Multidimensionale normalfordelinger

Normalfordeling i en dimension

Den endimensionale *normalfordeling* (eller Gauss-fordeling) med middelværdi $\mu \in \mathbb{R}$ og varians $\sigma^2 > 0$ har frekvensfunktionen:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Vi vil også her tænke på tilfældet hvor $\sigma^2 = 0$, altså hvor al vægten en samlet i μ som værende en normalfordeling. Dette tilfælde kaldes nogle gange for *degenereret*.

Hvis den stokastiske variabel X er normalfordelt med middelværdi μ og varians σ^2 skriver vi:

$$X \sim N(\mu, \sigma^2)$$

Multidimensional normalfordeling

Definition: En stokastisk vektor $X \in \mathbb{R}^n$ er en *multidimensional normalfordeling* (m.n.d.) hvis enhver linearkombination af elementer i X er en endimensional normalfordeling.

Sagt på en anden måske skal der for alle $a \in \mathbb{R}^n$ gælde at $a \cdot X = a^T X$ er normalfordelt.

Dette er en betydeligt stærkere betingelse end at de marginale fordelinger hver for sig er normalfordelte.

Definition: Hvis X er en multidimensional normalfordeling og $\mu \in \mathbb{R}^n$ hvor $\mu_i = E(X_i)$ og $C \in \mathbb{R}^{n \times n}$ hvor $C_{ij} = \mathrm{Cov}(X_i, X_j)$ skriver vi $X \sim N(\mu, C)$. μ kaldes middelværdien og C kovariansmatricen for fordelingen.

Da C er en symmetrisk matrix er den også positivt semidefinit.

Definition: En m.n.d. $X \sim N(\mu, C)$ kaldes *degenereret* hvis $\det C = 0$.

Intuition: En degenereret m.n.d. lever i et affint underrum af $\mathbb{R}^n$ med dimension mindre end n.

Momentgenererende funktion for multidimensional normalfordeling

Definition: For en stokastisk vektor $X \in \mathbb{R}^n$ er den momentgenererende funktion givet ved:

$$M_X \colon \mathbb{R}^n \to \mathbb{R}^n \colon M_X(t) = E[\exp(t^T X)]$$

Her er $t^T X$ altså lig med $t_1 X_1 + t_2 X_2 + \cdots t_n X_n$. For en m.v.n. er denne størrelse normalfordelt, så vi har:

Sætning: For en stokastisk vektor $X \in \mathbb{R}^n$ $X \sim N(\mu, C)$ er den momentgenererende funktion:

$$M_X(t) = \exp\left(t^T \mu + \frac{1}{2} t^T C t\right)$$

Bevis: Den momentgenererende funktion er pr. definition:

$$M_X(t) = E[\exp(t^T X)] = E[\exp(t_1 X_1 + t_2 X_2 + \cdots t_n X_n)]$$

Den momentgenererende funktion for en endimensional normalfordeling Y er:

$$M_Y(t) = \exp\left(E(tY) + \frac{1}{2}\mathrm{Var}(tY)\right)$$

Da indholdet af eksponentialfunktionen er normalfordelt har vi altså:

$$M_X(t) = \exp\left(E(t_1X_1 + t_2X_2 + \cdots t_nX_n) + \frac{1}{2}\mathrm{Var}(t_1X_1 + t_2X_2 + \cdots t_nX_n\right)$$

Brug de almindelige regneregler for forventningsværdi og varians:

$$E(t_1X_1 + t_2X_2 + \cdots t_nX_n) = E(t_1X_1) + E(t_2X_2) + \cdots + E(t_nX_n) =$$

$$t_1E(X_1) + t_2E(X_2) + \cdots + t_nE(X_n) = t_1\mu_1 + t_2\mu_2 + \cdots + t_n\mu_n = t^T\mu$$

$$\mathrm{Var}(t_1X_1 + t_2X_2 + \cdots + t_nX_n) = \sum_{i,j=1}^{n} t_it_j\mathrm{Cov}(X_i, X_j) = t^TCt$$

Sætningen følger nu umiddelbart. Bevis slut.

Korrelation vs. afhængighed for multidimensional normalfordeling

Sætning: Hvis $X \sim N(\mu, C)$ og $\mathrm{Cor}(X_k, X_l) = 0$ er X_k og X_l uafhængige.

Bevis: Den momentgenererende funktion er:

$$M_X(t) = \exp\left(t^T\mu + \frac{1}{2}t^TCt\right)$$

t^TCt kan skrives:

$$t^TCt = \sum_{i,j=1}^{n} t_it_j\mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} t_i^2\mathrm{Var}(X_i) + \sum_{i\neq j}^{n} t_it_j\mathrm{Cov}(X_i, X_j)$$

Da $\mathrm{Cor}(X_k, X_l) = 0$ er der ikke nogen led der indeholder t_jt_k i eksponenten, og de kan derfor separeres i hver sin sum. Dermed er de to tilsvarende stokastiske variable X_k og X_l uafhængige. Bevis slut.

Denne vigtige egenskab for m.d.v.'er bruges bl.a. i beviset for følgende:

Sætning: De stokastiske variable $X_1, X_2, \cdots, X_n$ er uafhængige og $X_i \sim N(\mu_i, \sigma_i^2)$ hvis og kun hvis den stokastiske vektor $X = (X_1, X_2, \cdots, X_n)^T \sim N(\mu, C)$, hvor $C = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$.

Bevis: "⇒": Hvis $X_1, X_2, \cdots, X_n$ er uafhængige og $X_i \sim N(\mu_i, \sigma_i^2)$ og $a \in \mathbb{R}^n$ er $a \cdot X$ en linearkombination af normfordelinger, hvilket igen er normalfordelt. Så X følger en m.n.d. middelværdien er oplagt μ, og da de forskellige komponenter er uafhængige er kovariansmatricen diagonal med de enkelte varianser i diagonalen.

"⇐": Antag tilsvarende, at $X \sim N(\mu, C)$. Ved at vælge $a = e_i$ ser man, at X_i er normalfordelt. Da $\mathrm{Cov}(X_i, X_j)$ er nul når $i \neq j$ betyder dette at X_i og X_j er ukorrelerede og dermed også uafhængige ifølge foregående sætning. Bevis slut.

Bemærk: Hvis $X_1, X_2, \cdots, X_n$ hver især er normalfordelte er X generelt ikke en m.d.n!

Frekvensfunktion

Frekvensfunktion for multidimensional standardnormalfordeling

Sætning: En m.d.n. $X \sim N(0, I)$, hvor I er enhedsmatricen i dimension n har fordelingsfunktionen:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}x^T x\right]$$

Bevis: Da kovariansmatricen er diagonal er de enkelte komponenter af X indbyrdes uafhængige. Hver af disse har både middelværdi 0 og varians 1. De er med andre ord alle standardnormalfordelte. Da de er uafhængige er den samlede frekvensfunktion blot produktet af de enkelte fordelingsfunktioner:

$$f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)\right] = f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left[-\frac{1}{2}x^T x\right]$$

Bevis slut.

En sådan m.d.n. kaldes for en *multidimensional standardnormalfordeling*.

Affin transformationsegenskab

Sætning: Lad $X \sim N(\mu, C)$ være en n-dimensional m.d.n. Lad $A \in \mathbb{R}^{m \times n}$ og $b \in \mathbb{R}^m$. Da er:

$$AX + b \sim N(A\mu + b, ACA^T)$$

Bevis: Den karakteristiske funktion for den transformerede variabel $Y = AX + b$ er:

$$M_Y(t) = E\left[\exp\left(t^T(AX + b)\right)\right] = E[\exp(t^T AX)\exp(t^T b)] = E[\exp(t^T AX)]\exp(t^T b)$$

Sæt nu $s = A^T t$. Så er $s^T = t^T A$, og dermed:

$$M_Y(t) = E[\exp(s^T X)]\exp(t^T b)$$

Vi ser $E[\exp(s^T X)] = M_X(s)$ og dermed:

$$M_Y(t) = \exp\left(s^T\mu + \frac{1}{2}s^T Cs\right)\exp(t^T b) = \exp\left(t^T A\mu - \frac{1}{2}t^T ACA^T t\right)\exp(t^T b) =$$

$$\exp\left(t^T(A\mu + b) - \frac{1}{2}t^T ACA^T t\right)$$

Dette er netop den momentgenererende funktion for en m.d.n. med middelværdi $A\mu + b$ og kovariansmatrix ACA^T. Bevis slut.

Fordelingsfunktion for generel multidimensional normalfordeling

Lad $C \in \mathbb{R}^{n \times n}$ være en positivt semidefinit, symmetrisk matrix. Da C er symmetrisk kan den diagonaliseres af en ortogonal matrix V, således at $C = VDV^T$, hvor D er en diagonalmatrix med C's egenværdier i diagonalen: $D = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. Da C er positivt demidefinit er alle disse egenværdier ikke-negative. Dermed eksisterer kvadratroden af alle disse. Sæt nu $D^{½} = \mathrm{diag}\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n}\}$ og definer:

$$C^{½} \equiv VD^{½}V^T$$

Man kan nu regne:

$$\left(C^{½}\right)^2 = VD^{½}V^TVD^{½}V^T = VD^{½}D^{½}V^T = VDV^T = C$$

Undervejs er det benyttet at $V^T = V^{-1}$, idet V er ortogonal. Det giver altså mening at tænke på $C^{½}$ som kvadratroden af C. Da transformationen V er entydigt bestemt er $C^{½}$ det også.

Der gælder fælgende sætninger om sådanne kvadratrødder.

Sætning: Givet en positivt semidefinit, symmetrisk matrix C, er $\left(C^{½}\right)^T = C^{½}$

Bevis: $C^{½} = VD^{½}V^T$, så $\left(C^{½}\right)^T = \left(VD^{½}V^T\right)^T = VD^{½}V^T = C^{½}$. Bevis slut.

Sætning: Givet en positivt semidefinit, symmetrisk matrix C, er en kvadratrod $C^{½}$ invertibel hvis og kun hvis C er invertibel. I dette tilfælde sætter vi $C^{-½} \equiv \left(C^{½}\right)^{-1}$.

Bevis: Hvis C har egenværdierne $\lambda_1, \lambda_2, \cdots, \lambda_n$ har $C^{½}$ egenværdierne $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n}$. En matrix er invertibel hvis og kun hvis den ikke har nul som egenværdi. Sætningen følger heraf. Bevis slut.

Sætning: Givet en positivt definit (og dermed invertibel), symmetrisk matrix C, er $\det\left(C^{-½}\right) = \frac{1}{\sqrt{\det C}}$

Bevis: Determinanten af $C^{½}$ findes ved at bruge reglen $\det(AB) = \det A \cdot \det B$:

$$\det C = \det\left(\left(C^{½}\right)^2\right) = \det C^{½} \cdot \det C^{½}$$

Da C er positivt definit er determinanten positiv, og der må derfor gælde $\det C^{½} = \sqrt{\det C}$. Idet $C^{-½} = \left(C^{½}\right)^{-1}$ følger sætningen af den sædvanlige regneregel for determinanter af inverse matricer: $\det C^{-½} = \frac{1}{\det C^{½}} = \frac{1}{\sqrt{\det C}}$. Bevis slut.

Vi kan nu endelig samle trådene og finde frekvensfunktionen for en generel m.n.d.:

Sætning: En m.d.n. $X \sim N(\mu, C)$ har en frekvensfunktion hvis og kun hvis $\det C \neq 0$, altså hvis fordelingen ikke er degenereret. I dette tilfælde er frekvensfunktionen givet ved:

$$f(x) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left[-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right]$$

Bevis: Betragt en standardnormalfordeling i n dimensioner. Altså $X \sim (0, I)$. Da C er en kovariansmatrix er den positivt semidefinit og symmetrisk. C har derfor en kvadratrod $C^{½}$. Betragt nu koordinattransformationen $Y = C^{½}X + \mu$. Ifølge den affine transformationsegenskab er:

$$Y \sim N\left(C^{½}0 + \mu, C^{½}I\left(C^{½}\right)^T\right) \sim N(\mu, C^{½}C^{½}) \sim N(\mu, C)$$

Vi har altså genskabt en vilkårlig m.d.n. ud fra en standardnormalfordeling. Frekvensfunktionen kan nu findes vha. transformationssætningen:

$$f_Y(y) = f_X(x)\left|\frac{\partial x}{\partial y}\right|$$

For at bruge sætningen skal X udtrykkes som funktion af Y. Sammenhængen er

$$Y = C^{½}X + \mu \Leftrightarrow C^{½}X = Y - \mu$$

Denne ligning har en løsning hvis og kun hvis $C^{½}$ er invertibel, og iflg. ovenstående sætning er $C^{½}$ invertibel hvis og kun hvis C er invertibel. Antag dette er tilfældet. Da gælder der $X = C^{-½}(Y - \mu)$. Den tilhørede Jacobiand-matrix er:

$$\left|\frac{\partial x}{\partial y}\right| = \det C^{-½} = \frac{1}{\sqrt{\det C}}$$

Her er ovenstående sætning benyttet. Fordelingsfunktionen er nu:

$$f_Y(y) = f_x\left(C^{-½}(y-\mu)\right)\frac{1}{\sqrt{\det C}} = \frac{1}{\sqrt{(2\pi)^n}}\exp\left[-\frac{1}{2}\left(C^{-½}(y-\mu)\right)^T\left(C^{-½}(y-\mu)\right)\right]\frac{1}{\sqrt{\det C}} =$$

$$\frac{1}{\sqrt{\det(2\pi C)}}\exp\left[-\frac{1}{2}(y-\mu)^T C^{-1}(y-\mu)\right]$$

Erstat nu y/Y med x/X og sætningen er bevist. Bevis slut.

Geometrisk intuition

Målteori

σ-algebraer

Definition: En σ-algebra på en mængde X er en brolægning $\mathbb{E} \subseteq \mathcal{P}(X)$ der opfylder:

1. $\mathbb{E}$ er ikke-tom.
2. $\mathbb{E}$ er lukket under komplementdannelse: $A \in \mathbb{E} \Rightarrow A^c \in \mathbb{E}$
3. $\mathbb{E}$ er lukket under tællelig forening: $\forall n \in \mathbb{N}$: $A_n \in \mathbb{E} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathbb{E}$

Kommentar: Bogstavet σ refererer her til tælleligheden i trin 3. En (mængde-)algebra er en brolægning opfylder 1 og 2, men kun 3 for endelige foreninger.

Sætning: Lad $\mathbb{E}$ være en σ-algebra på X. Da gælder der:

1. $X \in \mathbb{E}$
2. $\emptyset \in \mathbb{E}$
3. $A, B \in \mathbb{E} \Rightarrow A \cap B \in \mathbb{E}$
4. $A, B \in \mathbb{E} \Rightarrow A \backslash B \in \mathbb{E}$
5. $\forall n \in \mathbb{N}$: $A_n \in \mathbb{E} \Rightarrow \bigcap_{n\in\mathbb{N}} A_n \in \mathbb{E}$

Bevis:

1. Da $\mathbb{E}$ ikke er tom findes der et $A \in \mathbb{E}$. De $\mathbb{E}$ er lukket under komplementdannelse er $A^c \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under tællelig (og dermed endelig) forening er $A \cup A^c = X \in \mathbb{E}$.
2. Vi har lige vist $X \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under komplementdannelse er $X^c = \emptyset \in \mathbb{E}$.
3. Antag $A, B \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under komplementdannelse er $A^c, B^c \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under tællelig (og dermed endelig) forening er $A^c \cup B^c = A \cap B \in \mathbb{E}$.
4. Antag $A, B \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under komplementdannelse er $B^c \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under tællelig (og dermed endelig) forening er $A \cup B^c = A \backslash B \in \mathbb{E}$.
5. Antag $\forall n \in \mathbb{N}$: $A_n \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under komplementdannelse er $\forall n \in \mathbb{N}$: $A_n^c \in \mathbb{E}$. Da $\mathbb{E}$ er lukket under tællelig forening er $\bigcup_{n\in\mathbb{N}} A_n^c = \bigcap_{n\in\mathbb{N}} A_n \in \mathbb{E}$.

Bevis slut.

Bemærkning: Regel 1-4 gælder også for algebraer, mens 5 kun gælder for σ-algebraer. I beviset for 3 og 5 er De Morgans love benyttet.

Eksempler

- For en vilkårlig mængde X er $\mathcal{P}(X)$ trivielt en σ-algebra. Den størst mulige.
- Tilsvarende er $\{\emptyset, X\}$ også en σ-algebra. Den mindst mulige.
- Hvis $A \subseteq X$ er $\{\emptyset, A, A^c, X\}$ en σ-algebra.
- Hvis $A_1, A_2, \cdots, A_n$ er en klassedeling af X er $\mathbb{E} = \{\bigcup_{i\in I} A_i | I \subseteq \{1,2,\cdots,n\}$ en σ-algebra. Dette gælder, da: 1. $X \in \mathbb{E}$, så $\mathbb{E}$ er ikke tom. 2. Hvis $A \in \mathbb{E}$ findes der et $I \subseteq \{1,2,\cdots,n\}$ så $A = \bigcup_{i\in I} A_i$. Men så er $A^c = \bigcup_{i\in\{1,2,\cdots,n\}\backslash I} A_i \in \mathbb{E}$. 3. Lad $\forall n \in \mathbb{N}$: $B_n \in \mathbb{E}$. Dvs. $B_n = \bigcup_{i\in I_n} A_i$. Dermed er $\bigcup_{n\in\mathbb{N}} B_n = \bigcup_{i\in\bigcup_{n\in\mathbb{N}} I_n} A_i$. Da $\bigcup_{n\in\mathbb{N}} I_n \subseteq \{1,2,\cdots,n\}$ ligger foreningen i $\mathbb{E}$.
- Mængden $\mathbb{E} = \{B \subseteq \mathbb{R} | B$ er tællelig eller B^c er tællelig$\}$. I denne forbindelse regnes den tomme mængde som værende endelig. 1. Derfor er

Sætning: Lad $(\mathbb{E}_i)_{i\in I}$ være en ikke-tom familie af σ-agebraer på X. Da gælder der:

$$\bigcap_{i\in I} \mathbb{E}_i$$

er en σ-algebra på X.

Bevis: Det skal bevises at de tre betingelser er opfyldt:

1. Da $X \in \mathbb{E}_i$ for alle i er X også medlem af fællesmængden, der dermed ikke er tom
2. Lad $A \in \bigcap_{i\in I} \mathbb{E}_i$. Dvs. $\forall i \in I: A \in \mathbb{E}_i$. Da alle $\mathbb{E}_i$'erne er σ-algebraer må $\forall i \in I: A^c \in \mathbb{E}_i$. Derfor må også $A^c \in \bigcap_{i\in I} \mathbb{E}$.
3. Lad $\forall n \in \mathbb{N}: A_n \in \bigcap_{i\in I} \mathbb{E}_i$. Dvs. $\forall n \in \mathbb{N}, i \in I: A_n \in \mathbb{E}_i$. Da alle $\mathbb{E}_i$'erne er σ-algebraer må der gælde: $\forall i \in I: \bigcup_{n\in\mathbb{N}} A_n \in \mathbb{E}_i$, og dermed $\bigcup_{n\in\mathbb{N}} A_n \in \bigcap_{i\in I} \mathbb{E}_i$.

Bevis slut.

Sætningen kan benyttes til at bevis følgende:

Sætning: Lad X være en mængde og $D \subseteq \mathcal{P}(X)$ en vilkårlig delmængde af potensmængden. Da eksisterer der en mindste σ-algebra genereret af D - kaldet $\sigma(D)$ - i den forstand at den opfylder:

1. $D \subseteq \sigma(D)$.
2. For enhver σ-algebra over X, $\mathbb{E}$ gælder der: $D \subseteq \mathbb{E} \Rightarrow \sigma(D) \subseteq \mathbb{E}$.

Bevis: Betragt mængden:

$$\Sigma(D) = \{\mathbb{E} \subseteq \mathcal{P}(X) | \mathbb{E} \text{ er en } \sigma\text{-algebra over } X \text{ og } D \subseteq \mathbb{E}\}$$

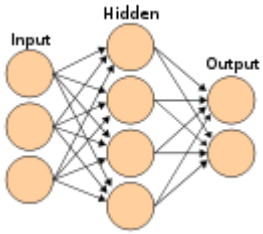
Det ses, at mængden ikke er tom, da $\mathcal{P}(X) \in \Sigma(D)$. Ifølge sætningen ovenfor er $\bigcap_{\Sigma\in\Sigma(D)} \Sigma$ en σ-algebra. Denne mængde opfylder pr. konstruktion punkt 1 og 2 ovenfor. Bevis slut.

Sætningen motiverer følgende definition:

Definition: Lad Lad X være en mængde og $D \subseteq \mathcal{P}(X)$. Da kaldes $\sigma(D)$ som defineret ovenfor *den af D frembragte σ-algebra*. D kaldes for et frembringersystem for en σ-algebra $\mathbb{E}$, såfremt $\sigma(D) = \mathbb{E}$. Hvis $\mathbb{E}$ har et tælleligt frembringersystem D kaldes for *tælleligt frembragt*.

Mål

Neurale netværk



Et *neuralt netværk* består af tre eller flere *lag*. Det første lag kaldes *input-laget*. Det sidste lag kaldes *output-laget*. Lagene imellem kaldes *skjulte*.

Figuren til højre illustrerer situationen hvor der er et enkelt skjult lag.

I hvert lag er der et antal *neuroner*, hver illustreret med en cirkel på figuren.

Neuronerne er forbundet med *axoner* – altså pilene på figuren. En pil betyder, at den ene neuron kan påvirker *aktivitetet* i den anden. Størrelsen af denne indflydelse kaldes *vægt*.

Derudover kan der være et konstantled i hvert lag undtagen output-laget. Dette kaldes for en *bias*-enhed.

Den samlede aktivitet i en neuron findes ved at indsætte resultatet af summen af alle axonernes påvirkning i en såkaldt *aktiveringsfunktion*. Vi vil her altid bruge sigmoid-funktionen.

Matematisk formulering

Betragt et neuralt netværk med L lag i alt. Input-laget, der altså er lag nr. 1, består af en observation $(x_1, x_2, \cdots, x_n)^T$. Observation j, foruden bias-enheden der svarer til $j = 0$ og dermed $x_0 = 1$, kan påvirke neuron nr. i i første skjulte lag med en vægt $\Theta_{ij}^{(1)}$. Dvs:

$$z_i^{(2)} = \sum_{j=0}^{m} \Theta_{ij}^{(i)} x_j$$

Eller på matrixform:

$$z^{(2)} = \Theta^{(1)} x$$

Her er $x \in \mathbb{R}^{n+1}$, $\Theta^{(1)} \in \mathbb{R}^{s_2 \times (n+1)}$ og $z^{(2)} \in \mathbb{R}^{s_2}$. Her angiver s_2 antallet af neuroner i andet lag (første skjulte lag). For at få aktiveringen for neuroner i dette lag anvendes sigmoid-funktionen:

$$a^{(2)} = g\left(z^{(2)}\right) = g\left(\Theta^{(1)} x\right)$$

Så $a^{(2)} \in \mathbb{R}^{s_2}$. Hertil føjes en bias-enhed, så $a^{(2)} \in \mathbb{R}^{s_2+1}$. Næste lag har så på helt analog vis aktiveringen:

$$a^{(3)} = g\left(z^{(3)}\right) = g\left(\Theta^{(2)} a^{(2)}\right)$$

Igen tilføjes en bias-enhed osv. Indtil output-laget nås:

$$h_\Theta(x) = a^{(L)} = g\left(\Theta^{(L-1)} a^{(L-1)}\right)$$

Likelihood-funktion

Vi kigger her kun på et enkelt datasæt (x, y), hvor x og y altså er vektorer. Tolkningen af værdien af det i'te element i $h_\Theta(x)$ er sandsynligheden for at y_i er lig 1. Altså et Bernoulli-eksperiment. Derfor er likelihood-funktionen:

$$L(\Theta) = \prod_{i=1}^{n} h_\Theta(x)^{y_i} \cdot \left(1 - h_\Theta(x)\right)^{(1-y_i)}$$

Den tilhørende log-likelihood er:

$$l(\Theta) = -\log\big(L(\Theta)\big) = -\sum_{i=1}^{n}[y_i \cdot \log\big(h_\Theta(x)\big) + (1-y_i) \cdot \log\big(1-h_\Theta(x)\big)]$$

Denne størrelse, som dybest set er det samme som omkostnings-funktionen $J(\Theta)$ - skal minimeres. Derfor ønsker vi at finde de afledede mht. de forskelle vægte i Θ. Dette gøres ved hjælp af den såkaldte *tilbagepropagerings*-algoritme (*back propagation*).

Afledte af omkostningsfunktionen – indledende forsøg

Vi er interesserede i at aflede $J(\Theta)$ efter vægtene i de forskellige lag:

$$\frac{\partial J}{\partial \Theta_{ij}^{(l)}} = -\sum_{i=1}^{m}\left[y_i \frac{1}{h_\Theta(x)}\frac{\partial h_\Theta(x)}{\partial \Theta_{ij}^{(l)}} + (1-y_i)\frac{1}{1-h_\Theta(x)}\left(-\frac{\partial h_\Theta(x)}{\partial \Theta_{ij}^{(l)}}\right)\right]$$

Tilbagepropagation (back-propagation)

Fejl i neurale netværk

Betragt igen et neuralt netværk med L lag. Vi kigger her igen kun på et enkelt datasæt (x, y). Forudsigelse for mulighed j i output-laget (lag L) er aktiveringen $a_j^{(L)}$, mens den sande værdi er y_j. "Fejlen" i dette lag er altså:

$$\delta_j^{(L)} = a_j^{(L)} - y_j^{(L)}$$

Alternativt kunne dette skrives:

$$\delta_j^{(L)} = \big(h_\Theta(x)\big)_j - y_j^{(L)}$$

Eller på vektorform:

$$\delta^{(L)} = a^{(L)} - y = h_\Theta(x) - y$$

Denne fejl *tilbagepropageres* nu til sidste skjulte lag vha. følgende formel:

$$\big(\delta^{(L-1)}\big)_j = \Big(\big(\Theta^{(L-1)}\big)^T \delta^{(L)}\Big)_j \, g'\left(z_j^{(L-1)}\right)$$

Her er z'erne inputværdierne for sidst skjulte lag. Da den afledte af sigmoid-funktionen er givet ved: $g'(z) = g(z)\big(1-g(z)\big)$ er dette det samme som:

$$\big(\delta^{(L-1)}\big)_j = \Big(\big(\Theta^{(L-1)}\big)^T \delta^{(L)}\Big)_j \, g\left(z_j^{(L-1)}\right)\left(1-g\left(z_j^{(L-1)}\right)\right) = \Big(\big(\Theta^{(L-1)}\big)^T \delta^{(L)}\Big)_j \, a_j^{(L-1)}\left(1-a_j^{(L-1)}\right)$$

Herefter kan man "fortsætte bagud" endnu et lag:

$$\big(\delta^{(L-2)}\big)_j = \Big(\big(\Theta^{(L-2)}\big)^T \delta^{(L-1)}\Big)_j \, a_j^{(L-2)}\left(1-a_j^{(L-2)}\right)$$

Osv. Indtil man har en fejlvektor for alle de skjulte lag. Inputlaget har ikke nogen fejlvektor.

Partielle afledte – et datasæt

De partielle afledede af omkostningsfunktionen – stadig kun for ét datasæt - er nu givet ved:

$$\frac{\partial J}{\partial \Theta_{ij}^{(l)}} = a_j^{(l)} \delta_i^{(l+1)}$$

Partielle afledte – alle datasæt

For at finde afledte for alle m datasæt skal denne beregning udføres for hver af de enkelte datasæt, summeres (og normaliseres ved at dividere med m som sædvanligt). I praksis gøres dette ved at bruge akkumulatorer $\Delta_{ij}^{(l)}$:

1. Sæt alle $\Delta_{ij}^{(l)} = 0$
2. Beregn alle a for et givent datasæt (x, y)
3. Brug tilbagepropagation til at beregne $\delta^{(l)}, \delta^{(l-1)}, \cdots, \delta^{(2)}$
4. Opdater akkumulatoren: Læg $a_j^{(l)} \delta_i^{(l+1)}$ til $\Delta_{ij}^{(l)}$
5. Gentag trin 2-4 for alle datasæt
6. Beregn de partielle afledede som $\frac{\partial J}{\partial \Theta_{ij}^{(l)}} = \frac{1}{m} \Delta_{ij}^{(l)}$

Regularisering

Når regularisering inkluderes kommer der et ekstra led på ikke-bias-elementerne:

$$D_{ij}^{(l)} = \Delta_{ij}^{(l)} + \lambda \Theta_{ij}^{(l)} (1 - \delta_{j0})$$

Her er $(1 - \delta_{j0})$-leddet blot en smart at undgå regularisering af bias-vægtene. De afledede beregnes nu:

$$\frac{\partial J}{\partial \Theta_{ij}^{(l)}} = \frac{1}{m} D_{ij}^{(l)}$$

Teoretisk udledning

Partielle afledte

Vi kigger igen på tilfældet hvor der kun er ét datasæt (x, y). Vi kan tænke på omkostningsfunktionen som en sammensat funktion:

$$J(\Theta) = J\left(a^{(L)}\left(z^{(L)}\left(\Theta^{(L-1)}\right)\right)\right)$$

Her er fokus på sammenhængen fra sidste skjulte lag til output-laget – vi tænker på alle andre vægte som værende konstanter. Lad os aflede efter vægtene i det sidste, skjulte lag:

$$\frac{\partial J}{\partial \Theta^{(L-1)}} = \frac{\partial J}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial \Theta^{(L-1)}}$$

(Bemærk at faktorerne alle er matricer). Hvis man er interesseret i at aflede efter vægtene i næstsidste, skjulte lag må man tænke på omkostningsfunktionen således:

$$J(\Theta) = J\left(a^{(L)}\left(z^{(L)}\left(a^{(L-1)}\left(z^{(L-1)}\left(\Theta^{(L-2)}\right)\right)\right)\right)\right)$$

Dermed kan de relevante afledte skrives:

$$\frac{\partial J}{\partial \Theta^{(L-2)}} = \frac{\partial J}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \frac{\partial z^{(L-1)}}{\partial \Theta^{(L-2)}}$$

Tidligere skjulte lag vil få tilføjet yderligere faktorer efter sammen mønster. Vi bemærker, at de to første faktorer er ens i alle udtrykkene:

$$\delta^{(L)} \equiv \frac{\partial J}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}}$$

Tilsvarende kan vi definere

$$\delta^{(L-1)} \equiv \delta^{(L)} \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}$$

Og mere generelt:

$$\delta^{(l-1)} \equiv \delta^{(l)} \frac{\partial z^{(l)}}{\partial a^{(l-1)}} \frac{\partial a^{(l-1)}}{\partial z^{(l-1)}}$$

Generelt kan de afledte nu skrives:

$$\frac{\partial J}{\partial \Theta^{(l)}} = \delta^{(l-1)} \frac{\partial z^{(L)}}{\partial \Theta^{(L-1)}}$$

Sidste skjulte lag

Lad os beregne de afledte mht. vægte i lag $L-1$:

$$\frac{\partial J}{\partial \Theta^{(L-1)}} = \frac{\partial J}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial \Theta^{(L-1)}} = \delta^{(L)} \frac{\partial z^{(L)}}{\partial \Theta^{(L-1)}}$$

Husk at omkostningsfunktionen på vektoriseret for er:

$$J(\Theta) = -\sum_{i=1}^{n} [y_i \cdot \log\left(a^{(L)}\right) + (1 - y_i) \cdot \log\left(1 - a^{(L)}\right)]$$

Derfor er:

$$\frac{\partial J}{\partial a_j^{(L)}} = -\sum_{i=1}^{n} \left[y_i \frac{1}{a_j^{(L)}} + (1 - y_i) \frac{1}{1 - a_j^{(L)}} (-1) \right] = \sum_{i=1}^{n} \left[\frac{1 - y_i}{1 - a_j^{(L)}} - \frac{y_i}{a_j^{(L)}} \right]$$

Aktiveringsfunktionen er sigmoid-funktionen, så:

$$\frac{\partial a_j^{(L)}}{\partial z_k^{(L)}} = a_j^{(L)} \left(1 - a_j^{(L)}\right) \delta_{jk}$$

Så:

$$\delta_j^{(L)} = \sum_{i=1}^{n} \left[\frac{1 - y_i}{1 - a_j^{(L)}} - \frac{y_i}{a_j^{(L)}} \right] a_j^{(L)} \left(1 - a_j^{(L)}\right) = \sum_{i=1}^{n} \frac{}{1 - a_j^{(L)}}$$

Pearsons χ^2-test

Goodness of fit

To mulige udfald

Betragt n ens og uafhængige Bernouilli-eksperimenter med sandsynlighed p. Altså et binomial-eksperiment. Det forventede antal succeser (mulighed 1) er da $E_1 = np$, mens det tilsvarende forventede antal fiaskoer (mulighed 2) er $E_2 = n(1-p)$. Variansen af begge er $\sigma_1^2 = \sigma_2^2 = np(1-p)$.

Hvis O_1 og O_2 er de observerede antal hhv. succeser og fiaskoer defineres teststørrelsen Q således:

$$Q \equiv \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Ved at indsætte fås:

$$Q = \frac{(O_1 - np)^2}{np} + \frac{\big(n - O_1 - n(1-p)\big)^2}{n(1-p)}$$

Her er brugt, at $O_1 + O_2 = n$. Reducer indholdet af anden tæller:

$$\frac{(O_1 - np)^2}{np} + \frac{(-O_1 + np)^2}{n(1-p)} = \frac{(O_1 - np)^2}{np} + \frac{(O_1 - np)^2}{n(1-p)} = \frac{\big((1-p)+p\big)(O_1 - np)^2}{np(1-p)} = \frac{(O_1 - np)^2}{np(1-p)}$$

Altså i alt:

$$Q = \left(\frac{O_1 - E_1}{\sigma_1}\right)^2$$

Når n er stor vil indholdet af parentesen iflg. den centrale grænseværdisætning (under de sædvanlige antagelser) tilnærmelsesvist være standardnormalfordelt. Derfor er Q tilnærmelsesvis χ^2-fordelt med 1 frihedsgrad.

m mulige udfald

Betragt n ens og uafhængige eksperimenter med m udfald (1, 2 og 3) med tilhørende sandsynligheder $p_1, p_2, \cdots, p_m$. Da den samlede sandsynlighed er 1 må der gælde $p_m = 1 - p_1 - p_2 - \cdots - p_m$. Den samlede fordeling af udfaldene udgør en multinomialfordeling. De forventede antal er $E_1 = np_1, E_2 = np_2, \cdots, E_m = np_m$. Varianserne for de tilhørende stokastiske variable er $\sigma_1^2 = np_1(1-p_1), \sigma_2^2 = np_2(1-p_2), \cdots, \sigma_m^2 = np_m(1-p_m)$.

For hvert udfald k vil vi desuden definere en stokastisk variabel svarende til det i'nde eksperiment. I_{ik} sættes til 1 hvis udfaldet af eksperiment i faktisk var k og 0 ellers. I_{ik} er dermed Bernoulli-fordelt med $p = p_k$. I_{ik} og I_{jl} er uafhængige når i og j er forskellige, men ellers ikke.

Lad nu O_1, O_2 og $O_3 = n - O_1 - O_2$ være de observerede hyppigheder af de tre udfald. Definer nu teststørrelsen:

$$Q \equiv \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_m - E_m)^2}{E_m}$$

Lad X_k være den stokastiske variabel der svarer til O_k. Ifølge den centrale grænseværdisætning er disse stokastiske variable tilnærmelsesvist standardnormalfordelte:

$$\frac{X_k - np_k}{\sqrt{np_k(1-p_k)}}$$

Alternativt kan man sige, at følgende stokastiske variable skal være normalfordelte med middelværdi 0 og varians $1 - p_k$:

$$Y_k = \frac{X_k - np_k}{\sqrt{np_k}}$$

Teststørrelsen (som stokastisk variabel) kan udtrykket vha. disse:

$$Q = \sum_{i=1}^{m} \frac{(O_1 - E_1)^2}{E_1} = \sum_{i=1}^{m} Y_k^2$$

Definer nu en stokastiske vektor $Y = (Y_1, Y_2, \cdots, Y_m)^T$. Lad desuden Z være en standardnormaltfordelt stokastisk vektor i m dimensioner. Sæt $p = \left(\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_m}\right)$. Definer nu flg. stokastiske vektor:

$$W = Z - V$$

Her er $V = (Z^T p)p$. Da $|p|^2 = 1$ kunne man også skrive $V = \frac{(Z^T p)}{|p|^2}p$. I denne formulering er det tydeligt, at V er projektionen af Z på p. W ligger dermed i det ortogonale komplement til underrummet udspændt af p. W har altså koordinater på formen $(0, W, W_3, \cdots, W_m)$ ifht. denne basis.

Suppler nu $e_1 = p$ op til en ortonormalbasis for $\mathbb{R}^m$: $e_1, e_2, \cdots, e_m$ og konstruer en matrix $A \in O(m)$ bestående af disse basisvektorer som søjler: $A = [e_1\ e_2 \cdots e_m]$. Sæt nu $Z' = AZ$. Z' er dermed standardnormaltfordelt i m dimensioner. $W' = AW$ er samme vektor, bortset fra at første koordinat er projiceret væk:

$$W' = (0, Z_2', Z_3', \cdots, Z_m')$$

Da A er ortogonal er $|W| = |W'|$, så:

$$|W| = (Z_2')^2 + (Z_3')^2 + (Z_m')^2$$

Dermed er $|W| \sim \chi^2_{n-1}$. Hvis vi kan vise $|Y| = |W|$ er vi færdige.

Lemma: Y og W som defineret ovenfor har kovariansmatrix, givet ved $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(W_i, W_j) = -\sqrt{p_i p_j}$ og $\mathrm{Var}(Y_i) = \mathrm{Var}(W_i) = 1 - p_i$.

Bevis: Til at starte med bemærkes, at begge matricer har identisk forventningsværdi, nemlig nulvektoren. Resultatet skal vises for begge matricer. Først Y. Her får vi brug for flg. forventningsværdi når $i \neq j$:

$$E(X_i, X_j) = E\left(\left(\sum_{k=1}^{n} I_{ki}\right)\left(\sum_{l=1}^{n} I_{lj}\right)\right) = E\left(\sum_{k=l} I_{ki} I_{lj}\right) + E\left(\sum_{k \neq l} I_{ki} I_{lj}\right)$$

Første led er altid nul, da udfaldet af hvert enkelt Bernoulli-eksperiment kun kan lande i én kategori. Tilbage er:

$$E(X_i, X_j) = E\left(\sum_{k\neq l} I_{ki}I_{lj}\right) = \sum_{k\neq l} E(I_{ki})E(I_{lj})$$

Her er uafhængigheden mellem de enkelte eksperimenter benyttet. De er i alt $n(n-1)$ led i summen:

$$E(X_i, X_j) = n(n-1)p_ip_j$$

Nu kan kovariansen beregnes (husk at alle forventningsværdier er nul):

$$\mathrm{Cov}(Y_i, Y_j) = E\left(\left(\frac{X_i - np_i}{\sqrt{np_i}}\right)\left(\frac{X_j - np_j}{\sqrt{np_j}}\right)\right) =$$

$$\frac{1}{n\sqrt{p_ip_j}}\left(E(X_iX_j) - np_jE(X_i) - np_iE(X_j) + n^2p_ip_j\right) =$$

$$\frac{1}{n\sqrt{p_ip_j}}\left(n(n-1)p_ip_j - np_jnp_i - np_inp_j + n^2p_ip_j\right) = \frac{1}{n\sqrt{p_ip_j}}\left(n(n-1)p_ip_j - n^2p_ip_j\right) =$$

$$\frac{1}{n\sqrt{p_ip_j}}\left(-np_ip_j\right) = -\sqrt{p_ip_j}$$

Varianserne er blot forventningsværdierne af kvadraterne:

$$\mathrm{Var}(Y_i) = E(Y_i^2) = E\left(\left(\frac{X_i - np_i}{\sqrt{np_i}}\right)^2\right) = \frac{1}{np_i}\left(E(X_i^2) - 2np_iE(X_i) + n^2p_i^2\right) =$$

Her skal vi igen bruge at $E(X_i) = np_i$, samt:

$$\mathrm{Var}(X_i) = np_i(1-p_i) = E(X_i^2) - E(X_i)^2 = E(X_i^2) - n^2p_i^2 \Leftrightarrow$$

$$E(X_i^2) = np_i(1-p_i) + n^2p_i^2$$

Så:

$$\mathrm{Var}(Y_i) = \frac{1}{np_i}\left(np_i(1-p_i) + n^2p_i^2 - 2n^2p_i^2 + n^2p_i^2\right) = 1 - p_i$$

Det samme skal vises for $W = Z - (Z^Tp)p$. Den i'te koordinat af denne vektor er altså:

$$W_i = Z_i - \sum_{k=1}^{m} Z_k\sqrt{p_k}\sqrt{p_i}$$

Kovarianserne bliver:

$$\mathrm{Cov}(W_i, W_j) = E\left(\left(Z_i - \sum_{k=1}^{m} Z_k\sqrt{p_k}\sqrt{p_i}\right)\left(Z_j - \sum_{l=1}^{m} Z_l\sqrt{p_l}\sqrt{p_j}\right)\right) =$$

$$E(Z_iZ_j) - \sum_{l=1}^{m}\sqrt{p_l}\sqrt{p_j}E(Z_iZ_l) - \sum_{k=1}^{m}\sqrt{p_k}\sqrt{p_i}E(Z_jZ_k) + \sum_{k=1}^{m}\sum_{l=1}^{m}\sqrt{p_k}\sqrt{p_i}\sqrt{p_l}\sqrt{p_j}E(Z_kZ_l) =$$

Da Z_i'erne er uafhængige og standardfordelte er $E(Z_i Z_j) = \delta_{ij}$. Derfor reducerer ovenstående til:

$$\mathrm{Cov}(W_i, W_j) = \delta_{ij} - \sqrt{p_i}\sqrt{p_j} - \sqrt{p_j}\sqrt{p_i} + \sum_{k=1}^{m} \sqrt{p_k}\sqrt{p_i}\sqrt{p_k}\sqrt{p_j} =$$

$$\delta_{ij} - 2\sqrt{p_i}\sqrt{p_j} + \sqrt{p_i}\sqrt{p_j}\sum_{k=1}^{m} p_k = \delta_{ij} - \sqrt{p_i}\sqrt{p_j}$$

For $i \neq j$ giver dette netop $-\sqrt{p_i p_j}$ og for $i = j$ giver dette $\mathrm{Var}(W_i) = 1 - p_i$. Bevis slut.

Y og W har altså samme forventningsværdi og kovariansmatrix. I grænsen hvor n er stor er Y tilnærmelsesvis normalfordelt, ligesom W. Altså må begge vektorer have samme længde i denne grænse.

Single Value Decomposition (SVD)

Spektralsætningen – Symmetriske matricer

Hvis en matrix $A \in \mathbb{R}^{n\times n}$ er symmetrisk, altså hvis $a_{ij} = a_{ji}$ kan den som bekendt diagonaliseres. Det betyder, at samtlige egenværdier for A er reelle, og at der findes en ortogonal matrix $V \in \mathbb{R}^{n\times n}$, så:

$$A = VDV^T$$

Her er $D = \text{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ en diagonalmatrix bestående af A's egenværdier og V består af søjler af tilsvarende egenvektorer. Da V er ortogonal følger, at $V^T = V^{-1}$.

Spørgsmålet er om man kan gøre noget tilsvarende, hvis A ikke er symmetrisk. Eller måske ikke engang en kvadratisk matrix. Det viser sig, at der er en generalisering, på engelsk kaldet *single value decomposition*.

Om AA^T og A^TA

Lad $A \in \mathbb{R}^{m\times n}$ uden yderligere restriktioner. Betragt nu matricen AA^T. $AA^T \in \mathbb{R}^{m\times m}$. Og der gælder:

$$(AA^T)_{ij} = \sum_{k=1}^{m} a_{ik}a_{jk} = \sum_{k=1}^{m} a_{jk}a_{ik} = (AA^T)_{ji}$$

Matricen er altså symmetrisk, og kan derfor diagonaliseres ifølge spektralsætningen. Lad nu $x \in \mathbb{R}^m$ vi ser:

$$x^TAA^Tx = (A^Tx)^T(A^Tx) = \|A^Tx\|^2 \geq 0$$

Her betegner dobbeltstregerne den sædvanlige norm for $\mathbb{R}^m$, og uligheden følger af de sædvanlige egenskaber for normer. Dermed ses det, at AA^T er positivt semidefinit.

Tilsvarende er $A^TA \in \mathbb{R}^{n\times n}$ og der gælder:

$$(A^TA)_{ij} = \sum_{k=1}^{n} a_{ki}a_{kj} = \sum_{k=1}^{n} a_{kj}a_{ki} = (A^TA)_{ji}$$

Så denne matrix er ligeledes diagonaliserbar. Også A^TA er positivt semidefinit. Lad nemlig $x \in \mathbb{R}^n$:

$$x^TA^TAx = (Ax)^T(Ax) = \|Ax\| \geq 0$$

Single value decomposition

For $A \in \mathbb{R}^{m\times n}$ er $A^TA \in \mathbb{R}^{n\times n}$ altså diagonaliserbar. Der eksisterer altså en ortonormal basis af egenvektorer $v_1, v_2, \cdots, v_n \in \mathbb{R}^n$ således at:

$$A^TAv_j = \lambda_j v_j$$

Gang nu med u_i^T på begge sider:

$$v_i^TA^TAv_j = \lambda_j v_i^Tv_j = \lambda_j\delta_{ij}$$

Da A^TA er positivt semidefinit er egenværdierne ikke-negative: $\lambda_j \geq 0$. For alle $\lambda_j > 0$ sættes:

$$u_j = \frac{Av_j}{\sqrt{\lambda_j}}$$

For to sådanne vektorer $u_i, u_j \in \mathbb{R}^m$ gælder der:

$$u_i^T u_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} v_i^T A^T A v_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} \lambda_j \delta_{ij} = \delta_{ij}$$

(Undervejs er den første identitet vi udledte brugt). Disse vektorer er altså ortonormale. Suppler nu op til en ortonormal basis for $\mathbb{R}^m$, i alt $u_1, u_2, \cdots, u_m$.

Definer nu $U \in \mathbb{R}^{m \times m}$ og $V \in \mathbb{R}^{n \times n}$ ved at bruge hhv. u'er og v'er som søjler:

$$U = [u_1 \quad u_2 \quad \cdots \quad u_m] \qquad \text{og} \qquad V = [v_1 \quad v_2 \quad \cdots \quad v_n]$$

Begge matricer er ortogonale, og der gælder derfor $U^{-1} = U^T$ og $V^{-1} = V^T$.

Beregn nu matricen $D \in \mathbb{R}^{m \times n}$:

$$D_{ij} \equiv (U^T A V)_{ij} = u_i^T A v_j = u_i^T \left(\sqrt{\lambda_j}\right) u_j = \sqrt{\lambda_j} \delta_{ij}$$

D er altså diagonal i den forstand at de eneste indgange der ikke er nul er af formen d_{ii}. Ligningen kan omskrives:

$$D = U^T A V \Leftrightarrow U D V^T = A$$

Dette er SVD.

Students t-fordeling

Beregning af frekvensfunktion

For at beregne frekvensfunktionen for t-fordelingen får vi brug for nogle resultater om frekvensfunktioner. Først et lemma:

Lemma: Lad $f(y)$ være kontinuert, $g(t)$ differentiabel og a et tal. Da gælder:

$$\frac{d}{dt}\int_a^{g(t)} f(y)\,dy = f\big(g(t)\big)g'(t)$$

Bevis: Da f er kontinuert har den en stamfunktion F. Så integralet er:

$$\int_a^{g(x)} f(x)\,dx = F\big(g(x)\big) - F(a)$$

Differentier nu efter t (brug kædereglen):

$$\frac{d}{dt}\int_a^{g(t)} f(x)\,dt = \frac{d}{dt}\big[F\big(g(x)\big) - F(a)\big] = f\big(g(x)\big)g'(x)$$

Bevis slut.

Lemmaet er ofte brugbart når man skal finde formler for frekvensfunktioner. F.eks. følgende:

Sætning: Lad X og Y være kontinuerte, stokastiske variable med fælles frekvensfunktion $f(x,y)$ og n et naturligt tal. Da har den stokastiske variabel $Z = \frac{Y}{\sqrt{X/n}}$ frekvensfunktionen (under "lette regularistionsantagelser"):

$$h(t) = \int_0^{\infty} f\left(x, t\sqrt{\frac{x}{n}}\right)\sqrt{\frac{x}{n}}\,dx$$

Bevis: Fordelingsfunktionen $H(t)$ beskriver sandsynligheden for at Z antager en værdi mindre end eller lig med t:

$$H(t) = P(Z \le t) = P\left(\frac{y}{\sqrt{x/n}} \le t\right)$$

Hvis x er givet betyder betingelsen kan betingelsen $\frac{y}{\sqrt{x/n}} \le t$ løses for y: $y \le t\sqrt{\frac{x}{n}}$. Derfor er fordelingsfunktionen:

$$H(t) = \int_0^{\infty}\int_{-\infty}^{t\sqrt{x/n}} f(x,y)\,dy\,dx$$

Under den "lette regularisationsantagelse" af, at differentation kan flyttes ind under integraltegnet giver ovenstående lemma:

$$h(t) = \frac{d}{dt}H(t) = \int_0^\infty \frac{d}{dt}\int_{-\infty}^{t\sqrt{x/n}} f(x,y)\,dy\,dx = \int_0^\infty f\left(x, t\sqrt{\frac{x}{n}}\right)\sqrt{\frac{x}{n}}\,dx$$

Bevis slut.

Korollar: Hvis X og Y er uafhængige med fordelingsfunktioner $f_X(x)$ og $f_Y(y)$ bliver fordelingsfunktionen for $Z = \frac{Y}{\sqrt{X/n}}$:

$$h(t) = \int_0^\infty f_X(x) f_Y\left(t\sqrt{\frac{x}{n}}\right)\sqrt{\frac{x}{n}}\,dx$$

Students t-fordeling

Definition: Lad X og Y være uafhængige så Y er standardnormalfordelt og X χ^2-fordelt med n frihedsgrader. Da kaldes fordelingen af den stokastiske variabel $T = \frac{Y}{\sqrt{X/n}}$ for t-fordelingen med n frihedsgrader.

Sætning: Frekvensfunktionen for t-fordelingen med n frihedsgrader er:

$$h(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\cdot\Gamma(n/2)}\cdot\frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}$$

Bevis: Vi kan anvende sætningen fra sidste sektion. Frekvensfunktionerne for X og Y er:

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(n/2)}e^{-\frac{x}{2}}x^{n/2-1}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

Dermed bliver fordelingsfunktionen:

$$h(t) = \int_0^\infty \frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{n/2-1}\frac{1}{\sqrt{2\pi}}e^{-\left(t\sqrt{x/n}\right)^2/2}\sqrt{\frac{x}{n}}\,dx =$$

$$\frac{1}{2^{n/2}\Gamma(n/2)}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{n}}\int_0^\infty e^{-x/2}x^{n/2-1}e^{-\left(t\sqrt{x/n}\right)^2/2}\sqrt{x}\,dx =$$

$$\frac{1}{\sqrt{n\pi}\cdot 2^{(2n+1)/2}\cdot\Gamma(n/2)}\int_0^\infty x^{(n-1)/2}\exp\left[-\frac{1}{2}x(1+t^2/n)\right]dx$$

Substituer nu $u = \frac{1}{2}x(1+t^2/n)$. Det medfører $x = \frac{2u}{1+t^2/n}$ og derfor $\frac{dx}{du} = \frac{2}{1+t^2/n}$. Så $dx = \frac{2}{1+t^2/n}du$. Integralet bliver dermed:

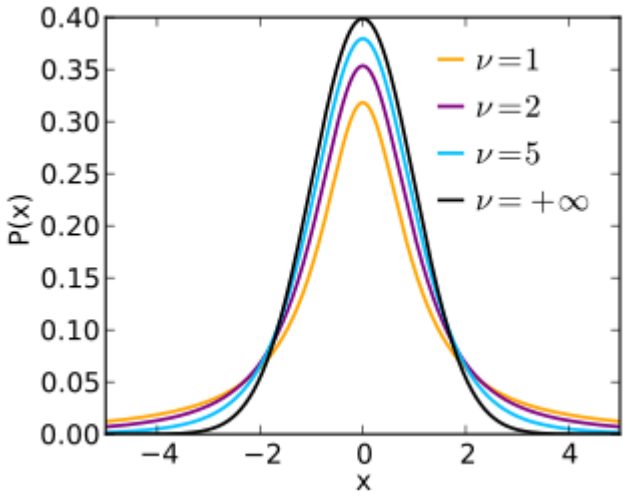
$$\int_0^\infty \left(\frac{2u}{1+t^2/n}\right)^{(n-1)/2} e^{-u}\frac{2}{1+t^2/n}du = \left(\frac{2}{1+t^2/n}\right)^{(n+1)/2}\int_0^\infty u^{(n-1)/2}e^{-u}\,du$$

Dette u-integral er netop $\Gamma((n+1)/2)$. Så i alt bliver frekvensfunktionen:

$$h(t) = \frac{1}{\sqrt{n\pi}\cdot 2^{(2n+1)/2}\cdot\Gamma(n/2)}\left(\frac{2}{1+t^2/n}\right)^{(n+1)/2}\Gamma((n+1)/2) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\cdot\Gamma(n/2)}\frac{1}{(1+t^2/n)^{(n+1)/2}}$$

Bevis slut.

Figuren til højre viser fordelingen for forskellige antal af frihedsgrader.



$n = 1$ – Cauchy-fordelingen

Når $n = 1$ får man:

$$h(t) = \frac{\Gamma(1)}{\sqrt{\pi}\cdot\Gamma(½)}\frac{1}{1+t^2} = \frac{1}{\pi}\frac{1}{1+t^2}$$

Her er $\Gamma(½) = \sqrt{\pi}$ benyttet. Dette er netop frekvensfunktionen for en Cauchy-fordeling.

n stor - Normalfordelingen

Som n vokser ligner t-fordelingen standardnormalfordelingen mere og mere.

Stikprøve med ukendt varians

t-fordelingen er først og fremmest vigtig pga. følgende sætning:

Sætning: Lad $X_1, X_2, \cdots, X_n$ være uafhængige med samme fordeling $X_i \sim N(\mu, \sigma^2)$, men hvor hverken middelværdi eller varians er kendte størrelser. Betragt følgende estimatorer:

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

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

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

Da er $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ og $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$ og de to er uafhængige. T er t-fordelt: $T \sim t_{n-1}$.

Bevis: Da $X_i \sim N(\mu, \sigma^2)$ er $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ ifølge de sædvanlige egenskaber ved normalfordelingen. Derfor er $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. For at vise den anden fordeling transformeres X_i-erne til $Y_i = \frac{X_i-\mu}{\sigma}$. Alle Y_i'er standardnormalfordelte: $Y_i \sim N(0,1)$. Da X_i'erne og dermed også Y_i'erne er uafhængige er $Y = (Y_1, Y_2, \cdots, Y_n)^T$ derfor en n-dimensional standardnormalfordeling. Sæt nu $e_1 = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}}\right)^T$ til en vektor i n dimensioner. Dette er en enhedsvektor. Suppler nu op til en ortonormalbasis $\{e_1, e_2, \cdots, e_n\}$ for $\mathbb{R}^n$. Saml disse vektorer i en ortogonal matrix $A = [e_1\ e_2 \cdots e_n] \in O(n)$. Ifølge den affine transformationsegenskab for multidimensionale normale fordelinger er også $Z \equiv AY$ standardnormalfordelt i n dimensioner. Specielt betyder det:

$$Z_1 = \frac{1}{\sqrt{n}}Y_1 + \frac{1}{\sqrt{n}}Y_2 + \cdots + \frac{1}{\sqrt{n}}Y_n = \frac{1}{\sqrt{n}}(Y_1 + Y_2 + \cdots + Y_n) = \frac{n}{\sqrt{n}}\bar{Y} = \sqrt{n}\bar{Y}$$

Regn nu:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{1}{\sigma}(\bar{X} - \mu) \Leftrightarrow \sigma\bar{Y} = \bar{X} - \mu$$

Så derfor er:

$$\sigma(Y_i - \bar{Y}) = X_i - \mu - (\bar{X} - \mu) = X_i - \bar{X}$$

Regn nu:

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

$$\sum_{i=1}^{n}Y_i^2 - n\,\bar{Y}^2 = |Y|^2 - Z_1^2 = |A^T Z|^2 - Z_1^2 = |Z|^2 - Z_1^2 = Z_2^2 + Z_3^2 + \cdots + Z_n^2$$

Dette er en sum kvadraterne på $n-1$ uafhængige standardnormalfordelinger. Størrelsen $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$ er altså χ^2_{n-1}-fordelt. Eller sagt på en anden måde: $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$. Da $\bar{X}$ kan udtrykkes vha. Z_1 og S^2 vha. $Z_2, Z_3, \cdots, Z_n$ er de to uafhængige. Hvilket bliver vigtigt når vi ser på T:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sqrt{S^2}/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S^2/\sigma}} = \frac{\sqrt{n}\bar{Y}}{\sqrt{S^2/\sigma^2}} = \frac{Z_1}{\sqrt{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}}$$

Tælleren er altså standardnormalfordelt, mens nævneren er fordelt som $\chi^2_{n-1}/\sqrt{n-1}$. Altså er hele størrelsen T t-fordelt med $n-1$ frihedsgrader. Bevis slut.

Integraler og beregninger

Gaussiske integraler

Det grundlæggende integral

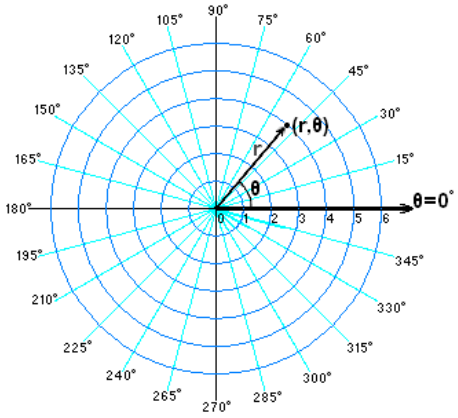
Vi er interesserede i værdien af følgende integral:

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx$$

Dette kan beregnes med et snedigt trick. Betragt i stedet I^2:

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2}\, dx \cdot \int_{-\infty}^{\infty} e^{-y^2}\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy$$

Der skiftes nu til polære koordinater, hvor r er radius og θ vinkel.



Transformationen kan udtrykkes som $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx\, dy = \int_0^{2\pi}\int_0^{\infty} r\, dr\, d\theta$. Så vi får:

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\cdot r\, dr\, d\theta = 2\pi\int_0^{\infty} e^{-r^2}\cdot r\, dr$$

Benyt nu substitutionen $t = r^2$, så $\frac{dt}{dr} = 2r \Leftrightarrow dr = \frac{1}{2r}dt$:

$$I^2 = 2\pi\int_0^{\infty} e^{-t}\cdot r\cdot\frac{1}{2r}dt = \pi\int_0^{\infty} e^{-t}\, dt = \pi[-e^{-t}]_0^{\infty} = \pi\big(0-(-1)\big) = \pi$$

Derfor må der gælde $I = \sqrt{\pi}$.

Gaussisk integral med skalering

Et relateret integral er:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{a}}\, dx$$

Dette beregnes nemt med følgende substitution: $t = \frac{x}{\sqrt{a}} \Rightarrow \frac{dt}{dx} = \frac{1}{\sqrt{a}} \Leftrightarrow dx = \sqrt{a}\, dt$. Så:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{a}}\, dx = \int_{-\infty}^{\infty} e^{-t^2}\sqrt{a}\, dt = \sqrt{\pi}\cdot\sqrt{a} = \sqrt{a\pi}$$

I tilfældet $a = 2$:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}$$

Gaussisk integral med parametrene μ og σ

Endnu et beslægtet integral er følgende:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

Dette kan beregnes med substitutionen $z = \frac{x-\mu}{\sigma} \Rightarrow \frac{dz}{dx} = \frac{1}{\sigma} \Leftrightarrow dx = \sigma\,dz$:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,\sigma\,dz = \sqrt{2\pi\sigma^2}$$

Vægtede gaussiske integraler

Vi vil også være interesseret i følgende interaler:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \cdot x^n\,dx$$

Når n er ulige er det tydeligt, at der er tale om en ulige integrand, hvorfor integralet må give 0. Hvad med n lige?

$n = 2$

Tilfældet $n = 2$ kan findes med følgende trick: Start med integralet

$$\int_{-\infty}^{\infty} e^{-\lambda x^2}\,dx = \sqrt{\frac{\pi}{\lambda}}$$

Værdien er fundet ved at bruge $a = \frac{1}{\lambda}$ og formlen fra sidste afsnit. Differentier nu integralet efter parameteren λ:

$$\frac{d}{d\lambda}\int_{-\infty}^{\infty} e^{-\lambda x^2}\,dx = \int_{-\infty}^{\infty} \frac{d}{d\lambda} e^{-\lambda x^2}\,dx = \int_{-\infty}^{\infty} e^{-\lambda x^2} \cdot (-x^2)\,dx = -\int_{-\infty}^{\infty} e^{-\lambda x^2} \cdot x^2\,dx$$

I alt har vi:

$$\int_{-\infty}^{\infty} e^{-\lambda x^2} \cdot x^2\,dx = -\sqrt{\pi}\frac{d}{d\lambda}\lambda^{-½} = -\sqrt{\pi}(-½)\lambda^{-\frac{3}{2}} = \frac{1}{2}\sqrt{\frac{\pi}{\lambda^3}}$$

Indsæt nu $\lambda = \frac{1}{2\sigma^2}$ for at få:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \cdot x^2\,dx = \frac{1}{2}\sqrt{\pi(2\sigma^2)^3} = \sqrt{2\pi}\sigma^3$$

$n = 4$

Stort set samme trick kan benyttes her, men denne gang skal man differentiere to gange:

$$\frac{d^2}{d\lambda^2}\int_{-\infty}^{\infty} e^{-\lambda x^2}\,dx = \int_{-\infty}^{\infty} \frac{d^2}{d\lambda^2} e^{-\lambda x^2}\,dx = \int_{-\infty}^{\infty} e^{-\lambda x^2}\cdot x^4\,dx$$

Integralet er derfor lig med:

$$\int_{-\infty}^{\infty} e^{-\lambda x^2}\cdot x^4\,dx = \sqrt{\pi}\frac{d^2}{d\lambda^2}\lambda^{-½} = -\frac{\sqrt{\pi}}{2}\frac{d}{d\lambda}\,\lambda^{-\frac{3}{2}} = -\frac{\sqrt{\pi}}{2}\left(-\frac{3}{2}\right)\lambda^{-\frac{5}{2}} = \frac{3\sqrt{\pi}}{4}\lambda^{-\frac{5}{2}}$$

Ved at indsætte $\lambda = \frac{1}{2\sigma^2}$ fås:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\cdot x^4\,dx = \frac{3}{4}\sqrt{\pi(2\sigma^2)^5} = 3\sqrt{2\pi}\sigma^5$$

Lige n generelt

Pr. induktion kan det vises, at den generelle formel for lige n er:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\cdot x^n\,dx = 1\cdot 3\cdot\cdots\cdot(n-1)\sqrt{2\pi}\sigma^{n+1}$$

Vægtede gaussiske integraler med både μ og σ

Med en simpel substitution $t = x - \mu$ ser man, at man får præcis de samme værdier for integraler af typen:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot(x-\mu)^n\,dx$$

Men hvad hvis vi er interesserede i følgende:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot x^n\,dx$$

Igen kan substitutionen $t = x - \mu \Leftrightarrow x = t + \mu$ hjælpe:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot x^n\,dx = \int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\cdot(t+\mu)^n\,dt$$

$(t+\mu)^n$ kan nu omskrives vha. binomialsætningen:

$$(t+\mu)^n = \sum_{k=0}^{n}\binom{n}{k}t^k\mu^{n-k}$$

Her er $\binom{n}{k}$ binomialkoefficienten givet ved $\frac{n!}{k!(n-k)!}$. Integralet kan nu skrives:

$$\int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\cdot(t+\mu)^n\,dt = \sum_{k=0}^{n}\binom{n}{k}\mu^{n-k}\int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\cdot t^k\,dt = \sqrt{2\pi}\sum_{k=0}^{n}\binom{n}{k}\mu^{n-k}\,c_k\sigma^{k+1}$$

Her er konstanten c_k lig 0 når k er ulige og $1\cdot 3\cdot\cdots\cdot(k-1)$ når k er lige. $c_0 = 1$.

For $n = 1$ betyder det:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot x \, dx = \sqrt{2\pi}(1 \cdot \mu^1 c_0 \sigma^1 + 1 \cdot \mu^0 c_1 \sigma^2) = \sqrt{2\pi}\mu\sigma$$

For $n = 2$:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot x^2 \, dx = \sqrt{2\pi}(1 \cdot \mu^2 c_0 \sigma^1 + 2 \cdot \mu^1 c_1 \sigma^2 + 1 \cdot \mu^0 c_2 \sigma^3) = \sqrt{2\pi}(\mu^2\sigma + \sigma^3)$$

For $n = 3$:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot x^3 \, dx = \sqrt{2\pi}(1 \cdot \mu^3 c_0 \sigma^1 + 3 \cdot \mu^2 c_1 \sigma^2 + 3 \cdot \mu^1 c_2 \sigma^3 + 1 \cdot \mu^0 c_3 \sigma^4) = \sqrt{2\pi}(\mu^3\sigma + 3\mu\sigma^3)$$

For $n = 4$ (hvor de ulige k'er nu udelades):

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot x^4 \, dx = \sqrt{2\pi}(1 \cdot \mu^4 c_0 \sigma^1 + 6 \cdot \mu^2 c_2 \sigma^3 + 1 \cdot \mu^0 c_4 \sigma^5) = \sqrt{2\pi}(\mu^4\sigma + 6\mu^2\sigma^3 + 3\sigma^5)$$

Og så videre.

Gaussisk integral med kvadratisk form

Betragt integralet:

$$\int_{-\infty}^{\infty} \exp[-a(x^2 + bx + c)] \, dx$$

Her skal a være større end 0. Betragt nu:

$$\left(x + \frac{b}{2}\right)^2 = x^2 + bx + \frac{b^2}{4} \Leftrightarrow x^2 + bx = \left(x + \frac{b}{2}\right)^2 - \frac{b^2}{4}$$

Derfor er:

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 - \frac{b^2}{4} + c$$

Integralet er derfor:

$$\int_{-\infty}^{\infty} \exp\left[-a\left(x + \frac{b}{2}\right)^2 + a\left(\frac{b^2}{4} - c\right)\right] dx = \exp\left[a\left(\frac{b^2}{4} - c\right)\right] \int_{-\infty}^{\infty} \exp\left[-a\left(x + \frac{b}{2}\right)^2\right] dx$$

Substituer nu $t = x + \frac{b}{2} \Rightarrow \frac{dt}{dx} = 1$, så $dx = dt$. Integralet er derfor:

$$\int_{-\infty}^{\infty} \exp[-at^2] dt = \sqrt{\frac{\pi}{a}}$$

Så i alt:

$$\int_{-\infty}^{\infty} \exp[-a(x^2+bx+c)]\,dx = \exp\left[a\left(\frac{b^2}{4}-c\right)\right]\sqrt{\frac{\pi}{a}}$$

Alternativ form

Betragt integralet:

$$\int_{-\infty}^{\infty} \exp[-(ax^2+bx+c)]\,dx$$

Her skal a være større end 0. Vi vil nu gerne omskrive eksponenten til et kvadrat plus et restled. Vi ser:

$$ax^2+bx+c = a\left(x^2+\frac{b}{a}x+\frac{c}{a}\right)$$

For at få noget der indeholder indmaden i parentesen beregnes følgende:

$$\left(x+\frac{b}{2a}\right)^2 = x^2+\frac{b}{a}x+\frac{b^2}{4a^2} \Leftrightarrow x^2+\frac{b}{a}x = \left(x+\frac{b}{2a}\right)^2-\frac{b^2}{4a^2}$$

Derfor er:

$$ax^2+bx+c = a\left[\left(x+\frac{b}{2a}\right)^2-\frac{b^2}{4a^2}+\frac{c}{a}\right] = a\left(x+\frac{b}{2a}\right)^2-\frac{b^2}{4a}+\frac{4ac}{4a} = a\left(x+\frac{b}{2a}\right)^2-\frac{d}{4a}$$

Her er $d = b^2-4ac$ diskriminanten. Integralet kan nu skrives:

$$\int_{-\infty}^{\infty} \exp\left[-a\left(x+\frac{b}{2a}\right)^2+\frac{d}{4a}\right]dx = e^{\frac{d}{4a}}\int_{-\infty}^{\infty}\exp\left[-a\left(x+\frac{b}{2a}\right)^2\right]dx$$

Substituer nu $t = x+\frac{b}{2a} \Rightarrow \frac{dt}{dx} = 1$, så $dx = dt$. Integralet er derfor:

$$e^{\frac{d}{4a}}\int_{-\infty}^{\infty} e^{-at^2}dt = \sqrt{\frac{\pi}{a}}e^{\frac{d}{4a}}$$

Kompleks funktionsteori – hurtigt resume

En kompleks funktion $f(z)$ er en funktion fra en delmængde af den komplekse plan $U \subseteq \mathbb{C}$ ind i $\mathbb{C}$. Differentiabilitet defineres på samme måde som for reelle funktioner, men her bliver betingelsen betydeligt stærkere: En funktion der er differentiabel én gang i z_0 er altid differentiabel et vilkårligt antal gange i z_0 og funktionen er lig med sin egen Taylorrække i en åben omegn af z_0.

En pol er et isoleret punkt c hvor f ikke er defineret. En differentiabel funktion der kun har et endeligt (eller tælleligt) antal af sådanne poler kaldes meromorf. I en omegn af en pol c kan funktionen skrives som en såkaldt Laurentrække:

$$f(z-c) = \sum_{-\infty}^{\infty} a_n(z-c)^n$$

Koefficienten a_{-1} kaldes for residuet af polen og betegnes Res_c eller $\mathrm{Res}(f, c)$. Hvis Det største tal n for hvilken $a_{-n} \neq 0$ kaldes for polens orden. En pol af orden 1 kaldes simpel. Hvis der ikke findes en sådant n kaldes polen for en essentiel singularitet. Residuet for en simpel pol kan findes som:

$$\mathrm{Res}_c = \lim_{z \to c}(z - c) \cdot f(z)$$

En kurve i planen der ikke skærer sig selv kaldes simpel. For en en simpel, lukket kurve γ der løber i positiv omløbsretning gælder Cauchys residuesætning:

$$\oint_\gamma f(z)\, dz = 2\pi i \sum_n \mathrm{Res}(f, c_n)$$

Her løber summen over samtlige poler der ligger inden for γ. Denne formel er ofte uhyre praktisk.

Lineær algebra

SVD (Singular Value Decomposition)

Vi får brug for følgende lemma:

Lemma: Lad A være en $n \times m$ matrix. Da gælder

- AA^{T} og $A^{\mathrm{T}}A$ er symmetriske og semi-positivt definitte.
- A, AA^{T} og $A^{\mathrm{T}}A$ har samme rang r.
- AA^{T} og $A^{\mathrm{T}}A$ har hver r egenværdier der ikke er nul, og disse er sammenfaldende.

Bevis: Lad os kalde $N = AA^{\mathrm{T}}$ og $M = A^{\mathrm{T}}A$. Der gælder nu om N's indgange: $n_{ij} = a_{ik}a_{jk} = a_{jk}a_{ik} = n_{ji}$ (her benyttes Einsteins summationskonvention), hvilket viser symmetrien. Tilsvarende for M. At matricerne er positivt definitte ses ved for en arbitrær vektor $\vec{x}$ at beregne:

$$\vec{x}^{\mathrm{T}}M\vec{x} = \vec{x}^{\mathrm{T}}A^{\mathrm{T}}A\vec{x} = (A\vec{x})^{\mathrm{T}}A\vec{x} = ||A\vec{x}||^{\,2} \geq 0$$

Tilsvarende for N.

Hvis A har rang r siger dimensionssætningen at A har nullitet $m - r$. A og A^{T} har samme rang, så her siger dimensionssætningen at A^{T} har nullitet $n - r$.

Der gælder følgende praktiske sætning:

Sætning (SVD): Lad A være en $n \times m$ matrix. Da findes der to ortogonale matricer U af dimension $n \times n$ og V af dimension $n \times n$ og D en diagonalmatrix af samme dimension som A således at:

$$A = UDV^{\mathrm{T}}$$

Bevis: Betragt matricen $N = AA^{\mathrm{T}}$. Den har dimension $n \times n$ og dens indgange er $n_{ij} = a_{ik}a_{jk} = a_{jk}a_{ik} = n_{ji}$ (her benyttes Einsteins summationskonvention). Altså er N symmetrisk og således diagonaliserbar med egenværdier $\lambda_1^{(N)}, \lambda_2^{(N)}, \cdots, \lambda_n^{(N)}$ og en tilhørende ortonormal basis af egenvektorer $e_1^{(N)}, e_2^{(N)}, \cdots, e_n^{(N)}$, der tilsammen udgør en ortogonal matrix P_N, så der gælder:

$$AA^{\mathrm{T}} = P_N^{-1}\,\mathrm{diag}\left(\lambda_1^{(N)}, \lambda_2^{(N)}, \cdots, \lambda_n^{(N)}\right) P_N$$

Der gælder noget helt tilsvarende for $M = A^{\mathrm{T}}A$, der altså har dimension $m \times m$:

$$A^{\mathrm{T}}A = P_M^{-1}\,\mathrm{diag}\left(\lambda_1^{(M)}, \lambda_2^{(M)}, \cdots, \lambda_m^{(M)}\right) P_M$$

AA^{T} og $A^{\mathrm{T}}A$ har de samme, positive egenværdier, bortset fra en evt. stribe af nuller, svarende til nulrummet. Hvis r er antallet af positive egenværdier har vi altså.

Stirlings formel

Stirlings formel giver en god approksimation til $n!$ for store værdier af n:

$$n! \approx \sqrt{2\pi n} \cdot n^n \cdot e^{-n}$$

Sætning (Stirlings formel): Der gælder følgende:

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n} \cdot n^n \cdot e^{-n}} = 0$$

Gamma-funktionen

Gamma-funktionen kan ses som en udvidelse af fakultetsfunktionen. For $x > 0$ er definitionen:

$$\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t}\ dt$$

Så der må gælde:

$$\Gamma(1) = \int_0^\infty t^{1-1} \cdot e^{-t}\ dt = \int_0^\infty e^{-t}\ dt = [-e^{-t}]_0^\infty = 0 - (-1) = 1$$

Ved hjælp af partiel integration ses desuden:

$$\Gamma(n+1) = \int_0^\infty t^n \cdot e^{-t}\ dt = [t^n \cdot (-e^{-t})]_0^\infty - \int_0^\infty n \cdot t^{n-1} \cdot (-e^{-t})\ dt$$

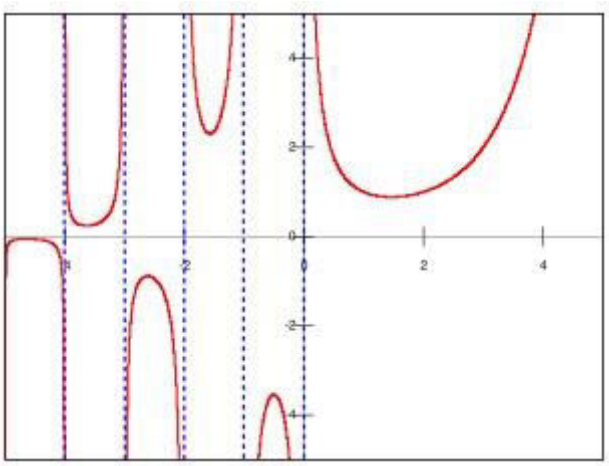
Det benyttes at e^{-t} aftager hurtigere end t^n:

$$0 - 0 + n \cdot \int_0^\infty t^{n-1} \cdot e^{-t}\ dt = n \cdot\ \Gamma(n)$$

Altså i alt:

$$\Gamma(n+1) = n \cdot\ \Gamma(n)$$

Det følger, at $\Gamma(n) = (n-1)!$ for heltallige n. Ligningen kan bruges til at udvide definitionsmængden til (ikke-heltallige) negative tal. Grafen for denne funktion er vist på figuren:



$\Gamma(½)$

Vi får brug for denne værdi senere:

$$\Gamma(½) = \int_0^\infty t^{½-1} \cdot e^{-t}\, dt = \int_0^\infty \frac{e^{-t}}{\sqrt{t}}\, dt$$

Foretag nu substitutionen $s = \sqrt{t}$, dvs. $t = s^2$ og $\frac{ds}{dt} = \frac{1}{2\sqrt{t}} \Leftrightarrow dt = 2\sqrt{t}\, ds = 2s\, ds$:

$$\int_0^\infty \frac{e^{-s^2}}{s} 2s\, ds = 2\int_0^\infty e^{-s^2}\, ds$$

Integralet er halvdelen af det velkendte $\int_{-\infty}^{\infty} e^{-s^2}\, ds = \sqrt{\pi}$, da integranden er en lige funktion. Så i alt:

$$\Gamma(½) = \sqrt{\pi}$$

Den afledte af gamma-funktionen

Vi kan også få brug for at differentiere gamma-funktionen:

$$\frac{d}{dx}\Gamma(x) = \int_0^\infty \frac{d}{dx}(t^{x-1} \cdot e^{-t})\, dt = \int_0^\infty \log(t) \cdot t^{x-1} \cdot e^{-t} dt$$

Her er log den naturlige logaritme.

Sandsynligheder

Sandsynligheder bruges når vi er situationer hvor en størrelse enten er bestemt ved en tilfældig proces, eller er ukendt og vi gisner om dens sande værdi. Sandsynligheder måles enten i procenter mellem 0% og 100% eller tilsvarende mellem 0 og 1.

Sandsynlighedsrum

Den generelle definition af hvad vi forstår ved et *sandsynlighedsrum* for et eksperiment er lidt teknisk. Der er tre komponenter:

1. Et udfaldsrum givet ved mængden Ω. Alle tænkelige udfald af eksperimentet er elementer i Ω.

En hændelse er defineret som en delmængde af Ω. Hvis det faktisk udfald af eksperimentet ligger i en hændelse $A \subseteq \Omega$ siger vi at hændelsen A er indtruffet. Vi er interesseret i at tildele sandsynligheder til en sådan hændelse. Generelt viser det sig dog, at det ikke er alle delmængder af Ω der kan tildeles en meningsfuld sandsynlighed! Rent teknisk vil vi kun tildele sandsynligheder til de delmængder der tilhører en såkaldt σ-algebra, der er defineret ved:

Definition: Givet en mængde Ω og lad σ være en delmængde af potensmængden af Ω (mængden af alle delmængder af Ω). $\mathcal{B}$ kaldes da en *σ-algebra* over Ω hvis den opfylder:

- $\mathcal{B}$ indeholder den tomme mængde: $\emptyset \in \mathcal{B}$
- $\mathcal{B}$ er lukket under komplement: $A \in \mathcal{B} \Rightarrow \ \Omega \backslash \mathrm{A} \in \mathcal{B}$
- $\mathcal{B}$ er lukket under tællelig foreningsmængde: $A_i \in \mathcal{B} \Rightarrow \bigcup_i A_i \in \mathcal{B}$, hvor $i \in \mathbb{N}$

I praksis vil vi ofte have *Borel-algebraen*, der består af den mindste σ-algebra der indeholder alle intervaller, i tankerne. Vores anden komponent er altså:

2. $\mathcal{B}$, en σ-algebra over Ω. Elementerne i $\mathcal{B}$ kalder vi *målelige hændelser*.

Endelig kommer vi til *sandsynlighedsmålet*. Til hver målelig hændelse A knyttes en sandsynlighed $P(A)$. For at stemme overens med vores intuition om sandsynligheder skal målet opfylde:

Definition: En afbildning $P\colon \mathcal{B} \to \mathbb{R}$, hvor $\mathcal{B}$ er en σ-algebra over Ω kaldes et sandsynlighedsmål hvis den opfylder:

- $P(A) \geq 0$ for alle $A \in \mathcal{B}$
- $P(\Omega) = 1$
- Hvis $A_1, A_2, \cdots \in \mathcal{B}$ er parvist disjunkte er $P(\bigcup_i A_i) = \sum_i P(A_i)$

Dette er vores tredje element:

3. P, et sandsynlighedsmål på $\mathcal{B}$.

Tilsammen udgør sættet $(\Omega, \mathcal{B}, P)$ sandsynlighedsrummet.

Ligefordelt udfaldsrum

Et vigtigt, grundlæggende eksempel er det *ligefordelte* udfaldsrum. Her tildeles hver af de individueller udfald i Ω lige stor sandsynlighed.

Endeligt udfaldsrum

For et endeligt udfaldsrum Ω med n elementer må der altså gælde følgende for alle $a \in \Omega$:

$$P(\{a\}) = \frac{1}{n}$$

Eksempler kun være en ærlig mønt, hvor $\Omega = \{\text{plat}, \text{krone}\}$ eller en ærlig terning hvor $\Omega = \{1,2,3,4,5,6\}$. Ofte skriver man blot f.eks. $P(3) = \frac{1}{6}$ i stedet for $P(\{3\}) = \frac{1}{6}$.

Kontinuert udfaldsrum

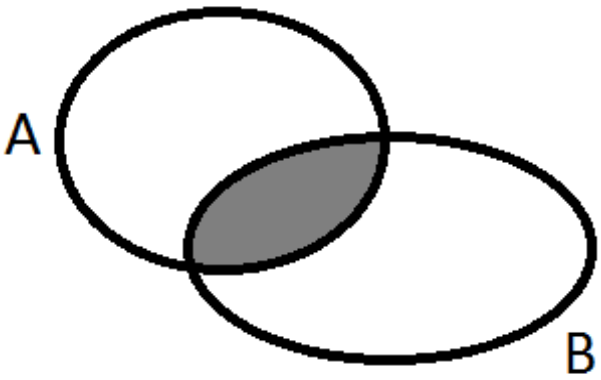
Hvis udfaldsrummet i stedet er et interval $[a, b]$ kan vi også konstruere en ligefordeling, men her vil sandsynligheden for enhver singleton være 0, hvilket måske virker kontraintuitivt. I stedet kan sandsynligheden for at få et udfald i et interval $[c, d] \subseteq [a, b]$ tillægges en sandsynlighed:

$$P([c, d]) = \frac{d - c}{b - a}$$

Betinget sandsynlighed og uafhængighed

Man kan spørge hvad sandsynligheden for at en hændelse A er indtruffet, hvis vi allerede ved at hændelsen B er indtruffet. Dette kaldes den *betingede sandsynlighed* $P(A|B)$.

Situationen er illustreret på figuren til højre.



Vi ved at B er indtruffet, så udfaldet ligger et sted i mængden. Fællesmængden $A \cap B$ (den grå mængde) angiver de udfald hvor både A og B er indtruffet. Sandsynligheden for at udfaldet også ligger heri, altså den betingede sandsynlighed må være:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' sætning

Heraf følger umiddelbart Bayes' sætning:

Sætning (Bayes): For to hændelser A og B gælder:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bevis: Der gælder pr. definition:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A \cap B) = P(A|B) \cdot P(B)$$

Men også:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Leftrightarrow P(A \cap B) = P(B|A) \cdot P(A)$$

Og derfor også:

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \Leftrightarrow P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bevis slut.1

Sætningen ser tilforladelig ud, men faktisk er der grundlagt en helt retning i statistikken hvis tolkninger er baserede på sætningen.

Uafhængighed

Hændelsen A kaldes uafhængig af hændelsen B, hvis det at B har indtruffet ikke ar nogen indflydelse på, om A er indtruffet. Med andre ord skal der gælde:

$$P(A) = P(A|B)$$

Pr. definition betyder dette:

$$P(A) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

Heraf ses, at uafhængighed svarer til, at man kan multiplicere sandsynligheder. Vi ser også, at hvis A er uafhængig af B, må B også være uafhængig af A.

Stokastiske variable

Definition

En *stokastisk variabel* (engelsk: random variable) er et tal eller en værdi der knyttes til hver af de enkelte udfald i et sandsynlighedsrum.

Definition: En stokastisk variabel X på et sandsynlighedsrum $(\Omega, \mathcal{B}, P)$ er en afbildning $X\colon \Omega \to \mathbb{R}$.

Eksempler med terninger

For et terningkast er antallet af øjne et stokastisk variabel. Hvis man kaster flere terninger kunne det samlede antal øjne eller gennemsnittet af øjne være eksempler på stokastiske variable.

Lidt om notation

Til enhver delmængde af $\mathbb{R}$ tilknyttes naturligt en hændelse:

Hvis $B \subseteq \mathbb{R}$ er $X^{-1}(B)$, altså urmængden af B en hændelse. Hvis $X^{-1}(B) \in \mathcal{B}$ kan vi altså naturligt definere:

$$P(x \in B) = P(X^{-1}(B))$$

Hvis B er en singleton kan man mere simpelt skrive:

$$P(X = x) = P(X^{-1}(\{x\}))$$

Tilsvarende er definitionen af udtryk som $P(X > x)$ etc. klar.

Eksempler med terninger

Hvis X angiver antal øjne ved kast med en ærlig terning er $P(X = 2)$ sandsynligheden for at slå en toer, altså $\frac{1}{6}$. $P(X \geq 5)$ er sandsynligheden for at slå fem eller mere, altså $\frac{2}{6} = \frac{1}{3}$.

Frekvens- og fordelingsfunktion for stokastisk variabel

Frekvensfunktion

Definition: For en stokastisk variabel X på et diskret udfaldsrum (ofte blot kaldet en diskret stokastisk variabel) er *frekvensfunktionen* defineret ved:

$$f_X(n) = P(X = n)$$

For en stokastiske variabel på et kontinuert udfaldsrum (ofte blot kaldet en kontinuert stokastisk variabel), angiver frekvensfunktionen $f_X(x)$ i stedet *sandsynlighedstætheden*, således at:

$$\int_{x_a}^{x_b} f_X(x)\, dx = P(x_a < X < x_b)$$

Frekvensfunktionen benævnes ofte pdf (probability density function) på engelsk.

Fordelingsfunktion

Definition: *Fordelingsfunktionen* for en stokastisk variabel er defineret ved:

$$F_X(x) = P(X \leq x)$$

For diskrete stokastiske variable er dette:

$$F_X(n) = \sum_{x \leq n} f_X(x)$$

For kontinuerte stokastiske variable kan man i stedet skrive:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

Heraf ses også, at $F_X{}'(x) = f_X(x)$ såfremt $F_X(x)$ er differentiabel.

Fordelingsfunktionen benævnes ofte cpf (cumulative probability function) på engelsk.

Eksempel med terninger

Hvis X angiver antal øjne ved kast med en ærlig terning er der tale om et diskret udfaldsrum, da der kun er seks mulige udfald. Da hver er lige sandsynlig er der tale om et ligefordelt udfaldsrum, så frekvensfunktionen er $f_X(n) = \frac{1}{6}$ for $n \in \{1,2,3,4,5,6\}$.

Fordelingsfunktionen fås ved at "tælle op":

$$F_X(n) = P(X \leq n) = \frac{n}{6}$$

Eksempel med tilfældig udvælgelse af tal

Lad X angive et tal, tilfældigt udtaget mellem 0 og 1 således, at alle udfald er lige sandsynlige. Der er et kontinuum af muligheder, så der er tale om et kontinuert udfaldsrum. Igen en ligefordeling, så frekvensfunktionen er konstant $f_X(x) = 1$, $x \in [0,1]$. Dette passer med, at sandsynligheden for at vælge et tal mellem a og b, hvor $0 \leq a < b \leq 1$ netop er $\int_a^b 1\, dx = b - a$.

Fordelingsfunktionen fås ved at integrere frekvensfunktionen:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt = \int_0^x 1\, dt = x - 0 = x$$

Undervejs er benyttet, af frekvensfunktionen pr. definition er sat til nul udenfor $[0,1]$.

Frekvens- og fordelingsfunktion for flere variable

Hvis vi har to stokastiske variable X_1 og X_2 kan man definere tilsvarende, samlede frekvens- og fordelingsfunktioner:

Definition: For to diskrete stokastiske variable X_1 og X_2 er den samlede frekvensfunktion:

$$f_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1 \text{ og } X_2 = x_2)$$

For kontinuerte variable skal der i stedet gælde:

$$P(x_{1a} < x_1 < x_{1b} \text{ og } x_{2a} < x_2 < x_{2b}) = \int_{x_{1a}}^{x_{1b}} \int_{x_{2a}}^{x_{2b}} f_{X_1,X_2}(x_1, x_2)\, dx_2\, dx_1$$

Eksempel med terninger

Lad X_1 være antallet af øjne ved kast med en ærlig terning. Lad X_2 være antallet af øjne i midten af samme terning, altså 0 for slagene to, fire og seks og 1 for slagene et, tre og fem.

Da er f.eks. $P(X_1 = 2, X_2 = 0) = \frac{1}{6}$ da der netop er et af de seks udfald der svarer til, mens $P(X = 5, X = 0) = 0$ da der ikke er nogen udfald der svarer til.

Definition: For to stokastiske variable X_1 og X_2 er den samlede fordelingsfunktion:

$$F_{X_1,X_2}(x_1, x_2) = P(X_1 \leq x_1 \text{ og } X_2 \leq x_2)$$

Uafhængighed mellem stokastiske variable

Ligesom hændelser kan være uafhængige kan det samme være tilfældet for to stokastiske variable X_1 og X_2.

Definition: To stokastiske variable X_1 og X_2 med frekvensfunktioner f_{X_1} og f_{X_2} kaldes *uafhængige*, såfremt der for alle x_1 og x_2 gælder:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$$

Eksempel med terninger

Vi ser, at X_1 og X_2 i eksemplet ovenfor *ikke* er uafhængige, da $f_{X_1}(x_1) = \frac{1}{6}$ for alle seks muligheder (et til seks) og $f_{X_2}(x_2) = \frac{1}{2}$ for begge muligheder (nul og et). Men vi så bl.a. at:

$$P(X_1 = 2, X_2 = 0) = \frac{1}{6}$$

Men:

$$P(X_1 = 2) \cdot P(X_2 = 0) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Regneregler for frekvensfunktioner

Frekvensfunktion for sum af stokastiske variable

Definition: Givet to reelle funktioner f og g, er *foldningen* mellem f og g – betegnet $f * g$ – defineret ved:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\,dx$$

Sætning: Hvis X og Y er uafhængige, kontinuerte stokastiske variable med tilhørende frekvensfunktioner f_X og f_Y, da er frekvensfunktionen for $X+Y$ lig foldningen mellem f_X og f_Y. For diskrete variable er den tilsvarende formel:

$$f_{X+Y}(t) = \sum_t f_X(x)f_Y(t-x)$$

Bevis: Da de to variable er uafhængige er den samlede frekvensfunktion netop produktet af de to frekvensfunktioner, altså $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Betragt nu fordelingsfunktionen for $X+Y$:

$$F_{X+Y}(t) = P(X+Y \leq t) = P(x \in \mathbb{R} \wedge Y \leq t-x) = \int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f_X(x)f_Y(y)\,dy\,dx$$

Foretag nu følgende substitution: $s = x+y$, dvs. $y = s-x$ og $dy = ds$:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{t} f_X(x)f_Y(s-x)\,ds\,dx$$

Det ses nu at $f_{X+Y}(t) = F'_{X+Y}(t) = \int_{-\infty}^{\infty} f(x)g(t-x)\,dx$. Tilsvarende for det diskrete tilfælde.

Bevis slut.

Frekvensfunktion for kvadrat af stokastisk variabel

Sætning: Hvis X er en kontinuert stokastisk variabel med frekvensfunktion f, da er frekvensfunktionen for X^2 givet ved:

$$f_{X^2}(t) = \frac{d}{dt}\int_{-\sqrt{t}}^{\sqrt{t}} f(x)\,dx$$

Bevis: Betragt fordelingsfunktionen for X^2:

$$F_{X^2}(t) = P(X^2 \leq t) = P\left(X \geq -\sqrt{t} \wedge X \leq \sqrt{t}\right) = \int_{-\sqrt{t}}^{\sqrt{t}} f(x)\,dx$$

Brug nu at frekvensfunktionen er den afledede af fordelingsfunktionen.

Bevis slut.

Normalfordelingen – en kort introduktion

Normalfordelingen

Normalfordelingen er på mange måder den vigtigste fordeling der findes i statistik og sandsynlighedsregning. Den beskriver en variabel der opfylder følgende tre betingelser:

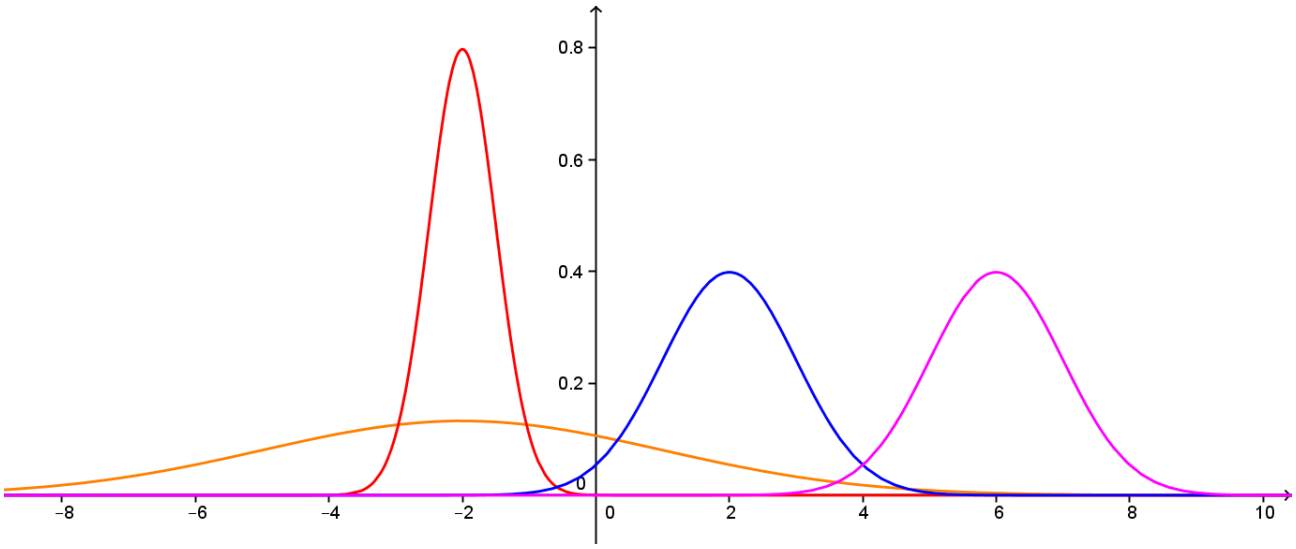
- Der er en tydelig middelværdi μ
- Observationerne er oftere tæt på μ end langt fra.
- Observationerne har lige stor chance for at ligge over som under μ.

Ud over middelværdien μ har normalfordelingen standardafvigelsen σ som parameter.

Definition: Normalfordelingen er givet ved følgende frekvensfunktion:

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

Hvorfor fordelingen præcis har denne form skal vi se nærmere på senere i forbindelse med en af statistikkens vigtigste sætninger: *den centrale grænseværdisætning*.



Figuren viser fire normalfordelinger: To med samme middelværdi, men med forskellige standardafvigelser: $\mu = -2$ og hhv. $\sigma = ½$ og $\sigma = 3$. Og to med samme standard afvigelse, men med forskellige middelværdien: $\sigma = 1$ og hhv. $\mu = 2$ og $\mu = 6$.

Standardnormalfordelingen

Alle normalfordelinger kan ses som en variabeltransformation af *standardnormalfordelingen*:

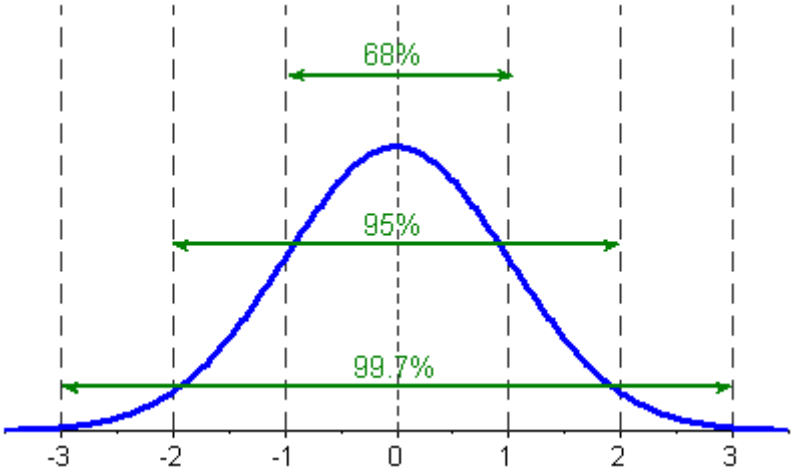
Definition: Normalfordelingen med $\mu = 0$ og $\sigma = 1$ kaldes standardnormalfordelingen. Den har altså frekvensfunktionen:

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

Empirisk regel for standardnormalfordelingen

For standardnormalfordelingen gælder der følgende såkaldte "empiriske regel":

- Ca. 68% af værdierne ligger mellem -1 og 1.
- Ca. 95% af værdierne ligger mellem -2 og 2.
- Ca. $99{,}7\%$ af værdierne ligger mellem -3 og 3.



Fordelingsfunktion for normalfordelinger

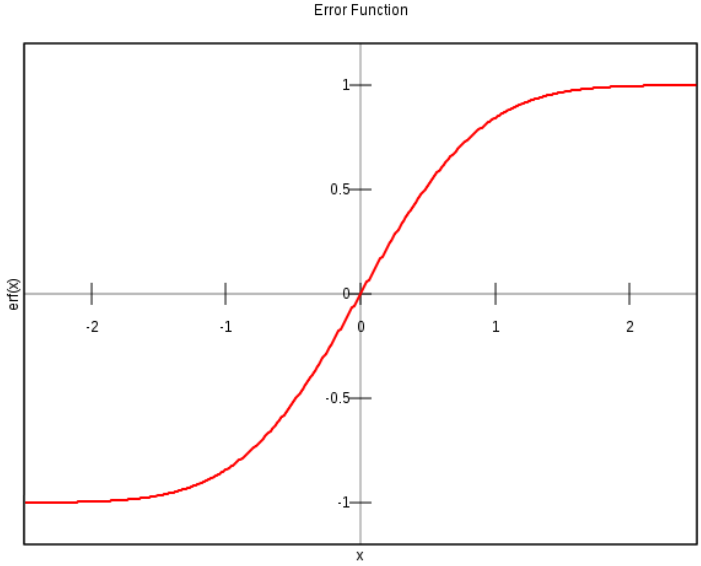
Standardnormalfordelingen

Fordelingsfunktionen F er som bekendt defineret på følgende måde:

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

For standardnormalfordelingen benyttes navnet Φ om denne funktion:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$



Dette integral kan generelt ikke beregnes eksakt. Men vi kan definere følgende:

Definition: Ved *fejl-funktionen* (engelsk: error function) forstås:

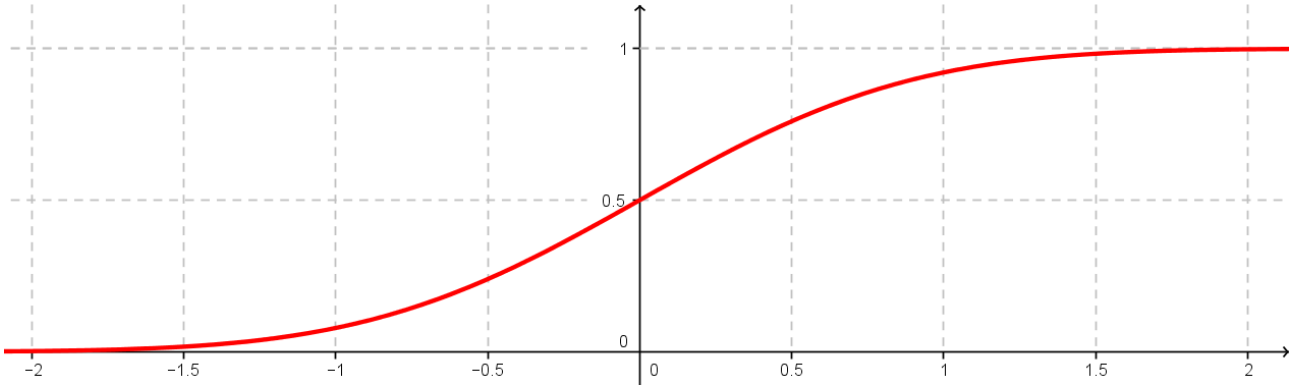
$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt$$

Figuren til venstre viser grafen for fejl-funktionen. Som man kan ane på figuren går funktionsværdien af $\mathrm{erf}(x)$ mod ± 1 når x går mod $\pm\infty$.

Derfor kan vi nu regne som følger ved at foretage substitutionen $s = \frac{t}{\sqrt{2}}$, hvilket medfører $\frac{ds}{dt} = \frac{1}{\sqrt{2}} \Leftrightarrow dt = \sqrt{2}\, ds$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}}\, ds = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}} \sqrt{2}\, ds = \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{0} e^{-\frac{s^2}{2}}\, ds + \int_{0}^{x} e^{-\frac{s^2}{2}}\, ds \right) =$$

$$\frac{1}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} \left(1 + \mathrm{erf}(x)\right) = \frac{1}{2} + \frac{1}{2} \mathrm{erf}(x)$$



Grafen for $\Phi(x)$ er vist ovenfor. Tabellen herunder viser nogle værdier af $\Phi(x)$:

x	0	1,282	1,645	1,960	2,576	3,291
$\Phi(x)$	0,5	0,9	0,95	0,975	0,995	0,9995

I praksis er det ofte *fraktiler* til fordelingsfunktionen der er interessante i anvendelser, hvorfor ovenstående tabel snarere skal læses nedefra og op end omvendt.

Generelle normalfordelinger

Man kan altid transformere en normalfordelt stokastisk variabel X om til en standardnormaltfordelt stokastisk variabel Z ved transformationen:

$$Z = \frac{X - \mu}{\sigma} \Leftrightarrow X = \sigma Z + \mu$$

Desuden gælder $\frac{dz}{dx} = \frac{1}{\sigma} \Leftrightarrow dx = \sigma\, dz$. For en normalfordelt variabel er fordelingsfunktionen:

$$F(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{(\sigma z + \mu - \mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{z^2}{2}} dx =$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\frac{t-\mu}{\sigma}} e^{-\frac{z^2}{2}} \sigma\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t-\mu}{\sigma}} e^{-\frac{z^2}{2}}\, dz = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

Med andre ord:

$$F(x) = \Phi(z)$$

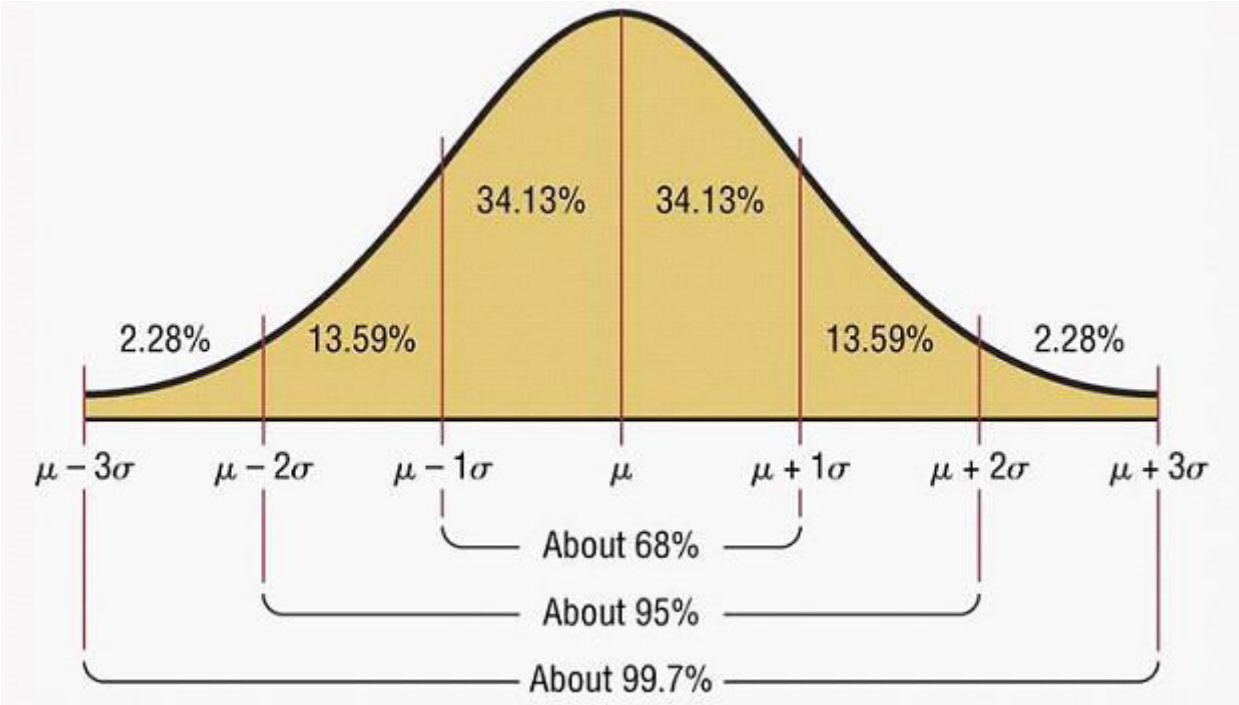
Empirisk regel for

Her er $z = \frac{x-\mu}{\rho}$. Størrelsen kaldes for *z-score* eller *standardscore*. Ved at finde z-score for en normaltfordelt stokastisk variabel kan man altså bruge alle resultater for standardnormalfordelingen. z-score angiver hvor mange standardafvigelser en observation ligger væk fra middelværdien.

Empirisk regel for generel normalfordeling

Den tilsvarende regel for en generel normalfordeling er som følger:

- Ca. 68% af værdierne ligger mellem $\mu - \sigma$ og $\mu + \sigma$, altså inden for en standardafvigelse af middelværdien.
- Ca. 95% af værdierne ligger mellem $\mu - 2\sigma$ og $\mu + 2\sigma$, altså inden for to standardafvigelser af middelværdien.
- Ca. $99{,}7\%$ af værdierne ligger mellem $\mu - 3\sigma$ og $\mu + 3\sigma$, altså inden for tre standardafvigelser af middelværdien.



Forventningsværdier og momenter

Forventningsværdier

Definition: *Forventningsværdien* $E[X]$ af en stokastisk variabel X med tilhørende frekvensfunktion $f(x)$ er for diskrete fordelinger givet ved:

$$E[X] = \sum_x f(x) \cdot x$$

Her forløber summen over hele udfaldsrummet. For kontinuerte fordelinger er den tilsvarende definition:

$$E[X] = \int_{-\infty}^{\infty} f(x) \cdot x \, dx$$

Sætning: Der gælder følgende regneregler for forventningsværdier:

a. $E[c] = c$
b. $E[X_1 + X_2] = E[X_1] + E[X_2]$
c. $E[c \cdot X_1] = c \cdot E[X_1]$

Her er c en konstant og X_1 og X_2 stokastiske variable med tilhørende frekvensfunktioner f og g.

Bevis: Beviset foregår for diskrete fordelinger, men forløber helt på samme måde for kontinuerte ditto:

a. $E[c] = \sum_x f(x) \cdot c = c \cdot \sum_x f(x) = c \cdot 1 = c$
b. Ved at bruge formlen for frekvensfunktionen for summen af to variable får man:

$$E[X_1 + X_2] = \sum_s \sum_t f(t)g(s-t) \cdot s = \sum_s \sum_t f(t)g(s-t) \cdot \big(t + (s-t)\big) =$$
$$\sum_s \sum_t f(t)\, g(s-t) \cdot t + \sum_s \sum_t f(t)\, g(s-t) \cdot (s-t)$$

Substitutionen $d = s - t$ foretages:

$$\sum_t f(t) \cdot t \sum_d g(d) + \sum_t f(t) \sum_d g(d) \cdot (d) = E[X_1] \cdot 1 + 1 \cdot E[X_2] = E[X_1] + E[X_2]$$

c. $E[c \cdot X_1] = \sum_x cx \cdot f(x) = c \sum_x x \cdot f(x) = c \cdot E[X_1]$

Bevis slut

Sætning: Hvis X_1 og X_2 er uafhængige stokastiske variable med tilhørende frekvensfunktioner f og g gælder:

$$E[X_1 X_2] = E[X_1]E[X_2]$$

Bevis: Beviset benytter at frekvensfunktionen for produktet af stokastiske variable for uafhængige variable blot er produktet af frekvensfunktionerne. Altså er:

$$E[X_1X_2] = \sum_{x_1}\sum_{x_2} f(x_1)g(x_2)x_1x_2 = \sum_{x_1} f(x_1)\sum_{x_2} g(x_2) = E[X_1]E[X_2]$$

Momenter og middelværdi

Definition: Hvis X er en stokastisk variabel kaldes forventningsværdien $m_n = E[X^n]$ – såfremt den eksisterer - for det n'te *moment* for X.

Definition: Det første moment for en stokastisk variabel kaldes variablens *middelværdi* μ.

Dette er blot det sædvanlige gennemsnit.

Varians og standardafvigelse

Definition: Hvis X er en stokastisk variabel kaldes forventningsværdien $\mu_n = E[(X-\mu)^n]$ – såfremt den eksisterer - for det n'te *centraliserede moment* for X.

Det ses umiddelbart fra regnereglerne, at det første centraliserede moment altid er lig med 0.

Definition: Det andet centraliserede moment for en stokastisk variabel kaldes – såfremt det eksisterer - *variansen* V. *Standardafvigelsen* eller *spredningen* defineres ved $\sigma = \sqrt{V}$. Variansen skrives derfor ofte også som σ^2.

Eksempel: Normalfordelingen med $\mu = 0$

Lad os betragte en normalfordeling med $\mu = 0$, der altså har frekvensfunktionen $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma}}$. Da funktionen er lige ser man umiddelbart, at alle ulige momenter er lig 0. Specielt er middelværdien og skævheden begge lig 0. Variansen, altså det andet moment er:

$$V = \mu_2 = \int_{-\infty}^{\infty} f(x)\cdot x^2\,dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma}}\cdot x^2\,dx$$

Fra sektionen om gaussiske integraler ved vi, at integralet giver $\sqrt{2\pi}\sigma^3$. Derfor er $V = \sigma^2$ og standardafvigelsen (heldigvis) lig med σ, så notationen er konsistent.

Fordelingens fjerde moment kan beregnes med en anden formel fra samme sektion:

$$\mu_4 = \int_{-\infty}^{\infty} f(x)\cdot x^4\,dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma}}\cdot x^4\,dx = \frac{3\sqrt{2\pi}\sigma^5}{\sqrt{2\pi}\sigma} = 3\sigma^4$$

Eksempel: Normalfordelingen med $\mu \neq 0$

For en normalfordeling, hvor $\mu \neq 0$ giver de centrale momenter samme værdier som ovenfor. Men hvad med de "almindelige" momenter? I afsnittet om gaussiske integraler har vi fundet en formel for følgende integral:

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\cdot x^n\,dx$$

De rå momenter er givet ved disse divideret med $\sqrt{2\pi}\sigma$. De første fire er:

m_1	m_2	m_3	m_4
μ	$\mu^2 + \sigma^2$	$\mu^3 + 3\mu\sigma^2$	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$

Regneregler for varianser og standardafvigelse

Sætning: For c konstant og X_1 og X_2 uafhængige stokastiske variable med tilhørende frekvensfunktioner f og g gælder:

a. $\sigma^2(X_1) = E[X_1^2] - E[X_1]^2$
b. $\sigma^2(c) = 0$
c. $\sigma^2(cX_1) = c^2\,\sigma^2(X_1)$, så $\sigma(cX_1) = c\,\sigma(X_1)$
d. $\sigma^2(X_1 + X_2) = \sigma^2(X_1) + \sigma^2(X_2)$
e. $\sigma^2(X_1X_2) = E[X_1^2]E[X_2^2] - E[X_1]^2E[X_2]^2$

Bevis: Undervejs bruges regneregler for forventningsværdier, samt tidligere beviste regler på listen

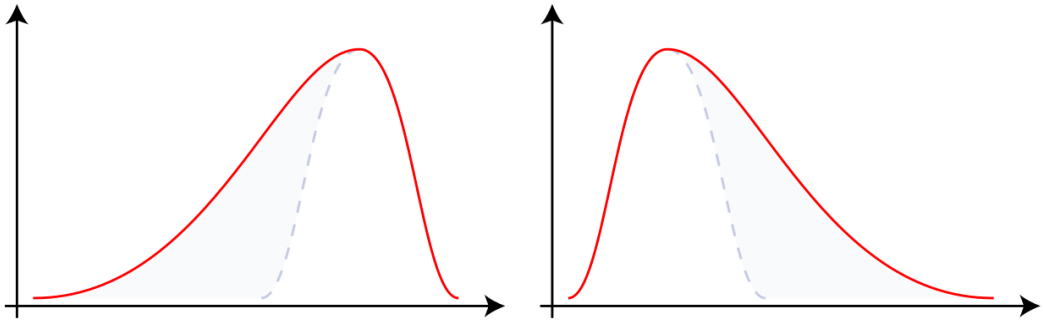
a. $\sigma^2(X_1) = E[(X_1 - \mu)^2] = \sum_x f(x)(x-\mu)^2 = \sum_x f(x)(x^2 - 2x\mu + \mu^2) =$
$\sum_x f(x)\cdot x^2 - 2\mu\sum_x f(x)\cdot x + \mu\sum_x f(x) = E(X_1^2) - 2\mu^2 + \mu\cdot 1 = E(X_1^2) - E(X_1)^2$
b. $\sigma^2(c) = E[c^2] - E[c]^2 = c^2 - c^2 = 0$
c. $\sigma^2(cX_1) = E[(cX_1)^2] - E[(cX_1)]^2 = c^2E[X_1^2] - c^2E[X_1]^2 = c^2(E[X_1^2] - E[X_1]^2) = c^2\sigma^2(X_1)$
d. $\sigma^2(X_1 + X_2) = E[(X_1 + X_2)^2] - E[X_1 + X_2]^2 = E[X_1^2 + 2X_1X_2 + X_2^2] - (E[X_1] + E[X_2])^2 =$
$E[X_1^2] + 2E[X_1X_2] + E[X_2^2] - (E[X_1]^2 + E[X_1]E[X_2] + E[X_2]^2) =$
$E[X_1^2] - E[X_1]^2 + E[X_2^2] - E[X_2]^2 + 2E[X_1X_2] - 2E[X_1]E[X_2] =$
$\sigma^2(X_1) + \sigma^2(X_2) + 2(E[X_1X_2] - E[X_1]E[X_2])$
Da de to stokastiske variable er uafhængige er indholdet af parentesen lig 0.
e. Igen benyttes det at de variable er uafhængige:
$\sigma^2(X_1X_2) = E[(X_1X_2)^2] - E[X_1X_2]^2 = E[X_1^2]E[X_2^2] - E[X_1]^2E[X_2]^2$

Skævhed og kurtosis

Definition: Ved det n'te *standardiserede moment* forstås (såfremt det eksisterer): $\frac{\mu_n}{\sigma^n}$

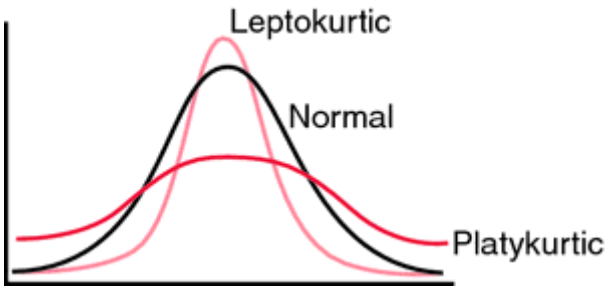
Det ses umiddelbart, at det første centraliserede moment altid er 0, og at det andet centraliserede moment altid er 1 (igen forudsat at disse eksisterer).

Definition: Det tredje standardiserede moment kaldes – såfremt det eksisterer – *skævheden* af fordelingen (på engelsk *skewness*). Fordelinger med positiv (negativ) skævhed er højreskæve (venstreskæve).



Definition: Det fjerde standardiserede moment kaldes – såfremt det eksisterer – for *kurtosis* eller *topstejlhed*. En fordeling med større (mindre) kurtosis end en normalfordeling kaldes *leptokurtisk* (*platykurtisk*). En fordeling med samme kurtosis som normalfordelingen kaldes *mesokurtisk*.

Populært sagt er en leptokurtisk fordeling "mere spids end en normalfordeling", mens en platykurtisk fordeling er "mindre spids end en normalfordeling".



Eksempel: Normalfordelingen

Normalfordelingen har $\mu_3 = 0$, hvorfor skævheden også er 0. Vi så, at $\mu_4 = 3\sigma^4$, så kurtosis for normalfordelingen er 3. Dette motiverer følgende definition:

Definition: For en fordeling kaldes $\frac{\mu_4}{\sigma^4} - 3$ (såfremt det eksisterer) for den *overskydende kurtosis*.

Der gælder nu oplagt:

Sætning: For en fordeling hvor den overskydende kurtosis eksisterer gælder der:

- Er den overskydende kurtosis positiv er fordelingen leptokurtisk.
- Er den overskydende kurtosis nul er fordelingen mesokurtisk.
- Er den overskydende kurtosis negativ er fordelingen platykurtisk.

Moment-genererende funktion (MGF)

Definition: For en stokastisk variabel X, hvor alle momenter $m_n, n \in \mathbb{N}$ eksisterer sættes den *moment-genererende funktion* (MGF) til:

$$M_X(t) = E(e^{tX})$$

Sætning: Den moment-genererende funktion for en stokastisk variabel X opfylder:

$$\left[\frac{d^n}{dt^n}M_X(t)\right]_{t=0} = m_n$$

Bevis: Beviset benytter Taylor-udviklingen for eksponentialfunktionen:

$$e^{tX} = 1 + \frac{1}{1!}tX + \frac{1}{2!}t^2X^2 + \cdots = \sum_{k=1}^{\infty}\frac{1}{k!}t^kX^k$$

$M_X(t)$ er forventningsværdien af denne:

$$M_X(t) = E\left(\sum_{k=1}^{\infty}\frac{1}{k!}t^kX^k\right) = \sum_{k=1}^{\infty}\frac{1}{k!}t^kE(X^k)$$

Når man tager den n'te-afledede af denne, er det kun det resulterende konstantled der giver et bidrag for $t = 0$. Altså leddet det n'te led. Da $\frac{d^n}{dt^n}t^k = n!$ betyder det netop, at $\left[\frac{d^n}{dt^n}M_X(t)\right]_{t=0} = E(X^k) = m_n$

Bevis slut

Sætningen viser også, at hvis man kender momenterne kan man skrive den moment-genererende funktion:

$$M_X(t) = \sum_{k=0}^{\infty}\frac{m_k}{k!}t^n$$

Eksempel: Normalfordelingen med $\mu = 0$

Vi betragter her en normalfordeling med $\mu = 0$. Som vi tidligere har set er $m_n = 0$ for n ulige. Fra afsnittet om gaussiske integraler ved vi at:

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \cdot x^n\, dx = 1 \cdot 3 \cdot \cdots \cdot (n-1)\sqrt{2\pi}\sigma^{n+1}$$

Derfor er $m_n = 1 \cdot 3 \cdot \cdots \cdot (n-1)\sigma^n$ for n lige. Vi kan nu beregne den moment-genererende funktion:

$$M_X(t) = 1 + \frac{1}{2!}\sigma^2t^2 + \frac{1 \cdot 3}{4!}\sigma^4t^4 + \frac{1 \cdot 3 \cdot\cdot 5}{6!}\sigma^6t^6 + \cdots =$$

$$1 + \frac{1}{2}(\sigma^2t^2) + \frac{1}{2 \cdot 4}(\sigma^2t^2)^2 + \frac{1}{2 \cdot 4 \cdot 6}(\sigma^2t^2)^3 + \cdots =$$

$$1 + \frac{\sigma^2t^2}{2} + \frac{1}{2!}\left(\frac{\sigma^2t^2}{2}\right)^2 + \frac{1}{3!}\left(\frac{\sigma^2t^2}{2}\right)^3 + \cdots = e^{\frac{\sigma^2t^2}{2}}$$

Eksempel: Normalfordelingen med $\mu \neq 0$

Nogle gange er det dog nemmere at bruge definitionen direkte. Her på en normalfordeling:

$$M_X(t) = E[e^{tX}] = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot e^{tx}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}+tx}\, dx$$

Substituer nu $s = x - \mu$, så $dt = dx$:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{s^2}{2\sigma^2}+t(s+\mu)}\, ds = \frac{1}{\sqrt{2\pi\sigma^2}} e^{t\mu}\int_{-\infty}^{\infty} e^{-\frac{s^2}{2\sigma^2}+ts}\, ds$$

Eksponenten kan skrives som:

$$-\frac{s^2}{2\sigma^2} + ts = -\frac{1}{2}\left(\frac{s^2}{\sigma^2} - 2ts\right)$$

Vi ser at:

$$\left(\frac{s}{\sigma} - t\sigma\right)^2 = \frac{s^2}{\sigma^2} - 2ts + t^2\sigma^2 \Leftrightarrow \frac{s^2}{\sigma^2} - 2ts = \left(\frac{s}{\sigma} - t\sigma\right)^2 - t^2\sigma^2$$

Så den moment-genererende funktion bliver:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{t\mu}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\left(\frac{s}{\sigma}-t\sigma\right)^2 - t^2\sigma^2\right)}\, ds = \frac{1}{\sqrt{2\pi\sigma^2}} e^{t\mu} e^{t^2\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{s}{\sigma}-t\sigma\right)^2}\, ds$$

Endelig substitueres $u = \frac{s}{\sigma} - t\sigma$, så $\frac{du}{ds} = \frac{1}{\sigma} \Leftrightarrow ds = \sigma\, du$:

$$M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{t\mu} e^{t^2\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\sigma\, du = e^{t\mu} e^{t^2\sigma^2}$$

Forskellen fra tilfældet hvor $\mu = 0$ er altså faktoren $e^{t\mu}$.

Moment-genererende funktion for sum af uafhængige stokastiske variable

Der gælder følgende praktiske sætning:

Sætning: Hvis X og Y er uafhængige stokastiske variable med moment-genererende funktioner $M_X(t)$ og $M_Y(t)$, da er den moment-genererende funktion for summen: $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$

Bevis: Vi regner:

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E[e^{tX} \cdot e^{tY}]$$

Da X og Y er uafhængige er dette lig produktet af de to forventningsværdier:

$$E[e^{tX}] \cdot E[e^{tY}] = M_X(t) \cdot M_Y(t)$$

Bevis slut.

Eksempel: Sum af normalfordelinger

Lad X være normalfordelt med parametrene μ_X og σ_X og Y normalfordelt med parametrene μ_Y og σ_Y. De tilsvarende moment-genererende funktioner er da $M_X(t) = \exp\left[\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right]$ og $M_Y(t) = \exp\left[\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right]$. Ifølge ovenstående sætning er den moment-genererende funktion for summen lig:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = \exp\left[\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right] \cdot \exp\left[\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right] =$$

$$\exp\left[(\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right]$$

Dette er netop den moment-genererende funktion for en normalfordeling med middelværdi $\mu = \mu_X + \mu_Y$ og varians $\sigma^2 = \sigma_X^2 + \sigma_Y^2$. Summen af to normalfordelinger er altså selv en normalfordeling.

Karakteristisk funktion

Den karakteristisk funktion for en fordeling er en ide der er tæt knyttet til den moment-genererende funktion:

Definition: For en stokastisk variabel X sættes *den karakteristiske funktion* til:

$$\varphi_X(t) = E\left(e^{itX}\right)$$

Her er i den imaginære enhed. For kontinuerte stokastiske variable er den karakteristiske funktion det samme som den inverse Fourier-transformation til frekvensfunktionen.

Den karakteristiske funktion har den fordel frem for den moment-genererende ditto, at den førstnævnte altid eksisterer[1], mens dette ikke altid er tilfældet for den sidstnævnte. Men hvis den moment-genererende funktion eksisterer kan man finde den karakteristiske funktion ved hjælp af variabelskiftet $t \to it$.

Eksempel: Normalfordelingen

Da normalfordelingens moment-genererende funktion eksisterer kan den karakteristiske funktion findes at skifte t ud med it:

$$M_X(t) = e^{\frac{\sigma^2 t^2}{2}} e^{-t\mu}, \text{ så } \varphi_X(t) = e^{\frac{\sigma^2 (it)^2}{2}} e^{-it\mu} = e^{-\frac{1}{2}\sigma^2 t^2 - i\mu t}$$

For standardnormalfordelingen gælder specielt:

$$\varphi_X(t) = e^{-\frac{1}{2}t^2}$$

Cauchy-fordelingen: En fordeling uden momenter

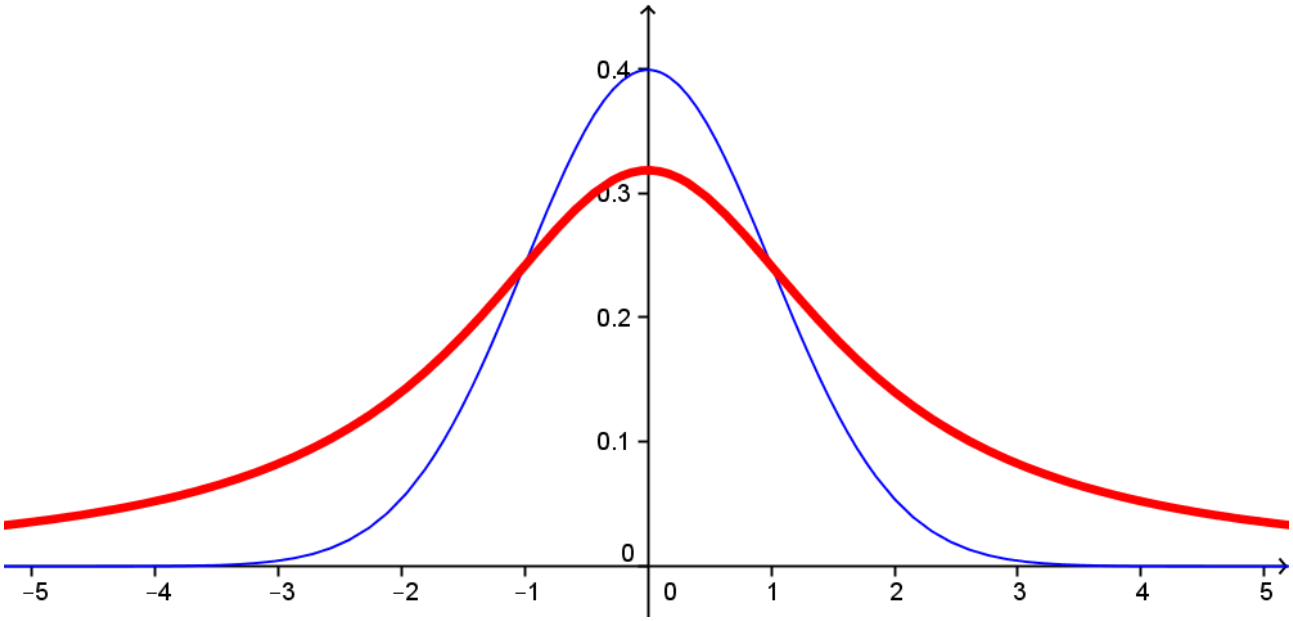
Et eksempel på en fordeling der ikke har en moment-genererende funktion er Cauchy-fordelingen, der har frekvensfunktionen:

---

[1] Dette følger af de sædvanlige resultater for Fourier-transformationer, da frekvensfunktionen (der er ikke-negativ) har et endeligt integral, nemlig 1.

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Grafen er vist med rødt på figuren herunder. Til sammenligning er standardnormalfordelingen vist i blåt:



Ved første øjekast ser det ud som om Cauchy-fordelingen dybest set er en bredere version af normalfordeligen. Men det viser sig, at fordelingen har så "fede haler", at ingen af dens momenter er veldefinerede!

Lad os først forsikre os om, at der rent faktisk er tale om en frekvensfunktion:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}\,dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \frac{1}{\pi}[\mathrm{Arctan}(x)]_{-\infty}^{\infty} = \frac{1}{\pi}\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = 1$$

Værre går det, hvis man forsøger at beregne forventningværdien. Ud fra grafen skulle man tro denne var nul, men:

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx$$

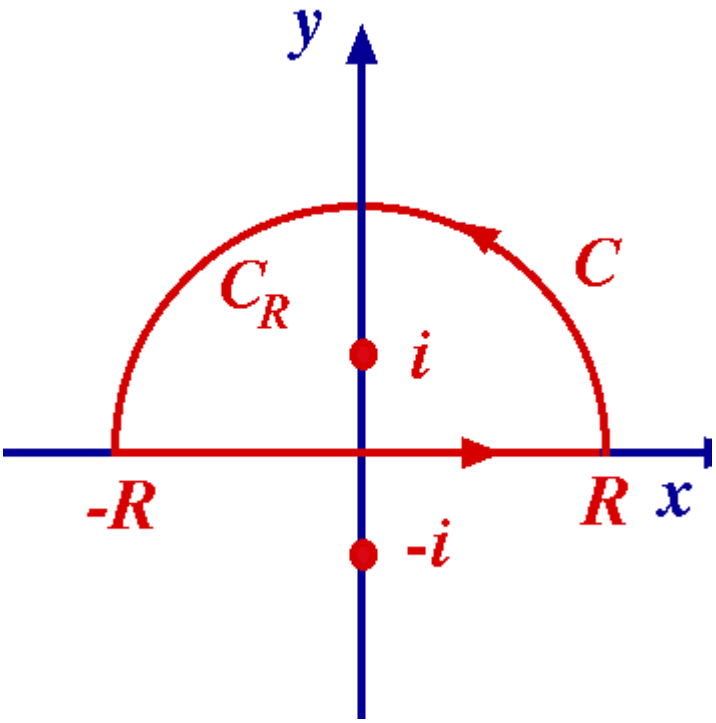
Det viser sig, at det ikke er muligt at tildele dette udtryk nogen veldefineret værdi! (Median og typetal er derimod begge 0). Det samme gør sig gældende for alle andre momenter. Derfor har Cauchy-fordelingen ikke noget moment-genererende funktion.

Derimod har fordelingen en karakteristisk funktion:

$$\varphi_X(t) = \int_{-\infty}^{\infty} \frac{e^{it}}{\pi(1+x^2)}\,dx$$

Denne beregnes nemmest ved at bruge kompleks funktionsteori. Betragt integranden som en kompleks funktion $f(z) = \frac{e^{iz}}{\pi(1+z^2)}$. Denne funktion har poler ved $z = \pm i$. Residuet for hver af disse findes:

$$\mathrm{Res}_{\pm i} = \lim_{z\to\pm i}(z \mp i)\cdot f(z) = \frac{1}{\pi}\lim_{z\to\pm i}(z \mp i)\frac{e^{it}}{(z+i)(z-i)} = \pm\frac{1}{2\pi i}e^{\mp t}$$



For $t > 0$ går $f(z)$ mod nul i den øvre halvdel af den komplekse plan. Derfor vil kurveintegralet af en halvcirkel med radius R går mod nul som R går mod uendelig. Så integralet af hele konturen vist på figuren går mod $\varphi_X(t)$. Men konturintegralet er også givet ved Cauchys residualsætning:

$$\varphi_X(t) = 2\pi i\mathrm{Res}_i = e^{-t}, \qquad t > 0$$

For $t < 0$ går $f(z)$ mod nul i nedre halvplan, så en tilsvarende kontur skal være spejlvendt i x-aksen. Da omløbsretningen dermed bliver negativ kommer der et minus på resultatet:

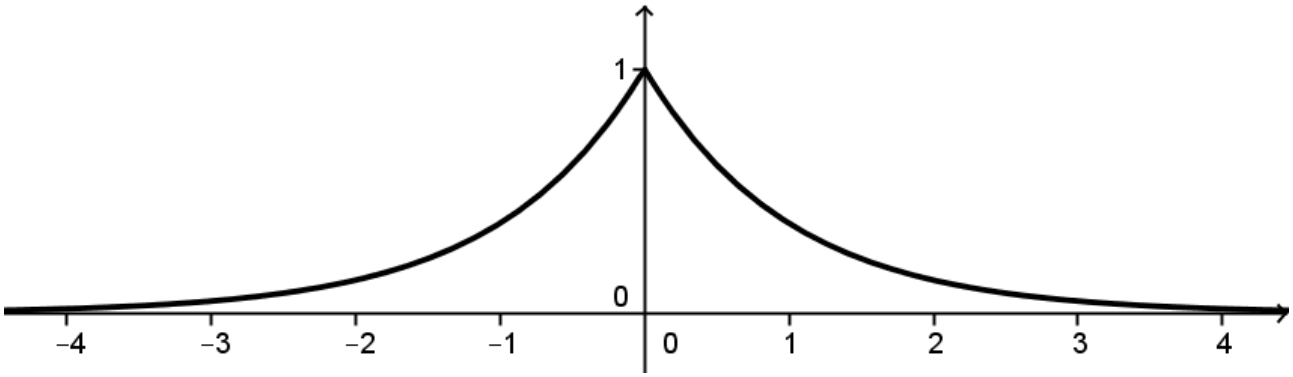
$$\varphi_X(t) = -2\pi i\mathrm{Res}_{-i} = e^{t}, \qquad t < 0$$

I alt bliver den karakteristiske funktion:

$$\varphi_X(t) = e^{-|t|}$$

(For $t = 0$ ved vi jo, at integralet skal give 1).

Figuren viser grafen for den karakteristiske funktion:



Pga. den absolutte værdi i forskriften bliver funktionen ikke-differentiabel for $t = 0$, hvilket er et tegn på, at momenterne ikke eksisterer.

Kumulanter

I stedet for at beskrive en fordeling i termer af dens momenter kan man i stedet bruge de såkaldte *kumulanter*.

Definition: For en stokastisk variabel X med en moment-genererende funktion sættes den kumulant-genererende funktion $g_X(t)$ til logaritmen af denne:

$$g_X(t) = \log E[e^{tX}]$$

Kumulanterne κ_n defineres nu helt analogt til momenterne:

$$\kappa_n = \left[\frac{d^n}{dt^n} g_X(t)\right]_{t=0}$$

Eksempel: Normalfordelingen

Da den moment-genererende funktion for en normalfordelt stokastisk variabel X er $M_X(t) = e^{\frac{\sigma^2 t^2}{2}} e^{-t\mu}$, bliver den kumulant-genererende funktion:

$$g_X(t) = \frac{1}{2}\sigma^2 t^2 + t\mu$$

Dermed bliver kumulanterne:

$$\kappa_1 = \left[\frac{d}{dt}\left(\frac{1}{2}\sigma^2 t^2 + t\mu\right)\right]_{t=0} = [\sigma^2 t + \mu]_{t=0} = \mu$$

$$\kappa_2 = \left[\frac{d}{dt}(\sigma^2 t + \mu)\right]_{t=0} = [\sigma^2]_{t=0} = \sigma^2$$

For $n \geq 3$ er $\kappa_n = 0$. Der er altså sammenfald mellem centrale momenter og kumulanter for for $n = 1$ og $n = 2$. Dette viser sig at gælde generelt:

Sætning: For en stokastisk variabel X for hvilken den moment-genererende funktion eksisterer er g_X er $\kappa_1 = \mu$ og $\kappa_2 = \sigma^2$.

Bevis: Den moment-genererende funktion kan skrives:

$$M_X(t) = 1 + \frac{m_1}{1!}t + \frac{m_2}{2!}t^2 + \cdots = 1 + m_1 t + \frac{m_2}{2}t^2 + \cdots$$

Den kumulant-genererende funktion kan tilsvarende skrives som en potensrække for logaritmer:

$$g_X(t) = \log(1 + (M_X(t) - 1)) = M_X(t) - 1 - \frac{1}{2}(M_X(t) - 1)^2 - \cdots$$

Hvis vi kun beholder op til andenordensled er dette:

$$g_X(t) = m_1 t + \frac{m_2}{2}t^2 - \frac{m_1^2}{2}t^2 + \cdots = m_1 t + \frac{m_2 - m_1^2}{2}t^2 + \cdots$$

Heraf ses, at $\kappa_1 = m_1 = E[X] = \mu$ og $\kappa_2 = m_2 - m_1^2 = E[X^2] - E[X]^2 = \sigma^2$.

Bevis slut.

Sætning: For to uafhængige stokastiske variable X og Y gælder der:

$$g_{X+Y}(t) = g_X(t) + g_Y(t)$$

Bevis: Ifølge definitionen gælder der:

$$g_{X+Y}(t) = \log E\left[e^{t(X+Y)}\right] = \log E[e^{tX}e^{tY}]$$

Da X og Y er uafhængige er forventningsværdien af produktet lig produktet af forventningsværdierne:

$$\log(E[e^{tX}]E[e^{tY}]) = \log(E[e^{tX}]) + \log(E[e^{tY}]) = g_X(t) + g_Y(t)$$

Ikke-uafhængige variable og kovarians

I flere sætninger indtil videre har uafhængighed mellem stokastiske variable været en forudsætning. Men hvad gør man hvis dette ikke er tilfældet? Det viser sig, at graden af afhængighed til dels kan beskrives med størrelsen kovarians.

Definition af kovarians

Definition: For to stokastiske variable X og Y med middelværdier μ_X og μ_Y sættes *kovariansen* – der skrives $\mathrm{Cov}(X,Y)$ - til:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)\cdot(Y-\mu_Y)]$$

Hvis $f_{X,Y}(x,y)$ er den samlede frekvensfunktion for X og Y kan dette skrives som:

$$\mathrm{Cov}(X,Y) = \sum_x \sum_y f(x,y)\cdot(x-\mu_X)\cdot(y-\mu_Y)$$

For kontinuerte variable bliver summene til integraler.

Definition: Hvis $\mathrm{Cov}(X,Y) = 0$ kaldes X og Y *ukorrelerede*.

Sætning: Hvis X og Y er uafhængige er de også ukorrelerede.

Bevis: Hvis X og Y er uafhængige kan den samlede frekvensfunktion faktoriseres:

$$f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y)$$

Her er f_X og $f_Y(y)$ frekvensfunktionerne for hhv. X og Y. Så kan vi regne:

$$\mathrm{Cov}(X,Y) = \sum_x \sum_y f_X(x)\cdot f_Y(y)\cdot(x-\mu_X)\cdot(y-\mu_Y) =$$

$$\sum_x f_X(x)\cdot(x-\mu_X)\sum_y f_Y(y)\cdot(y-\mu_Y) =$$

$$\left(\sum_x f_X(x)\cdot x - \mu_X\sum_x f_X(x)\right)\cdot\left(\sum_y f_Y(y)\cdot y - \mu_Y\sum_y f_Y(y)\right) = (\mu_X-\mu_X\cdot 1\cdot)\cdot(\mu_Y-\mu_Y\cdot 1) = 0$$

Bevis slut.

Bemærk, at det modsatte *ikke* gør sig gældende! Der findes sæt af ukorrelerede stokastiske variable, der *ikke* er uafhængige.

Eksempel: Ukorreleret men ikke uafhængig

Lad X være et tilfældigt valgt tal mellem -1 og 1. Lad $Y = X^2$. De to stokastiske variable er tydeligvis ikke uafhængige, da Y kan bestemmes ud fra X. Da X er ligefordelt er $f_X(x) = ½$. Fordelingsfunktionen for Y findes:

$$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} ½\, dx = [½x]_{-\sqrt{y}}^{\sqrt{y}} = \sqrt{y}$$

Frekvensfunktionen er den afledede af fordelingsfunktionen:

$$f_Y(y) = \frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}$$

Middelværdierne findes:

$$\mu_X = \int_{-1}^{1} ½\, x\, dx = \left[\frac{1}{4}x^2\right]_{-1}^{1} = 0$$

$$\mu_Y = \int_{0}^{1} \frac{1}{2\sqrt{y}} y\, dy = \left[\frac{1}{2}\cdot\frac{2}{3}y\sqrt{y}\right]_{0}^{1} = \frac{1}{3}$$

Kovariansen beregnes

$$\mathrm{Cov}(X,Y) = \int_{-1}^{1}\int_{0}^{1} f(x,y)\cdot(x-0)\cdot\left(y-\frac{1}{3}\right) dy\, dx$$

$f(x,y)$ er kun forskellig fra nul når $y = x^2$. I disse tilfælde er frekvensfunktionen det samme som for X. Så:

$$\mathrm{Cov}(X,Y) = \int_{-1}^{1} ½\cdot x\cdot\left(x^2-\frac{1}{3}\right) dx = \int_{-1}^{1} \frac{1}{2}x^3 - \frac{1}{6}x\, dx = 0$$

Sidste lighedstegn skyldes at integranden er en ulige funktion. X og Y er altså ikke uafhængige, men alligevel ukorrelerede.

Kovarians og varians

Kovarians er forbundet med varians gennem følgende sætning:

Sætning: Lad X være en stokastisk variabel. Da gælder der:

$$\sigma^2(X) = \mathrm{Cov}(X,X)$$

Bevis: Der må gælder at $f_{X,X} = f_X$, så kovariansen af X med sig selv bliver:

$$\mathrm{Cov}(X,X) = E[(X-\mu_X)\cdot(X-\mu_X)] = E[(X-\mu_X)^2] = \sigma^2(X)$$

Bevis slut.

Generalisering af sætninger for uafhængige stokastiske variable

Kovariansen viser sig at blive praktisk når vi ønsker at generalisere de sætninger fra sidste afsnit der forudsatte uafhængighed.

Middelværdi af produkt

I sidste afsnit så vi, at der for uafhængige variable gælder $E[XY] = E[X]E[Y]$. Generelt gælder der:

Sætning: For stokastiske variable X og Y gælder der:

$$E[XY] = E[X]E[Y] + \text{Cov}(X, Y)$$

Bevis: Af definitionen af kovarians følger:

$$\text{Cov}(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)] = E[XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y] =$$

$$E[XY] - \mu_X E[Y] - E[X]\mu_Y + \mu_X\mu_Y = E[XY] - 2E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y] \Leftrightarrow$$

$$E[XY] = E[X]E[Y] + \text{Cov}(X, Y)$$

Bevis slut.

Varians af sum og differens

For uafhængige variable er varians additiv: $\sigma^2(X \pm Y) = \sigma^2(X) \pm \sigma^2(X)$. Generelt gælder der:

Sætning: Hvis X og Y er stokastiske variable, er variansen af sum/differens givet ved:

$$\sigma^2(X \pm Y) = \sigma^2(X) \pm \sigma^2(X) + 2\,\text{Cov}(X, Y)$$

Bevis: Vi regner:

$$\sigma^2(X \pm Y) = E[(X \pm Y)^2] - E[X \pm Y]^2 = E[X^2 \pm 2XY + Y^2] - (E[X] \pm E[Y])^2 =$$

$$E[X^2] \pm 2E[XY] + E[Y^2] - (E[X]^2 \pm 2E[X]E[Y] + E[Y]^2) =$$

$$E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \pm 2E[XY] \mp 2E[X]E[Y] = \sigma^2(X) \pm \sigma^2(X) \pm 2E[XY] \mp 2E[X]E[Y]$$

Brug nu ovenstående sætning:

$$\sigma^2(X \pm Y) = \sigma^2(X) \pm \sigma^2(X) \pm 2\big(E[X]E[Y] + \text{Cov}(X, Y)\big) \mp 2E[X]E[Y] =$$

$$\sigma^2(X) \pm \sigma^2(X) + 2\text{Cov}(X, Y)$$

Bevis slut.

Varians og kovarians for stikprøver

Når man beskæftiger sig med en stikprøve, altså et endeligt datasæt frem for hele fordelingen viser det sig, at man bliver nødt til at definere begreberne en lille smule anderledes.

Vi forestiller os, at vi har n sammenhørende værdier af de stokastiske variable X og Y, altså $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ med tilhørende gennemsnit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ og $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Da sættes *stikprøvekovariansen* $s_{X,Y}$ mellem x'er og y'er til:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$$

Det følger, at *stikprøvevarianserne* s_x^2 og s_y^2 er:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{og} \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

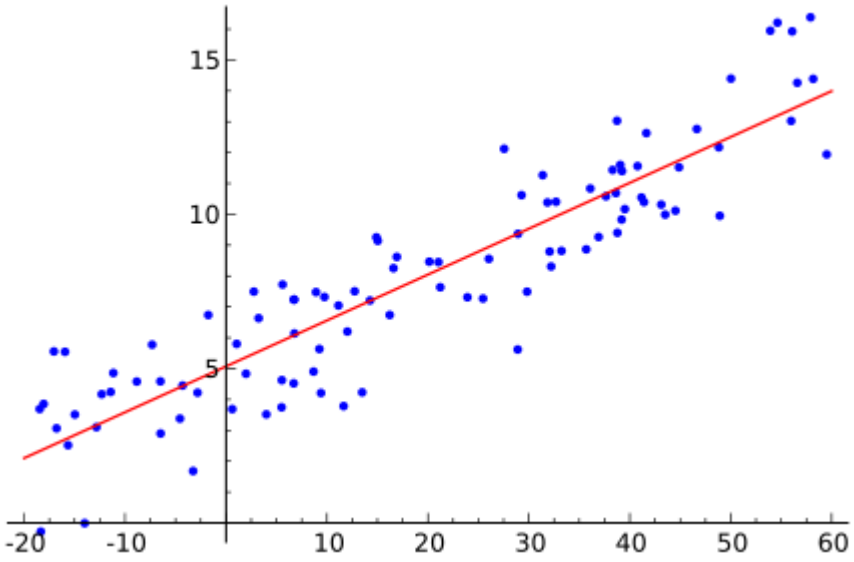
Der flere detaljer om stikprøver og om faktoren $\frac{1}{n-1}$ i afsnittet om stikprøver.

Regression

Hvad gør man, hvis man har et sæt af punkter, og man ønsker at bestemme en matematisk model der beskriver punkterne bedst muligt? Dette spørgsmål beskæftiger regressionsanalysen sig med.

Lineær regression – At finde den bedste rette linje

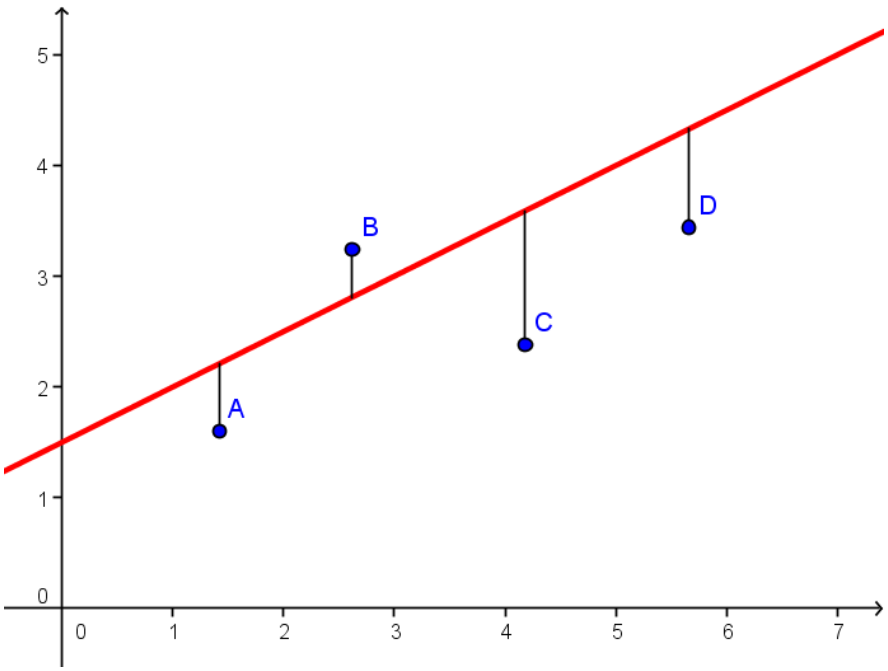
Billedet til højre viser en lang række datapunkter (blå). Selvom det er tydeligt, at de ikke alle kan ligge perfekt på samme rette linje er der stadig en meget tydelig lineær tendens i måden de fordeler sig på.



Den røde linje er et bud på en lineær model der beskriver datasættet godt. Spørgsmålet er, hvordan man finder denne bedste rette linje. Dette kaldes *lineær regression*.

De mindste kvadraters metode

Det mest almindelige princip der benyttes lineær regression er *de mindste kvadraters metode*.



På figuren er vist fire punkter: A, B, C og D. Indtegnet er også et gæt på hvad der kunne være den bedste rette linje.

For at finde den linje der passer bedst, vil vi gerne have, at afstandene til linjen er lille. Så man kunne søge at minimere den samlede afstand til linjen.

Hvis linjen har forskriften $f(x) = ax + b$ og punkterne koordinaterne $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ kan denne samlede afstand skrives:

$$\sum_{i=1}^{n} |ax_i + b - y_i|$$

Dette er imidlertid upraktisk at arbejde med, da vi gerne vil differentiere udtrykke og sætte den afledede lig nul, for at finde en minimumsværdi. I stedet vil vi arbejde med summen af kvadraterne på afstandene:

$$Q(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Her tænker vi på punkternes koordinater som konstanter og a og b som variable. At der arbejdes med kvadrater har, ud over det regnetekniske, den fordel, at store afvigelser "straffes hårdere" end små.

Formler for a og b

Vi ønsker nu at minimere størrelsen Q og differentierer derfor efter hhv. a og b og sætter lig nul:

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{n} 2(ax_i + b - y_i)\, x_i = 0$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} 2(ax_i + b - y_i)\, 1 = 0$$

Divider med $2n$ på hver side og split summerne op:

$$a\frac{1}{n}\sum_{i=1}^{n} x_i^2 + b\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 0$$

$$a\frac{1}{n}\sum_{i=1}^{n} x_i + b - \frac{1}{n}\sum_{i=1}^{n} y_i = 0$$

Indfør nu betegnelserne $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ og $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$:

$$a\overline{x^2} + b\bar{x} - \overline{xy} = 0$$

$$a\bar{x} + b - \bar{y} = 0 \Leftrightarrow b = \bar{y} - a\bar{x}$$

Sidste ligning indsættes i første:

$$a\overline{x^2} + (\bar{y} - a\bar{x})\bar{x} - \overline{xy} = 0 \Leftrightarrow a\left(\overline{x^2} - (\bar{x})^2\right) = \overline{xy} - \bar{x}\cdot\bar{y} \Leftrightarrow a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

Ved at skrive definitionerne ud fås:

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Ovenstående formel kan skrives simplere ved at benytte stikprøvevarians og –kovarians:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i \cdot \bar{x} + \bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) =$$

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

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y} - n \cdot \bar{y} \cdot \bar{x} + n \cdot \bar{x} \cdot \bar{y}\right) =$$

$$\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}n\bar{x} \cdot n\bar{y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)$$

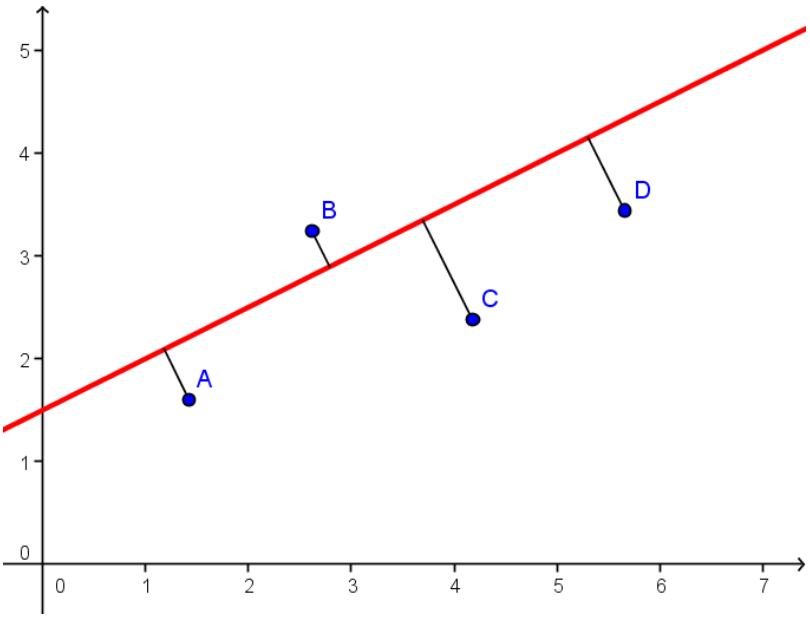
Heraf ses det, at:

$$a = \frac{s_{xy}}{s_x^2}$$

For b gælder der stadig:

$$b = \bar{y} - a\bar{x}$$

Deming-regression



I det ovenstående er det hele tiden antaget, at x-værdierne lå fast, mens det var y-værdierne der havde en afvigelse. Hvad nu, hvis det var tilfældet for begge?

Situationen er illustreret på figuren til venstre i det tilfælde hvor spredningen (variansen) af x'er og y'er er ens. Nu er det i stedet den vinkelrette afstand til linjen der ønskes minimeret (situationen kaldes da også *ortogonal regression*). Eller rettere: summen at kvadraterne på de vinkelrette afstande.

I dette tilfælde bliver forholdet mellem varianserne af x'er og y'er vigtig. Denne størrelse kaldes f:

$$f = \frac{\sigma_y^2}{\sigma_x^2}$$

(For ortogonal regression er $f = 1$). Formlen for a bliver i dette tilfælde:

$$a = \frac{s_y^2 - f s_x^2 + \sqrt{\left(s_y^2 - f s_x^2\right)^2 + 4 f s_{xy}}}{2 s_{xy}}$$

Formlen for b er uændret.

Eksponentiel regression og potensregression

Det er muligt at lave regression til eksponentielle modeller, ved først at tage logaritmerne til y-værdierne og derefter udføre lineær regression. Da gælder der nemlig:

$$\log y = ax + b \Leftrightarrow y = e^{ax+b} = e^b \cdot (e^a)^x$$

Tilsvarende kan man lave potensregression ved at tage logaritmen til både x'er og y'er og derefter udføre lineær regression. Da gælder der nemlig:

$$\log y = a \log x + b \Leftrightarrow y = e^{a \log x + b} = e^b \cdot x^a$$

Pearsons produkt-moment-korrelationskoefficient

Kan man sige noget om hvor godt den lineære model (eller eksponentielle model/potensmodellen) passer? Svaret er (i hvert fald til dels) ja, idet man kan beskrive dette med *korrelationskoefficienten*.

Definition: Korrelationskoefficienten (eller mere præcist Pearsons produk-moment-korrelationskoefficient eller blot PPMCC) mellem x og y i et datasæt $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ er givet ved:

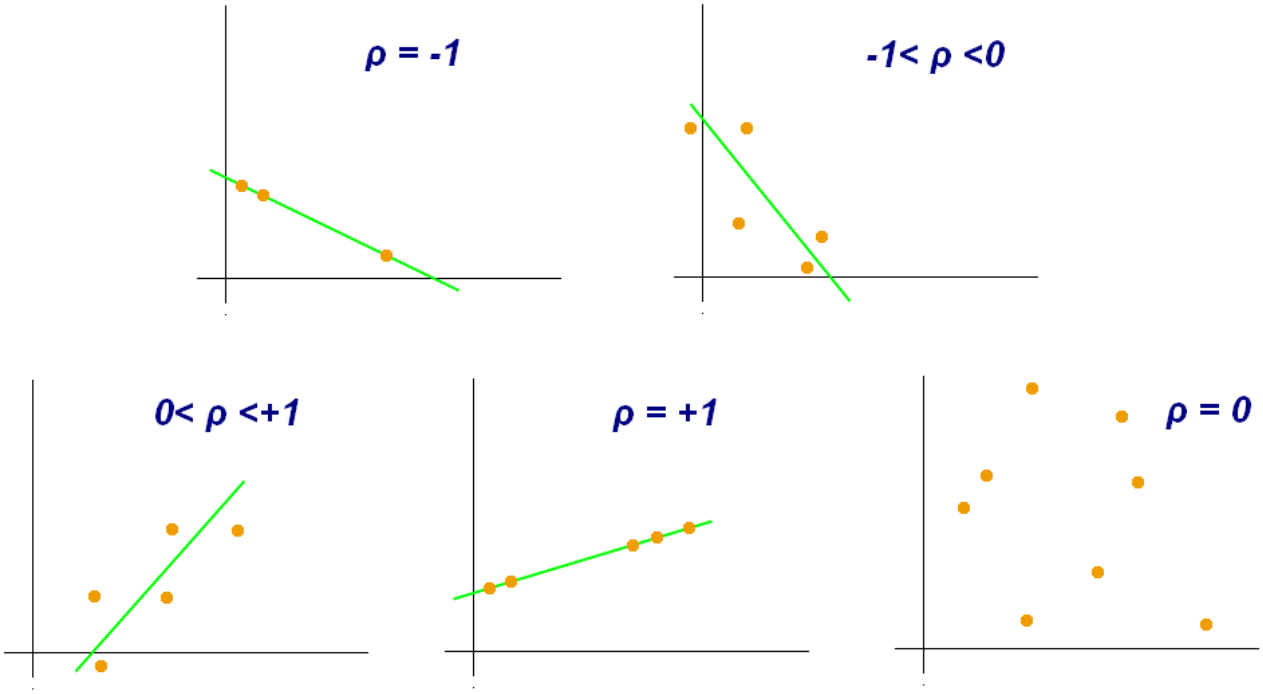
$$\rho_{xy} = \frac{s_{xy}}{s_x^2 s_y^2}$$

Definitionen kan også benyttes mere generelt mellem stokastiske variable:

Definition: Korrelationskoefficienten mellem to stokastiske variable X og Y er givet ved:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

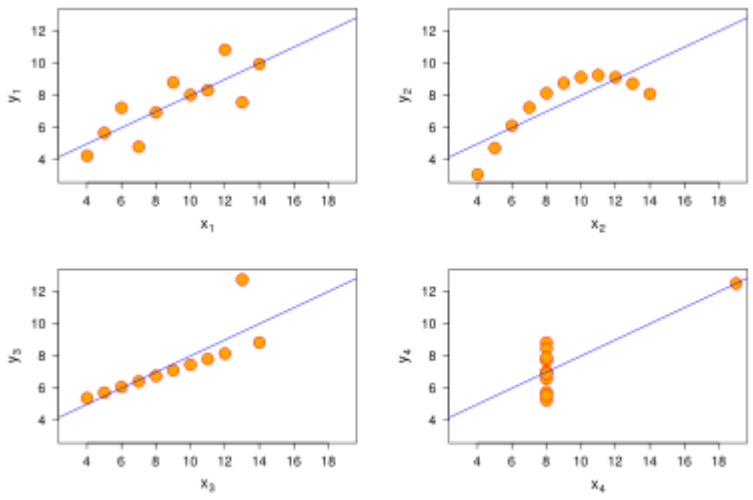
Korrelationskoefficienten antager altid værdier mellem -1 og 1. Fortegnet angiver, om der er om voksende eller aftagende tendens mellem x og y. Jo større numerisk værdi, desto mere korrelerede er de to variable. For en korrelationskoefficient på nul er der ingen korrelation. Figuren viser eksempler på forskellige værdier af ρ_{xy}:



I regressionsanalyse er denne størrelse kendt som r eller R, og kvadratet r^2 eller R^2 som *forklaringsgraden*. Den siger i en vis forstand noget om hvor meget af data der der beskrives med den givne model.

Anscombes kvartet

Ofte sætter man en (noget arbitrær) grænse for, om der er tale om en god model ved værdien $R^2 = 0{,}95$. Men det er farligt kun at kigge på denne værdi, som et berømt eksempel viser.



Anscombes kvartet er de fire datasæt der er vist til højre. De har den særlige egenskab, at stort set alle almindelige deskriptorer en ens for dem! Således også forklaringsgraden $R^2 = 0{,}816$.

Moralen er, at man altid skal kigge på plottet af sit datasæt inden man konkluderer hvorvidt en given model er god eller ej. Dette er en naturlig ting at gøre med de computere vi har til rådighed i dag, men på Anscombes tid fik man ofte blot spyttet en række tal deskriptorer ud.

Kovariansmatrix

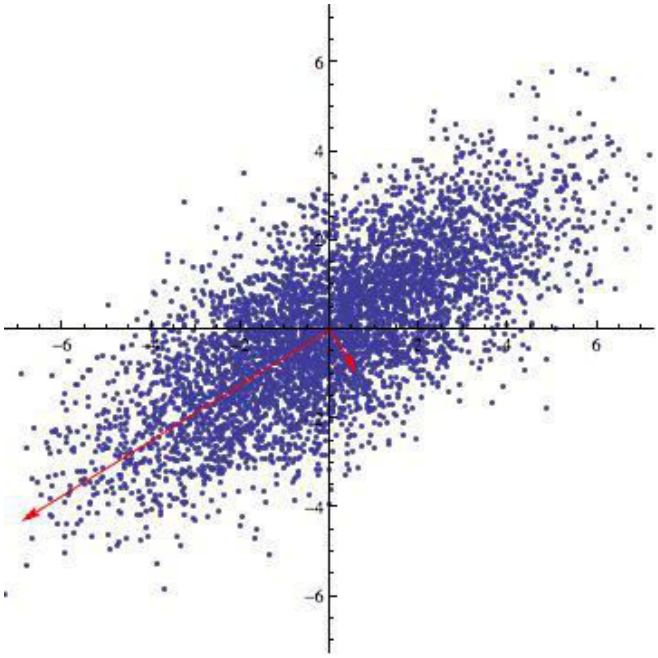
Kovariansmatricen benyttes når man er interesseret i korrelationerne mellem mere end to koordinater/stokastiske variable. Definitionen for stokastiske variable er som følger (den tilsvarende definition for datasæt er forhåbentlig klar):

Definition: Givet stokastiske variable $X_1, X_2, \cdots, X_n$ defineres kovariansmatricen som den $n \times n$-matrix Σ, hvor $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$

Da $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$ er Σ en symmetrisk matrix. Da kovarianserne tillige er reelle er Σ selvadjungeret og dermed diagonaliserbar med en ortogonal basis af egenvektorer. Da Σ er semi-positiv definit kan egenværdierne aldrig være negative.

Visualisering i to dimensioner

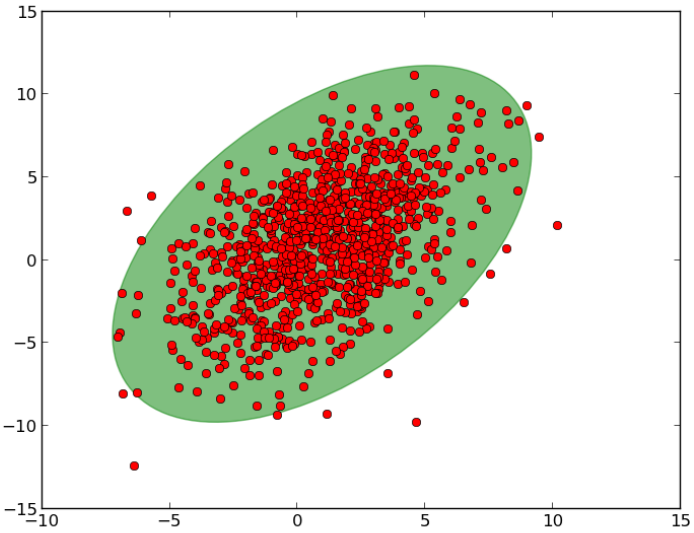
I to dimensioner kan et datasæt visualiseres på samme måde som på figurerne ovenfor. Man kan forestille sig noget tilsvarende i tre (og i princippet flere) dimensioner, men her er det særligt nemt at visualisere situationen.



Figuren viser et datasæt hvor der er en tydelig korrelationstendens. Egenvektorerne er indtegnet på figuren, således at den største egenværdi har størst længde. Det er tydeligt, at den største egenværdi siger mest om tendensen i datasættet.

Man kan tænke på dette, som at egenvektorerne udspænder den ellipse, der beskriver datasættets tendens. I flere dimensioner er dette en (hyper-)ellipsoide[2]. Figuren viser ideen (for et andet, men lignende datasæt):

[2] I tre dimensioner er egenvektorer og -værdier ækvivalente med principalakser og -momenter for inertimoment af et stift legeme.



Kovariansmatrix i to dimensioner

Her bliver kovariansmatricen mellem x'er og y'er simpelthen:

$$\Sigma = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

Egenværdierne findes som rødderne i det karakteristiske polynomium:

$$(\lambda - s_x^2)(\lambda - s_y^2) - (s_{xy})^2 = 0$$

$$\lambda^2 - (s_x^2 + s_y^2)\lambda + s_x^2 s_y^2 - (s_{xy})^2 = 0$$

Diskriminanten for andengradspolynomiet bliver:

$$d = \left(-(s_x^2 + s_y^2)\right)^2 - 4 \cdot 1 \cdot \left(s_x^2 s_y^2 - (s_{xy})^2\right) =$$

$$(s_x^2)^2 + (s_y^2)^2 + 2s_x^2 s_y^2 - 4s_x^2 s_y^2 + 4(s_{xy})^2 =$$

$$(s_x^2)^2 + (s_y^2)^2 - 2s_x^2 s_y^2 + 4(s_{xy})^2 = (s_x^2 - s_y^2)^2 + 4(s_{xy})^2$$

Egenværdierne bliver dermed:

$$\lambda_{\pm} = \frac{s_x^2 + s_y^2 \pm \sqrt{(s_x^2 - s_y^2)^2 + 4(s_{xy})^2}}{2}$$

Her betegner fortegnet hhv. stor og lille egenværdi.En egenvektor $\vec{v}$ til den største egenværdi λ_+ skal opfylde:

$$\Sigma\vec{v} = \lambda_+\vec{v}$$

Under antagelse af, at egenrummet ikke er sammenfaldende med y-aksen kan vi antage $\vec{v}$ kan skrives på formen $\vec{v} = \begin{pmatrix} 1 \\ a \end{pmatrix}$. Nu bliver ovenstående ligning:

$$\begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} \begin{pmatrix} 1 \\ a \end{pmatrix} = \lambda_+ \begin{pmatrix} 1 \\ a \end{pmatrix} \Leftrightarrow$$

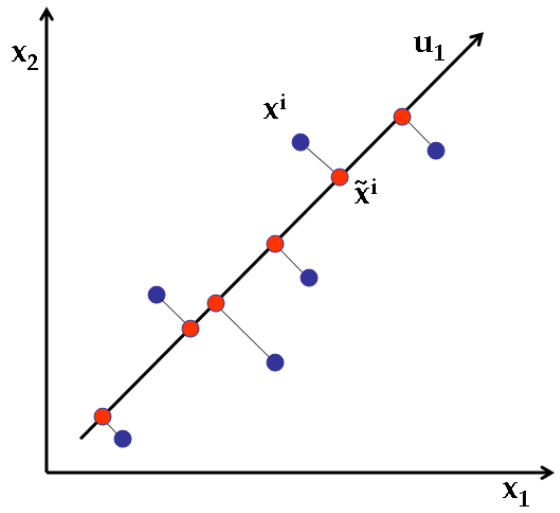
$$s_x^2 + a s_{xy} = \lambda_+ \quad \text{og} \quad s_{xy} + a s_y^2 = a\lambda_+$$

Den første ligning er den nemmeste at løse her:

$$a = \frac{\lambda_+ - s_x^2}{s_{xy}} = \frac{\left(\frac{s_x^2 + s_y^2 + \sqrt{\left(s_x^2 - s_y^2\right)^2 + 4\left(s_{xy}\right)^2}}{2} - s_x^2 \right)}{s_{xy}} = \frac{s_y^2 - s_x^2 + \sqrt{\left(s_y^2 - s_x^2\right)^2 + 4\left(s_{xy}\right)^2}}{2 s_{xy}}$$

Denne egenvektor $(1, a)$ svarer til en linje med hældningkoefficient a, hvilket netop var resultatet for Deming-regression i tilfældet $f = 1$.

Principalkomponentanalyse (PCA) og dimensionsreduktion



Ovenstående viser, at den største variation i datasættet foregår langs den vektor der svarer til den numerisk største egenværdi. Hvis man projicerer data ind på den tilsvarende egenvektor har man reduceret datasættet fra mange dimensioner til en enkelt[3].

Figuren til venstre viser dette, hvor egenvektoren er betegnet med $\overrightarrow{u_1}$. De originale data (blå punkter) projiceres ind på aksen (røde punkter).

Der går naturligvis noget information tabt ved en sådan reduktion, men ofte vil man fange noget essentielt ved datasættet på denne måde.

I stedet for at nøjes med en enkelt egenværdi/vektor kan man vælge at tage de n største egenværdier (numerisk set) og projicere datasættet ind på de underrummet udspændt at de tilsvarende egenvektorer. At finde den rette balance mellem simplicitet (lavt n) og bibeholdelse af relevant information uden overfitting (højt n) er kunstarten i dimensionsreduktion.

[3] Med mindre egenværdien er degenereret.

De store tals lov

Markovs ulighed

Sætning (Markovs ulighed): Lad X være en stokastisk variabel der kun antager ikke-negative værdier. Hvis X har en endelig forventningsværdi μ gælder der for alle $a > 0$:

$$P(X \geq a) \leq \frac{\mu}{a}$$

Bevis: Definer en stokastisk variabel X_a ved at sætte $X_a(x) = 0$ for $x < a$ og a ellers. Det er klart, at der gælder $X_a(x) \leq X(x)$ for alle $x \geq 0$. Tag nu forventningsværdien på hver side af denne ulighed for at få:

$$a \cdot P(X \geq a) \leq \mu \Leftrightarrow P(X \geq a) \leq \frac{\mu}{a}$$

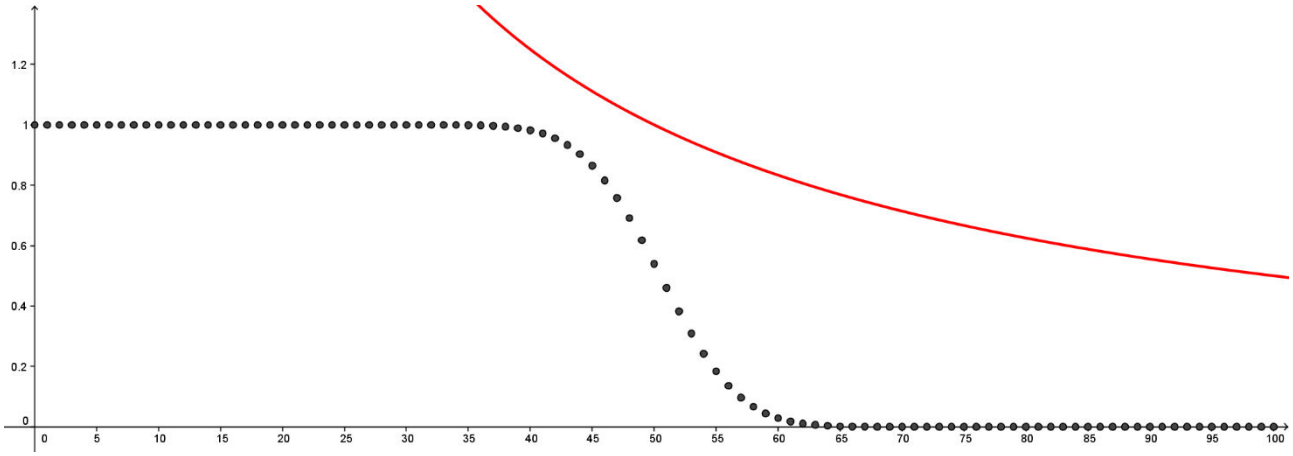
Bevis slut.

Eksempel med møntkast

En ærlig mønt kastes 100 gange. Med Markovs ulighed kan vi sætte en grænse for, hvad sandsynligheden for at slå et vist antal krone eller mere. Sæt X til antallet af kronekast. Middelværdien af X er 50. Sandsynligheden for at få n eller flere kronekast opfylder ifølge Markovs ulighed:

$$P(X \geq n) \leq \frac{50}{n}$$

Man ser straks, at der ikke er nogen ny information når n er mindre end 50. Dette gælder generelt: Markovs ulighed er triviel for a mindre end μ. Grafen viser grænsen uligheden sætter sammenlignet med de faktiske sandsynligheder:



Som man ser af figuren, er grænsen sat af Markovs ulighed ikke specielt prangende, men man skal huske at den gælder generelt.

Chebyshevs ulighed

Markovs ulighed kan bruges til at bevise følgende ulighed:

Sætning (Chebyshevs ulighed): Lad X være en stokastisk variabel med endelig middelværdi μ og endelig standardafvigelse σ. Da gælder:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Bevis: Brug Markovs ulighed på den stokastiske variabel $Y = (X-\mu)^2$. Dette giver:

$$P((X-\mu)^2 \geq a) \leq \frac{E[(X-\mu)^2]}{a} = \frac{\sigma^2}{a}$$

Da $(X-\mu)^2 = |X-\mu|^2$ er $P((X-\mu)^2 \geq a) = P(|X-\mu|^2 \geq a) = P\left(|x-\mu| \geq \sqrt{a}\right)$, så:

$$P\left(|x-\mu| \geq \sqrt{a}\right) \leq \frac{\sigma^2}{a}$$

Sæt nu $a = k^2\sigma^2$ for at få:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Bevis slut.

k	Min % within k standard deviations of mean	Max % beyond k standard deviations from mean
1	0%	100%
$\sqrt{2}$	50%	50%
1.5	55.56%	44.44%
2	75%	25%
3	88.8889%	11.1111%
4	93.75%	6.25%
5	96%	4%
6	97.2222%	2.7778%
7	97.9592%	2.0408%
8	98.4375%	1.5625%
9	98.7654%	1.2346%
10	99%	1%

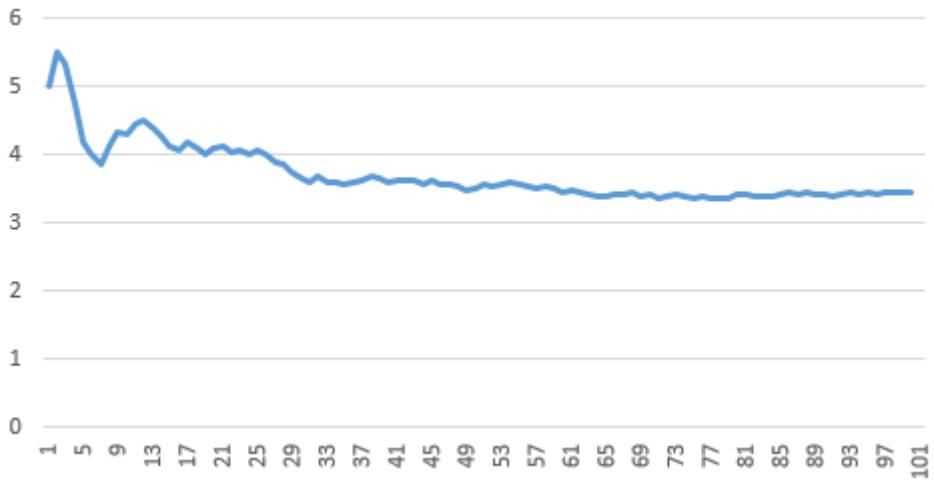
Uligheden sætter grænser for, hvor meget af vægten af en sandsynlighedsfordelingfordeling der kan ligge inden for et vist antal standardafvigelser. For $k \leq 1$ er uligheden triviel, men for $k > 1$ giver uligheden os information, som vist i tabellen til venstre.

Disse grænser er relativt dårlige, da de gælder for generelle fordelinger. Sammenlign f.eks. med den empiriske regel for normalfordelinger, hvor ca. 68% ligger inden for 1 standardafvigelse, ca. 95% ligger inden for 2 standardafvigelser, og ca. $99{,}7\%$ ligger inden for 3 standardafvigelser. Disse ligger et pænt stykke over tabellens hhv. $0\%, 75\%$ og $88{,}9\%$.

Så generelt får man selvfølgelig bedre grænser hvis man ved noget mere om den pågældende fordeling.

De store tals lov

De store tals lov udtaler sig om, hvordan gennemsnittet af en stor række gentagelser af samme eksperiment opfører sig. Hvis vi f.eks. kaster en ærlig terning gentagende gange forventer vi selvfølgelig ikke at slå det gennemsnitlige antal øjne hver gange, specielt ikke, da gennemsnittet (altså forventningsværdien) er lig $3{,}5$. Men hvis vi tager gennemsnittet over alle kast indtil videre, forventer vi umiddelbart, at vi kommer tæt på de $3{,}5$. Grafen viser resultatet af et sådant eksperiment. Her ser vi netop denne opførsel som antallet af samlede slag bliver stort.



De store tals lov udtaler sig netop om denne type situation. Den findes i to versioner, en "svag" og en "stærk". Man skelner, fordi der er flere måder at tænke på konvergens af en følge af fordelinger.

Konvergens af stokastiske variable

Vi får i det kommende brug for at skelne mellem forskellige typer af konvergens for følger af stokastiske variable. Vi starter her med den svageste betingelse og går mod de stærkere.

Definition: En følge af stokastiske variable $X_1, X_2, \cdots$ siges at *konvergere i fordeling* mod den stokastiske variabel X, hvis der for alle x gælder:

$$\lim_{n\to\infty} F_n(x) = F(X)$$

Her er F_n fordelingsfunktionen for X_n og F fordelingsfunktionen for X. Sagt med andre ord skal fordelingsfunktionen konvergere punktvist.

Definition: En følge af stokastiske variable $X_1, X_2, \cdots$ siges at *konvergere i sandsynlighed* mod den stokastiske variabel X, hvis der for alle $\varepsilon > 0$ gælder:

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 0$$

Definition: En følge af stokastiske variable $X_1, X_2, \cdots$ siges at *konvergere næsten sikkert* eller *næsten overalt* mod den stokastiske variabel X, hvis der gælder:

$$P\left(\forall x\colon \lim_{n\to\infty} F_n(x) = F(x)\right) = 1$$

Som sagt er disse tre definitioner progressivt stærkere udsagn. Der gælder altså:

Sætning: $X_1, X_2, \cdots$ konvergerer næsten sikkert mod $X \Longrightarrow X_1, X_2, \cdots$ konvergerer i sandsynlighed mod $X \Longrightarrow X_1, X_2, \cdots$ konvergerer i fordeling mod X.

Den svage version af de store tals lov

Sætning (de store tals lov, svag): For en række identisk fordelte, uafhængige stokastiske variable $X_1, X_2, \cdots, X_n$ med endelig middelværdi μ og endelig standardafvigelse $\sigma > 0$ gælder der, at gennemsnittet $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergerer i sandsynlighed mod μ.

Bevis: Det er klart, at forventningværdien for $\bar{X}$ er lig μ. Variansen kan nemt beregnes, da X_i'erne alle er uafhængige:

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

Lad $\varepsilon > 0$ være givet. Ifølge Chebyshevs ulighed gælder der:

$$P\left(|\bar{X} - \mu| \geq k\frac{\sigma}{n}\right) \leq \frac{1}{k^2}$$

Sæt nu $k = \frac{n\varepsilon}{\sigma}$. Da fås:

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n^2\varepsilon^2}$$

Når n går mod uendelig går dette mod nul.

Bevis slut.

Den stærke version af de store tals lov

Sætning (de store tals lov, stærk): For en række identisk fordelte, uafhængige stokastiske variable $X_1, X_2, \cdots, X_n$ med endelig middelværdi μ og endelig standardafvigelse σ gælder der, at gennemsnittet $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergerer næsten sikkert mod μ.

Beviset er indviklet og springes over.

Tolkning af sandsynligheder: frekventisme vs. bayesianisme

De store tals lov danner basis for *frekventisttolkningen* af sandsynligheder. Her forstås sandsynligheden for en hændelse netop som andelen af gangene hændelsen indtræffer hvis man gentager det relevante eksperiment et stort antal gange. Sandsynligheden er grænseværdien af denne andel når antallet af gentagelser går mod uendeligt.

Dette er umiddelbart en intuitivt klar tolkning, men der er nogle problemer forbundet med den. En ting er, at et eksperiment kan være svært at gentage i praksis, men som tankeeksperiment er det ikke et problem. Di løber dog ind i problemer i det tilfælde, hvor sandsynligheden snarere er et udtryk for vores usikkerhed

om parameteren. Hvis man rent faktisk kunne udføre eksperimentet mange gange ville man få samme resultat hver gang. Her er en *Bayesisk* tilgang til sandsynlighederne mere frugtbar. Bayesisk statistik bygger på Bayes sætning, som vi så tidligere. Den grundlæggende ide er, at vi har en forudgående vurdering af, hvis en given sandsynlighed er. Efterhånden som vi indsamler flere informationer ændrer vi vores værdi af hvad sandsynligheden er. Den præcise sammenhæng mellem sandsynligheden før (*prior*) og efter (*posterior*) er netop givet ved Bayes sætning. Mere om bayesisk statistik senere.

Bernoulli- og binomialfordelingerne

Bernoulli-fordelingen

Et Bernoulli-eksperiment har to udfald, ofte kaldet succes og fiasko. Sandsynligheden for succes kaldes p. Sandsynligheden for fiasko er således $1 - p$.

Definition: En stokastisk variabel X der antager værdien 1 ved succes i et Bernoulli-eksperiment og værdien 0 ved fiasko siges at være *Bernoulli-fordelt*. Der gælder altså $f(1) = p$ og $f(0) = 1 - p$

Middelværdi, momenter og standardafvigelse for Bernoulli-fordelingen

Ved at bruge resultaterne fra sidste afsnit ser vi:

$$\mu = E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$$

Generelt er det n'te moment:

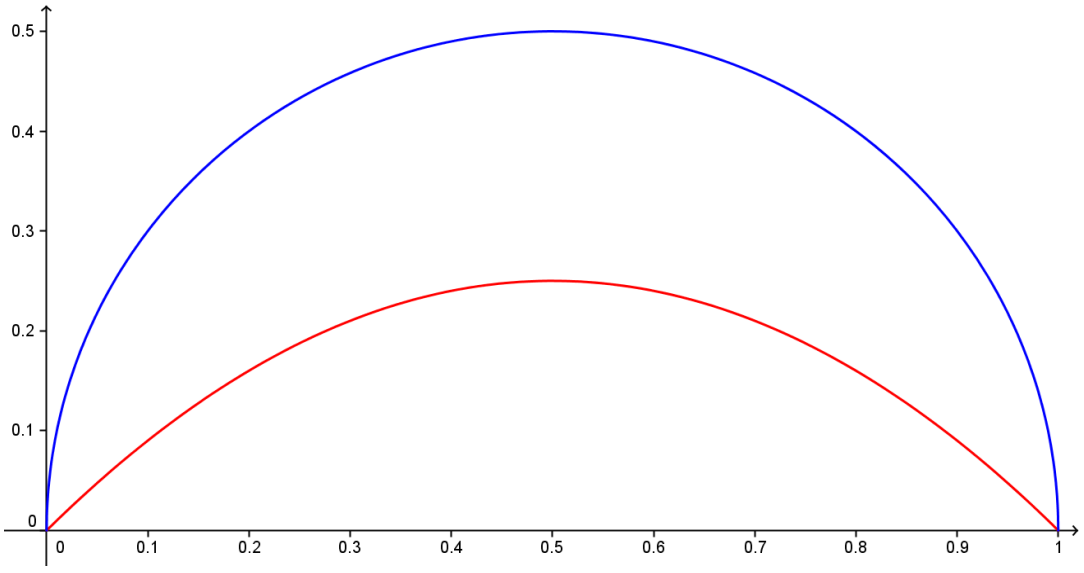
$$m_n = E[X^n] = \ p \cdot 1^n + (1 - p) \cdot 0^n = p$$

Således specielt $m_2 = E[X^2] = p$. Variansen er:

$$\sigma^2(X) = E[X^2] - \mu^2 = p - p^2 = p(1 - p)$$

Og dermed:

$$\sigma(X) = \sqrt{p - p^2}$$



Figuren viser hvordan hhv. variansen og standardafvigelsen varierer som funktion af p. Begge har maksimum ved $p = ½$.

Centraliserede momenter, skævhed og kurtosis for Bernoulli-fordelingen

De centraliserede momenter er:

$$\mu_n = E[(X-\mu)^n] = p\cdot(1-p)^n + (1-p)\cdot(0-p)^n = p\cdot(1-p)^n + (1-p)\cdot(-p)^n$$

Det tredje centraliserede moment findes ved brug af binomialsætningen:

$$\mu_3 = p\cdot(1-p)^3 - (1-p)\cdot p^3 = p\cdot(1-3p+3p^2-p^3) - p^3 + p^4 =$$

$$p - 3p^2 + 3p^3 - p^4 - p^3 + p^4 = 2p^3 - 3p^2 + p$$

Dermed bliver skævheden, det tredje standardiserede moment:

$$\frac{\mu_3}{\sigma^3} = \frac{2p^3 - 3p^2 + p}{\big(p(1-p)\big)^{3/2}}$$

Det fjerde centraliserede moment findes tilsvarende:

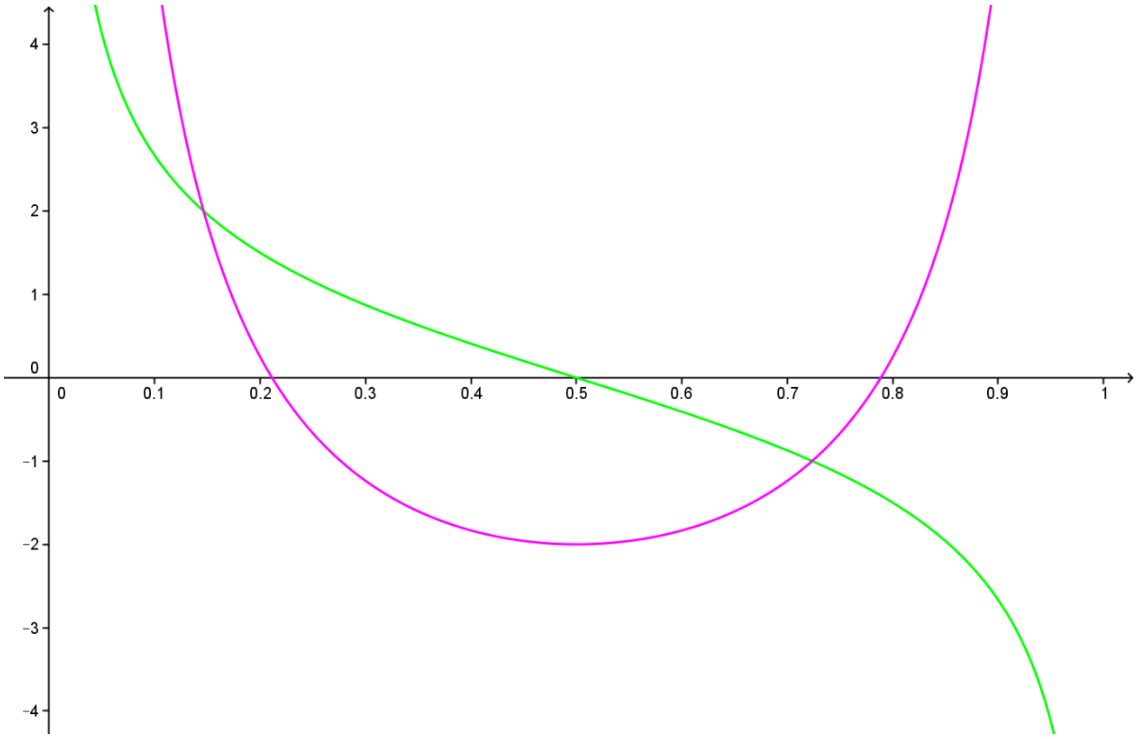
$$\mu_4 = p\cdot(1-p)^4 + (1-p)\cdot p^4 = p\cdot(1-4p+6p^2-4p^3+p^4) + p^4 - p^5 =$$

$$p - 4p^2 + 6p^3 - 4p^4 + p^5 + p^4 - p^5 = -3p^4 + 6p^3 - 4p^2 + p$$

Dermed bliver kurtosis, det fjerde standardiserede moment:

$$\frac{\mu_4}{\sigma^4} = \frac{-3p^4 + 6p^3 - 4p^2 + p}{p^2(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} - 3$$

Ved at trække 3 fra dette opnår man den overskydende kurtosis: $\frac{1}{p} + \frac{1}{1-p} - 6$



Figuren viser skævheden og overskydende kurtosis som funktion af p. For $p = ½$ har fordelingen mimimal overskydende kurtosis -2 (der er det mindst mulige for fordelinger generelt) og er altså platykurtisk. For $p = \frac{3\pm\sqrt{3}}{6} \approx \begin{cases} 0{,}79 \\ 0{,}21 \end{cases}$ er fordelingen mesokurtisk, og for p tæt på 0 og 1 leptokurtisk. Den overskydende kurtosis går faktisk mod uendelig her.

Moment-genererende funktion for Bernoulli-fordelingen

Den moment-genererende funktion bliver:

$$M_X(t) = E[e^{tX}] = (1-p)\cdot e^{0\cdot t} + p\cdot e^{1\cdot t} = 1 - p + p\cdot e^t$$

Dermed bliver den karakteristiske funktion:

$$\varphi_X(t) = 1 - p + p\cdot e^{it}$$

Og den kumulant-genererende funktion:

$$g_X(t) = \log\big(M_X(t)\big) = \log(1 - p + p\cdot e^t)$$

Binomialfordelingen

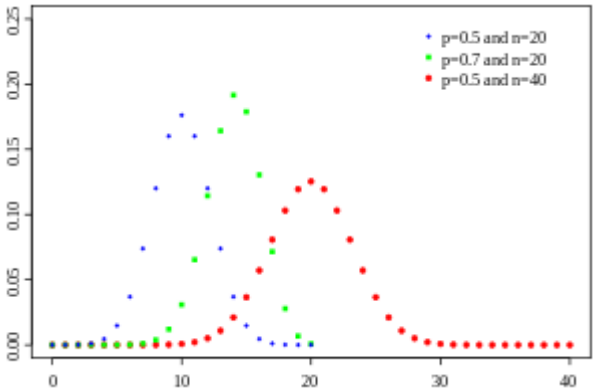
Definition: Der udføres n uafhængige Bernoulli-eksperimenter med samme p. Den stokastiske variabel X der angiver antallet af succeser siges at være *binomialfordelt*.

Sætning: Binomialfordelingen med n Bernoulli-eksperimenter med parameter p har frekvensfunktionen:

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Bevis: For at få præcis k succeser ud af n mulige skal man først udvælge k ud af n muligheder. Dette kan gøres på $\binom{n}{k}$ måder. Hver af disse har en sandsynlighed på $p^k(1-p)^{n-k}$, da der jo skal være k succeser og dermed $n-k$ fiaskoer, og eksperimenterne er uafhængige. Dermed er formlen bevist.

Bevis slut



Figuren til venstre viser binomialfordelingen for $n = 20$ og forskellige værdier af p. Man bemærker, at fordelingerne har den karakteristiske klokkeform vi kender fra normalfordelingen.

Faktisk viser det sig, at binomialfordelingen nærmer sig en normalfordeling med samme middelværdi og standardafvigelse når n bliver meget stor.

For at vise dette har vi først brug for at finde μ og σ:

Sætning: Binomialfordelingen har middelværdien $\mu = np$ og varians $\sigma^2 = np(1-p)$.

Bevis: Hvis de uafhængige stokastiske variable $X_1, X_2, \cdots, X_n$ er Bernoulli-fordelte med parameter p, da er $X = X_1 + X_2 + \cdots + X_n$ netop binomialfordelt. Ved at bruge sætningerne om additivitet af hhv. middelværdi og varians får vi:

$$\mu_X = \mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n} = p + p + \cdots + p = np$$

$$\sigma_X^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2 = p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)$$

Bevis slut

De Moivre-Laplaces sætning

Nu er vi klar til at vise det, der kaldes de Moivre-Laplaces sætning.

Sætning (de Moivre-Laplace): For store værdier af n gælder der, at binomialfordelingen med middelværdi μ og standardafvigelse σ er tilnærmelsesvis lig med normalfordelingen med samme parametre.

Bevis: Sætningen bygger på Stirlings formel, der siger at for store n gælder:

$$n! \approx \sqrt{2\pi n} \cdot n^n \cdot e^{-n}$$

Så for store n, k og $n-k$ kan binomialkoefficienten approksimeres[4]:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k!)} \approx \frac{\sqrt{2\pi n} \cdot n^n \cdot e^{-n}}{\sqrt{2\pi k} \cdot k^k \cdot e^{-k} \cdot \sqrt{2\pi(n-k)} \cdot (n-k)^{n-k} \cdot e^{-(n-k)}} =$$

$$\frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}} \cdot e^{-n+k+(n-k)} = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}}$$

Binomialfordelingen kan altså tilnærmes til:

$$\frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}} \cdot p^k (1-p)^{n-k} = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot n^n \cdot \left(\frac{p}{k}\right)^k \left(\frac{1-p}{n-k}\right)^{n-k}$$

Da $n^n = n^k \cdot n^{n-k}$ kan dette yderligere simplificeres til:

$$\frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \left(\frac{k}{np}\right)^{-k} \left(\frac{n-k}{n(1-p)}\right)^{-(n-k)}$$

Her er det sidste to brøker "vendt på hovedet" ved at vende fortegnet på eksponenten.

Vi ønsker nu at lave k om til hvad der svarer til x i en standardnormalfordeling med samme middelværdi og standardafvigelse som binomialfordelingen. Med andre ord skal der gælde:

[4] k og $n-k$ kan ikke antages at være store i halerne for fordelingen, hvorfor konvergensen her vil være langsommmere.

$$x = \frac{k-\mu}{\sigma} \Leftrightarrow \sigma x = k - \mu \Leftrightarrow k = \mu + \sigma x = np + \sqrt{np(1-p)}x$$

Derfor er

$$\frac{k}{np} = \frac{np + \sqrt{np(1-p)}x}{np} = 1 + x\frac{\sqrt{1-p}}{\sqrt{np}}$$

Og tilsvarende:

$$\frac{n-k}{n(1-p)} = \frac{n - np - \sqrt{np(1-p)}x}{n(1-p)} = 1 - x\frac{\sqrt{p}}{\sqrt{n(1-p)}}$$

Så udtrykket kan skrives:

$$\frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{n}{k(n-k)}} \cdot \left(1 + x\frac{\sqrt{1-p}}{\sqrt{np}}\right)^{-k} \left(1 - x\frac{\sqrt{p}}{\sqrt{n(1-p)}}\right)^{-(n-k)}$$

Udtrykket i kvadratroden omskrives:

$$\frac{n}{k(n-k)} = \frac{\frac{1}{n}}{\frac{k}{n} \cdot \frac{n-k}{n}} = \frac{\frac{1}{n}}{\frac{k}{n}\left(1 - \frac{k}{n}\right)}$$

Da $p \approx \frac{k}{n}$ for store n er dette omtrent[5]:

$$\frac{1}{np(1-p)} = \frac{1}{\sigma^2}$$

Så i alt:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left(1 + x\frac{\sqrt{1-p}}{\sqrt{np}}\right)^{-k} \left(1 - x\frac{\sqrt{p}}{\sqrt{n(1-p)}}\right)^{-(n-k)}$$

For at undersøge de to sidste faktorers opførsel for store n tages den naturlige logaritme:

$$\log\left[\left(1 + x\frac{\sqrt{1-p}}{\sqrt{np}}\right)^{-k} \left(1 - x\frac{\sqrt{p}}{\sqrt{n(1-p)}}\right)^{-(n-k)}\right] =$$

$$-k \cdot \log\left(1 + x\frac{\sqrt{1-p}}{\sqrt{np}}\right) - (n-k) \cdot \log\left(1 - x\frac{\sqrt{p}}{\sqrt{n(1-p)}}\right)$$

Taylorudviklingen til anden orden giver $\log(1+x) \approx x - \frac{x^2}{2}$, så dette er approksimativt lig:

[5] Igen kun så længe vi er tæt på middelværdien. I halerne af fordelingen er dette mere tvivlsomt.

$$-k\left(x\frac{\sqrt{1-p}}{\sqrt{np}}-x^2\frac{1-p}{2np}\right)-(n-k)\left(-x\frac{\sqrt{p}}{\sqrt{n(1-p)}}-x^2\frac{p}{2n(1-p)}\right)$$

Brug nu at $k=np+\sqrt{np(1-p)}x$. Første led:

$$-\left(np+\sqrt{np(1-p)}x\right)\left(x\frac{\sqrt{1-p}}{\sqrt{np}}-x^2\frac{1-p}{2np}\right)=$$

$$-x\sqrt{np(1-p)}+x^2\frac{1-p}{2}-x^2(1-p)+\cdots=-x\sqrt{np(1-p)}-x^2\frac{1-p}{2}+\cdots$$

Her er kun led op til anden orden medtaget. Andet led:

$$-\left(n-np-\sqrt{np(1-p)}x\right)\left(-x\frac{\sqrt{p}}{\sqrt{n(1-p)}}-x^2\frac{p}{2n(1-p)}\right)=$$

$$-\left(n(1-p)-\sqrt{np(1-p)}x\right)\left(-x\frac{\sqrt{p}}{\sqrt{n(1-p)}}-x^2\frac{p}{2n(1-p)}\right)=$$

$$x\sqrt{np(1-x)}+x^2\frac{p}{2}-x^2p+\cdots=x\sqrt{np(1-x)}-x^2\frac{p}{2}+\cdots$$

I alt til anden orden:

$$-x\sqrt{np(1-p)}-x^2\frac{1-p}{2}+x\sqrt{np(1-x)}-x^2\frac{p}{2}=-\frac{x^2}{2}$$

Da dette var logaritmen kan selve udtrykket skrives $e^{-\frac{x^2}{2}}$. I alt kan binomialfordelingen approksimeres til:

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2}}$$

Dette er netop standardnormalfordelingen.

Bevis slut.

Det viser sig, at de Moivre-Laplaces sætning kun er et specialtilfælde af den mere generelle *centrale grænseværdisætning*, der dybest set siger, at alle fordelinger vil nærme sig en normalfordeling under lignende omstændigheder. Ligesom fodnoterne i dette bevis antyder at konvergensen af normalfordelingens haler her vil foregå langsommere vil dette også være tilfældet mere generelt.

(Lidt mere) styr på halerne

Problemet med konvergens af halerne kan man tænke på på følgende måde: normalfordelingen har en uendelig lang hale, mens binomialfordelingen i sagens natur kun har en endelig. I grænsen $n\to\infty$ stemmer disse overens, men i praksis skal der "være plads til halerne" for at approksimationen er god:

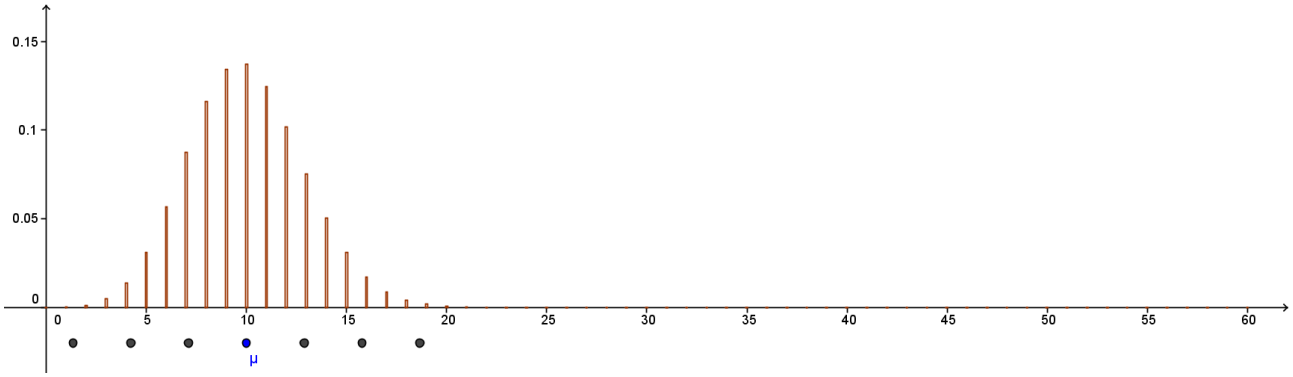
Når man bruger normalfordelingen som tilnærmelse for binomialfordelingen er den samlede fordeling altid en lille smule mindre end 1. Vi kan minimere denne fejl, hvis vi går ud fra der er plads til "hele" fordelingen, hvilket vi i praksis har set betyder inden for ca. 3 standardafvigelser.

Eksempel: Plads til halerne

Vi kaster en ærlig terning $n = 60$ gange og tæller antallet af seksere vi får. Altså er $p = \frac{1}{6}$.

Forventningsværdien er $\mu = np = 60 \cdot \frac{1}{6} = 10$ og standardafvigelsen $\sigma = \sqrt{np(1-p)} = \sqrt{60 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 2{,}89$. Fordelingen af sandsynligheder er vist på figuren nedenfor.

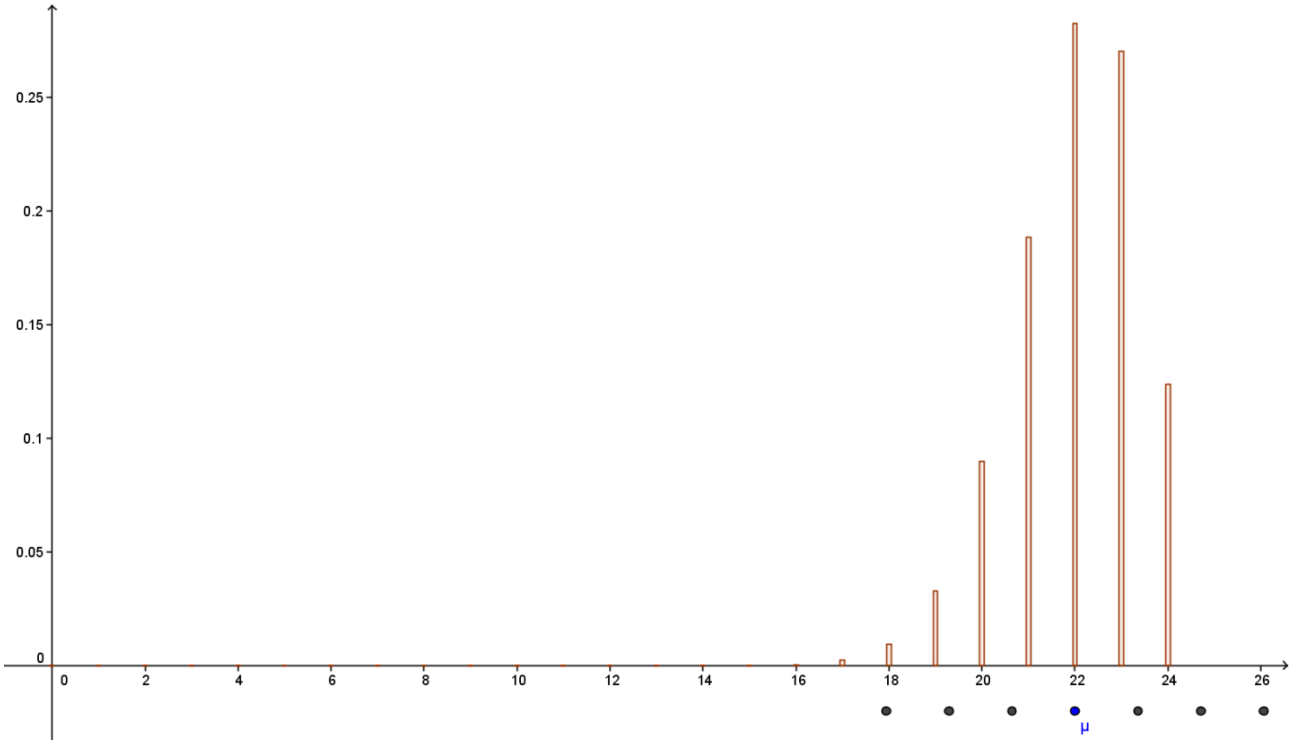
Her angiver den blå prik positionen af middelværdien, og de grå prikker hele standardafvigelser herfra. Det ses, at begge haler ligger fint inden for udfaldene mellem 0 og 60 seksere.



I dette tilfælde vil tilnærmelsen altså passe fint.

Eksempel: Ikke plads til halerne

To ærlige terninger kastes $n = 24$ gange, og antallet af slag med samlet øjental på 4 eller over tælles. Med andre ord er $p = \frac{33}{36} = \frac{11}{12}$. Forventningsværdien er derfor $\mu = np = 24 \cdot \frac{11}{12} = 22$ og standardafvigelsen er $\sigma = \sqrt{np(1-p)} = \sqrt{24 \cdot \frac{11}{12} \cdot \frac{1}{12}} \approx 1{,}35$. Fordelingen af sandsynligheder er vist på figuren:



Her ser man, at der ikke er plads til højre side af halen, da kun ca. 1½ halvanden standardafvigelse ligger inden for udfaldsområdet mellem 0 og 24. Her vil tilnærmelsen altså ikke være god.

Kriterier for n og p

At der er plads til både højre og venstre hale i udfaldsrummet kan udtrykkes i to uligheder. Hvis der skal være plads til venstre hale, skal middelværdien være større end omtrent 3 standardafvigelser:

$$\mu = np > 3\sigma = 3\sqrt{np(1-p)}$$

Her er det underforstået, at ulighedstegnet er omtrentligt. Dette kan omskrives til:

$$\sqrt{np} > 3\sqrt{1-p}$$

Da $1 - p < 1$ betyder dette:

$$\sqrt{np} > 3$$

Hvis man kvadrerer får man $np > 9$, hvilket oftest rundes op til:

$$np \geq 10$$

Hvis der tilsvarende skal være plads til højre hale skal der gælde:

$$\mu = np < n - 3\sigma = n - 3\sqrt{np(1-p)}$$

Dette kan omskrives til:

$$n(1-p) > 3\sqrt{np(1-p)}$$

$$\sqrt{n(1-p)} > 3\sqrt{p}$$

Da $p < 1$ betyder det $n(1-p) > 9$, hvilket også oftest rundes op:

$$n(1-p) \geq 10$$

Man kan indvende, at vi "både går med livrem og seler" ved at runde op to steder. Nogle steder sættes grænsen således til 5 i stedet for 10.

Kontinuetetskorrektion

Hvordan kan man starte med en diskret fordeling (binomial) og pludselig have en kontinuert fordeling (normal)? Forklaringen er, at normalfordelingen i virkeligheden angiver sandsynlighedstætheden i et lille interval omkring et punkt, altså ca. $f(x) \cdot \Delta x$, hvor Δx er intervallets størrelse. I approksimationen i de Moivre-Laplaces sætning er $\Delta x = 1$, da der jo altid er netop 1 mellem forskellige værdier af k.

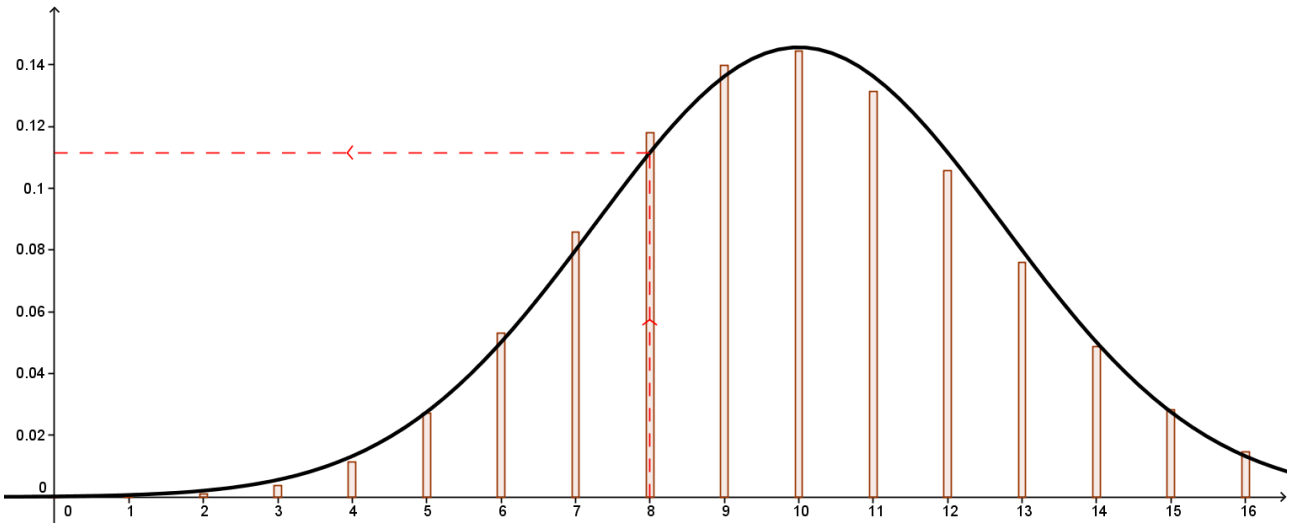
Eksempel: Møntkast

En skæv mønt kastes $n = 40$ gange. Sandsynligheden for at slå kroner er $p = \frac{1}{4}$. Hvad er sandsynligheden for at få præcis 8 gange krone? Den korrekte sandsynlighed kan beregnes til:

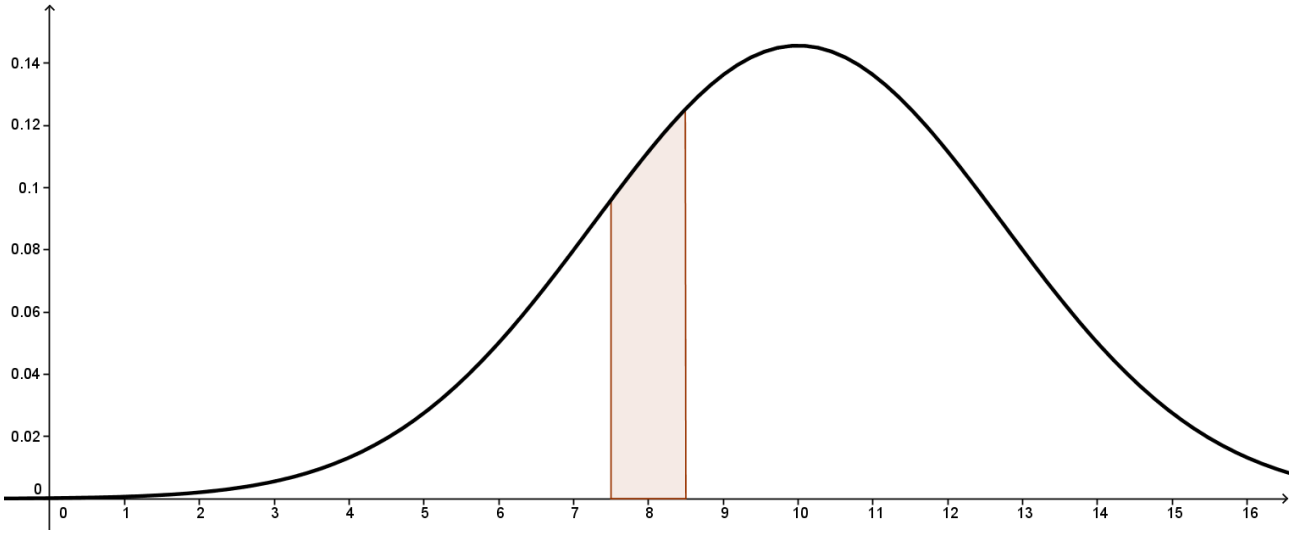
$$\binom{40}{8}\left(\frac{1}{4}\right)^8\left(\frac{3}{4}\right)^{32} \approx 0{,}1179$$

Altså ca. 11,8%.

Figuren nedenfor viser både binomialfordelingen og den tilsvarende normalfordelingen, der altså har middelværdi $\mu = np = 40 \cdot \frac{1}{4} = 10$ og standardafvigelse $\sigma = \sqrt{np(1-p)} = \sqrt{40 \cdot \frac{1}{4} \cdot \frac{3}{4}} \approx 2{,}74$. Da $np = 10$ og $n(1-p) = 30$ burde normalapproksimationen være ok.



Alligevel giver approksimationen 0,1116 her. Måske ikke milevidt fra det korrekte svar, men alligevel en vis afvigelse. Afvigelsen skyldes til dels, at vi i stedet bør integrere over alle observationer "i nærheden af 8", hvilket i praksis betyder mellem 7,5 og 8,5:



Dette areal er $0{,}1113$, altså noget tættere på det korrekte svar.

Poisson- og eksponentialfordelingerne

Beskrivelse af forsøget

I løbet af et bestemt tidsrum forestiller vi os, at der i gennemsnit sker λ hændelser. Bortset fra dette sker hændelserne tilfældigt og uafhængige af hinanden. Den stokastiske variabel der angiver antallet af hændelser kalder vi X.

Vi forstiller os nu, at vi deler tidsrummet op i n lige store tidsrum, altså hver med længden $\frac{1}{n}$. Hvis n er stor nok, er vi praktisk taget sikre på, der højst sker én hændelse i tidsrummet. Sandsynligheden for dette må være $p = \frac{\lambda}{n}$. Hvert tidsrum kan nu ses som en Bernoulliproces.

Fra binomial til Poisson

I løbet af hele tidsrummet sker der n Bernoulliprocesser med parameter p. Altså må det samlede antal hændelser være binomialfordelt. Sandsynligheden for præcis k hændelser er således:

$$\binom{k}{n} p^k (1-p)^{n-k} = \frac{n!}{k!\,(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

Vi er interesseret i hvad der sker, når n går mod uendelig. Lad os derfor starte med at sætte faktorer der ikke afhænger af n udenfor:

$$\frac{1}{k!}\,\lambda^k \cdot \lim_{n\to\infty} \frac{n!}{(n-k)!} \cdot \frac{1}{n^k} \cdot \left(1-\frac{\lambda}{n}\right)^n \cdot \left(1-\frac{\lambda}{n}\right)^{-k}$$

Lad os kigge på grænsen i tre dele. Først:

$$\frac{n!}{(n-k)!} \cdot \frac{1}{n^k} = \frac{1 \cdot 2 \cdot \cdots \cdot n}{1 \cdot 2 \cdot \cdots \cdot (n-k) \cdot \underbrace{n \cdot n \cdot \cdots \cdot n}_{k\ n\text{'er}}} = \frac{1}{1} \cdot \frac{2}{2} \cdot \ldots \cdot \frac{n-k}{n-k} \cdot \underbrace{\frac{n-k+1}{n} \cdot \frac{n-k+2}{n} \cdot \ldots \cdot \frac{n}{n}}_{k \text{ faktorer}}$$

De første $n-k$ faktorer er alle 1. De sidste k faktorer går mod 1 når $n \to \infty$, så hele udtrykket går mod 1.

Anden del er $\left(1-\frac{\lambda}{n}\right)^n$. Det er velkendt, at når $n \to \infty$ går dette mod $e^{-\lambda}$.

Tredje del er $\left(1-\frac{\lambda}{n}\right)^{-k}$. Når $n \to \infty$ går indmaden af parentesen mod 1. k er et fast tal, så hele parentesen går dermed også mod 1.

Alt i alt går hele udtrykket altså mod:

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Dette er frekvensfunktionen for *Poissonfordelingen* med parameter λ.

Fra binomial til geometrisk og eksponentiel

Lad os sige, at der netop er sket en hændelse. Nu kan vi spørge, hvor mange af de små tidsrum af længden $\frac{1}{n}$ der går inden den næste hændelse. Sandsynligheden for, at der ikke går nogen tidsrum, altså at næste hændelse optræder umiddelbart efter, er $p = \frac{\lambda}{n}$. Sandsynligheden for at vente et tidsrum må være $(1-p)p$, for det betyder jo netop, at vi skal vente 1 tidsrum før der indtræffer en hændelse.

Tilsvarende må sandsynligheden for at vente k tidsrum være:

$$(1-p)^k p$$

Den stokastiske variabel $X = k$ siges at være *geometrisk fordelt* med parameter p. Der indsættes nu, at $p = \frac{\lambda}{n}$:

$$\left(1-\frac{\lambda}{n}\right)^k \frac{\lambda}{n} = \left(1-\frac{\lambda}{n}\right)^k \cdot \frac{1}{1-\frac{\lambda}{n}} = \left(1-\frac{\lambda}{n}\right)^k \cdot \frac{n}{n-\lambda}$$

Den samlede ventetid er givet ved $t = k \cdot \frac{1}{n} \Leftrightarrow k = nt$, altså har vi:

$$\left(1-\frac{\lambda}{n}\right)^{nt} \cdot \frac{n}{n-\lambda}$$

Vi er nu interesseret i grænsen hvor n bliver uendelig stor. $\frac{n}{n-\lambda}$ går mod 1, mens første faktor kan omskrives til:

$$\left[\left(1-\frac{\lambda}{n}\right)^n\right]^t$$

I grænsen $n \to \infty$ går indmaden som bekendt mod $e^{-\lambda}$, hvorfor vi ender med følgende frekvensfunktion:

$$f(t) = \lambda e^{-\lambda t}$$

Dette er frekvensfunktionen for *eksponentielfordelingen* med parameter λ.

Den centrale grænseværdisætning

Dette er et af da vigtigste resultater i sandsynlighedsregningen. Sætningen tillader blandt andet at udtale sig på baggrund af en stikprøve selv om man ikke ved noget om hvilken fordeling den underliggende population følger.

Sætningen gælder under meget generelle betingelser. Men vi skal her først og fremmest kigge på den simpleste version.

Normalfordelingen – hvorfor er den så almindelig?

Rigtig mange datasæt fra virkeligheden viser sig at være normalfordelt. Hvorfor er netop denne fordeling så almindelig? de Moivre-Laplaces sætning giver os et hint: der så vi nemlig, at binomialfordelingen bliver til normalfordeling for store værdier af n. Dette viser sig at være et specialtilfælde at den sætning afsnittet handler om: den centrale grænseværdisætning.

I sidste afsnit så vi, at Poisson-fordelingen ligeledes ligner normalfordelingen for store værdier af parameteren λ. Igen er det ikke et tilfælde. Det viser sig nemlig, at uanset hvilket fordeling vi starter med vil normalfordelingen altid dukke op i sidste ende, hvis der er mange tilfældige bidrag.

Dette er essensen af den centrale grænseværdisætning: gennemsnittet af et stort antal observable er altid normalfordelt uanset hvilken underliggende fordeling de følger.

Konvergens for funktioner

Sætningen udtaler sig om en grænseværdi af en funktion, her frekvensfunktionen. Generelt er dette et område hvor man skal være lidt varsom. Der findes to forskellige typer konvergens for funktioner: punktvis og uniform konvergens.

Punktvis konvergens

En følge af funktioner $f_1, f_2, f_3, \cdots$ siges at *konvergere punktvist* mod f hvis der for alle x gælder

$$f_i(x) \to f(x) \text{ for } i \to \infty$$

Eller i kvantorform med ε'er:

$$\forall x \in \mathbb{R}\ \forall \varepsilon > 0\ \exists n > 0\ \forall N > n: |f_N(x) - f(x)| < \varepsilon$$

Ved punktvis konvergens konvergerer hver x-værdi så at sige i "sit eget tempo": ε vælges efter x.

Uniform konvergens

En følge af funktioner $f_1, f_2, f_3, \cdots$ siges at *konvergere uniformt* mod f hvis der gælder:

$$\forall \varepsilon > 0\ \forall x \in \mathbb{R}\ \exists n > 0\ \forall N > n: |f_N(x) - f(x)| < \varepsilon$$

Her vælges x før ε. Derfor konvergerer alle x så at sige "i samme tempo".

Uniform konvergens er altså et stærkere kriterie end punktvis konvergens. Som vi bemærkede i beviset for deMoivre-Laplaces sætning var konvergensen af halerne i fordelingen ikke så hurtig som konvergensen af midten. Med andre ord var der tilsyneladende "kun" tale om punktvis konvergens. Dette kommer til at gælde generelt for den centrale grænseværdisætning: den kan "kun" garantere punktvis konvergens.

Lévys kontinuitetssætning

Denne sætning (hvis bevis overspringes) udtaler sig om konvergens af frekvensfunktioner i termer af karakteristiske funktioner for stokastiske variable.

Sætning (Lévys kontinuitetssætning): Lad $X_1, X_2, X_3, \cdots$ være en følge af stokastiske variable med tilhørende karakteristiske funktioner $\varphi_{X_1}, \varphi_{X_2}, \varphi_{X_3}, \cdots$, altså $\varphi_{X_n} = E[e^{itX_n}]$. Lad ligeledes X være en stokastisk variabel med karakteristisk funktion $\varphi = E[e^{itX}]$. Hvis de karakteristiske funktioner konvergererpunktvist mod φ, da konvergerer frekvensfunktionen for de stokastiske variable i fordeling mod frekvensfunktionen for X.

Den centrale grænseværdisætning

Sætning (den centrale grænseværdisætning): Lad $X_1, X_2, \cdots, X_n$ være uafhængige, ens fordelte stokastiske variable med samme endelige middelværdi μ og endelige standardafvigelse $\sigma > 0$. Summen af de variable er selv en stokastisk variabel: $S_n = X_1 + X_2 + \cdots + X_n$. Når $n \to \infty$ gælder nu:

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \to N(0,1)$$

Her er der tale om konvergens i fordeling.

Bevis: Definer en række nye stokastiske variable ved:

$$Y_i = \frac{X_i - \mu}{\sigma}$$

Da gælder der $E[Y_i] = 0$ og $E[Y_i^2] = 1$. Derfor er den karakteristiske funktion for Y_i til anden orden:

$$\varphi_{Y_i}(t) = 1 - \frac{t^2}{2} + \cdots$$

Definer nu følgende stokastiske variabel:

$$U_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Vi kan nu bestemme den karakteristiske funktion for U_n:

$$\varphi_{U_n}(t) = E[e^{itU_n}] = E\left[\exp\left(\frac{it}{\sqrt{n}} \sum_{i=1}^{n} Y_i\right)\right]$$

Da alle Y_i'erne er uafhængige er dette lig ed:

$$\prod_{i=1}^{n} E\left[\exp\left(\frac{it}{\sqrt{n}}Y_i\right)\right] = \prod_{i=1}^{n} \varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right)$$

Da Y_i'erne er identisk fordelt har de alle samme karakteristiske funktion

$$\varphi_{Y_i}(t) = 1 - \frac{t^2}{2} + \cdots$$

Så vi har:

$$\varphi_{U_n}(t) = \left(1 - \frac{t^2}{2n} + \cdots\right)^n$$

Når n går mod uendelig går dette punktvist mod $e^{-\frac{t^2}{2}}$. Dette er netop den karakteristiske funktion for standardnormalfordelingen. Ifølge Lévys kontinuitetssætning konvergerer frekvensfunktionerne derfor i fordeling (punktvist) mod standardnormalfordelingen.

Bevis slut.

Af sætningen følger umiddelbart, at gennemsnittet af n uafhængige, identisk fordelte stokastiske variable tilnærmelsesvis følger en normalfordeling med middelværdi μ og standardafvigelse $\frac{\sqrt{n\sigma^2}}{n} = \frac{\sigma}{\sqrt{n}}$ når n er stor.

Konvergens mod normalfordelingen

Så længe de stokastiske variable er uafhængige og følger samme fordeling med endelig middelværdi og endelig standardafvigelse større end nul, vil summen altså altid konvergere mod en normalfordeling. Spørgsmålet er hvor hurtigt? Hvis denne konvergens går for langsomt er den praktiske værdi af den centrale grænseværdisætning lav.

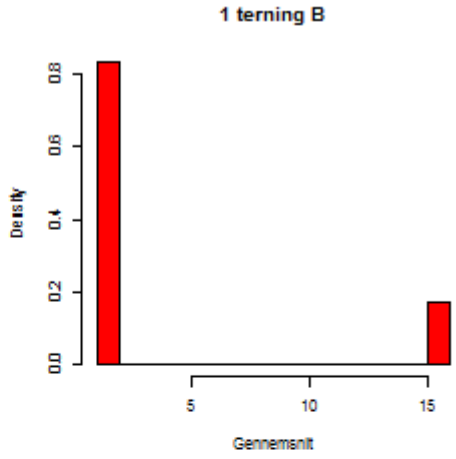
Lad os se på et praktiske eksempel på forskellen i konvergenshastighed.

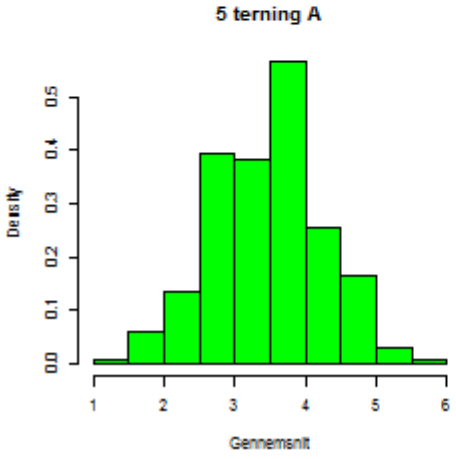
Eksperiment med terninger

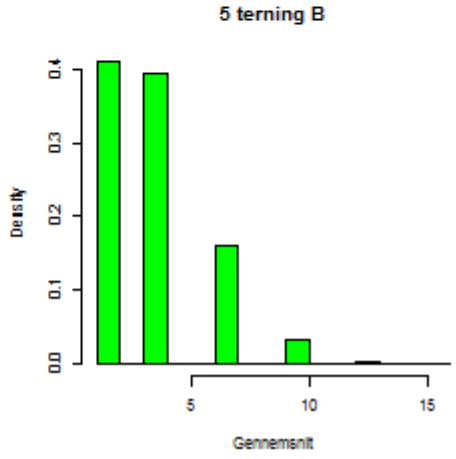
Terning A er en helt almindelig terning, mens terning B har tallet 1 på fem af siderne og tallet 16 på den sjette side. For begge terninger gælder, at forventningsværdien af resultatet er 3,5, men terning B har tydeligvis en større skævhed i resultaterne.

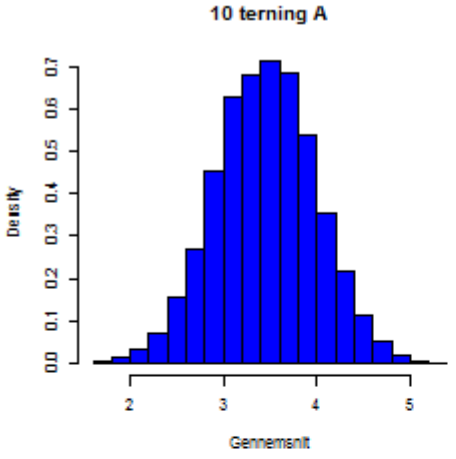
Figuren nedenfor viser fordelingen af gennemsnittet af hhv. 1, 5, 10 og 100 kast med hver terning, hver simuleret 10000 gange.

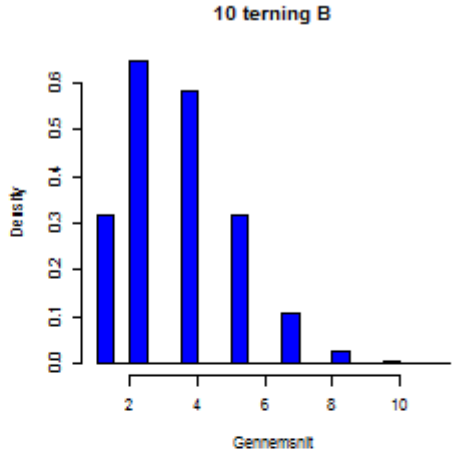
For begge terninger er konvergensen mod normalfordeling tydelig, men det står også klart, at dette går noget langsommere for terning B.

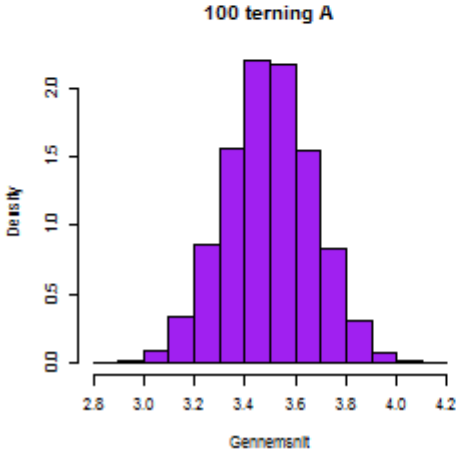


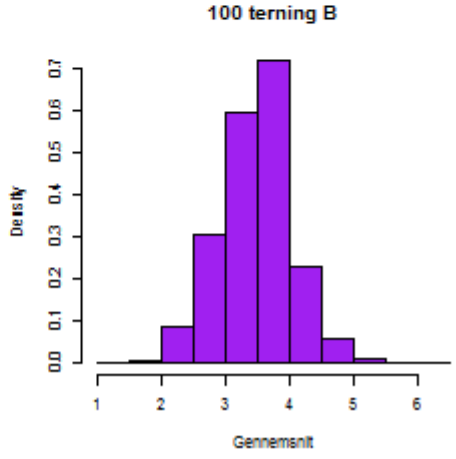












Vi har to terninger.

Set i lyset heraf, er følgende sætning om konvergensen af gennemsnittet intuitivt rimelig:

Sætning (Berry-Esseen): Lad $X_1, X_2, \cdots, X_n$ være uafhængige, ens fordelte stokastiske variable med samme middelværdi $\mu = 0$, endelige standardafvigelse $\sigma > 0$ og $\rho = E[|X_1|^3] < \infty$. Da findes der et $C > 0$, så der for fordelingsfunktionen F_n for den stokastiske variabel $\frac{1}{n}(X_1 + \cdots + X_n) \cdot \frac{\sqrt{n}}{\sigma}$ gælder:

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

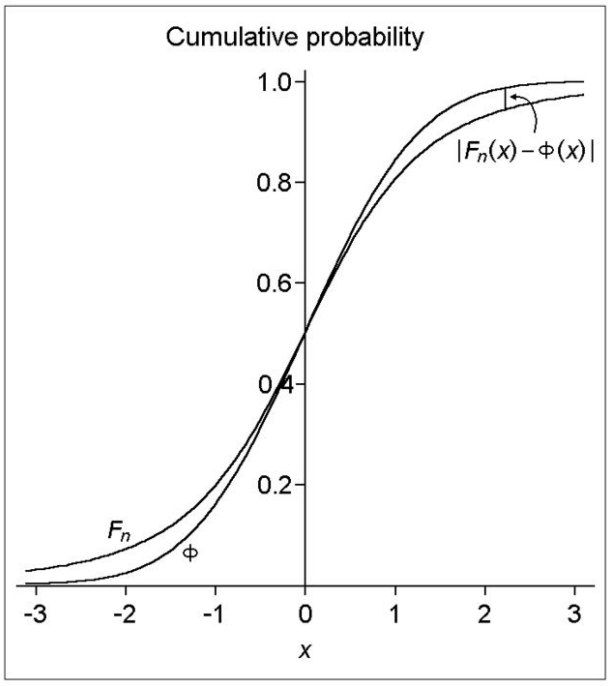
Beviset overspringes.

Konstanten ρ er tydeligvis relateret til skævheden af fordelingen, da den dybest set er en absolut version af det tredje moment.

Uligheden er illustreret på figuren til højre: Der er altså en grænse for, hvor stort stykket mellem de to grafer kan blive, og denne grænse varierer som $n^{-½}$.

Størrelsen af C er blevet forbedret i tidens løb. Det foreløbigt bedste bud (fra 2012) er $C = 0{,}4748$.

Hvis middelværdien af de stokastiske variable ikke er nul, generaliserer sætningen umiddelbart til et tilsvarende resultat, hvor der i stedet skal beregnes et centralt, absolut tredjemoment.



Eksempel: Bernoullifordeling

Lad os se hvad Berry-Esseen siger om Bernouillifordelingen. Her er det absolutte, centraliserede tredjemoment:

$$E[|X - p|^3] = p \cdot |1 - p|^3 + (1 - p) \cdot |0 - p|^3 = p(1 - p)^3 + (1 - p)p^3 = p(1 - p)((1 - p)^2 + p^2)$$

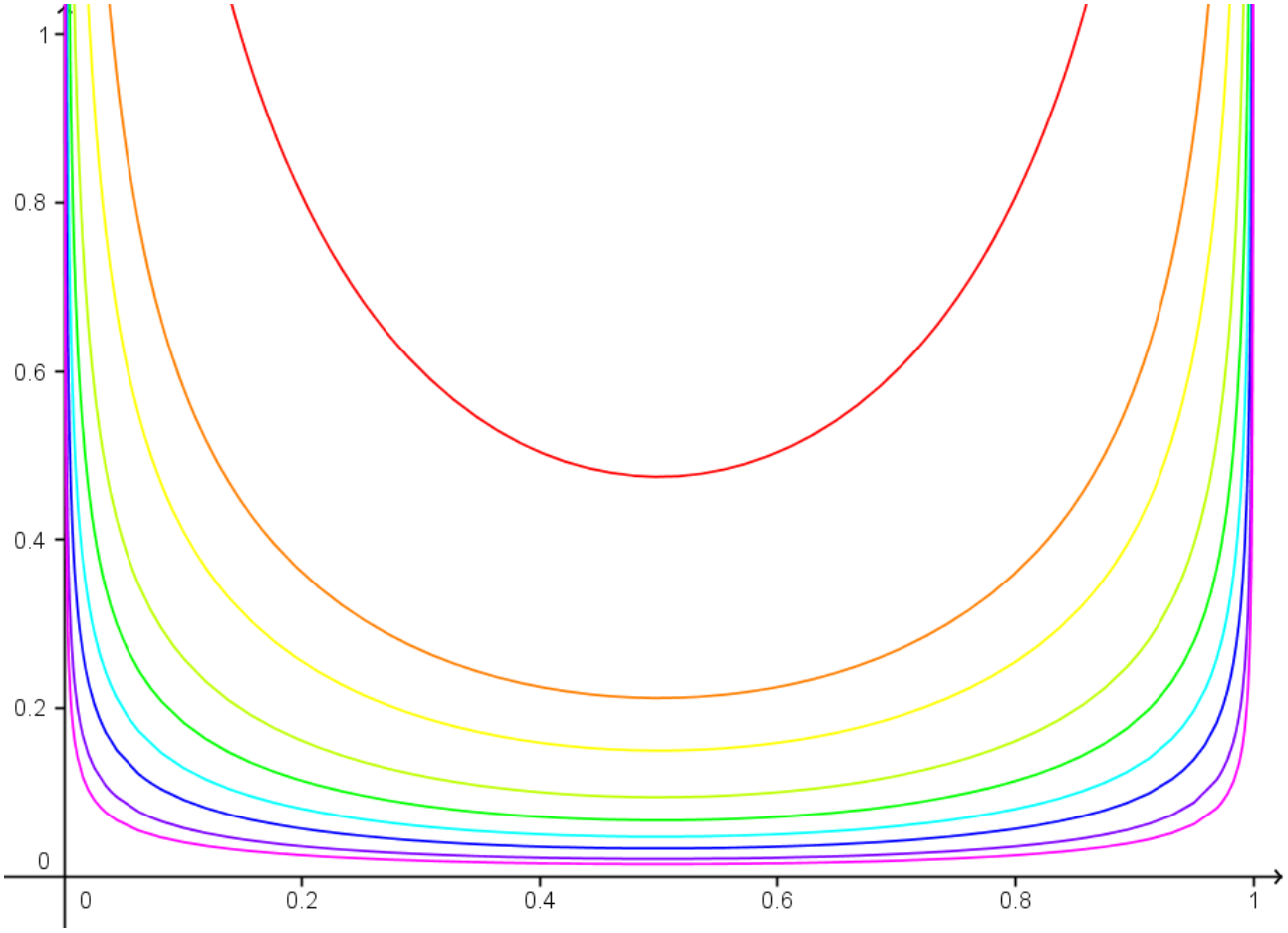
Forskellen mellem den normaliserede fordelingsfunktion for binomialfordelingen og standardnormalfordelingen opfylder altså:

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}} = \frac{0{,}4748\, p(1-p)(p^2 + (1-p)^2)}{\big(p(1-p)\big)^{3/2}\sqrt{n}} = \frac{0{,}4748(p^2 + (1-p)^2)}{\sqrt{np(1-p)}}$$

Figuren nedenfor viser graferne for funktionen:

$$f_n(p) = \frac{0{,}4748(p^2 + (1-p)^2)}{\sqrt{np(1-p)}}$$

Som n stiger bliver fejlmargen altså lavere og lavere (den lilla graf har $n = 1000$, så det går ikke specielt stærkt).



Det ses, at den mindste afvigelser sker for $p = ½$. Hvis vi sætter en tolerancetærskel δ for afvigelsen leder vi altså efter n der opfylder:

$$\frac{0{,}4748(p^2 + (1-p)^2)}{\sqrt{np(1-p)}} < \delta \Leftrightarrow \frac{0{,}4748(p^2 + (1-p)^2)}{\delta\sqrt{p(1-p)}} < \sqrt{n} \Leftrightarrow n > \frac{0{,}4748^2(p^2 + (1-p)^2)^2}{\delta^2 p(1-p)}$$

Hvis f.eks. $p = 0{,}4$ og vi ønsker en øvre grænse på $\delta = 0{,}01$ bliver kriteriet $n > 6009$.

Kontinuitetskorrektion kan dog forbedre dette betydeligt.

En generalisering af den centrale grænseværdisætning

Man kan nu spørge sig selv, hvad der sker med gennemsnittet af en række identisk fordelte, uafhængige stokastiske variable, hvis de ikke opfylder kriterierne i den centrale grænseværdisætning. Med andre ord: hvad sker der, hvis forventningsværdien og/eller standardafvigelsen ikke eksisterer?

Det viser sig, at der findes en hel familie af funktioner – kaldet *stabile fordelinger* - der kan være mulige grænseværdier i denne situation. Normalfordelingerne er kun en af disse familier.

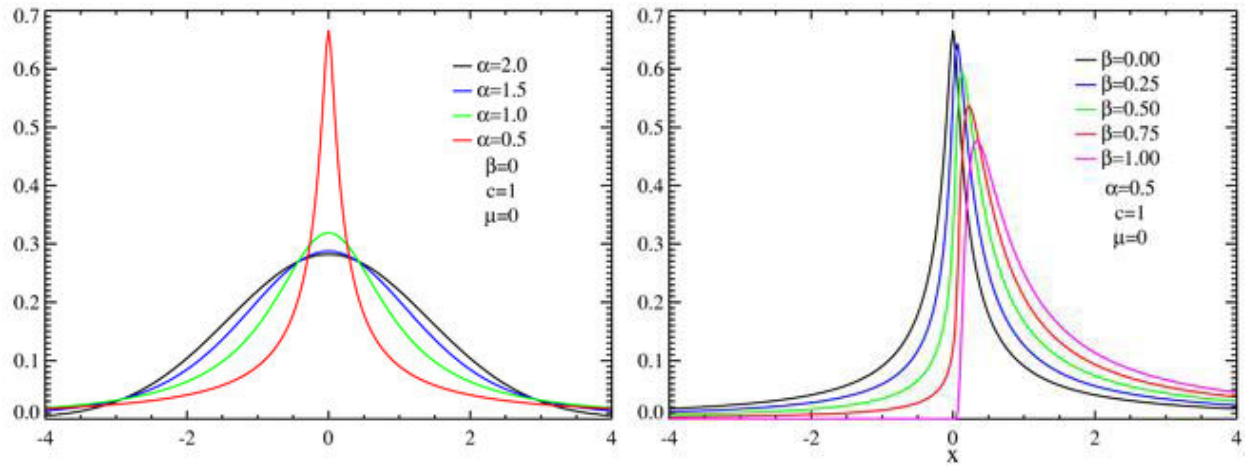
Stabile fordelinger

Definition: Lad X_1 og X_2 være uafhængige stokastiske variable med samme fordeling som X. X siges da at være en stabil fordeling, hvis der for alle $a, b > 0$ gælder at alle $aX_1 + bX_2$ har samme fordeling som $cX + d$, for et eller andet $c > 0$ og d. Hvis $d = 0$ kaldes fordelingen *strikst* stabil.

Det viser sig, at alle sådanne stabile fordelinger kan parametriseres ved fire konstanter μ, c, α, β. For en given værdi af de fire parametre svarer den karakteristiske funktion:

$$\varphi(t;\ \mu, c, \alpha, \beta) = \exp(it\mu - |ct|^{\alpha}(1 - i\beta \mathrm{sgn}(t) \tan(\pi\alpha/2))$$

Her angiver μ placering (ikke nødvendigvis middelværdi, da denne ikke altid er defineret), c skala og α (koncentration) og β (skævhed) form af fordelingen. Billederne viser nogle eksempler på hvordan de tilsvarende frekvensfunktioner ser ud.



For $\alpha = 2$ får man netop normalfordelinger med $\sigma^2 = 2c^2$ (β gør ingen forskel i dette tilfælde).

Den centrale grænseværdisætning for fordelingen med "fede haler"

Vi har set, at de fordelinger der ikke opfylder betingelserne i den centrale grænseværdisætning, og altså mangler middelværdi og/eller standardafvigelse har "fede" haler. Cauchy-fordelingen, f.eks.

Hvis de ens fordelte, uafhængige variable har sådanne fede hale der aftager som $|x|^{-\alpha-2}$, hvor $0 < \alpha < 2$, da vil gennemsnittet af fordelingerne konvergere mod en stabil fordeling med samme α.

Resampling

Resampling er en betegnelse for processer hvorved man kan lave empiriske overslag over variationen af givne parametre ved hjælp af stikprøven. Vi skal her kigge på to metoder: *jackknife resampling* og *bootstrapping*.

Jackknife resampling

Ideen i jackknife resampling er undlade et enkelt element fra stikprøven og foretage det relevante punktestimat på den resulterende del-stikprøve. Hvis den oprindelige stikprøves størrelse er N kan dette gøres på N måder, der hver giver et punktestimat. Fordelingen af disse estimater

Bootstrapping

Skriv ligningen her.

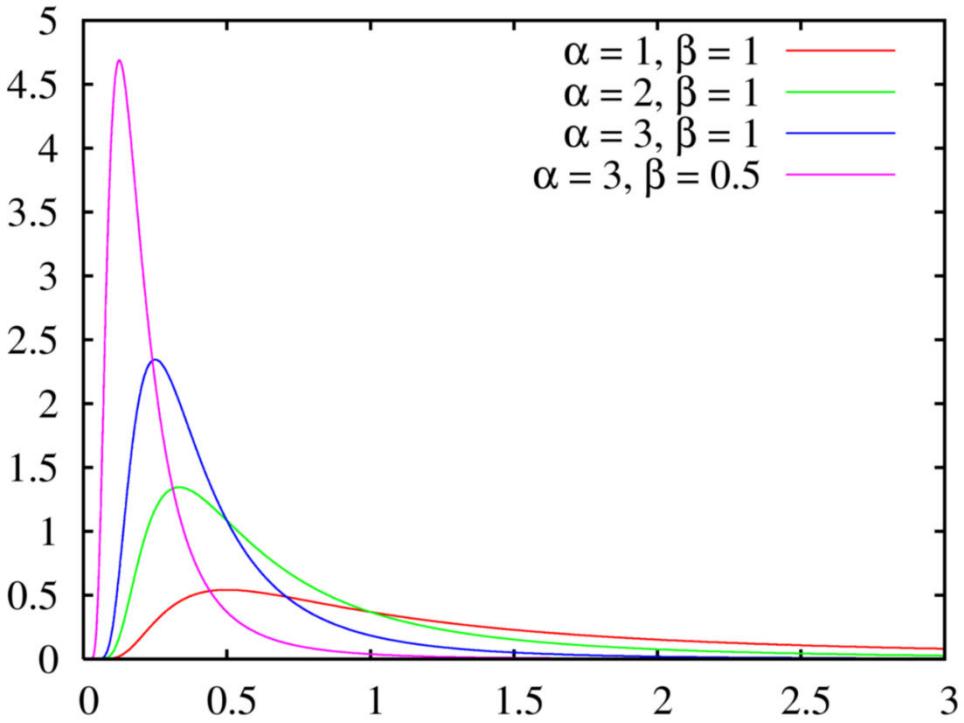
Gamma- og χ^2-fordelingerne

Gammafordelingen

Definition: *Gammafordelingen* med *formparameter* α og *skalaparameter* β er givet ved frekvensfunktionen:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x > 0$$

Figuren viser nogle eksempler på hvordan fordelingen kan se ud:



Frekvensfunktion?

At der virkelig er tale om en frekvensfunktion ses ved hjælp af følgende sætning:

Sætning:

$$\int_0^{\infty} x^{\alpha-1} e^{-\frac{x}{\beta}}\, dx = \beta^{\alpha}\, \Gamma(\alpha)$$

Bevis: Benyt substitutionen $t = x/\beta$ benyttet, hvilket betyder $x = \beta t$ og $dx = \beta\, dt$:

$$\int_0^{\infty} x^{\alpha-1} e^{-\frac{x}{\beta}}\, dx = \int_0^{\infty} \beta^{\alpha-1} t^{\alpha-1} e^{-t} \cdot \beta\, dt = \beta^{\alpha} \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt = \beta^{\alpha}\, \Gamma(\alpha)$$

Bevis slut.

Nu er det nemt at se, at integralet af gammafordelingen er 1, og at der derfor er tale om en frekvensfunktion:

$$\int_0^\infty f(x)\,dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\frac{x}{\beta}}\,dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \beta^\alpha \Gamma(\alpha) = 1$$

Fordelingsfunktion for sum af gammafordelinger

Sætning: Lad X_1 og X_2 være gammafordelte stokastiske variable med formparametre hhv. α_1 og α_2, men med samme skalaparameter β. Da er $X_1 + X_2$ gammafordelt med formparameter $\alpha_1 + \alpha_2$ og skalaparameter β.

Bevis: Ifølge sætningen om fordelingsfunktionen for sum af stokastiske variable er fordelingsfunktionen for $X_1 + X_2$ givet ved:

$$(f_1 * f_2)(x) = \int_0^x \frac{1}{\beta^{\alpha_1}\Gamma(\alpha_1)} t^{\alpha_1-1} e^{-\frac{t}{\beta}} \frac{1}{\beta^{\alpha_2}\Gamma(\alpha_2)} (x-t)^{\alpha_2-1} e^{-\frac{(x-t)}{\beta}}\,dt$$

Her er det benyttet, at fordelingsfunktionerne kun er forskellige fra nul for positive værdier. Faktorer der ikke afhænger af t sættes uden for integralet og der reduceres:

$$\frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\frac{x}{\beta}} \int_0^x t^{\alpha_1-1}(x-t)^{\alpha_2-1}\,dt$$

Benyt nu substitutionen $s = \frac{t}{x}$. Da er $t = sx$ og $ds = x\,dt$. Integralet i ovenstående bliver:

$$\int_0^1 s^{\alpha_1-1} x^{\alpha_1-1} (x - sx)^{\alpha_2-1}\, x\,ds = x^{\alpha_1+\alpha_2-1} \int_0^1 s^{\alpha_1-1}(1-s)^{\alpha_2-1}\,ds$$

I alt:

$$(f_1 * f_2)(x) = \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\frac{x}{\beta}} x^{\alpha_1+\alpha_2-1} \int_0^1 s^{\alpha_1-1}(1-s)^{\alpha_2-1}\,ds$$

Da foldningen er en frekvensfunktion må integralet være lig 1:

$$\int_0^\infty (f_1 * f_2)(x)\,dx = 1 \iff$$

$$\frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 s^{\alpha_1-1}(1-s)^{\alpha_2-1}\,ds \cdot \int_0^\infty e^{-\frac{x}{\beta}} x^{\alpha_1+\alpha_2-1}\,dx = 1$$

Det sidste integral er lig med $\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)$ ifølge sætningen fra sidste sektion. Dvs:

$$\frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 s^{\alpha_1-1}(1-s)^{\alpha_2-1}\,ds \cdot \beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2) = 1 \iff$$

$$\frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 s^{\alpha_1-1}(1-s)^{\alpha_2-1}\,ds = \frac{1}{\Gamma(\alpha_1+\alpha_2)}$$

Dermed reducerer fordelingsfunktionen for $X_1 + X_2$ til:

$$\frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)} e^{-\frac{x}{\beta}} x^{\alpha_1+\alpha_2-1}$$

Dette er netop en gammafordeling med formparameter $\alpha_1 + \alpha_2$ og skalaparameter β.

Bevis slut.

Der følger nu umiddelbart pr. induktion:

Korollar: Hvis n stokastiske variable $X_1, X_2, \cdots, X_n$ alle er gammafordelte med formparametre hhv. $\alpha_1, \alpha_2, \cdots, \alpha_n$ og med samme skalaparameter β, da er $X_1 + X_2 + \cdots X_n$ gammafordelt med formparameter $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ og skalaparameter β.

χ^2-fordelingen

Definition: Hvis er $X_1, X_2, \cdots, X_n$ alle er standardnormalfordelt, altså med middelværdi 0 og spredning 1 sættes χ^2-fordelingen med n frihedsgrader til frekvensfunktionen for følgende stokastiske variabel:

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

For at finde formen for denne fordeling varmer vi op med følgende sætning:

Sætning: Hvis er X er standardnormalfordelt er X^2 gammafordelt med formparameter $\alpha = \frac{1}{2}$ og skalaparameter $\beta = 2$.

Bevis: Standardnormalfordelingen har frekvensfunktionen:

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

Fordelingsfunktionen for X^2 er givet ved:

$$F(x) = P(X^2 < x) = P\left(-\sqrt{x} < X < \sqrt{x}\right) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = \frac{1}{\sqrt{2\pi}} 2 \int_0^{\sqrt{x}} e^{-\frac{t^2}{2}}\, dt$$

I sidste trin er det benyttet, at frekvensfunktionen er lige. Vi foretager nu substitutionen $s = t^2$, hvilket medfører $t = \sqrt{s}$ og $\frac{dt}{ds} = \frac{1}{2\sqrt{s}} \Leftrightarrow dt = \frac{1}{2\sqrt{s}} ds$. Det giver:

$$F(x) = \frac{1}{\sqrt{2\pi}} 2 \int_0^x e^{-\frac{1}{2}s} \cdot \frac{1}{2\sqrt{s}} ds = \frac{1}{\sqrt{2\pi}} \int_0^x s^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}s}\, ds$$

Frekvensfunktionen for X^2 er den afledte af denne fordelingsfunktion, altså:

$$\frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}x}$$

Da $\Gamma(½) = \sqrt{\pi}$ kan dette skrives:

$$\frac{1}{2^{½}\Gamma(½)} x^{(½-1)x} \cdot e^{-\frac{x}{2}}$$

Dette er netop fordelingsfunktionen for en gammafordeling med formparameter $\alpha = \frac{1}{2}$ og skalaparameter $\beta = 2$.

Bevis slut.

Sætning: χ^2-fordelingen med n frihedsgrader har fordelingsfunktionen:

$$f(x) = \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

Bevis: Pr. definition er χ^2-fordelingen med n frihedsgrader lig frekvensfunktionen for

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

Her er $X_1, X_2, \cdots, X_n$ alle standardnormalfordelt. Men hvert af disse led er gammafordeling med formparameter $\alpha = \frac{1}{2}$ og skalaparameter $\beta = 2$. Af ovenstående korollar følger, at Y er gammafordeling med formparameter $\alpha = \frac{n}{2}$ og skalaparameter $\beta = 2$. Dette svarer netop til ovenstående frekvensfunktion.

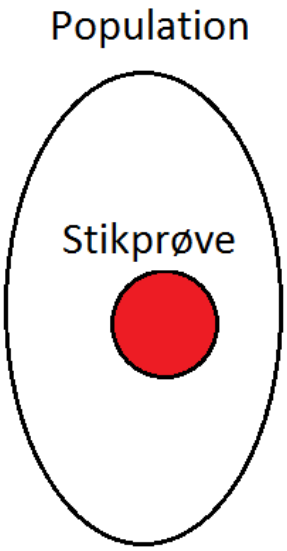
Bevis slut.

Cochrans sætning

Lad $X_1 = \{x_{11}, x_{12}, \cdots, x_{1k}\}, X_2 = \{x_{21}, x_{22}, \cdots, x_{2k}\}, \cdots, X_n = \{x_{n1}, x_{n2}, \cdots, x_{nk}\}$ være stokastiske variable, der alle er multinomialfordelt med parameter 1

Stikprøver og estimation

Når man er interesseret i egenskaber ved en stor gruppe er det sjældent praktisk muligt at undersøge hele gruppen. I stedet undersøger man en lille delmængde af gruppen. Dette kaldes en stikprøven. Afsnittet handler dels om hvordan man udtager en sådan stikprøve på hensigtsmæssig vis, og hvordan man drager konklusioner på baggrund af stikprøven. Sidste punkt er det der kalder statistik.



Population og stikprøve

Populationen er den gruppe vi ønsker at udtale os om. Enkelte gange, som ved folketingsvalg eller folketællinger (engelsk: census) er det muligt at undersøge hele populationen, men som oftest udtager man en *stikprøve*, der er en delmængde af populationen.

Stikprøvetagning med eller uden tilbagelæggelse

Stikprøver kan foretages både med og uden *tilbagelæggelse*. Med tilbagelæggelse kan et element fra populationen altså forekomme flere gange i stikprøven. For store populationer gør det ikke den store forskel om man skelner mellem de to. Vi vil her mest kigge på stikprøver uden tilbagelæggelse.

Ordnet eller uordnet stikprøve

Stikprøvens elementer kan være ordnede eller uordnede. Vi vil her mest kigge på uordnede stikprøver.

Eksempler på parametre

Der kan være mange forskelle parametre ved populationen man er interesseret i, og dermed prøver at afdække med stikprøven. Her er nogle almindelige eksempler:

- μ, altså middelværdien af en bestemt egenskab. F.eks. gennemsnitsvægt.
- σ, altså standardafvigelsen af en bestemt egenskab. F.eks. standardafvigelsen for højde.
- p, populationsandelen for en bestemt egenskab. F.eks. andelen af venstrehåndede.
- λ, intensiteten af en bestemt hændelse, hvilket betyder gennemsnitlig forekomst pr. tidsenhed. F.eks. antallet af sygedage pr. år.

Repræsentativitet

En stikprøve siges at være *repræsentativ* for populationen, hvis den giver samme estimat for den parameter som man er interesseret i, som hvis man havde målt på hele befolkningen. Repræsentativitet afhænger altså strengt taget af hvilken parameter man er interesseret i at bestemme.

Hvis stikprøven ikke er repræsentativ kan man ikke generalisere fra stikprøven til hele populationen!

Udtagelse af stikprøver

Målet er altså, at stikprøven er repræsentativ. Hvordan sikrer man dette? Herunder er nogle forskellige strategier med forskellige fordele og ulemper både i forhold til repræsentativitet og den praktiske udtagelse af stikprøven.

Stikprøveusikkerhed

Når en stikprøve udtages kan der ske forskellige typer fejl, forstået som afvigelser fra repræsentativitet. Dem vil vi kalde *stikprøveusikkerhed* (engelsk: sample error). Her skelnes mellem *systematiske afvigelser* (også kaldet *bias*), og rent *tilfældige* fejl.

Eksempel: Simpel, tilfældig udvælgelse

En *simpel, tilfældig udvælgelse* er den situation, hvor stikprøven udtrækkes tilfældigt blandt alle elementerne i populationen, med lige stor sandsynlighed. I praksis er dette ofte meget svært at gøre nemt og effektivt.

- Fordele: Den tilfældige udvælgelse sikrer, at der ikke begås systematiske afvigelser.
- Ulemper: Upraktisk. Sandsynligheden for tilfældige fejl er stor, da man "skyder i blinde".

Eksempel: Systematisk udvælgelse

I en *systematisk udvælgelse* tildeles hvert element i populationen et nummer (eller en lignende ordning). Herefter udtager man f.eks. hver tiende element startende fra et tilfældigt valgt element indtil man har den ønskede stikprøvestørrelse. Hvis tildelingen af nummer er tilfældig (det ideelle tilfælde) kaldes dette for en *systematisk, tilfældig udvælgelse*.

Et eksempel på en ikke-tilfældig, systematisk udvælgelse kunne være en kundeundersøgelse i en butik, hvor man udvælger f.eks. hver tyvende kunde til at deltage i undersøgelsen.

- Fordele: Relativt simpel at udføre. Hvis tilfældig begås der ikke systematiske afvigelser.
- Ulemper: Hvis der er mønstre i nummereringen kan der opstå systematiske afvigelser. Hvis tilfældig er der i sagens natur chance for tilfældige fejl.

Eksempel: Stratificeret udvælgelse

Ordet stratificeret kommer af ordet strata, der betyder lag. En *stratificeret udvælgelse* er altså en lagdelt udvælgelse. Lagene er her de parametre man forventer vil være udslagsgivende i forhold til undersøgelsen. Eksempler kunne være køn, alder og indkomst. Når man har valgt de parametre man vil stratificere efter, skal stikprøven udvælges, så den har den samme fordeling parametrene har i populationen.

- Fordele: Sikrer mod systematiske afvigelser, i hvert fald for de stratificerede variable.
- Ulemper: Kan ikke bruges, hvis der ikke er klare strata i populationen. Udvælgelses af parametre er ikke altid oplagt. Kan være besværlig/dyr at udføre i praksis.

Eksempel: Klyngeudvælgelse

Ofte vil populationen være inddelt i naturlige *klynger*. Hvis populationen f.eks. er alle danske skolebørn vil hver skole udgøre en naturlig klynge. *Klyngeudvælgelse* er når stikprøven foretages på klyngeniveau. Udtagelsen af klynger kan i sig selv følge en af ovenstående metoder.

- Fordele: Nem at foretage.
- Ulemper: Øger sandsynligheden for tilfældige fejl og kræver derfor større stikprøve, overordnet set.

Ikke-stikprøveusikkerheder

Denne type fejl er uafhængig af den valgte type udvælgelse, men bunder i problemer med selve designet af undersøgelsen eller efterbehandlingen (engelsk: non-sample error). Typiske fejl er:

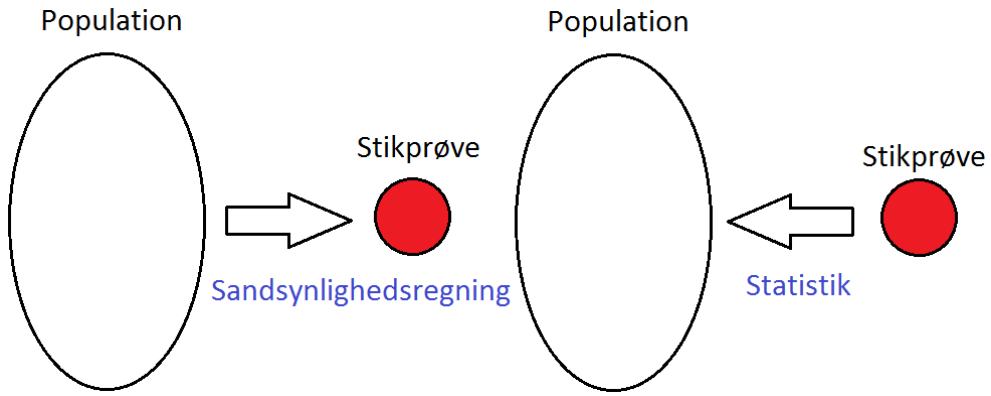
- Overdækning: Stikprøven indeholder elementer der ikke er med i populationen.
- Underdækning: Stikprøven underrepræsenterer bestemte dele af populationen.
- Målefejl: Kan være knyttet til et forsøg. Eller en respondent kan misforstå et spørgsmål.
- Procesfejl: Fejl i efterbehandlingen af data.
- Manglende svar: Der mangler svar/data for nogle elementer af stikprøven.

Sandsynlighedsregning vs. statistik

Hvis man ved hvordan hele populationen er fordelt kan man bruge *sandsynlighedsregning* til at udtale sig om, hvordan en stikprøve sandsynligvis ser ud.

I praksis står man som regel med det modsatte problem: Man har givet en stikprøve og skal på baggrund af den udtale sig om fordelingen af hele populationen. Dette kaldes *(inferentiel) statistik*.

De to situationer er illustreret på figuren nedenfor.



Estimation

Ved *estimation* forstås et skøn over hvad den parameter er interesseret i populationen, baseret på stikprøven. Her skal man skelne mellem to grundlæggende typer af estimater:

Punktestimat

Et punktestimat af en parameter består af en enkelt værdi, altså den værdi der vurderes at være størst sandsynlighed for at parameteren har for populationen.

Konfidensinterval

Et *konfidensinterval* er et interval med en tilhørende sandsynlighed - ofte 95%. Parameteren for populationen vil da med denne sandsynlighed ligge i det relevante interval.

Begge typer estimater bygger i praksis på den centrale grænseværdisætning.

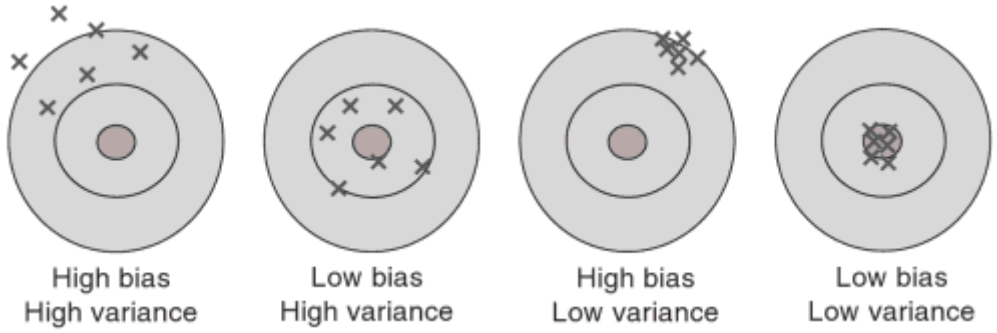
Egenskaber ved estimatorer

En *estimator* $\hat{\theta}$ er en funktion af stikprøvens data, der kan forstås som en række stokastiske variable, her kollektivt blot kaldet X. $\hat{\theta}$ er altså selv en stokastisk variabel. $\hat{\theta}$ søger at estimere en bestemt parameter for populationen, der har den sande værdi θ. For en given stikprøve $X = x$ punktestimerer $\hat{\theta}$ altså parameteren θ til at være $\hat{\theta}(x)$.

Definition: For en estimator $\hat{\theta}$ sættes *skævheden* eller *bias* til forventningsværdien af afvigelsen fra θ:

$$B(\hat{\theta}) = E[\hat{\theta}(X)] - \theta$$

Hvis $B(\hat{\theta}) = 0$ kaldes $\hat{\theta}$ for *ikke-skæv* eller *ikke-biased*. Det er klart, at en fornuftig estimator bør have denne egenskab. Variansen af estimatoren defineres på almindelig vis.



Billedet viser eksempler på høj/lav bias og varians for en estimator der prøver at ramme bullseye. Selvom en lav varians umiddelbart er ønskværdigt skal man huske, at estimatoren stadig kan have høj bias, og dermed typisk ramme langt ved siden af målet.

Hvis fordelingen af en ikke-biased estimator $\hat{\theta}$ er kendt, kan dens fraktiler benyttes til at bestemme konfidensintervaller for θ. Dette er generelt ikke nemt, da vi umiddelbart ikke har gjort nogen antagelser om fordelingen af fordelingen af observationer. Heldigvis kommer den centrale grænseværdisætning os til undsætning.

Estimation af middelværdi μ

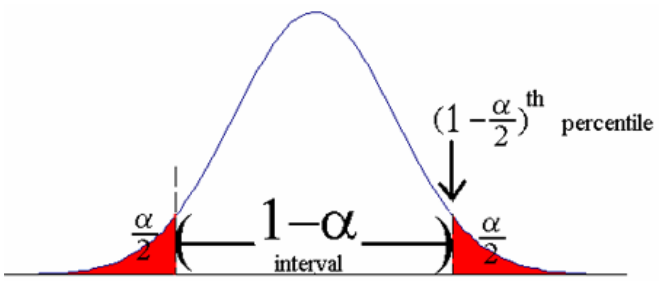
Givet en stikprøve af størrelse n med observationer $X_1, X_2, \cdots, X_n$ forstået som stokastiske variable ønsker vi at estimere populationens middelværdi μ. Et oplagt bud på en estimator er middelværdien af stikprøven:

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

Ifølge den centrale grænseværdisætning er denne størrelse normalfordelt med middelværdi μ og standardafvigelse $\frac{\sigma}{\sqrt{n}}$, i hvert fald når n ikke er alt for lille. Senere vil vi kigge på hvad man gør, hvis dette ikke er tilfældet. Her vil vi i første omgang antage, at n er stor, hvilket ofte i praksis antages at betyde 30 eller større. Der er dog intet magisk ved dette antal – visse fordelinger vil kræve større n, men for

fordelinger der ikke er voldsomt skæve vil dette være nok. Vi vil også antage, at stikprøvens størrelse er lille i forhold til hele populationen, i praksis mindre end 10% af populationens størrelse.

Sætningen viser umiddelbart, at $\bar{X}$ har bias 0 og varians $\frac{\sigma^2}{n}$. Store stikprøver giver altså – ganske intuitivt - bedre vurderinger af populationens middelværdi.



Resultatet kan også bruges til at give et konfidensinterval for μ: Hvis vi ønsker et interval med tilhørende sandsynlighed $1-\alpha$ leder vi efter værdier, der gør arealet under hver hale for normalfordlingen til $\frac{\alpha}{2}$. På den måde bliver arealet af det røde område på figuren til venstre netop lig α. Vi har altså brug for at finde $\left(1-\frac{\alpha}{2}\right)$-fraktilen.

For $\alpha = 5\%$ er denne $\Phi^{-1}(0.975) = 1{,}96$ (eller blot 2 i en håndevending). Derfor er konfidensintervallet her:

$$\left[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$$

Estimation af varians σ^2

Givet samme betingelser som i sidste sektion skulle man tro, at et godt bud på en estimator for populationens varians σ^2 er:

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

Det viser sig imidlertid, at denne størrelse er biased! Vi laver følgende snedige omskrivning:

$$B[s_n^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] - \sigma^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2\right] - \sigma^2 =$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2\,(\bar{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\; + (\bar{X} - \mu)^2\right] - \sigma^2 =$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2\,(\bar{X} - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right] - \sigma^2 =$$

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

Husk, at selvom forventningsværdien af $\bar{X} - \mu$ er nul, er dette ikke sandt for kvadratet. Så $\bar{V}$ estimerer generelt variansen for lavt. For at finde en passende justering skal denne evalueres:

$$E[(\bar{X}-\mu)^2] = \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

Så:

$$B[s_n^2] = E[s_n^2] - \sigma^2 = -\frac{\sigma^2}{n} \Leftrightarrow E[s_n^2] = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(1-\frac{1}{n}\right)$$

Hvis vi i stedet ønsker at definere en ny, ikke-biased estimator s^2 der er lig s_n^2 gange en konstant k (kaldet *Bessels rettelse*) skal der gælde:

$$B[s^2] = B[k \cdot s_n^2] = E[k \cdot s_n^2] - \sigma^2 = k \cdot E[s_n^2] - \sigma^2 = k\sigma^2\left(1-\frac{1}{n}\right) - \sigma^2 = 0 \Leftrightarrow$$

$$k\sigma^2\left(1-\frac{1}{n}\right) = \sigma^2 \Leftrightarrow k\left(\frac{n-1}{n}\right) = 1 \Leftrightarrow k = \frac{n}{n-1}$$

Det betyder:

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

Denne størrelse er

Små og store stikprøver

Hvad gør man, hvis en stikprøve ikke er stor nok til at sikre rimelig konvergens mod et normaltfordelt gennemsnit? Eller hvis stikprøven er så stor, at dens størrelse ikke længere er ubetydelig i forhold til populationens størrelse? Det skal vi se på i dette afsnit.

Små stikprøver

Students t-fordeling

Definition: Lad Z være en standardnormalfordelt stokastisk variabel, og V en χ^2-fordelt stokastisk variabel med n frihedsgrader, hvor Z og V er uafhængige. Vi definerer nu en ny stokastisk variabel:

$$T = \frac{Z}{\sqrt{V/n}}$$

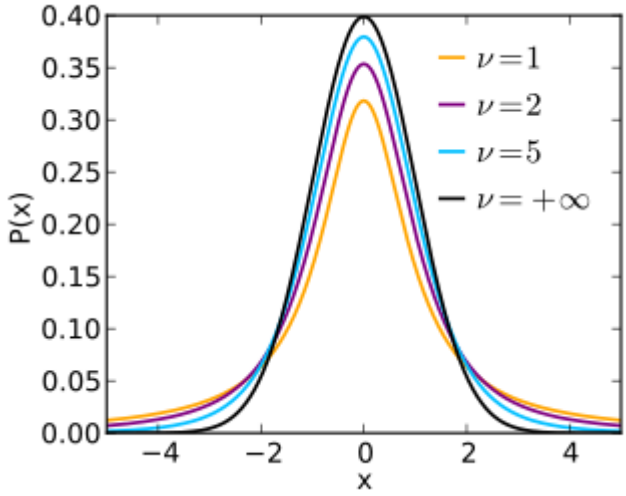
T siges at være *t-fordelt* med n frihedsgrader.

Sætning: En t-fordelt variabel med n frihedsgrader har frekvensfunktionen:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \cdot \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n-1}{2}}$$

Bevis: Lad g og h være frekvensfunktionerne for hhv. standardnormalfordelingen og en χ^2-fordeling med n frihedsgrader, altså:

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad h(v) = \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} v^{\frac{n}{2}-1} e^{-\frac{v}{2}}$$



Store stikprøver

For stikprøver der er store i forhold til populationen (underforstået uden tilbagelæggelse) bliver antagelsen om uafhængighed mellem de enkelte observationer udfordret.

Likelihood og deskriptorer

Likelihood

"Almindelig sandsynlighed" er sandsynligheden for, at et bestemt udfald af et eksperiment vil ske, givet parametrene i en fordeling.

Likelihood[6] er en betegnelse for sandsynligheden for, at en fordeling har et givet sæt parametre på baggrund af et bestemt udfald af et eksperiment. Så selv om de to udtryk matematisk set er ens, er det altså i en vis forstand det modsatte af ovenstående "almindelige sandsynlighed".

Definition: Likelihood-funktionen er givet ved:

$$\mathcal{L}(x|\theta) = P(\theta|x)$$

Her er x et bestemt udfald i udfaldsrummet og θ et sæt af parametre til den underliggende fordeling. Funktionen skrives ofte $p_\theta(x)$ for diskrete udfaldsrum og $f_\theta(x)$ for kontinuerte udfaldsrum.

Likelihood-funktionen spiller en stor rolle i inferentiel statistik.

Likelihood-funktion for uafhængige stokastiske variable

For uafhængige stokastiske variable er sandsynligheder multiplikative, så hvis $X_1, X_2, \cdots, X_n$ er uafhængige med frekvensfunktioner $f_1, f_2, \cdots, f_n$ bliver likelihood-funktionen:

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f_i(x_i)$$

θ er her en forkortelse for samtlige parametre i frekvensfunktionerne. Hvis alle variable har samme frekvensfunktion f reducerer dette til:

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i)$$

Log-likelihood

Ofte er det mere praktisk at arbejde med logaritmen af likelihood-funktion. Denne kaldes log-likelihood:

Definition: Log-likelihood defineres ved:

$$l(x|\theta) = \log\big(\mathcal{L}(x|\theta)\big)$$

Log betegner her den naturlige logaritme, ikke ti-talsditto.

Det ses umiddelbart, at der for uafhængige, identisk fordelte stokastiske variable gælder:

[6] Der er ikke noget tilsvarende udtryk på dansk.

$$l(x_1, x_2, \cdots, x_n|\theta) = \sum_{i=1}^{n} \log\big(f(x_i)\big)$$

Eksempel: Normalfordelingen

Normalfordelingen har frekvensfunktionen $f(x) = (2\pi\sigma^2)^{-½} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Likelihood-funktionen for n uafhængige observationer bliver derfor:

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2]$$

Log-likelihood bliver derfor:

$$l(x_1, x_2, \cdots, x_n|\theta) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Eksempel: Binomialfordelingen

Binomialfordelingen har frekvensfunktionen $f(k) = \binom{k}{n} p^k (1-p)^{n-k}$, hvorfor likelihood-funktionen for m uafhængige observationer bliver:

$$\mathcal{L}(k_1, k_2, \cdots, k_m|\theta) = \binom{k_1}{n}\binom{k_2}{n}\cdots\binom{k_m}{n} p^{k_1+k_2+\cdots+k_m}(1-p)^{mn-k_1-k_2-\cdots-k_n}$$

Log-likelihood for en enkelt observation er:

$$l(k|\theta) = \log\left(\binom{k}{n}\right) + k \cdot \log(p) + (n-k) \cdot \log(1-p)$$

Da binomialkoefficienten ikke afhænger af p er den sjældent vigtig. Log-likelihood for m uafhængige observationer bliver:

$$l(x_1, x_2, \cdots, x_n|\theta) = \sum_{i=1}^{m} \log\left(\binom{k_i}{n}\right) + \left(\sum_{i=1}^{m} k_i\right) \cdot \log(p) + \left(mn - \sum_{i=1}^{m} k_i\right) \cdot \log(1-p) =$$

$$\sum_{i=1}^{m} \log\left(\binom{k_i}{n}\right) + m\bar{k} \cdot \log(p) + m(n-\bar{k}) \cdot \log(1-p)$$

Her er $\bar{k}$ gennemsnittet af de observerede k-værdier.

Eksempel: Poissonfordelingen

Poissonfordelingen har frekvensfunktionen $f(k) = \frac{\lambda^k}{k!}e^{-\lambda}$, hvorfor likelihood-funktionen for n uafhængige observationer bliver:

$$\mathcal{L}(k_1, k_2, \cdots, k_n|\theta) = \frac{\lambda^{k_1+k_2+\cdots k_n}}{k_1!\,k_2!\cdots k_n!}e^{-n\lambda}$$

Så log-likelihood bliver:

$$l(k_1, k_2, \cdots, k_n|\theta) = \left(\sum_{i=1}^{n} k_i\right)\log(\lambda) - \sum_{i=1}^{n}\log(k_i!) - n\lambda$$

Eksempel: Eksponentialfordeling

Eksponentialfordelingen har frekvensfunktionen: $f(x) = \lambda \cdot e^{-\lambda x}$. Likelihood-funktionen for n uafhængige observationer bliver derfor:

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = \lambda^n \cdot \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

Log-likelihood bliver:

$$l(x_1, x_2, \cdots, x_n|\theta) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

Eksempel: Gammafordeling

Gammafordelingen har frekvensfunktion $f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$, hvorfor likelihood-funktionen for n uafhængige observationer bliver:

$$\mathcal{L}(x_1, x_2, \cdots, x_n|\theta) = \frac{1}{\beta^{n\alpha}\big(\Gamma(\alpha)\big)^n}(x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\frac{x_1+x_2+\cdots+x_n}{\beta}}$$

Den tilhørende log-likelihood er:

$$l(x_1, x_2, \cdots, x_n|\theta) = -n\big(\alpha \cdot \log(\beta) + \log\big(\Gamma(\alpha)\big)\big) + (\alpha - 1) \cdot \left(\sum_{i=1}^{n}\log(x_i)\right) - \frac{1}{\beta}\sum_{i=1}^{n} x_i$$

Maksimum Likelihood Estimation (LME)

Dette princip siger, at givet et bestemt udfald x, er det bedste estimat for parametrene θ i modellen givet ved de værdier af parametrene, der har størst likelihood-funktion. Da logaritmefunktionen er voksende, kan man i stedet finde maksimum af log-likelihood.

Eksempel: Normalfordelingen

For at estimere μ og σ differentieres log-likelihood efter begge. Først μ:

$$\frac{\partial l}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(x_i - \mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i - n\mu$$

Ved maksimum skal den afledede være 0:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i - n\mu = 0 \Leftrightarrow n\mu = \sum_{i=1}^{n} x_i \Leftrightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Det bedste estimat for μ ifølge MLE er altså gennemsnittet af observationerne. For σ får vi:

$$\frac{\partial l}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Ved maksimum skal den afledede være 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \Leftrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = n \Leftrightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

Det bedste estimat for variansen σ^2 ifølge MLE er altså den sædvanlige (ikke-justerede) stikprøvevarians.

Eksempel: Binomialfordelingen

For at estimere p for kendt n differentieres log-likelihood efter p:

$$\frac{\partial l}{\partial p}\left(\sum_{i=1}^{m}\log\left(\binom{k_i}{n}\right) + m\bar{k}\cdot\log(p) + m\left(n-\bar{k}\right)\cdot\log(1-p)\right) = \frac{m\bar{k}}{p} - \frac{m(n-\bar{k})}{1-p}$$

Ved maksimum skal den afledede være 0:

$$\frac{m\bar{k}}{p} - \frac{m\left(n-\bar{k}\right)}{1-p} = 0 \Leftrightarrow \frac{\bar{k}}{p} - \frac{n-\bar{k}}{1-p} = 0 \Leftrightarrow \frac{\bar{k}(1-p)}{p(1-p)} - \frac{p\left(n-\bar{k}\right)}{p(1-p)} = 0 \Leftrightarrow$$

$$\bar{k}(1-p) - p\left(n-\bar{k}\right) = 0 \Leftrightarrow \bar{k} + p\left(-\bar{k}-n+\bar{k}\right) = 0 \Leftrightarrow p = \frac{-\bar{k}}{-n} = \frac{\bar{k}}{n}$$

Altså er det bedste estimat for p ifølge MLE gennemsnittet af de observerede k over antallet af Bernoulli-eksperimenter.

Eksempel: Poissonfordelingen

For at estimere λ differentieres log-likelihood:

$$\frac{\partial l}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left[\left(\sum_{i=1}^{n}k_i\right)\log(\lambda) - \sum_{i=1}^{n}\log(k_i!) - n\lambda\right] = \frac{\sum_{i=1}^{n}k_i}{\lambda} - n$$

Ved maksimum skal den afledede være 0:

$$\frac{\sum_{i=1}^{n}k_i}{\lambda} - n = 0 \Leftrightarrow \frac{\sum_{i=1}^{n}k_i}{\lambda} = n \Leftrightarrow \lambda = \frac{\sum_{i=1}^{n}k_i}{n} = \bar{k}$$

Det bedste estimat for λ ifølge MLE er altså gennemsnittet af de observerede k.

Eksempel: Eksponentialfordeling

For at estimere λ differentieres log-likelihood:

$$\frac{dl}{d\lambda} = \frac{d}{d\lambda}\left(n\log(\lambda) - \lambda\sum_{i=1}^{n}x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^{n}x_i$$

Ved maksimum skal den afledede være 0:

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \frac{n}{\lambda} = \sum_{i=1}^{n} x_i \Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Det bedste estimat for λ ifølge MLE er altså gennemsnittet af observationerne.

Eksempel: Gammafordeling

For at estimere α og β differentieres log-likelihood efter begge:

$$\frac{\partial}{\partial \alpha}\left[-n\big(\alpha \cdot \log(\beta) + \log\big(\Gamma(\alpha)\big)\big) + (\alpha - 1) \cdot \left(\sum_{i=1}^{n} \log(x_i)\right) - \frac{1}{\beta}\sum_{i=1}^{n} x_i\right] =$$

$$-n\left(\log(\beta) + \frac{1}{\Gamma(\alpha)}\int_0^{\infty} \log(t) \cdot t^{\alpha-1} \cdot e^{-t} dt\right) + \sum_{i=1}^{n} \log(x_i)$$

$$\frac{\partial}{\partial \beta}\left[-n\big(\alpha \cdot \log(\beta) + \log\big(\Gamma(\alpha)\big)\big) + (\alpha - 1) \cdot \left(\sum_{i=1}^{n} \log(x_i)\right) - \frac{1}{\beta}\sum_{i=1}^{n} x_i\right] =$$

$$-n\alpha\frac{1}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i$$

Ved maksimum skal begge disse være 0. At sætte $\frac{\partial l}{\partial \beta}$ lig 0 giver:

$$-n\alpha\frac{1}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \frac{1}{\beta}\sum_{i=1}^{n} x_i = n\alpha \Leftrightarrow \beta = \frac{\sum_{i=1}^{n} x_i}{n\alpha} = \frac{\bar{x}}{\alpha}$$

Hvis α er kendt giver dette det bedste estimat af β ifølge MLE. Hvis ikke må der indsættes i ligningen $\frac{\partial l}{\partial \alpha} = 0$, der så generelt løses skal numerisk.

Statistiske deskriptorer

Definition: En *deskriptor* er en størrelse der benyttes til at beskrive et datasæt eller stikprøve. Så hvis alle data samles i betegnelsen $\boldsymbol{X}$, kan en deskriptor T forstås som en funktion $T = t(\boldsymbol{X})$.

Definition: En deskriptor T kaldes *tilstrækkelig* (engelsk: sufficient) for en parameter θ, hvis al den information der er om θ i datasættet $\boldsymbol{X}$ er indeholdt i T. Med andre ord må den betingede likelihood $\mathcal{L}(\boldsymbol{X}|T)$ ikke afhænge af θ.

Lineær regression

Et datasæt $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$, hvor $x^{(i)} \in \mathbb{R}^n$ og $y^{(i)} \in \mathbb{R}$ søges forklaret vha. en lineær sammenhæng:

$$h_\Theta(x) = \Theta_0 + \Theta_1 x_1 + \cdots + \Theta_n x_n$$

Her er $\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_n \end{bmatrix}$, mens h står for *hypotese*. Vi leder altså efter det $\Theta \in \mathbb{R}^{n+1}$ der giver den hypotese der – i en eller anden forstand - passer bedst med observationerne.

Vektorisering

Mange formler kan simplificeres ved at skrives dem i vektor/matrixform. Med vektor af dimension n menes her en søjlevektor, altså et element i $\mathbb{R}^{n\times 1}$, med mindre andet er nævnt.

Prikproduktet mellem to vektorer af dimension n, v og w, kan f.eks. skrives:

$$v \cdot w = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n = [v_1\ v_2 \cdots v_n] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = v^T w$$

Dette kan benyttes til at finde afledte:

$$\frac{\partial}{\partial x_i}(a \cdot x) = \frac{\partial}{\partial x_i}(a^T x) = a_i$$

Eller på vektorform:

$$\nabla(a^T x) = a$$

Hvis a er en matrix i stedet for en vektor holder formlen stadig, da relationen gælder søjle for søjle.

En kvadratisk form kan skrives:

$$x^T A x = \sum_{i,j=1}^{n} x_i a_{ij} x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i \neq j} 2 a_{ij} x_i x_j$$

Her er A en symmetrisk $n \times n$ matrix. Hvis man afleder får man:

$$\frac{\partial}{\partial x_i}(x^T A x) = 2 a_{ii} x_i + \sum_{i \neq j} 2 a_{ij} x_j = 2 \sum_{j=1}^{n} a_{ij} x_j = 2(Ax)_i$$

Eller på vektorform:

$$\nabla(x^T A x) = 2Ax$$

Lineær hypotese på vektorform

Den lineære hypotese kan nu skrives:

$$h_\Theta(x) = [1\ x_1\ x_2 \cdots x_n] \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_n \end{bmatrix}$$

Hvis x redefineres, så $x_0 = 1$ kan dette kort skrives:

$$h_\Theta(x) = x^T\Theta$$

Designmatrix X

Til et givet observationssæt defineres *designmatricen* $X \in \mathbb{R}^{m\times(n+1)}$ ved:

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

Ved at gange med Θ kan denne matrix benyttes til at lave *prædiktion*, altså forudsige værdierne af y under den givne hypotese:

$$X\Theta = \begin{bmatrix} h_\Theta(x^{(1)}) \\ h_\Theta(x^{(2)}) \\ \vdots \\ h_\Theta(x^{(m)}) \end{bmatrix}$$

Hvis man ønsker hypoteser fra en anden familie af funktioner kan designmatricen ændres derefter.

Mindste kvadraters metode: Normalligningen

Givet en hypotese $h_\Theta(x)$ kan man vurdere hvor godt denne beskriver de observerede data. Dette kan gøres på flere måder, men her vil vi benytte *mindste kvadraters metode*. Vi søger her den hypotese der gør følgende *omkostnings-funktion* (en såkaldt *cost function*) mindst mulig:

$$J(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\Theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Summen kan tænkes på som kvadraten på længden af en vektor, der iflg. ovenstående må være givet ved $X\Theta - y$. Da $|v|^2 = v \cdot v$ kan vi igen bruge en formel fra ovenstående afsnit og få:

$$J(\Theta) = \frac{1}{2m}(X\Theta - y)^T(X\Theta - y) =$$

$$\frac{1}{2m}\left(\Theta^{\mathrm{T}}X^TX\Theta - \Theta^{\mathrm{T}}X^Ty - y^TX\Theta + y^2\right)$$

Vi ønsker at finde den Θ der gør $J(\Theta)$ mindst mulig. Derfor differentierer vi efter Θ ved at bruge resultaterne fra sidste afsnit:

$$\nabla_\Theta J(\Theta) = \frac{1}{2m}\nabla_\Theta\left(\Theta^{\mathrm{T}}X^TX\Theta - \Theta^{\mathrm{T}}X^Ty - y^TX\Theta + y^2\right) =$$

$$\frac{1}{2m}(2X^TX\Theta - X^Ty - X^Ty) = \frac{1}{m}(X^TX\Theta - X^Ty)$$

I minimum skal gradienten være nul:

$$\frac{1}{m}(X^T X\Theta - X^T y) = 0 \Leftrightarrow X^T X\Theta - X^T y = 0 \Leftrightarrow X^T X\Theta = X^T y \Leftrightarrow \Theta = (X^T X)^{-1} X^T y$$

Dette kaldes for *normalligningen*. Hvis $X^T X$ ikke er invertibel taler man om *perfekt multikollinearitet*.

Eksempler på brug af normalligningen

Kun konstantled

Her har hypotesefunktionen den simple form $h(x) = \Theta_0$. Designmatricen X er dermed blot en søjlevektoren af dimension $m \times 1$ fyldt med 1-taller. En sådan vektor vil vi benævne J_n (ikke at forveksle med kostfunktionen). Dermed bliver $X^T X$ simpelthen lig med en sum over n 1-taller, dvs:

$$X^T X = [n]$$

Og dermed:

$$(X^T X)^{-1} = \left[\frac{1}{m}\right]$$

Tilsvarende bliver $X^T y$ simpelthen en sum over alle y-værdier, altså $X^T y = [y_\bullet]$. Så i alt $\Theta_0 = \bar{y}$. Altså simpelthen gennemsnittet af y-værdierne.

Et lineært led, ingen konstantled – Et datapunkt

Her har hypotesefunktionen formen $h(x) = \Theta_1 x$, altså en ligefrem proportionalitet. Datapunktet vil vi blot kalde (x, y). I dette tilfælde er der ikke nogen søjle med 1-taller i designmatricen X. I stedet er $X = [x]$, så $X^T X = [x^2]$ og dermed er:

$$(X^T X)^{-1} = \left[\frac{1}{x^2}\right]$$

Derfor er $\Theta_1 = \frac{1}{x^2} xy = \frac{y}{x}$ som forventet. Her opstår der multikollinearitet hvis $x = 0$.

Et lineært led – to datapunkter

Her er hypotesefunktionen $h(x) = \Theta_0 + \Theta_1 x$. De to punkter vil vi betegne (x_1, y_1) og (x_2, y_2). Her bliver designmatricen:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

Dermed er:

$$X^T X = \begin{bmatrix} 1 & 1 \\ x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} = \begin{bmatrix} 2 & x_1 + x_2 \\ x_1 + x_2 & x_1^2 + x_2^2 \end{bmatrix}$$

For at invertere denne matrix beregnes determinanten:

$$\det(X^T X) = 2(x_1^2 + x_2^2) - (x_1 + x_2)^2 = 2x_1^2 + 2x_2^2 - x_1^2 - x_2^2 - 2x_1 x_2 = x_1^2 + x_2^2 - 2x_1 x_2$$

Med den sædvanlige notation $\Delta x = x_2 - x_1$ betyder det $\det(X^T X) = (\Delta x)^2$. Der er altså kollinearitet når $\Delta x = 0$. Nu kan den inverse skrives op:

$$(X^TX)^{-1} = \frac{1}{(\Delta x)^2}\begin{bmatrix} x_1^2 + x_2^2 & -(x_1 + x_2) \\ -(x_1 + x_2) & 2 \end{bmatrix}$$

Vi regner:

$$(X^TX)^{-1}X^T = \frac{1}{(\Delta x)^2}\begin{bmatrix} x_1^2 + x_2^2 & -(x_1 + x_2) \\ -(x_1 + x_2) & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ x_1 & x_2 \end{bmatrix} =$$

$$\frac{1}{(\Delta x)^2}\begin{bmatrix} x_1^2 + x_2^2 - x_1^2 - x_1x_2 & x_1^2 + x_2^2 - x_1x_2 - x_2^2 \\ -x_1 - x_2 + 2x_1 & -x_1 - x_2 + 2x_2 \end{bmatrix} = \frac{1}{(\Delta x)^2}\begin{bmatrix} x_2\Delta x & -x_1\Delta x \\ -\Delta x & \Delta x \end{bmatrix} = \frac{1}{\Delta x}\begin{bmatrix} x_2 & -x_1 \\ -1 & 1 \end{bmatrix}$$

Dermed bliver Θ givet med:

$$\Theta = (X^TX)^{-1}X^Ty = \frac{1}{\Delta x}\begin{bmatrix} x_2 & -x_1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{\Delta x}\begin{bmatrix} x_2y_1 - x_1y_2 \\ -y_1 + y_2 \end{bmatrix}$$

Heraf ses umiddelbart den kendte formel $\Theta_1 = \frac{y_2 - y_1}{x_2 - x_1}$. Det er sværere at genkende formlen for Θ_0. Vi forventer:

$$\Theta_0 = y_1 - \Theta_1x_1 = y_1 - \frac{y_2 - y_1}{\Delta x}x_1 = \left(1 + \frac{x_1}{\Delta x}\right)y_1 + \left(-\frac{x_1}{\Delta x}\right)y_2$$

Regn på første parentes:

$$1 + \frac{x_1}{\Delta x} = \frac{\Delta x + x_1}{\Delta x} = \frac{x_2}{\Delta x}$$

Udtrykket stemmer altså overens med de sædvanlige formler for to punkter.

Lineære transformationer i 2D

Rotationer

En rotation på θ er repræsenteret ved flg. matrix:

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Denne matrix er ikke symmetrisk, så egenværdierne er ikke nødvendigvis reelle:

$$\det(R_\theta - \lambda I) = 0 \Leftrightarrow (\cos\theta - \lambda)^2 + (\sin\theta)^2 = 0 \Leftrightarrow$$

$$(\cos\theta - \lambda)^2 = -(\sin\theta)^2 \Leftrightarrow \cos\theta - \lambda = \pm i\sin\theta \Leftrightarrow \lambda = \cos\theta \pm i\sin\theta = e^{\pm i\theta}$$

De tilhørende egenvektorer

Logistisk regression

Et datasæt $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$, hvor $x^{(i)} \in \mathbb{R}^n$ og $y^{(i)} \in \{0,1\}$ søges forklaret vha. følgende sammenhæng:

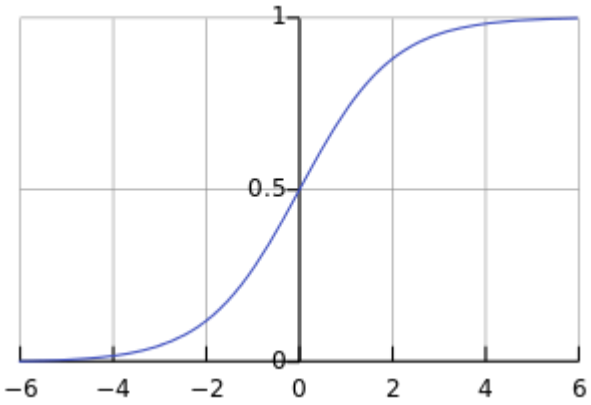
$$h_\Theta(x) = g(\Theta_0 + \Theta_1 x_1 + \cdots + \Theta_n x_n) = g(\Theta^T x)$$

Her sættes $x_0 = 1$ og g er den såkaldte *sigmoid-funktion*:

$$g(t) = \frac{1}{1 + e^{-t}}$$

(Funktionen kaldes også den *logistiske* funktion).

Der er altså tale om et klassifikationsproblem. Værdien af hypotesefunktionen kan tænkes som sandsynligheden for et givent x klassificeres som værdien 1.



Sigmoid-funktionen

Funktionen er defineret ved:

$$g(t) = \frac{1}{1 + e^{-t}}$$

Den afledte funktion er:

$$g'(t) = \frac{(1)' \cdot (1 + e^{-t}) - 1 \cdot (1 + e^{-t})'}{(1 + e^{-t})^2} = \frac{e^{-t}}{(1 + e^{-t})^2} = g(x) \cdot \frac{e^{-t}}{1 + e^{-t}}$$

Der gælder:

$$1 - g(t) = \frac{1 + e^{-t}}{1 + e^{-t}} - \frac{1}{1 + e^{-t}} = \frac{e^{-t}}{1 + e^{-t}}$$

Så i alt:

$$g'(t) = g(t) \cdot \big(1 - g(t)\big)$$

Likelihoodfunktion

Hvert y kan tænkes på som udfaldet af et Bernoulli-eksperiment. Ifølge hypotesen er p for dette eksperiment lig $g(\Theta^T x)$. Den samlede likelihood-funktion er altså:

$$L(\Theta) = \prod_{i=1}^{m} g\big(\Theta^T x^{(i)}\big)^{y^{(i)}} \cdot \Big(1 - g\big(\Theta^T x^{(i)}\big)\Big)^{1 - y^{(i)}}$$

Den tilhørende log-likelihood er:

$$l(\Theta) = -\log\big(L(\Theta)\big) = -\sum_{i=1}^{m} (y^{(i)} \cdot \log[\, g\big(\Theta^T x^{(i)}\big)] + (1 - y^{(i)}) \cdot \log[1 - g(\Theta^T x^{(i)})]$$

MLE-estimation

Vi søger nu den Θ der minimerer $l(\Theta)$. Ofte bruges omkostnings-funktionen $J(\Theta)$, der dybest set er samme funktion:

$$J(\Theta) = \frac{1}{m} l(\Theta)$$

For at bestemme minimum afledes funktionen:

$$\nabla_\Theta J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \cdot \frac{1}{g(\Theta^T x^{(i)})}\, g(\Theta^T x^{(i)}) \cdot \left(1 - g(\Theta^T x^{(i)})\right) x^{(i)} +$$

$$\left(1 - y^{(i)}\right) \cdot \frac{1}{1 - g(\Theta^T x^{(i)})} \left(1 - g(\Theta^T x^{(i)})\right) \cdot g(\Theta^T x^{(i)})(-x^{(i)})] =$$

$$-\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \cdot \left(1 - g(\Theta^T x^{(i)})\right) x^{(i)} - \left(1 - y^{(i)}\right) \cdot g(\Theta^T x^{(i)}) x^{(i)}] =$$

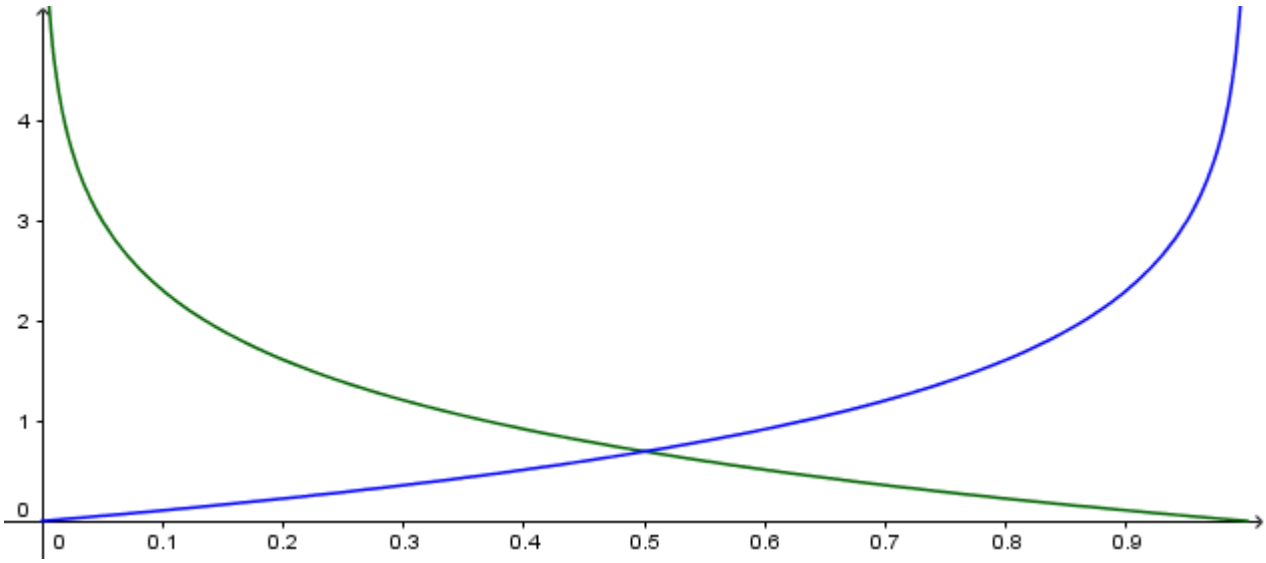
$$-\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\left(1 - h_\Theta(x^{(i)})\right) - \left(1 - y^{(i)}\right) h_\Theta(x^{(i)})\right) x^{(i)} =$$

$$-\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - h_\Theta(x^{(i)})\right) x^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\Theta(x^{(i)}) - y^{(i)}\right) x^{(i)}$$

Så ligningen der skal opfyldes ved minimum er:

$$\sum_{i=1}^{m}\left(h_\Theta(x^{(i)}) - y^{(i)}\right) x^{(i)} = 0$$

Konveksitet af $J(\Theta)$



De to funktioner $-\log(x)$ og $-\log(1-x)$ (se graferne ovenfor) er begge konvekse i intervallet $]0,1[$:

$$\frac{d^2}{dx^2}(-\log(x)) = \frac{d}{dx}\left(-\frac{1}{x}\right) = \frac{1}{x^2}$$

$$\frac{d^2}{dx^2}(-\log(1-x)) = \frac{d}{dx}\left(-\frac{1}{1-x}\cdot(-1)\right) = \frac{1}{(1-x)^2}$$

Begge udtryk er positive i dette interval, hvorfor funktionerne er konvekse. Hvis variablen i en konveks funktion transformeres lineært får man igen en af konveks funktion. Da $J(\Theta)$ er en linearkombination af sådanne transformationer af disse to funktioner med positive koefficienter er $J(\Theta)$ også konveks. Ud over at vise at løsningen til normalligningen virkelig er et minimum, gør dette problemet velegnet til at løse numerisk, f.eks. ved gradient descent.