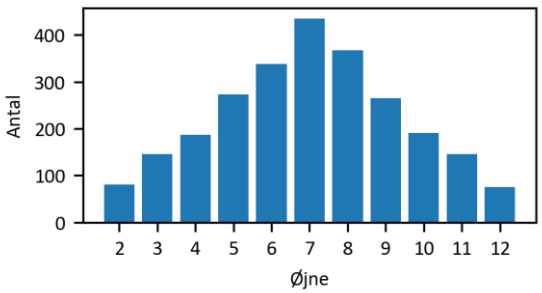
Hvis vi tager udgangspunkt i RAND() og kigger på fordelingen af RAND() + RAND(), kan vi først og fremmest se, intervallet nu er [0,2[(summen af 0+0 og 1+1, hhv. de laveste og højeste værdier begge fordelinger kan antage).

Man kunne nu forestille sig, vi igen ville have en uniform sandsynlighedsfordeling i intervallet [0,2[, men det er ikke tilfældet! I Figur 5 ses resultatet af at summere to sæt af 2500 tilfældigt genererede tal. Her ser vi en tydelig ændring fra fordelingen i Figur 2 med langt større sandsynlighed for at få et tal i midten af intervallet end i yderpunkterne (den såkaldte trekantsfordeling). Hvordan kan vi forstå det? For at få et tal, der ligger i f.eks. den høje ende af intervallet, kræver det, at *begge* tal på samme tid er høje, hvor det samme gør sig gældende i den lave ende. Omvendt kan vi ramme midten med flere forskellige kombinationer. Et dagligdags eksempel på dette kender vi fra et slag med to sekssidede terninger. De fleste har oplevet, man langt oftere får et slag i omegnen af syv end f.eks. to seksere. Da udfaldet fra en sekssidet terning er tilfældigt, burde summen af to sekssidede terninger altså give samme sandsynlighedsfordeling som den, vi ser i Figur 5.



Figur 5: Fordelingen af summen af to sæt af 2500 tilfældige tal.

Eksperimentet er igen nemt at lave med tilfældigt genererede tal. Ved at følge samme procedure som for én terning og summere to sådanne sæt hver bestående af 2500 tal kommer vi frem til en fordeling som vist i Figur 6. Vi ser ikke overraskende en slående lighed med trekantsfordelingen, der fremkom ved bare at summere to tilfældige tal mellem 0 og 1.



Figur 6: Fordelingen af 2500 simulerede kast med to sekssidede terninger.

Vi kan også se, at fordelingen forekommer mere trekantet end i Figur 5. Dette skyldes afrundingen, som giver en større søjlebredde end før. Fordi vi samler flere tal i hver søjle, får vi lavere usikkerhed, og vi får derfor en "pænere" trekant. Man skal dog generelt være påpasselig med at bruge for få (eller for mange) søjler i histogrammer, da dette kan give misvisende repræsentation og fortolkning af datasættet.

2.4 Den centrale grænseværdisætning

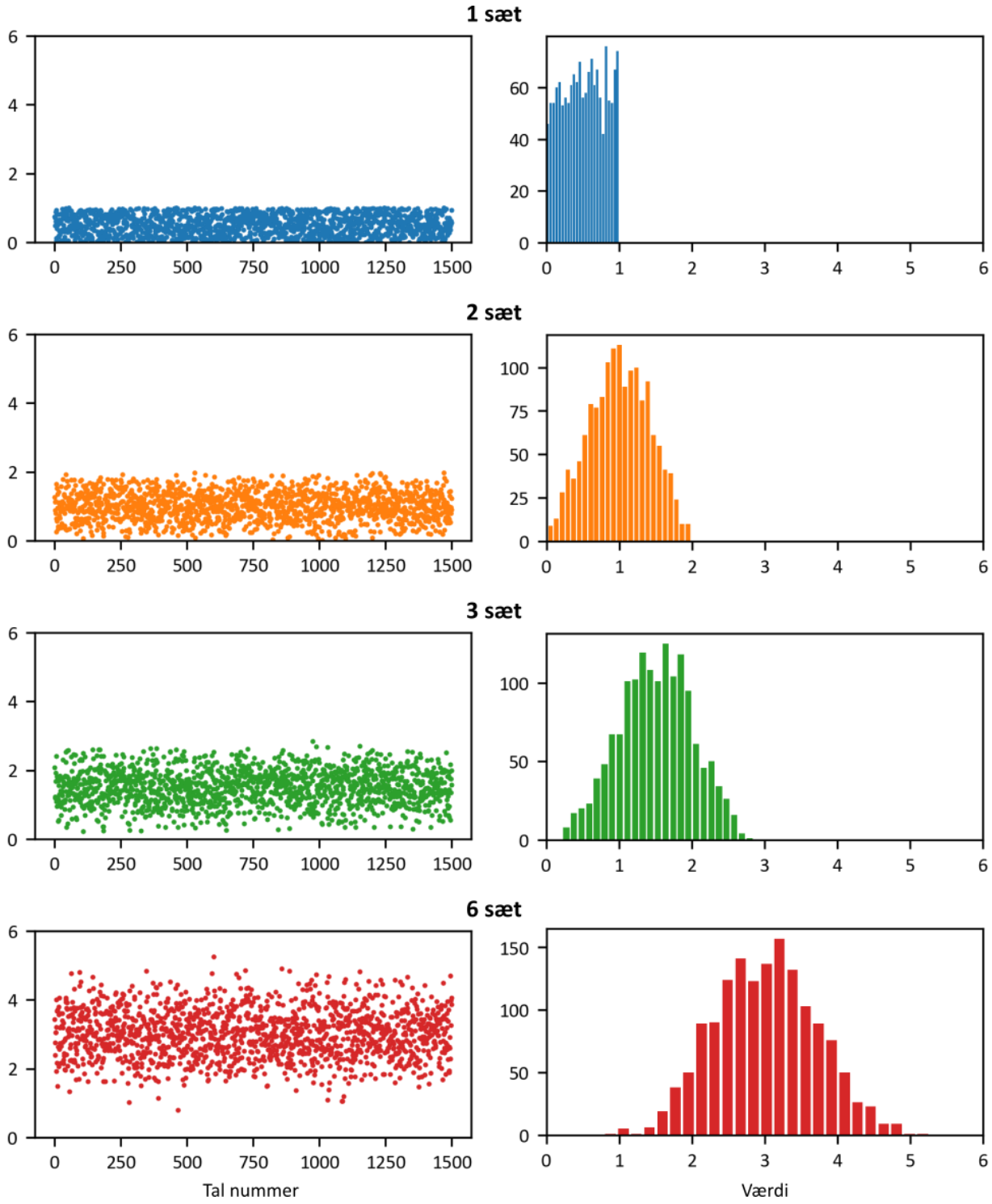
Man kunne spørge sig selv, hvad der sker, hvis vi bliver ved med at summere flere og flere tilfældigt genererede tal. Hvilken fordeling ender vi med at have?

Svaret kan vi til dels finde ved at kigge på Figur 7. Øverst kan vi se værdien af 1500 tal genereret med RAND() (til venstre) samt fordelingen af disse tal (til højre), som vi også så det i Figur 1 og 2. Disse 1500 tal udgør et talsæt. Genererer vi endnu et af disse talsæt og lægger til det første, har vi altså igen 1500 tal, som hver især er summen af to tal, og sådan kan vi blive ved. Kigger vi ned gennem Figur 7 ser vi, hvad der sker, når vi lægger hhv. 2, 3 og 6 sæt af tilfældige tal

sammen. For 2 sæt genfinder vi resultatet fra det foregående afsnit, nemlig trekantsfordelingen. For 3 sæt ser vi en begyndende tendens til, at sandsynligheden for at finde tal længere væk fra midten af fordelingen falder hurtigere end den lineært aftagende i trekantsfordelingen. Dette flugter med ræsonnementet fra de to talsæt: Hvis vi skal have et lavt tal, skal alle tre tal på samme tid være lave, hvilket er endnu mindre sandsynligt end for to. Denne tendens ses forstærket, når vi lægger endnu flere tal sammen som her ved 6 talsæt. Der er nu et tydeligt og hurtigt fald ud mod siderne, men samtidig også en større spredning af punkterne. Det sidste skyldes, at spredningen af den resulterende fordeling bliver større, efterhånden som vi lægger flere og flere af fordelinger sammen. Vi begynder dog uanset hvad at kunne ane en normalfordeling af tallene omkring en central værdi.

Dette resultat er opsummeret i *den centrale grænseværdisætning.* Den siger, at summen af tilstrækkeligt mange uafhængige, tilfældige tal fra samme fordeling (her den uniforme) vil normalfordele sig omkring en given middelværdi. Overraskende nok gælder den centrale grænseværdisætning uanset hvilken fordeling, tallene oprindeligt kommer fra.

Den centrale grænseværdisætning er et væsentligt resultat i statistisk teori, fordi den gør, at man kan udtale sig kvantitativt om sit datasæt, på trods af man ikke kender den underliggende fordeling, det stammer fra. Som vi skal se senere, er dette ganske anvendeligt til f.eks. at estimere den sande værdi af resultatet af et eksperiment.



Figur 7: Udviklingen i fordelingen af tilfældigt genererede tal fra den uniforme fordeling, når flere og flere talsæt lægges sammen. Hvert talsæt består af 1500 tal, og antallet af summerede talsæt er hhv. 1, 2, 3 og 6 fra øverst til nederst. Venstre side viser, hvordan de enkelte værdier for de 1500 talsæt fordeler sig. I højre side er talsættenes værdier vist i histogrammer med antallet i hver søjle indikeret på y-aksen.

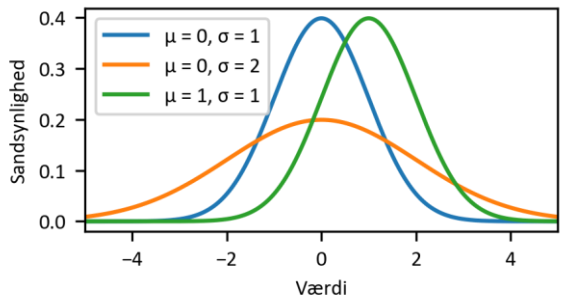
3 Normalfordelingen

Med den centrale grænseværdisætning i bagagen er det næste naturlige skridt at kigge på den fordeling, som summen af tilfældige tal nærmer sig: normalfordelingen (undertiden også kaldet gaussfordelingen). Den er matematisk beskrevet ved

Ligning 1

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

hvor μ er middelværdien og σ er standardafvigelsen. I sin rolle som sandsynlighedsfordeling beskriver udtrykket $f(x)$ således sandsynligheden for at få værdien x. Normalfordelingen findes mange steder grundet den centrale grænseværdisætning. Den er til dels ansvarlig for, f.eks. højden af en befolkningsgruppe har en tendens til at normalfordele sig omkring en middelværdi, fordi mange egenskaber i sidste ende styres af en sum af mange tilfældige tal. Normalfordelingen er derfor helt central i statistisk analyse.



Figur 8: Eksempler på normalfordelinger med forskellige middelværdier og standardafvigelser.

Tre eksempler på normalfordelinger er vist i Figur 8, hvor vi ser indflydelsen af de to parametre μ og σ. Ændring i middelværdien (blå til grøn) flytter hele kurven, fordi normalfordelingen er symmetrisk omkring sin middelværdi. Ændringer i standardafvigelsen (blå til orange) styrer derimod bredden af fordelingen, hvilket intuitivt giver god mening. Jo større standardafvigelse, des større interval kan værdierne falde indenfor, og des bredere "vinger" har sandsynlighedsfordelingen.

3.1 Simulering af normalfordeling

I stedet for blot at generere tal fra en uniform sandsynlighedsfordeling, som vi gjorde det tidligere, er det også muligt at gøre det fra en normalfordeling. Dette gøres med funktionen NORM.INV(RAND(), μ, σ), hvor μ og σ er hhv. den ønskede middelværdi og standardafvigelse. Kaldet til RAND() giver det tilfældigt genererede tal, som NORM.INV derefter konverterer til det tilsvarende tal fra normalfordelingen.

Opgave:

Åben Excel-dokumentet *introduktionTilStatistik.xlsx* og prøv at ændre på de forskellige parametre (middelværdi, standardafvigelse og antal målinger). Opfører fordelingen af de tilfældige tal sig som forventet? Hvor mange målinger skal medtages, før det tydeligt kan ses, de er normalfordelte?

Opgave:

I samme dokument er fordelingen af de udregnede middelværdier fra stikprøverne også vist. Øg gradvist "Antal stikprøver" og se, hvordan fordelingen udvikler sig. Normalfordeler værdierne sig omkring den valgte middelværdi?

4 Statistiske begreber

Der er tre grundlæggende begreber, vi skal have defineret, inden vi kan gå videre til at anvende statistikken til at foretage test på data: middelværdien, standardafvigelsen, og spredningen på middelværdien. Det er muligt at udregne disse for et vilkårligt datasæt, men i visse tilfælde kan de også bruges som *estimatorer* for parametrene i den underliggende fordeling. En estimator er altså et udtryk, man bruger til at udregne (estimere) en parameter.

I alle tilfældene tager vi udgangspunkt i et datasæt bestående af n målinger med værdierne $x_1, x_2, \ldots, x_n$.

4.1 Middelværdi

Med middelværdien mener vi den aritmetiske middelværdi (gennemsnittet) af datasættet. Den er defineret som

Ligning 2

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

For normalfordelingen er gennemsnittet $\bar{x}$ af datasættet lig med middelværdien μ, og gennemsnittet her defineret er derfor den bedste estimator for μ.

I Excel kan middelværdien af et datasæt findes med kaldet AVERAGE(data) (på dansk GENNEMSNIT).

4.2 Standardafvigelse

Standardafvigelsen angives ofte med SD (standard deviation), og den er et udtryk for spredningen af datasættet (den kaldes derfor undertiden også blot for spredningen). Den er givet ved

Ligning 3

$$\mathrm{SD}(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2}.$$

Kogt ned er standardafvigelsen altså bare den gennemsnitlige, kvadrerede afvigelse af datapunkterne fra middelværdien, som vi efterfølgende tager kvadratroden af. En nært tilknyttet parameter er variansen, som er givet ved $\mathrm{Var}(x) = \mathrm{SD}(x)^2$, svarende til udtrykket under kvadratroden. For normalfordelingen svarer SD ovenfor defineret til fordelingens σ, og det ovenstående udtryk er derfor den bedste estimator for at bestemme en normalfordelings standardafvigelse.

I Excel beregnes standardafvigelsen af et datasæt med kommandoen STDEV.S(data). Der findes flere forskellige til mere specifik brug, men STDEV.S er i langt de fleste tilfælde den rigtige at bruge (på dansk STDAFVS).

4.3 Spredning på middelværdien

Hvis man gentager et forsøg mange gange, svarer det til at tage flere stikprøver fra den samme underliggende fordeling. Beregner man middelværdien for hvert af disse forsøg, vil de pga. den statistiske usikkerhed ikke være identiske, men vil i sig selv vise en spredning. Denne spredning på middelværdierne gør det svært at vide, hvor sikkert én enkelt stikprøves beregnede estimat for μ egentligt er. Desuden kender vi ikke denne spredning, da det jo vil fordre, at vi skal gentage forsøget i det uendelige.

Imidlertid giver den enkelte stikprøves egen spredning (SD) et fingerpeg om, hvor stor spredning der vil være på middelværdierne, hvis vi gentager forsøget en masse gange. Jo større spredning, der er i stikprøven, des større spredning forventer vi i den underliggende fordeling, og des mere vil en masse stikprøvers middelværdier derfor sprede sig. Desuden giver stikprøvestørrelsen også et fingerpeg om sikkerheden på estimatet. Jo mindre stikprøve, des større risiko er der for, at man har udtaget et ikke repræsentativt udsnit af den underliggende fordeling.

Ud fra stikprøvens SD og n kan man derfor estimere hvor meget spredning der er på alle de middelværdier, man vil opnå, hvis man gentager forsøget i det uendelige. Dette estimat kaldes standard error on mean (SEM), og den kan udregnes ved

Ligning 4

$$\mathrm{SEM}(x) = \frac{\mathrm{SD}(x)}{\sqrt{n}}.$$

Som vi kan se, bliver spredningen mindre, efterhånden som vores estimat baseres på flere og flere målinger, hvilket også intuitivt er den vej, det bør gå. Jo flere målinger, des mere sikre kan vi være på vores resultat, hvilket igen afspejles i usikkerheden.

Der er ingen direkte kommando til at udregne SEM i Excel; dette skal gøres manuelt baseret på antallet af målinger og den beregnede standardafvigelse med STDEV.S(data).

Opgave:

I *introduktionTilStatistik.xlsx* udregnes middelværdien, standardafvigelsen og spredningen på middelværdien (SEM) automatisk for stikprøven genereret med tilfældige tal. Indtast forskellige antal målinger og noter værdierne. Plot dem herefter i et separat ark. Hvordan udvikler de sig, når n bliver større? Passer det med forventningen? Tag ligeledes et antal stikprøver for f.eks. $n = 5$ og $n = 50$ og se, hvordan spredningen i værdierne ændrer sig med n.

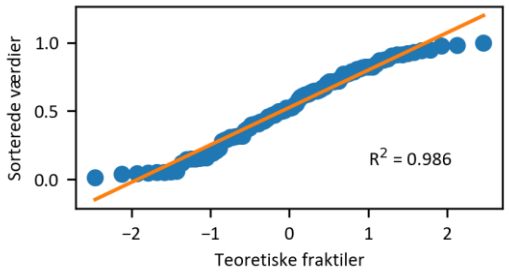
5 Statistiske test

I dette afsnit gives der en kort introduktion til forskellige statistiske test. Der er dog langt mere teori og nuancer i de forskellige test, end vi har plads til her. Den følgende tekst skal derfor kun ses som et hurtigt overblik over testenes metode og basale teori; for mere information og baggrundsviden henvises til egentlige lærebøger om statistik.

5.1 QQ-plot

Et fraktil-fraktil-plot (quantile-quantile, qq) kan bruges til at teste, om ens data følger den fordeling, man forventer. Den bruges ofte til at teste, om ens data er normalfordelt, da dette er et krav til at kunne anvende f.eks. en t-test.

Matematisk udføres den ved, man sorterer sit data i stigende orden og udregner, hvilken fraktil hvert målepunkt tilhører. Ud fra fraktilen kan man udregne en tilhørende værdi, som svarer til den fordeling, man gerne vil teste for. Hvis man derefter plotter sine sorterede målepunkter som funktion af de udregnede værdier, kan man se, om de falder på en ret linje. Hvis de gør, er der en god sandsynlighed for, målingerne stammer fra den antagne fordeling.



Figur 9: QQ-plot af 100 tilfældige tal genereret fra en uniform fordeling og testet mod en normalfordeling.

Et eksempel herpå kan ses i Figur 9, hvor værdierne fra det 1. sæt i Figur 7 er brugt ($n = 1500$). De teoretiske fraktiler er beregnet på baggrund af en normalfordeling. Regressionslinjen $f(x) = ax + b$ giver resultaterne $a = 0.273$ og $b = 0.526$ med et r-kvadrat tæt på 1. Om end r-kvadratet siger, der er god overensstemmelse med en ret linje, kan vi ved inspektion af plottet se en systematisk afvigelse i den karakteristiske s-form. Dette leder os til konklusionen, at vores data ikke stammer fra en normalfordeling – hvilket det jo heller ikke gør!

Opgave:

Åben Excel-dokumentet *introduktionTilStatistik.xlsx*, hvor de genererede tal i QQ-plottene sammenlignes med en normalfordeling med $\mu = 0$ og $\sigma = 1$. Ændr antallet af målinger og se, hvordan QQ-plottet ændrer sig for de to fordelinger. Hvor mange punkter skal der til, før vi kan se forskel på de to? Prøv desuden at tilføje regressionslinjer til plottene (højreklik på et punkt → Add Trendline). Er der stor forskel på r-kvadratet for de to ved få målinger? Kun for normalfordelingen: Er der en sammenhæng mellem skæringspunkt og hældning af regressionslinjen og de to underliggende parametre?

Ud over den grafiske metode i QQ-plottet findes der også statistiske metoder til at teste den underliggende fordeling. Disse er bl.a. en chi i anden- eller en Shapiro-Wilkes-test. Som vi lige har set, kan det nogle gange være svært ud fra en grafisk tilgang at bedømme, om ens data kommer fra f.eks. en normalfordeling. Der kan man med fordel anvende en statistisk test som supplement til undersøgelsen.

5.2 *T*-fordelingen

5.2.1 Matematisk baggrund

Hvis man fra en given normalfordeling udtager mange stikprøver med et stort antal i hver, vil stikprøvernes middelværdier i sig selv følge en normalfordeling. Hvis stikprøverne derimod er små, vil fordelingen af disses middelværdier brede sig mere ud til siderne, idet der er større risiko for, at små stikprøver udelukkende består af ekstreme tal. Jo mindre stikprøve, des bredere fordeling.

T-fordelingen kan bruges til at forudsige, hvor langt en stikprøves udregnede middelværdi vil kunne befinde sig fra den sande middelværdi baseret på stikprøvestørrelsen, n. Den er, modsat normalfordelingen, kun karakteriseret ved en enkelt parameter, ν, der angiver antallet af frihedsgrader, som igen afspejler stikprøvens størrelse. Har man f.eks. taget n datapunkter fra en normalfordeling, kan middelværdien estimeres ved at bruge en *t*-fordeling med $\nu = n - 1$ frihedsgrader. Matematisk er den givet ved

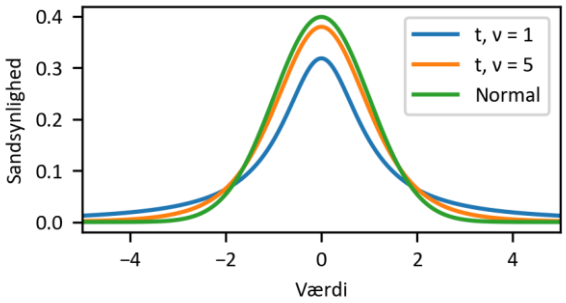
Ligning 5

$$f(x;\nu) = \frac{\Gamma\big((\nu+1)/2\big)}{\sqrt{\pi\nu}\,\Gamma(\nu)}(1 + x^2/\nu)^{-(\nu+1)/2},$$

hvor $\Gamma(x)$ er Gamma-funktionen, og ν er antallet af frihedsgrader. Fordelingen kan ses i Figur 10 for $\nu = 1$ og $\nu = 5$. Som antallet af frihedsgrader stiger, kan vi se, at vingerne på fordelingen bliver mindre, hvilket er en indikation på, at usikkerheden falder. For tilstrækkeligt høje ν nærmer *t*-fordelingen sig normalfordelingen, som også er vist i Figur 10 med $\mu = 0$ og $\sigma = 1$. De bredere vinger i *t*-fordelingen udtrykker

altså den større sandsynlighed for at få ekstreme værdier, der forekommer, når en stikprøves størrelse er lille.

Selvom normal- og t-fordelingen med $\nu = 5$ ser ud til



Figur 10: Sammenligning af t-fordelingen med forskellige frihedsgrader ν. Det ses, at t-fordelingen for større ν nærmer sig normalfordelingen.

at være ganske tæt på hinanden, siger en tommelfingerregel, at man skal op på $\nu \approx 100$, før de to fordelinger er tilpas ens til, man kan bruge en normalfordeling. Ellers er der for meget vægt i vingerne af t-fordelingen, hvilket gør sandsynligheden for ekstreme yderpunkter for stor.

5.2.2 Statistisk *t*-test mod kendt værdi

En t-test bruges til at undersøge, om middelværdien af to målinger (f.eks. en referenceværdi og en stikprøve) afviger signifikant fra hinanden. Nulhypotesen er ofte, at der ikke er en afvigelse, og dette kan testes ved at udregne

Ligning 6

$$t = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}},$$

hvor μ er den referenceværdi, man tester imod. Ud over det udregnede t skal man bruge en kritisk t-værdi. Den kritiske værdi bestemmes ved hjælp af signifikansniveauet P (ofte 5 %, altså $P = 0.05$) og antallet af frihedsgrader. Den kritiske værdi er defineret som den værdi på x-aksen, hvor arealet under kurven af t-fordelingen i intervallet $[-t, t]$ udgør $1 - P$ af det samlede areal (så 95 % med $P = 0.05$).

Hvis absolutværdien af ens udregnede t er større end den kritiske værdi, er forskellen mellem middelværdierne for stor til at kunne forklares med tilfældige fejl, og nulhypotesen om ens middelværdier forkastes. Fordi t-fordelingen har bredere vinger for lavere antal frihedsgrader, vil vi også få en højere kritisk t-værdi. Den større usikkerhed fra det lavere antal målinger gør det altså mere sandsynligt, vi ikke kan se en signifikant forskel på de to værdier.

Et vigtigt krav til at kunne udføre en t-test er, at ens data kommer fra en normalfordeling. Som vi lige har set, kan dette undersøges med et QQ-plot. I tilfælde af ens data ikke er normalfordelt, kommer den centrale grænseværdisætning dog til undsætning. En middelværdi er en sum af uafhængige, tilfældige tal, og vi ved derfor, at middelværdierne fra flere forsøg vil normalfordele sig omkring den sande værdi. En t-test er derfor generelt robust over for afvigelser fra normalfordelingen for tilpas store datasæt.

Opgave:

Åben Excel-dokumentet *introduktionTilStatistik.xlsx* og start med at kigge på t-testen af det normalfordelte datasæt. Hvordan ændrer t-fordelingen sig, når antallet af målinger ændres? Opdater udregningen og se, hvor ofte nulhypotesen afvises for et givent signifikansniveau. Passer det med det forventede? Ændrer det sig, når antallet af målinger ændres?

Opgave:

Sammenlign nu t-testen foretaget på både den uniforme fordeling og normalfordelingen. Er der forskel på de udregnede t-værdier? Har det nogen konsekvens for konklusionerne af en t-test, hvad den underliggende fordeling er?

5.2.3 Variationer af *t*-testen

Ud over at teste mod en kendt værdi, kan man også bruge t-testen til at sammenligne middelværdien for to fordelinger. Dette er som oftest den anvendte variant, fordi en referencemåling også har en vis usikkerhed og derfor er repræsenteret af en fordeling i stedet for blot ét tal. Der findes forskellige varianter af udtrykket for t-værdien, afhængigt af om stikprøvernes størrelse og varians er ens eller ej (Wikipedia har en glimrende oversigt). Den simpleste form er, når både størrelse og varians er ens. Her udregnes t-værdien som

Ligning 7

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sigma_{1,2}/\sqrt{n_{1,2}}},$$

hvor antallet af frihedsgrader er givet ved $2n_{1,2} - 2$. Hvis både stikprøvernes størrelse og varians er forskellige, skal man ud over et modificeret udtryk for *t*-værdien også udregne et vægtet antal frihedsgrader, der altid vil være lig med eller lavere end $2n_{1,2} - 2$. Vi ser altså, at den ekstra viden, vi besidder, har direkte indvirkning på resultatet. Et lavere antal frihedsgrader giver en højere kritisk *t*-værdi, som dermed igen influerer den konklusion, vi kan drage.

Grundet disse forskellige variationer af *t*-testen, er det meget vigtigt at bruge den korrekte. Om man har samme stikprøvestørrelse ved man i sagens natur godt. Omvendt er det sværere at vurdere, om varianserne er signifikant forskellige.

Til at undersøge dette bruger man en *F*-test, der i sin natur minder meget om *t*-testen. På samme måde som med *t*-testen kan man udregne en *F*-værdi for sin nulhypotese (typisk at varianserne ikke er forskellige), og man kan ligeledes finde en kritisk *F*-værdi fra en bestemt fordeling. Hvis den udregnede værdi overstiger den kritiske, vil man ligesom med *t*-testen forkaste nulhypotesen. Tematisk minder en *F*-test altså meget om *t*-testen; det er blot en anden værdi, man udregner, og en anden fordeling, man får sin kritisk værdi fra. For at lære mere om, hvordan man udfører denne vigtige test, henvises til lærebøger i statistik.

Opgave:

Åben Excel-dokumentet *normalfordelinger.xlsx* og kig på de to *t*-test for ens og uens varians. Ændr parametrene (middelværdi, standardafvigelse og antallet af målinger) for de to fordelinger og undersøg, hvordan resultatet af de to *t*-test reagerer. Hvad er forskellen(e)? Bliver de mere eller mindre tydelige, når mange målinger medtages? Er *t*-testen bedre til at skelne mellem fordelingerne, end du kan gøre grafisk fra plottet?

Til slut kan vi notere, at *t*-testen kun virker mellem to datasæt. Man kunne forestille sig, man blot kunne foretage flere på hinanden følgende *t*-test imellem de forskellige datasæt, men det er ikke tilfældet. Fejlene fra de enkelte test ville blive værre for hver gang, og man kan derfor ende i nogle meget forkerte slutninger, hvis man bruger denne fremgangsmåde. Vil man teste flere forskellige datasæt, skal man derfor have fat i mere avancerede statistiske værktøjer som f.eks. variansanalyse (analysis of variance, ANOVA), der kan sammenligne mange datasæt på tværs af flere parametre.