Sandsynlighedsregning
13. forelæsning

Bo Friis Nielsen

Informatik og Matematisk Modellering

Danmarks Tekniske Universitet

2800 Kgs. Lyngby – Danmark

Email: bfn@imm.dtu.dk

Sandsynlighedsregningens grundregler

- $0 \leq P(A) \leq 1 \qquad P(\Omega) = 1 \qquad P(\emptyset) = 0$
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- Komplementær regel p.21: $P(A^c) = 1 - P(A)$
- Differens regel p.22, for $A \subset B$: $P(B \cap A^c) = P(B) - P(A)$
- Inkusion, eksklusion for 2 hændelser p.22:
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Betinget sandsynlighed for A givet B (partiel information) p.36: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Multiplikationsreglen p.37: $P(A \cap B) = P(B)P(A|B)$
- Gennemsnit af betingede sandsynligheder (loven om den totale sandsynlighed) p.41: (B_i er en partitionering), $P(A) = \sum_i P(B_i)P(A|B_i)$
- Bayes sætning p.49: (B_i er en partitionering): $P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$
- Uafhængighed p.42: $P(A|B) = P(A|B^c) \qquad (P(A \cap B) = P(A)P(B))$

Afsnit 2.1, 2.2 og 2.5

- Binomialfordelingen p.81 : $P(k\text{successer}) = \binom{n}{k} p^k q^{n-k}$
- Normalfordelingsapproksimationen p.99

$$P(\text{mellem } a \text{ og } b) = \Phi\left(\frac{b + \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right) - \Phi\left(\frac{a - \frac{1}{2} - n \cdot p}{\sqrt{n \cdot p \cdot q}}\right)$$

- Hypergeometrisk fordeling - stikprøve uden tilbagelægning - p.125

$$P(g \text{ gode og } b \text{ dårlige}) = \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

Afsnit 3.1 og 3.2

- Stokastiske variable: *udfald identificeres med relle tal*
  - ◇ Fordeling $P(X = x) \qquad \sum_x P(X = x) = 1$
  - ◇ Simultan(joint) fordeling $P(X = x, Y = y)$
  - ◇ Marginal fordeling $P_X(X = x) = \sum_y P(X = x, Y = y)$
  - ◇ Betinget fordeling $P(Y = y|X = x) = \frac{P(X=x,Y=y)}{P_X(X=x)}$
  - ◇ Uafhængighed $P(Y = y, X = x) = P_X(X = x)P_Y(Y = y)$
- Multinomialfordelingen
- Middelværdi $E(X) = \sum x \cdot P(X = x)$
- Forventningsværdi mere generelt $E(g(X)) = \sum g(x) \cdot P(X = x)$
- Linearitet $E(aX + bY + c) = aE(X) + bE(Y) + c$

Afsnit 3.3 og 3.4

- Varians/standardafvigelse
$$\mathrm{Var}(X) = E[(X - E(X))^2] = \sum_{\text{alle } x} (x - E(X))^2 P(X = x)$$
- Normalfordelingsapproksimation/den Centrale Grænseværdisætning
$$P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a)$$
- Markovs og Chebychevs uligheder for ekstreme udfald
$$P(X \geq a) \leq \frac{E(X)}{a} \qquad P(|X - E(X)| \geq k\mathrm{SD}(X)) \leq \frac{1}{k^2}$$
- Et udpluk af diskrete fordelinger
$$P(T = i) = (1-p)^{i-1}p \qquad P(T_r = i) = \binom{x+r-1}{r-1} p^{r-1}(1-p)^i p$$
- Indikatorfunktioner $I_A = \begin{cases} 1 & \text{hvis } A \text{ indtræffer} \\ 0 & \text{hvis } A^c \text{ indtræffer} \end{cases}$

Afsnit 3.5 og 4.1

Poissonfordelingen
$$P(X = x) = \frac{\mu^x}{x!}e^{-\mu}$$

- Kontinuerte stokastiske variable
  - ◇ Tæthed (density):
    $f(x) \geq 0, \quad \int f(x)\mathrm{d}x = 1, \quad P(X \in \mathrm{d}x) = f(x)\mathrm{d}x$
  - ◇ Middelværdi, varians (momenter): $\mathsf{E}(X) = \int x f(x)\mathrm{d}x$
    $\mathsf{E}(g(X)) = \int g(x)f(x)\mathrm{d}x$, $\mathsf{E}(X^k) = \int x^k f(x)\mathrm{d}x$
- Normalfordelingen: $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad Z = \frac{X-\mu}{\sigma}$
  - ◇ Den centrale grænseværdisætning

Afsnit 4.2, 4.3 og 4.4

- Poissonprocessen/eksponentialfordelingen $f(x) = \lambda e^{-\lambda x}$
  - ◇ Rate λ Overlevelsesfunktion: $G(x) = e^{-\lambda x}$
  - ◇ Gammafordelingen $f(x) = \lambda\frac{(\lambda x)^{r-1}}{\Gamma(r)}e^{-\lambda x}$ hvor $\Gamma(r) = (r-1)!$
    for r heltallig. For r heltallig $P(X \leq x) = 1 - \sum_{i=0}^{r-1}\frac{(\lambda x)^i}{i!}e^{-\lambda x}$
- Hazard-funktionen (hazard rate)
$$\lambda(t) = \lim_{\Delta t \to 0}\frac{P(T \in (t, t+\Delta t)|T > t)}{\Delta t} = \frac{f(t)}{G(t)}$$
- Variabelskift, $P(X \in [x, x+\mathrm{d}x]) = f_X(x)\mathrm{d}x$, $Y = g(X)$
$$P(Y \in [y, y+\mathrm{d}y]) = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|}$$

Afsnit 4.5 og 4.6

- (Kumulerede) fordelingsfunktion $P(X \leq x) = F(x)$
  - ◇ Fraktiler $F(x_p) = p$ - (specielt median $F(x_m) = 0.5$)
- Fordeling for maksimum og minimum $F_{max}(x) = \prod_{i=1}^n F_i(x)$ og $F_{min}(x) = 1 - \prod_{i=1}^n (1 - F_i(x))$
- Fordelinger af ordnede variable
$$f_{(k)}(x) = nf(x)\binom{n-1}{k-1}(F(x))^{k-1}(1-F(x))^{n-k}$$
- Betafordelingen
$$f(x) = \frac{1}{B(r,s)}x^{r-1}(1-x)^{s-1}$$
  med $B(r,s) = \int_0^1 x^{r-1}(1-x)^{s-1}\mathrm{d}x = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}\left(= \frac{(r-1)!(s-1)!}{r+s-1)}, \text{ for } (r,s) \text{ heltallige}\right)$

Dagens emner 5.1 og 5.2

- Ligefordeling med to variable (5.1): $P\left((x,y)\in C\right)=\frac{A(C)}{A(D)}$
- Simultane tætheder (5.2)

  $f(x,y)\mathrm{d}x\mathrm{d}y = P(x\le X\le x+\mathrm{d}x, y\le Y\le y+\mathrm{d}y), f(x,y)\ge 0$
- Simultan fordelingsfunktion (5.2)

$$F(x,y)=P(X\le x,Y\le y)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(x,y)\mathrm{d}x\mathrm{d}y$$

- Repetition af maximum/minimum - 4.5

  $F_{max}(x)=\prod_{i=1}^{n}F_i(x)\ G_{min}(x)=\prod_{i=1}^{n}G_i(x)$
- Simultan fordeling af minimum og maximum
- Repetition af "order statistics" - 4.6

$$f_{(k)}(x)=nf(x)\begin{pmatrix} n-1\\ k-1\end{pmatrix}(F(x))^{k-1}(1-F(x))^{n-k}$$

- Simultan fordeling af "order statistics" (Example 3. p.352-355)

$$f(x,y)=5!x(y-x)(1-y)\quad ,\qquad \text{for}\ \ 0<x<y<1$$

Afsnit 5.3 og 5.4

- Simultane kontinuerte fordelinger

  $P(X\in \mathrm{d}x, Y\in \mathrm{d}y)=f(x,y)\mathrm{d}x\mathrm{d}y$
- Uafhængige normalfordelte variable ($\sim normal(\mu,\sigma^2)=N(\mu,\sigma^2)$)
  - ◇ $f(x,y)=\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}$
  - ◇ $F_R(r)=P(\sqrt{X^2+Y^2}\le r)=1-e^{-\frac{1}{2}r^2}$ (Rayleigh)
  - ◇ $X\sim N(\lambda,\sigma^2), Y\sim N(\mu,\tau^2)\Rightarrow X+Y\sim N(\lambda+\mu,\sigma^2+\tau^2)$
  - ◇ For $X_i\sim N(0,1)$ er $Y=\sum_{i=1}^{n}X_i^2\quad \chi^2$-fordelt (gamma).
- Operationer med stokastiske variable
  - ◇ $f_Z(z)\mathrm{d}z=P(Z=X+Y\in \mathrm{d}z)=\int_{-\infty}^{\infty}f(x,z-x)\mathrm{d}x\ \ \mathrm{d}z$
  - ◇ $f_Z(z)\mathrm{d}z=P\left(Z=\frac{Y}{X}\in \mathrm{d}z\right)=\int_{-\infty}^{\infty}|x|f(x,zx)\mathrm{d}x\ \ \mathrm{d}z$

Afsnit 6.1 og 6.2

- Betingede diskrete fordelinger $P(Y=y|X=x)=\frac{P(X=x,Y=y)}{P(X=x)}$
  - ◇ $P(Y=y)=\sum_x P(X=x)P(Y=y|X=x)$
- Betinget middelværdi (det diskrete tilfælde)
  - ◇ $E(Y|X=x)=\sum_y yP(Y=y|X=x)$
  - ◇ $E(Y|X)$ en stokastisk variabel
  - ◇ $E(Y)=E_X(E_Y(Y|X))=\sum_x P(X=x)E(Y|X=x)$
  - ◇ $E(g(X,Y)|X=x)=E(g(x,Y)|X=x)$
  - ◇ $E(XY|X)=XE(Y|X)$

Afsnit 6.3 og 6.4

- Betingede kontinuerte fordelinger $f(y|X=x)=\frac{f(x,y)}{f_X(x)}$
- Integralformel for gennemsnit af betingede sandsynligheder

  $P(A)=\int P(A|X=x)f_X(x)\mathrm{d}x$
- Betinget forventning $E(Y)=E(E(Y|X))$
- Kovarians/korellation

$$Cov(X,Y)=E[(X-E(X))(Y-E(Y))]=E(XY)-E(X)E(Y)$$

$$Corr(X,Y)=\frac{Cov(X,Y)}{SD(X)SD(Y)}$$

- (X,Y) uafhængige $\Rightarrow Corr(X,Y)=0$
- Varians af sum af variable

  $Var\left(\sum_{k=1}^{N}X_k\right)=\sum_{k=1}^{n}Var(X_k)+2\sum_{1\le j<k\le n}Cov(X_j,X_k)$
- Bilinearitet af kovariansen

  $Cov\left(\sum_{i=1}^{n}a_iX_i,\sum_{j=1}^{m}b_jY_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jCov(X_i,Y_j)$

Afsnit 6.5

- Den standardiserede bivariate normalfordeling

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

$$Y = \rho X + \sqrt{1-\rho^2} Z, \qquad X, Z \text{ uafhængige standard NF}$$

- $V = \mu_V + \sigma_V X$, $W = \mu_W + \sigma_W Y$ er bivariat normalfordelt med $E(V) = \mu_V$, $E(W) = \mu_W$, $Var(V) = \sigma_V^2$, $Var(W) = \sigma_W^2$, samt $Cov(V,W) = \rho\sigma_V\sigma_W$.

$$V = \sum_i a_i Z_i, \quad W = \sum_i b_i Z_i \text{ med } Z_i \sim normal(\mu_i, \sigma_i^2) \text{ uafhængige}$$

$$\Rightarrow V \sim normal\left(\sum a_i\mu_i, \sum a_i^2\sigma_i^2\right), W \sim normal\left(\sum b_i\mu_i, \sum b_i^2\sigma_i^2\right),$$

$$Cov(V,W) = \sum_i a_i b_i \sigma_i^2 \qquad (V,W) \text{ bivariat normalfordelte}$$