INSTITUT FOR MATEMATISKE FAG

AALBORG UNIVERSITET
FREDRIK BAJERS VEJ 7 G ▪ 9220 AALBORG ØST
URL: www.math.aau.dk
E-mail: bjh@math.aau.dk

Tlf.: 96 35 89 27
Fax: 98 15 81 29

Dataanalyse

Sandsynlighed og stokastiske variable

Udfaldsrum S: mængden af alle mulige udfald for et eksperiment.

Hændelse A: delmængde af S.

Sandsynlighedsmål $\mathbb{P}$: Afbilder hændelser ind i $[0, 1]$. Opfylder

$$0 \leq \mathbb{P}(A) \leq 1$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B),\ A \cap B = \emptyset$$
$$\mathbb{P}(S) = 1$$

Det giver blandt andet følgende resultater

$$\mathbb{P}(\emptyset) = 0$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

Betinget sandsynlighed:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},\ \mathbb{P}(B) > 0$$

A og B er uafhængige hvis

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Uafhængighed medfører

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

En klassedeling $E_1, E_2, \ldots, E_k$ opfylder $E_i \cap E_j = \emptyset$ og $E_1 \cup E_2 \cup \cdots \cup E_k = S$.

Regel om total sandsynlighed

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|E_i)\mathbb{P}(E_i).$$

Bayes' formel:

$$\mathbb{P}(E_j|A) = \frac{\mathbb{P}(A|E_j)\mathbb{P}(E_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|E_j)\mathbb{P}(E_j)}{\sum_{i=1}^{k} \mathbb{P}(A|E_i)\mathbb{P}(E_i)}$$

Stokastisk variabel

$X : S \mapsto \mathbb{R}$
X diskret: $X \in D = \{x_1, x_2, \ldots, x_n, \ldots\} \subset \mathbb{R}$
X kontinuert: $X \in C \subseteq \mathbb{R}$ (sammenhængende delmængder)

Diskret stokastisk variabel

Sandsynlighedsfunktion $p(x)$:

$$p(x) = \mathbb{P}(X = x), x \in D$$

Fordelingsfunktion $F(x)$:

$$F(x) = \mathbb{P}(X \leq x) = \sum_{y \in D; y \leq x} p(y),\ x \in \mathbb{R}$$

Middelværdi $\mathbb{E}[X]$ (massemidtpunkt for sandsynlighedsmassen)

$$\mathbb{E}[X] = \sum_{x \in D} xp(x) = \mu_X$$

X siges at have endelig middelværdi, hvis

$$\mathbb{E}[X] = \sum_{x \in D} |x|p(x) < \infty.$$

Forventet værdi af funktion $g(X)$:

$$\mathbb{E}[g(X)] = \sum_{x \in D} g(x)p(x)$$

Variansen af X:

$$Var(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2 = \sigma_X^2$$

Spredningen/standardafvigelsen for X:

$$\sigma_X = \sqrt{Var(X)}.$$

Eksempel:**Uniform fordeling** (Terningkast). $D = \{1, 2, 3, 4, 5, 6\}$, $p(x) = 1/6, x \in D$. $\mathbb{E}[X] = 3/7$, $Var(X) = 35/12$

Eksempel: **Binomialfordelingen** n uafhængige forsøg. p er sandsynlighed for succes i et forsøg. X antal succeser. $D = \{0, 1, \ldots, n\}$,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$\mathbb{E}[X] = np$, $Var(X) = np(1-p)$.

Eksempel: **Poisson fordeling** Gennemsnitlig antal hændelser i vilkårligt område O er $\lambda|O|$. X er antal hændelser et givent område af størrelse 1.

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$\mathbb{E}[X] = \lambda$, $Var(X) = \lambda$

Kontinuert stokastisk variabel

$X \in C \subseteq \mathbb{R}$ kan altid udvides til $X \in \mathbb{R}$ ved at sætte $\mathbb{P}(X \notin C) = 0$. For X kontinuert kan fordelingsfunktionen skrives som

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

hvor $f(y)$ er tæthedsfunktion og opfylder:

$$f(y) \geq y$$

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

Hvis $X \in C$ sættes $f(y) = 0$, $y \notin C$ og så er

$$\int_{-\infty}^{\infty} f(y)dy = \int_{C} f(y)dy = 1$$

$$\mathbb{P}(X = x) = 0, x \in \mathbb{R}$$

Middelværdien givet ved

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \mu_X$$

og X har endelig middelværdi hvis

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} |x|f(x)dx < \infty.$$

og for generel funktion $g(x)$:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

Eksempel: **Uniform fordeling** $X \in [a, b]$, $f(y) = c, y \in [a, b]$:

$$\int_{a}^{b} cdy = c(b - a) \Rightarrow c = 1/(b - a)$$

$$\mathbb{E}[X] = \frac{b + a}{2}, Var(X) =$$

Eksempel: **Normalfordelingen** $X \sim N(\mu, \sigma^2)$, hvis $X \in \mathbb{R}$ og tæthedsfunktionen er

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

$$\mathbb{E}[X] = \mu, Var(X) = \sigma^2$$

Flerdimensionelle fordelinger

Diskrete variable

$(X, Y) \in D = D_X \times D_Y \subset \mathbb{R}^2$. Simultan sandsynlighedsfunktion

$$p(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \cap Y = y), x \in D_X, y \in D_Y$$

Marginale sandsynlighedsfunktioner

$$p_X(x) = P(X = x) = \sum_{y \in D_Y} p(x, y), p_Y(y) = \sum_{x \in D_X} p(x, y)$$

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y) \in D} g(x, y)p(x, y)$$

Eksempelvis

$$\mathbb{E}[g(X)] = \sum_{(x,y) \in D} g(x)p(x, y) = \sum_{x \in D_x} g(x) \sum_{y \in D_Y} p(x, y) = \sum_{x \in D_X} p_X(x)$$

som vi ved fra det en-dimensionelle.

X og Y er uafhængige hvis $p(x, y) = p_X(x)p_Y(y)$ for alle $(x, y) \in D$

Eksempel: Slag med to terninger X maximum øjne, Y summen af øjnene. $D_X = \{1,2,3,4,5,6\}$, $D_Y = \{2,3,4,\ldots,11,12\}$

$$p(1,2) = 1/36, p(1,y) = 0, y > 2$$

$$p(2,2) = 0, p(2,3) = 2/36, p(2,4) = 1/36, p(2,y) = 0, y > 4$$

$$p(3,2) = p(3,3) = 0, p(3,4) = 2/36, p(3,5) = 2/36, p(3,6) = 1/36, p(3,y) = 0, y > 6$$

Find selv resten. Er X og Y uafhængige?

Kontinuerte variable

$(X,Y) \in C \subseteq \mathbb{R}^2$. Simultan fordelingsfunktion

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \int\int_{(u,v)\in C: u\leq x, v\leq y} f(u,v)dudv$$

$f(u,v)$ er den simultane tæthedsfunktion. Marginale tætheder findes ved

$$f_X(x) = \int_{v\in C_x} f(x,v)dv, f_Y(y) = \int_{u\in C_y} f(u,y)du$$

X og Y er uafhængige hvis tæthederne opfylder $f(x,y) = f_X(x)f_Y(y)$.

Eksempel: Lad $f(u,v) = 4u^2, 0 \leq u \leq 1, 0 \leq v \leq x$. Så ses, at

$$f_X(x) = \int_0^x 4x^2 dv = 4x^3, 0 \leq x \leq 1$$

geometriske overvejelser giver

$$f_Y(y) = \int_y^1 4u^2 du = \frac{4}{3}(1-y^3), 0 \leq y \leq 1$$

X og Y er oplagt afhængige.

Middelværdi for en funktion $g(X,Y)$ findes ved

$$\mathbb{E}[g(X,Y)] = \int\int_C g(u,v)f(u,v)dvdu$$

Højere dimensioner

Alt ovenstående kan udvides simpelt fra to dimensioner til n dimensioner, hvor vi ser på stokastiske vektor $X_1, X_2, \ldots X_n$.

Hovedsætnigen for middelværdi er følgende:

Sætning 1 *Lad $(X_1, X_2, \ldots, X_n)$ være en stokastisk vektor med simultan ssh. fkt. $p(x_1, \ldots, x_n)$. Lad $g : \mathbb{R}^n \mapsto \mathbb{R}$. Så*

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \sum_{(x_1,\ldots x_n)} g(x_1, \ldots, x_n)p(x_1, \ldots, x_n)$$

og er endelig, hvis $\mathbb{E}[|g(X_1, \ldots, X_n)|] < \infty$. Hvis $(X_1, X_2, \ldots, X_n)$ er kontinuert med tæthedsfunktion $f(x_1, \ldots, x_n)$ er den givet ved

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n)f(x_1, \ldots, x_n)dx_1 \cdots dx_n$$

Middelværdi, varians og korrelation

Regneregler for middelværdi

- $\mathbb{P}(X = c) = 1$ medfører $\mathbb{E}[X] = c$.
- $\mathbb{E}[1_A(X)] = \mathbb{P}(X \in A)$, hvor $1_A(X) = 1$, hvis $X \in A$ og 0 ellers.
- $X_1, \ldots, X_n$ stok. var. og $a_1, \ldots, a_n$ reelle konstanter medfører

$$\mathbb{E}[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

- $X_1, \ldots, X_n$ uafhængige stok. var.

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$$

- $X_1 \leq X_2$ medfører $\mathbb{E}[X_1] \leq \mathbb{E}[X_2]$.
- $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

Momenter

Det k'te moment for en stok. var. er $\mathbb{E}[X^k]$. Det k'te centrale moment er $\mathbb{E}[(X - \mu_X)^k]$, hvor $\mu_X = \mathbb{E}[X]$.

Varians

Varians er navnet for det andet centrale moment, som beskriver hvor meget en stok variabel spreder sig omkring sin middelværdi

$$Var(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2 \equiv \sigma_X^2$$

hvoraf det ses at variansen er endelig hvis $\mathbb{E}[X^2] < \infty$. *Spredningen* eller *standard afvigelsen* er givet ved $\sqrt{Var(X)} = \sigma_X$. Vi har følgende resultater:

- $\mathbb{P}(X = c) = 1$ er ækvivalent med $Var(X) = 0$.
- a, b reelle konstanter: $Var(a + bX) = b^2 Var(X)$
- $X_1, \ldots, X_n$ uafhængige, så gælder

$$Var(X_1 \pm X_2 \pm \cdots \pm X_n) = \sum_{i=1}^{n} Var(X_i)$$

Kovarians

Kovariansen mellem to stokastiske variable X, Y er givet ved

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y.$$

Kovariansen er defineret, hvis begge variable har endelig varians. Vi har følgende regneregler:

- $$Var(X_1 + \cdots + X_n) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} Cov(X_i, X_j).$$

- $Cov(a + bX, c + dY) = bdCov(X, Y)$
- $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$
- $Cov(X, Y) = Cov(Y, X)$
- X, Y uafhængige medfører $Cov(X, Y) = 0$.

Korrelation

Korrelationen er en enhedsfri udgave af kovariansen, eller sagt på en anden måde kovariansen mellem de standardisered variable. Den er givet ved

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

- $-1 \leq Corr(X, Y) \leq 1$
- $Corr(X, Y) = \pm 1$ medfører der eksisterer α, β, så $X = \alpha + \beta Y$, hvor $sign(\beta) = sign(Corr(X, Y))$.

Med venlig hilsen
Bjarne Højgaard