Sandsynlighedsregning og statistisk

J. C. F. Gauss
(1777 – 1855)

Peter Haremoës
Niels Brock

9. april 2013

1 Indledning

Dette hæfte er lavet som supplement til 2. udgave af bogen Mat B. Der er lagt vægt på at give en bedre forståelse for de metoder, der benyttes i deskriptiv statistik på Mat C niveau. Endvidere er der lagt vægt på at teorien for kontinuerte fordelinger kan ses som en anvendelse af B- og A-niveauets differential- og integralregning.

2 Integraler over ubegrænsede intervaller

I det integralregning vi stiftede bekendtskab med i Mat A-bogen, blev alle bestemte integraler taget over begrænsedede intervaller. Man kan imidlertid ofte også tage integraler over ubegrænsede intervaller.

Eksempel 1 *Lad $t > 1$ være et reelt tal. Da et*

$$\begin{aligned}\int_1^t \frac{1}{x^2}\,\mathrm{d}x &= \left[-\frac{1}{x}\right]_1^t \\ &= \left(-\frac{1}{t}\right) - \left(-\frac{1}{1}\right) \\ &= 1 - \frac{1}{t}.\end{aligned}$$

Vi ser at $1 - 1/t$ er en voksende funktion og at $1 - 1/t \to 1$ for $t \to \infty$. Vi skriver derfor

$$\int_1^\infty \frac{1}{x^2}\,\mathrm{d}x = 1.$$

Definition 2 *Lad f være en kontinuert funktion. Hvis $\int_a^b f(x)\,\mathrm{d}x$ har en grænseværdi for b gående mod uendelig, så betegnes denne grænseværdi*

$$\int_a^\infty f(x)\,\mathrm{d}x.$$

Tilsvarende defineres $\int_{-\infty}^b f(x)\,\mathrm{d}x$ som den eventuelle grænseværdi af $\int_a^b f(x)\,\mathrm{d}x$ for a gående mod $-\infty$. Hvis $\int_{-\infty}^b f(x)\,\mathrm{d}x$ er defineret og har en grænsevær for b gående mod uendelig, så betegnes denne grænseværdi med $\int_{-\infty}^\infty f(x)\,\mathrm{d}x$.

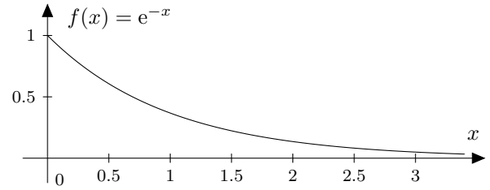
3 Kontinuerte fordelinger

Definition 3 *Lad X være en stokastisk variabel. Da er* fordelingsfunktionen *F for X defineret ved*

$$F(x) = P(X \leq x).$$

Fordelingsfunktionen svarer til de sumkurver vi har tegnet i deskriptiv statistik.

2: Heltal
Returnerer kommandoen randInt(som fungerer som på TI-89.
5: Sandsynlighed >
4: Tilfældig >
4: Normal
Returnerer kommandoen randNorm(som fungerer som på TI-89.



Figur 1: Tæthedsfunktion for eksponentialfordelingen.

Eksempel 4 *En stokastisk variabel X siges at være eksponentialfordelt med middelværdi λ dersom dens fordelingsfunktion er givet ved*

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ 1 - \mathrm{e}^{-x/\lambda} & \text{for } x > 0. \end{cases}$$

En sådan eksponentialfordeling giver f.eks. en god beskrivelse for ventetiden for et radioaktivt henfald af et atom.

Vi lægger mærke til at fordelingsfunktionen er en voksende funktion og at

$$\lim_{x \to -\infty} F(x) = 0,$$
$$\lim_{x \to \infty} F(x) = 1.$$

Hvis vi kender fordelingsfunktionen for en stokastisk variabel, kan vi beregne sandsynligheden for at den stokastiske variabel ligger i et vilkårligt interval, idet der gælder at

$$P(a < X \leq b) = F(b) - F(a).$$

Definition 5 *Hvis fordelingsfunktionen F for en stokastisk variabel X er en kontinuert funktion, så siges X at være en* kontinuert variabel. *Hvis F er differentiabel, så kaldes funktionen $f(x) = F'(x)$ for den stokastiske variabels* tæthedsfunktion.

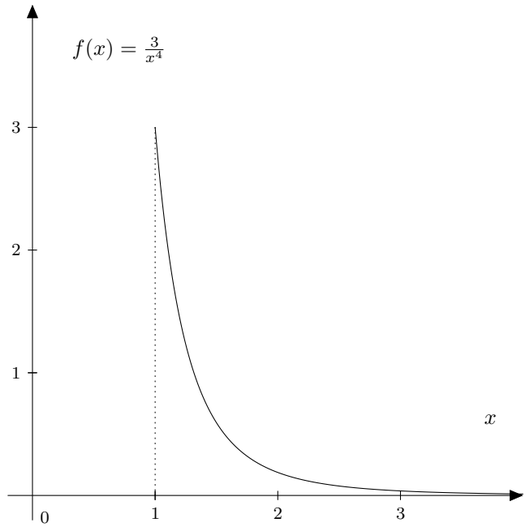
Tæthedsfunktionen svarer til de pinde- og søjlediagrammer vi har tegnet i deskriptiv statistik.

Eksempel 6 *Tæthedsfunktionen for en eksponentialfordeling er givet ved*

$$\begin{aligned} f(x) &= F'(x) \\ &= \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{1}{\lambda} \cdot \mathrm{e}^{-x/\lambda} & \text{for } x > 0. \end{cases} \end{aligned}$$

Hvis f er tæthed for en stokastisk variabel med for deling F, så er F stamfunktion til f og der gælder at

$$F(t) = \int_{-\infty}^{t} f(x) \, \mathrm{d}x.$$



Figur 2: Tæthed for en Pareto-fordeling.

Endvidere gælder der, at

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a) \\ &= \int_a^b f(x)\ \mathrm{d}x. \end{aligned}$$

Sandsynligheden for at $a < X \leq b$ svarer derfor til arealet under grafen for f mellem a og b. For at en funktion f kan være en tæthedsfunktion skal der gælde, at $f(x) \geq 0$ og at $\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = 1$.

De fleste kontinuerte fordelinger er defineret ud fra deres tæthedsfunktion.

Eksempel 7 *Ved en ligefordeling i intervallet $[a;b]$ forstå en fordeling med tæthed*

$$f(x) = \begin{cases} 0 & \text{for } x \notin [a;b], \\ \frac{1}{b-a} & \text{for } x \in [a;b]. \end{cases}$$

Vi checker, at der rent faktisk er tale om en sandsynlighedsfordeling ved at udregne

$$\int_a^b \frac{1}{b-a}\ \mathrm{d}x = \left[\frac{x}{b-a}\right]_a^b = 1.$$

Når vi tegner søjlediagrammet for grupperede data, antager vi faktisk, at data er ligefordelt i hvert delinterval. Ligesom for diskrete variable kan man beregne middelværdi og varians for kontinuerte fordelinger. Dette sker ved at erstatte summer med integraler.

Et vindue kommer frem, hvor man indtaster x, μ (middelværdi) og σ (standardafvigelse). Et nyt vindue kommer frem med angivelse af værdien af tæthedsfunktionen.

4:Normal Cdf...

Et vindue kommer frem, hvor man indtaster Upper value og Lower value (intervalendepunkterne), μ (middelværdi) og σ (standardafvigelse). I stedet for $-\infty$ og ∞ kan man bruge -10^{99} og 10^{99}. Et nyt vindue kommer frem med angivelse af sandsynligheden for at en normalfordelt variabel med de angivne parametre ligger i intervallet.

Tilfældige tal kan genereres ved at taste [MATH] **7:Probability**

4:rand

Returnerer et tilfældigt helt tal hvis en størsteværdi angives eller et ligefordelt decimaltal fra hvis intet argument intastes.

Syntax:

rand()

rand(største tal)

6:randNorm(

Returnerer et antal tilfældige normalfordelte tal.

Syntax:

randNorm(antal tilfældige tal, middelværdi, standardafvigelse)

6.5 TI-nspire

I beregningsdelen trykkes på menuknappen. Her kan blandt andet vælges:

5: Sandsynlighed

5: Fordelinger

1: Normal Pdf

Et vindue kommer frem, hvor man indtaster x-værdi, μ (middelværdi) og σ (standardafvigelse). Et nyt vindue kommer frem med angivelse af værdien af tæthedsfunktionen.

5: Sandsynlighed >

5: Fordelinger >

2: Normal Cdf

Et vindue kommer frem, hvor man indtaster Nedre grænse og Øvre grænse (intervalendepunkter), samt μ (middelværdi) og σ (standardafvigelse). Uendelig kan indtastes ved at hente tegnet fra listen af specialtegn. Et nyt vindue kommer frem med angivelse af sandsynligheden for at en normalfordelt variabel med de angivne parametre ligger i intervallet.

5: Sandsynlighed >

5: Fordelinger >

3: Invers normal

Et vindue kommer frem, hvor man indtaster Areal (sandsynlighed), μ (middelværdi) og σ (standardafvigelse). Et nyt vindue kommer frem med angivelse af den tilsvarende fraktil.

5: Sandsynlighed >

4: Tilfældig >

1: Tal

Returnerer kommandoen rand(som fungerer som på TI-89.

5: Sandsynlighed >

4: Tilfældig >

normalpdf(x)
normalpdf (x, middelværdi, standardafvigelse)

2: normalcdf(

Returnerer værdien af fordelingsfunktionen i et givet punkt. Man kan vælge både at angive en nedre og en øvre grænse. I stedet for -∞ og ∞ kan man bruge -10^{99} og 10^{99}

Syntax:
normalcdf(x)
normalcdf(x, middelværdi, standardafvigelse)
normalnormalcdf (nedre grænse, øvre grænse, middelværdi, standardafvigelse)

3: invNorm(

Returnerer fraktilen svarende til et et tal mellem 0 og 1.

Syntax:
invNorm(sandsynlighed)
invNorm(sandsynlighed, middelværdi, standardafvigelse)

Der er følgende kommandoer til at generere tilfældige tal. Tast MATH > PRB

1: rand

Returnere et ligefordelt tal mellem i $[0; 1]$

Syntax:
rand

randNorm(

Returnerer et tilfældige normalfordelte tal.

Syntax:
randNorm(middelværdi, standardafvigelse, antal tilfældige tal)

randInt

Returnerer et tilfældigt helt tal.

Syntaks:
randInt(mindste tal, største tal)

6.4 TI-89/Voyage 200

Man kan kalde kommandoer svarende til kommandoerne i TI-83+/TI-84+ ved hente dem fra kataloget eller skrive henholdsvis:
tistat.normpdf(
tistat.normcdf(
tistat.invNorm(
Alternativt kan man starte applicationen list/stat og vælge F5 Distr

1:Shade

1:Shade Normal

Et vindue kommer frem, hvor man indtaster Upper value og Lower value (intervalendepunkterne), μ (middelværdi) og σ (standardafvigelse). En graf bliver vist med en markering af det areal under kurven man har angivet.

2:Inverse \blacktriangleright

1:Inverse Normal...

Et vindue kommer frem, hvor man indtaster Area (sandsynlighed), μ (middelværdi) og σ (standardafvigelse). Et nyt vindue kommer frem med angivelse af den tilsvarende fraktil.

3:Normal Pdf...

Eksempel 8 *Tæthedsfunktionen*

$$f(x) = \begin{cases} 0 & \text{for } x < 1, \\ \frac{3}{x^4} & \text{for } x \geq 1, \end{cases}$$

definerer en såkaldt Pareto-fordeling. Vi checker at det rent faktisk er en tæthedsfunktion ved at integrere

$$\begin{aligned} \int_1^\infty \frac{3}{x^4}\ \mathrm{d}x &= \left[\frac{\text{-1}}{x^3}\right]_1^\infty \\ &= \lim_{x\to\infty} \frac{\text{-1}}{x^3} - (\text{-1}) \\ &= 1. \end{aligned}$$

Definition 9 *Lad X være en stokastisk variabel med tæthedsfunktion f. Da defineres middelværdien af X ved*

$$E[X] = \int_{\text{-}\infty}^{\infty} x \cdot f(x)\ \mathrm{d}x.$$

Hvis den stokastiske variabel X har middelværdi μ, så er variansen af X defineret ved

$$Var(X) = \int_{\text{-}\infty}^{\infty} (x-\mu)^2 f(x)\ \mathrm{d}x.$$

Standardafvigelsen er givet ved

$$\sigma(X) = (Var(X))^{1/2}.$$

Standardafvigelsen kaldes også standardafvigelsen.

Eksempel 10 (Kræver kendskab til partiel integration) *Eksponentialfounktionen med tæthed $\frac{\mathrm{e}^{\text{-}x/\mu}}{\mu}$ for $x \geq 0$ har middelværdi*

$$\begin{aligned} \int_{\text{-}\infty}^{\infty} x \cdot f(x)\ \mathrm{d}x &= \int_{\text{-}\infty}^{0} x \cdot 0\ \mathrm{d}x + \int_0^\infty x \cdot \frac{\mathrm{e}^{\text{-}x/\mu}}{\mu}\ \mathrm{d}x \\ &= 0 + \mu \int_0^\infty \frac{x}{\mu} \cdot \mathrm{e}^{\text{-}x/\mu} \cdot \frac{1}{\mu}\ \mathrm{d}x. \end{aligned}$$

Her laves substitution $t = x/\mu$, hvilket ved brug af partiel integration giver

$$\begin{aligned} \mu \cdot \int_0^\infty \frac{x}{\mu} \cdot \mathrm{e}^{\text{-}x/\mu} \cdot \frac{1}{\mu}\ \mathrm{d}x &= \mu \cdot \int_0^\infty t \cdot \mathrm{e}^{\text{-}t}\ \mathrm{d}t \\ &= \mu \cdot \left(\left[t \cdot \left(\mathrm{e}^{\text{-}t}\right)\right]_0^\infty - \int_0^\infty \left(\mathrm{e}^{\text{-}t}\right)\ \mathrm{d}t \right) \\ &= \mu \cdot \left(0 + \int_0^\infty \mathrm{e}^{\text{-}t}\ \mathrm{d}t \right) \\ &= \mu \cdot \left[\mathrm{e}^{\text{-}t}\right]_0^\infty \\ &= \mu. \end{aligned}$$

For at beregne variansen laves igen substitutionen $t = x/\mu$, hvilket giver

$$\int_0^\infty (x-\mu)^2 \frac{\mathrm{e}^{-x/\mu}}{\mu}\,\mathrm{d}x = \int_0^\infty (\mu t-\mu)^2 \mathrm{e}^{-t}\,\mathrm{d}t$$
$$= \mu^2 \int_0^\infty (t-1)^2 \mathrm{e}^{-t}\,\mathrm{d}t.$$

Det sidste integral beregnes ved at lave partiel integration 2 gange:

$$\begin{aligned}\int_0^\infty (t-1)^2 \mathrm{e}^{-t}\,\mathrm{d}t &= \left[(t-1)^2\left(-\mathrm{e}^{-t}\right)\right]_0^\infty - \int_0^\infty 2(t-1)\left(-\mathrm{e}^{-t}\right)\,\mathrm{d}t\\ &= 0 + 2\int_0^\infty (t-1)\mathrm{e}^{-t}\,\mathrm{d}t\\ &= 2\left(\left[(t-1)\left(-\mathrm{e}^{-t}\right)\right]_0^\infty - \int_0^\infty \left(-\mathrm{e}^{-t}\right)\,\mathrm{d}t\right)\\ &= 2\left(1+\int_0^\infty \mathrm{e}^{-t}\,\mathrm{d}t\right)\\ &= 2\left(1+\left[-\mathrm{e}^{-t}\right]_0^\infty\right)\\ &= 2(1+1)\\ &= 4.\end{aligned}$$

Derfor er variansen $4\mu^2$, og standardafvigelsen er 2μ.

Øvelse 11 *Beregn middelværdi, varians og standardafvigelsen af en ligefordeling.*

Øvelse 12 *Reregn middelværdi, varians og standardafvigelse for Pareto-fordelingen fra Eksempel 8.*

Øvelse 13 (Kræver kendskab til partiel integration) *En stokatisk variabel med sandsynlighedstæthed $x\mathrm{e}^{-x}$ for $x \geq 0$ siges at være Gammafordelt.*

Eksempel 14 a *Vis at dette er en sandsynlighedstæthed.*

b *Bestem middelværdien af denne Gammafordeling.*

c *Bestem varians og standardafvigelse af denne Gammafordeling.*

Det kan vises at $\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x = (2\pi)^{1/2}$. Derfor er

$$\phi(x) = \frac{\mathrm{e}^{-\frac{x^2}{2}}}{(2\pi)^{1/2}}$$

en tæthedsfunktion. Den tilsvarende fordeling kaldes en standard-normalfordeling. Det kan vises, at den har middelværdi 0 og varians 1. Fordelingsfunktionen for standard normalfordelingen betegnes Φ. Det ikke er muligt at opskrive et beregningsudtryk for Φ, så værdier af Φ kan kun beregnes tilnærmelsesvis ved hjælp af såkaldt numerisk integration. Hvis tæthedsfunktionen i stedet er

$$\frac{\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}}{(2\pi)^{1/2}\sigma},$$

så er der tale om en normalfordeling med middelværdi μ og standardafvigelse σ.

Konfidensintervaler Disse beregnes under forudsætninger af at data antages at være normalfordelt eller tilnærmelsesvis normalfordelt. Hvis stikprøven er tilstrækkelig stor kan middelværdien altid antages at være normalfordelt. For en dataliste vælges 4: Statis...>3: Konfidensintervaller... Herefter vælges 1: z-interval...hvis standardafvigelsen af normalfordelingen kendes. Hvis standardafvigelsen ikke kendes men skal estimeres vha. datasættets standardafvigelse, så vælges 2: t-interval... Hvis konfidensintervallet for succes-sandsynligheden i en binomialfordeling skal beregnes, vælger man 4: Statis...>3: Konfidensintervaller...> 5: 1-Prop z-interval...

Regression Hvis man skal undersæge om to størrelser i et datasæt kan beskrives men en lineær eller eksponentiel funktion, så skal de først indlæses som 2 kolonner i et regneark. Her efter vælges 4: Statis...>1: Stat beregning... hvoerefter man vælger 3: Lineær regression (mx+b)... eller A: Eksponentiel regression... Bemærk at man også kan lave regression ved først at lave datalister, så lave et xy-plot med applikationen 5: Tilføj Data og Statistik hvorefter man vælger 4: Analys...>6: Regression , hvorefter man vælger 1: Vis lineær (mx+b) eller 8: Vis eksponentiel .

6.2 Fordelinger

I løbet af kurset har vi beskæftiget os med 3 forskellige fordelingstyper: Normalfordelinger, binomialfordelinger og χ^2-fordelinger. Beregninger vedr. disse fordelinger kan laves ved at vælge 4: Statis...>2: Stat-fordelinger... Dem vi kan få brug for er:

1: Normal Pdf... giver sandsynlighedstætheden i et punkt for en normalfordeling.

2: Normal Cdf... giver sandsynligheden for et interval for en normalfordelt stokastisk variabel.

3: Invers normal... giver fraktilen svarende til en bestemt sandsynlighed, som vi kan opfatte som en procentdel. I TI-nspire skal sandsynligheden intastes i feltet "Areal".

7: χ^2 Pdf... giver sandsynlighedstætheden i et punkt for en χ^2-fordeling

8: χ^2Cdf... giver sandsynligheden for et interval for en χ^2-fordelt stokastisk variabel.

9: Invers χ^2... giver fraktilen svarende til en bestemt sandsynlighed for en χ^2-fordelt stokastisk variabel.

D: Binom Pdf... giver punktsandsynligheden for en binomialfordelt stokastisk variabel.

E: Binom Cdf... giver sandsynligheden for et interval for en binomialfordelt stokastisk variabel.

6.3 TI-83+/84+

Menuen for normalfordelinger kan findes under DISTR (2nd VARS). Bemærk at middelværdi og standardafvigelse har defaultværdier 0 og 1 svarende til en stadard-normalfordeling.

1: normalpdf(

Returnerer sandsynlighedstætheden i et givet punkt.

Syntax:

Bevis. Vi vil antage at normalfordelingen har middelværdi 0 og varians σ^2. Da gælder

$$\begin{aligned} E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= \frac{1}{n-1}\sum_{i=1}^{n}E\left[\left(X_i-\bar{X}\right)^2\right] \\ &= \frac{n}{n-1}E\left[\left(X_1-\bar{X}\right)^2\right] \\ &= \frac{n}{n-1}E\left[X_1^2+\bar{X}^2-2X_1\bar{X}\right] \\ &= \frac{n}{n-1}\left(E\left[X_1^2\right]+E\left[\bar{X}^2\right]-2E\left[X_1\bar{X}\right]\right). \end{aligned}$$

Vi benytter nu at $E\left[X_1^2\right]=\sigma^2$ og $E\left[\bar{X}^2\right]=\sigma^2/n$ samt at $\bar{X}=n^{-1}\sum_{i=1}^{n}X_i$ til at få

$$\begin{aligned} &\frac{n}{n-1}\left(E\left[X_1^2\right]+E\left[\bar{X}^2\right]-2E\left[X_1\bar{X}\right]\right) \\ &= \frac{n}{n-1}\left(\sigma^2+\frac{\sigma^2}{n}-2E\left[X_1\frac{1}{n}\sum_{i=1}^{n}X_i\right]\right) \\ &= \frac{n}{n-1}\left(\sigma^2+\frac{\sigma^2}{n}-\frac{2}{n}\sum_{i=1}^{n}E\left[X_1X_i\right]\right) \\ &= \frac{n}{n-1}\left(\sigma^2+\frac{\sigma^2}{n}-\frac{2}{n}\left(E\left[X_1^2\right]+\sum_{i=2}^{n}E\left[X_1\right]E\left[X_i\right]\right)\right) \\ &= \frac{n}{n-1}\left(\sigma^2+\frac{\sigma^2}{n}-\frac{2}{n}\left(\sigma^2+0\right)\right) \\ &= \frac{n}{n-1}\left(\sigma^2-\frac{\sigma^2}{n}\right)=\sigma^2. \end{aligned}$$

■

6 Statistik med TI-nspire

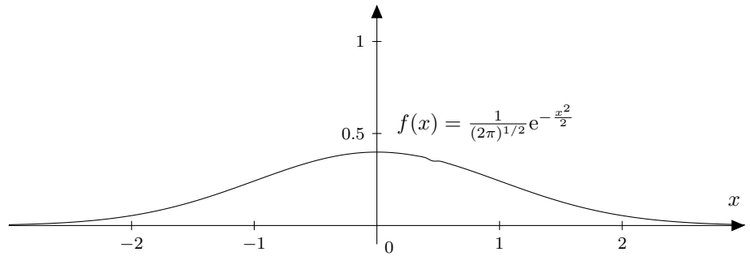
Af de mange statistikfunktioner, som findes i TI-nspire CAS, er det kun nogle få vi bruger. Her er en oversigt.

6.1 Undersøgelse af datasæt

Uafhængighedstest Bruges til at test om to størrelser eller hændelser er uafhængige ud fra en tabel med to inddelingskriterier. Man samler data i en matrix og vælger 4: Statis...> 4: Stat-tests...>8: χ^2 2-vejstest...

Goodness-of-fittest Bruges til at teste om en størrelse eller hændelse følger en bestemt fordeling. De observerede og de forventede hyppigheder skrives som kolonner i et regneark hvorefter man vælger 4: Statis... > 4: Stat-tests...> 7: χ^2 GOF...

Deskriptorer For at bestemme diverse deskriptorer for et datasæt skrives værdierne som en kolonne i et regneark. Man kan evt. tilføje en hyppighedsliste. Herfter vælges 4: Statis...>1: Stat beregning...> 1: Statistik med én variabel...



Figur 3: Tæthedsfunktion for standardnormalfordelingen

4 Middelværdi og varians

Uden bevis nævner vi, at hvis X_1 og X_2 er to stokastiske variable, så gælder der at

$$E\left[X_1+X_2\right]=E\left[X_1\right]+E\left[X_2\right].$$

Hvis endvidere X_1 og X_2 er uafhængige så gælder

$$E\left[X_1\cdot X_2\right]=E\left[X_1\right]\cdot E\left[X_2\right].$$

Sætning 15 *Lad X_1 og X_2 være uafhængige stokastiske variable. Da gælder at*

$$Var\left(X_1+X_2\right)=Var\left(X_1\right)+Var\left(X_2\right).$$

Bevis. Lad μ_1 og μ_2 betegne middelværdierne af X_2 og X_2. Da er middelværdien af X_1+X_2 lig $\mu_1+\mu_2$. Derfor gælder

$$\begin{aligned} &Var\left(X_1+X_2\right) \\ &= E\left[\left(\left(X_1+X_2\right)-\left(\mu_1+\mu_2\right)\right)^2\right] \\ &= E\left[\left(\left(X_1-\mu_1\right)+\left(X_2-\mu_2\right)\right)^2\right] \\ &= E\left[\left(X_1-\mu_1\right)^2+\left(X_2-\mu_2\right)^2+2\left(X_1-\mu_1\right)\left(X_2-\mu_2\right)\right] \\ &= E\left[\left(X_1-\mu_1\right)^2\right]+E\left[\left(X_2-\mu_2\right)^2\right]+E\left[2\left(X_1-\mu_1\right)\left(X_2-\mu_2\right)\right]. \end{aligned}$$

Da X_1 er uafhængig af X_2 er $X_1-\mu_1$ uafhængig af $X_2-\mu_2$ og der gælder at

$$\begin{aligned} E\left[2\left(X_1-\mu_1\right)\left(X_2-\mu_2\right)\right] &= 2E\left[X_1-\mu_1\right]\cdot E\left[X_2-\mu_2\right] \\ &= 2\left(E\left[X_1\right]-E\left[\mu_1\right]\right)\cdot\left(E\left[X_2\right]-E\left[\mu_2\right]\right) \\ &= 2\left(\mu_1-\mu_1\right)\cdot\left(\mu_2-\mu_2\right) \\ &= 0. \end{aligned}$$

Derfor er

$$\begin{aligned} Var\left(X_1+X_2\right) &= E\left[\left(X_1-\mu_1\right)^2\right]+E\left[\left(X_2-\mu_2\right)^2\right] \\ &= Var\left(X_1\right)+Var\left(X_2\right). \end{aligned}$$

■

5 Estimation

Antag af vi om nogle data (en stikprøve) ved at de er normalfordelte med standardafvigelse 2 men vi ikke kender normalfordelingens middelværdi. Opgaven er ud fra data at give et bud på værdien af normalfordelingens middelværdi.

Definition 16 *Et estimat er en funktion, der til en vilkårlig stikprøve knytter et reelt tal. Et estimat er med andre ord en stokastisk variabel defineret ud fra en stikprøve.*

Om et estimat er godt eller skidt er en anden sag. Hvis vi f.eks. skal estimere middelværdien af en normalfordeling, kan vi bruge stikprøvens median. Hvis stikprøven ellers er stor, vil medianen ligge tæt på middelværdien, så medianen er en udemærket estimator for middelværdien. I stedet for medianen kunne man tage den største værdi i stikprøven. Denne vil oplagt give et dårligt estimat af middelværdien, og jo større stikprøven er jo dårligere vil estimatet være.

Definition 17 *Et estimat siges at være* centralt *dersom middelværdien af estimatet er den sande værdi. Hvis et estimat ikke er centralt, siges det at være* skævt.

Medianen er et centralt estimat af middelværdien, mens maksimum er et skævt estmat, idet maksimum i middel giver en for høj værdi.

Sætning 18 *Stikprøvens gennemsnit giver et centralt estimat af normalfordelingens middelværdi.*

Bevis. Lad $(X_1, X_2, \ldots, X_n)$ betegne en stikprøve. Da er

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

og

$$\begin{aligned} E\left[\bar{X}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

■

Vi kan udregne variansen af gennemsnittet. Antag at den stokastiske variabel har middelværdi 0. Så gælder at

$$\begin{aligned} Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2} \cdot Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} \cdot \sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2} \cdot n\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Derfor er gennemsnittets standardafvigelse $\sigma/n^{1/2}$.

Det kan vises at stikprøvens gennemsnit er det centrale estimat, som har den mindste varians. Derfor vil gennemsnittet være vores foretrukne estimat for middelværdien.

Hvis man ved at en normalfordeling har middelværdi μ og skal estimere dens varians på grundlag af en stikprøve, så kan man bruge estimatet

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \mu\right)^2.$$

Dette estimat er centralt. Hvis man hverken kender en normalfordelings middelværdi eller varians kunne man tage stikprøvens varians

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2,$$

som estimat for den ukendte varians. Det viser sig imidlertid, at dette er et skævt estimat, som er systematisk for lille. Hvis stikprøvestørrelsen f.eks. er $n = 1$, så vil $\bar{X} = X_1$ og så bliver

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 &= \frac{1}{1}\left(X_1 - X_1\right)^2 \\ &= 0. \end{aligned}$$

Sætning 19 *Et centralt estimat af variansen af en normalfordeling med ukendt middelværdi er givet ved*

$$\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

for $n \geq 2$.