Course 02403 Introduktion til Matematisk Statistik

Forelæsning 3: Regneregler, uafhængighed og simulation

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.dk

Oversigt

Oversigt

Tæthedsfunktion (probability density function (pdf))

- Tæthedsfunktionen for en stokastisk variabel betegnes ved $f(x)$
- $f(x)$ siger noget om hyppigheden af udfaldet x for den stokastiske variabel X.
- Der gælder:

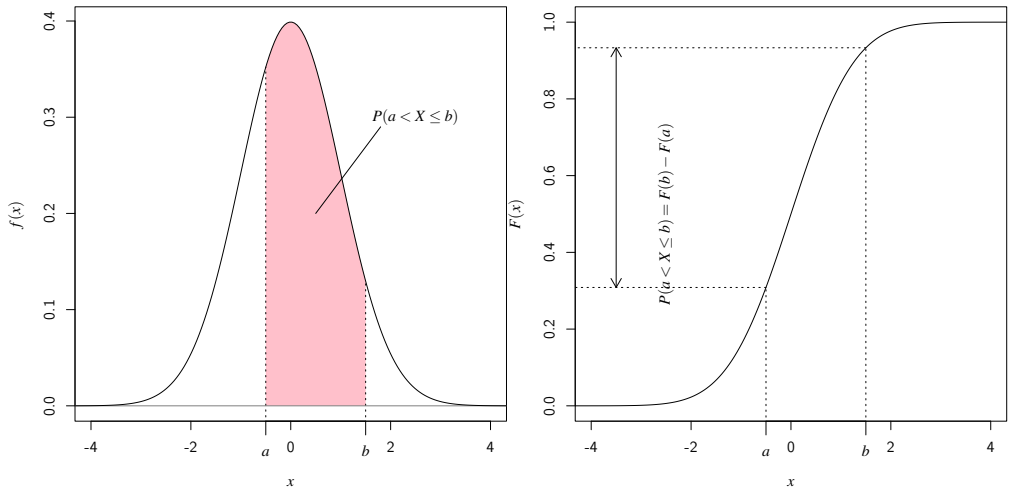
$$f(x) \geq 0 \quad \text{for alle mulige } x$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Fordelingsfunktionen svarer til den kumulerede tæthedsfunktion:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

Tæthedsfunktion for en kontinuert variabel



Middelværdi (mean) af en kontinuert stokastisk variabel

Middelværdien af en kontinuert stokastisk variabel

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Variansen af en kontinuert stokastisk variabel:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

Oversigt

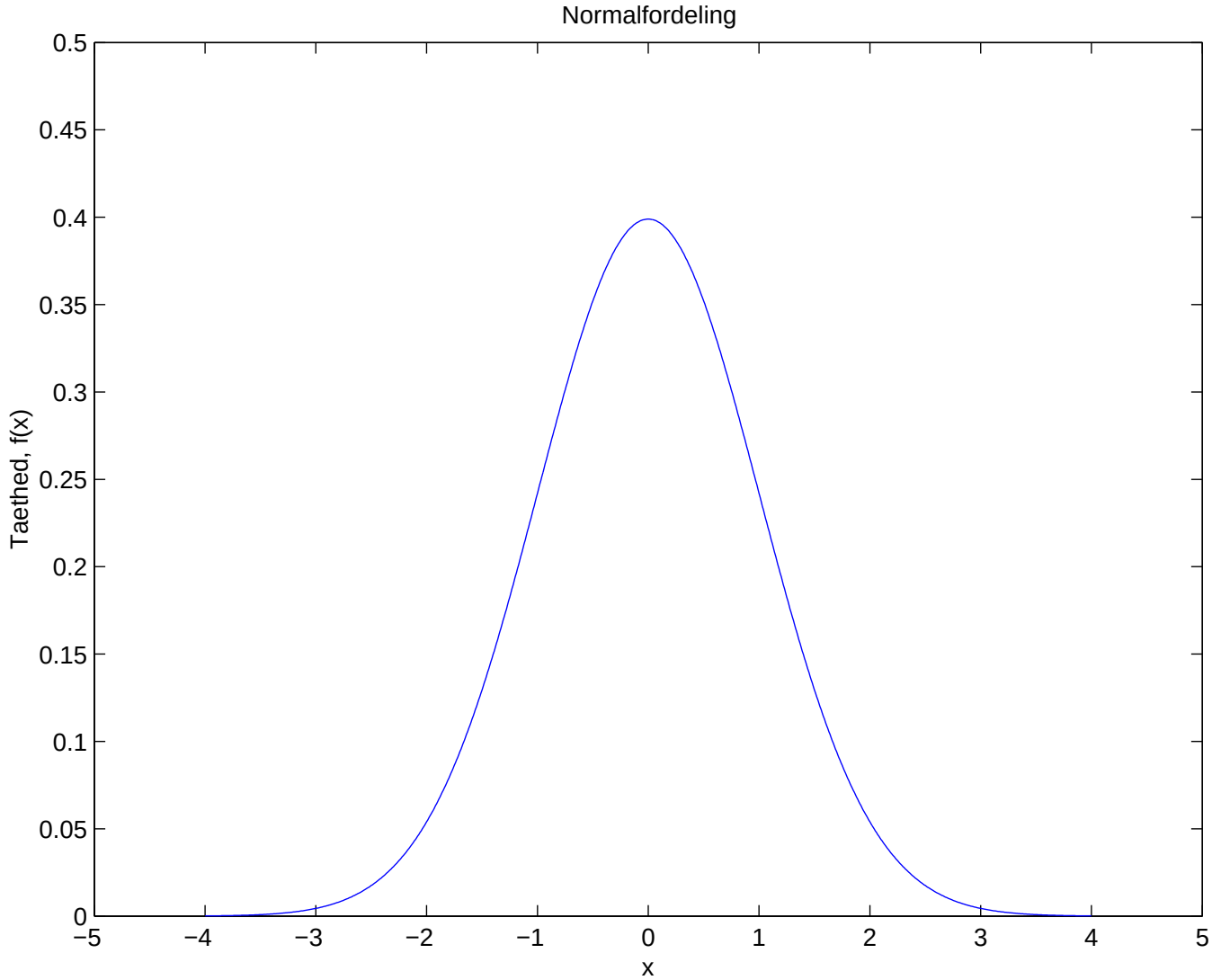
Konkrete statistiske fordelinger

- Der findes en række statistiske fordelinger, som kan bruges til at beskrive og analysere forskellige problemstillinger med

Vi har set en række kontinuerte fordelinger

- Uniform fordelingen
- Eksponential fordelingen
- Normal fordelingen
- Log-Normal fordelingen

Normalfordelingen



Normal fordelingen

Skrivemåde:

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

Tæthedsfunktion:

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

Middelværdi:

$$\mu = \mu$$

Varians:

$$\sigma^2 = \sigma^2$$

Normal fordelingen

En standard normal fordeling:

$$Z \sim N(0, 1^2)$$

En normalfordeling med middelværdi 0 og varians 1.

Standardisering:

En vilkårlig normal fordelt variabel $X \sim N(\mu, \sigma^2)$ kan standardiseres ved at beregne

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

Linear kombination af Normale Stokastiske variable

Hvis $X_i \sim N(\mu_i, \sigma_i^2)$, $(i = 1, \ldots, n)$ er uafhængige stokastiske variable så er

$$Y = a_0 + a_1 X_1 + \cdots + a_n X_n$$

også normal fordelt.

Kontinuerte fordelinger i R

R	Betegnelse
`norm`	Normalfordelingen
`unif`	Den uniforme fordeling
`lnorm`	Log-normalfordelingen
`exp`	Exponentialfordelingen

- `d` Tæthedsfunktion $f(x)$ (probability density function).
- `p` Fordelingsfunktion $F(x)$ (cumulative distribution function).
- `q` Fraktil (quantile) i fordeling.
- `r` Tilfældige tal fra fordelingen.

Oversigt

Regneregler for stokastiske variable

(Gælder BÅDE kontinuert og diskret)

X er en stokastisk variabel.

Vi antager at a og b er konstanter Da gælder:

Middelværdi-regel:

$$E(aX+b) = aE(X)+b$$

Varians-regel:

$$Var(aX+b) = a^2 Var(X)$$

Eksempel 1

X er en stokastisk variabel.

En stokastisk variabel X har middelværdi 4 og varians 6.

Spørgsmål:

Beregn middelværdi og varians for $Y = -3X + 2$

Regneregler for stokastiske variable

$X_1, \ldots, X_n$ er stokastiske variable

Da gælder:

Middelværdi-regel:

$$E(a_1X_1 + a_2X_2 + .. + a_nX_n)$$
$$= a_1E(X_1) + a_2E(X_2) + .. + a_nE(X_n)$$

Hvis X_i og X_j $(i \neq j)$ er uafhængige gælder desuden

Varians-regel:

$$Var(a_1X_1 + a_2X_2 + .. + a_nX_n)$$
$$= a_1^2Var(X_1) + .. + a_n^2Var(X_n)$$

Eksempel 2

Flypassager-planlægning

Vægten af passagerer på en flystrækning antages normalfordelt $X \sim N(70, 10^2)$.

Et fly, der kan tage 55 passagerer, må max. lastes med 4000 kg (kun passageres vægt betragtes som last).

Spørgsmål:

Beregn sandsynligheden for at flyet bliver overlastet

Eksempel 3

- Lad $X_1, \ldots, X_n$ være uafhængige identisk fordelte (i.i.d.) stokastiske variable, hvad er $E(\overline{X})$ og $Var(\overline{X})$?
- Hvis yderligere X_i er normalfordelte hvilken fordeling følger da $\overline{X}$?

Oversigt

Motivation

- Mange relevant beregningsstørrelser ("computed features") har komplicerede samplingfordelinger:
  - Et trimmed gennemsnit
  - Medianen
  - Fraktiler generelt, dvs. f.eks. også IQR$= Q_3 - Q_1$
  - Variationkoefficienten
  - Enhver ikke-lineær funktion af en eller flere input variable
  - (Spredningen)
- Data/populations fordelingen kan være ikke-normal, hvilket komplicerer den statistiske teori for selv en simpel gennemsnitsberegning
- Vi kan HÅBE på the magic of CLT (Central Limit Theorem)
- MEN men: Vi kan aldrig være helt sikre på om det er godt nok - simulering kan gøre os mere sikre!
- Kræver: Brug af computer - R er et super værktøj til dette!

Hvad er simulering egentlig?

- (Pseudo)tilfældige tal genereret af en computer
- En tilfældighedsgenerator er en algoritme der kan generere x_{i+1} ud fra x_i

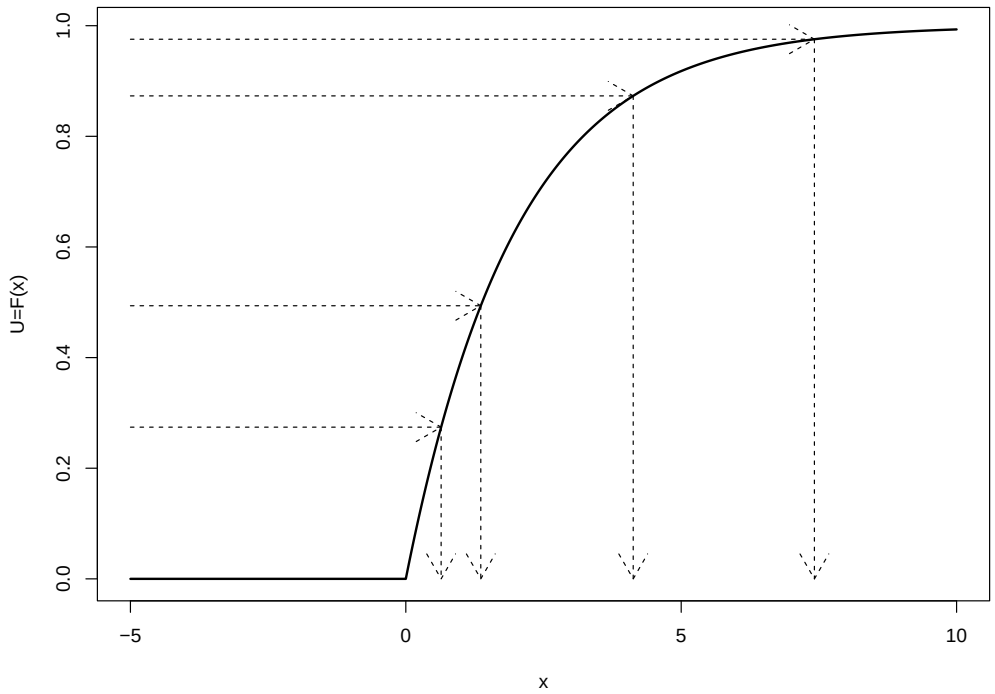
$$x_{i+1} = f(x_i)$$

- Sekvensen af tal "ser tilfældige ud"
- Kræver en "start" - kaldet "seed" .(Bruger typisk uret i computeren)
- Grundlæggende simuleres den uniforme fordeling, og så bruges:

Hvis $U \sim \text{Uniform}(0,1)$ og F er en fordelingsfunktion for en eller anden sandsynlighedsfordeling, så vil $F^{-1}(U)$ følge fordelingen givet ved F

Eksempel: Exponentialfordelingen med $\lambda = 0.5$:

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



I praksis i R

De forskellige fordelinger er gjort klar til simulering:

`rbinom`	Binomialfordelingen
`rpois`	Poissonfordelingen
`rhyper`	Den hypergeometriske fordeling
`rnorm`	Normalfordelingen
`rlnorm`	Lognormalfordelingen
`rexp`	Eksponentialfordelingen
`runif`	Den uniforme(lige) fordeling
`rt`	t-fordelingen
`rchisq`	χ^2-fordelingen
`rf`	F-fordelingen

Eksempel 4: Vejen til arbejde

Jeg cykler på arbejde, turen kan deles op i nogle enkeltdele

- Cykeltid (hvor cyklen bevæger sig) $X_c \sim N(16,1)$
- Aflevere min datter i skole $X_d \sim Exp(1/2)$ $(E[X_d] = 2)$
- Der er 4 lyskryds $X_{l,i} \sim Exp(2)$ $(E[X_{l,i}] = 1/2)$
- På 1 ud af 30 ture punkterer jeg
  - Hvis jeg punkterer er lappetiden fordelt som $X_{la} \sim N(8,2)$

Hvad er

- Den samlede middeltid i minutter?
- Variansen af den samlede tid?
- Hvad er sandsynligheden for at det tager mere en 30 minutter at komme på arbejde?
- Hvad er den samlede fordeling?

The simulation approach has a number of crucial advantages:

1. It offers a simple tool to compute many other quantities than just the standard deviation (the theoretical derivations of such other quantities could be much more complicated than what was shown for the variance here)
2. It offers a simple tool to use any other distribution than the normal, if we believe such better reflect reality.
3. It does not rely on any linear approximations of the true non-linear relations.

Oversigt

2-dimensionale diskrete stokastiske variable

Tæthedsfunktionen for en 2-dimensional stokastisk variabel $[X, Y]$ er givet

$$f(x,y) = P(X = x, Y = y)$$

Der gælder at

$$f(x,y) \geq 0; \quad \sum_x \sum_y f(x,y) = 1$$

Desuden bruger vi

$$f_X(x) = \sum_y f(x,y); \quad f_Y(y) = \sum_x f(x,y)$$

f_X og f_Y kaldes de marginale fordelinger.

Eksempel 5:

To kast med en ærlig mønt (dvs. $p = \frac{1}{2}$), definer den stokastiske variable X_i ved $X_i = 1$ $i = \{1,2\}$ hvis krone og $X_i = 0$ ellers, vi har nu udfaldene

$$(X_1, X_2) = \begin{cases} (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{cases}$$

hver med sandsynligheden $P(X_1 = l, X_2 = k) = \frac{1}{4}$, $l,k = \{0,1\}$. De marginale tætheder bliver

$$f_{X_1}(x) = \sum_{x_2=0}^{1} P(X_1 = x_1, X_2 = x_2) = \frac{1}{2}; \quad x = \{0,1\}$$

2-dimensionale kontinuerte stokastiske variable

Tæthedsfunktionen for en 2-dimensional kontinuert stokastisk variabel $[X, Y]$ er givet

$$f(x,y)$$

Der gælder at

$$f(x,y) \geq 0; \quad \int\int f(x,y)dxdy = 1$$

Desuden bruger vi

$$f_X(x) = \int f(x,y)dy; \quad f_Y(y) = \int f(x,y)dx$$

f_X og f_Y kaldes de marginale fordelinger.

2-dimensionale kontinuerte stokastiske variabele

For en kontinuert stokastisk vektor $[X, Y]$ findes middelværdi og varians for X ved

$$\mu_X = \int\int x f(x,y) dxdy$$
$$\sigma_X^2 = \int\int (x-\mu_X)^2 f(x,y) dxdy$$

Kovariansen og korrelation mellem X og Y er givet ved

$$\begin{aligned} Cov(X,Y) &= \int\int (x-\mu_X)(y-\mu_Y) f(x,y) dxdy \\ &= E[(x-\mu_X)(y-\mu_Y)] \\ Cor(X,Y) &= \frac{Cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \end{aligned}$$

Regneregler for stokastiske variable

(Gælder BÅDE kontinuert og diskret) Lad X og Y være stokastiske variable med $Z_1 = a_0 + a_1 X + a_2 Y$ og $Z_2 = b_0 + b_1 X + b_2 Y$ da gælder

$$\begin{aligned} E(Z_1) &= a_0 + a_1 E(X) + a_2 E(Y) \\ Cov(Z_1, Z_2) &= a_1 b_1 V(X) + a_2 b_2 V(Y) + (a_1 b_2 + a_2 b_1) Cov(X, Y) \\ V(Z_1) &= a_1^2 V(X) + a_2^2 V(Y) + 2 a_1 a_2 Cov(X, Y) \end{aligned}$$

Regneregler for stokastiske variable

$X_1, \ldots, X_n$ er stokastiske variable

Da gælder:

Varians-regel:

$$
\begin{aligned}
Z =& a_1 X_1 + a_2 X_2 + .. + a_n X_n \\
Var(Z) =& a_1^2 Var(X_1) + .. + a_n^2 Var(X_n) + 2a_1 a_2 Cov(X_1, X_2) + ... + \\
& 2a_1 a_n Cov(X_1, X_n) + 2a_2 a_3 Cov(X_2, X_3) + ... + \\
& 2a_{n-1} a_n Cov(X_{n-1}, X_n)
\end{aligned}
$$

Eksempel 6

Hvis $X \sim N(2,3)$ og $Y \sim N(0,1)$ (X og Y uafhængige), hvad er så $\mathrm{E}(Z)$, $\mathrm{V}(Z)$ og $\mathrm{Cov}(X,Z)$ når $Z = X + 2Y$?

Uafhængighed

De diskrete stokastiske variable (X og Y) er uafhængige hvis

$$f(x,y) = P(X = x, Y = y) = P(X = x)P(Y = y)$$

De kontinuerte stokastiske variable (X og Y) er uafhængige hvis

$$f(x,y) = f_X(x)f_Y(y)$$

Hvis 2 stokastiske variable er uafhængige er de også ukorrelerede ($\mathrm{Cor}(X,Y) = 0$), men det modsatte er ikke nødvendigvis tilfældet.

Eksempel 5 forts.:

To kast med en ærlig mønt (dvs. $p = \frac{1}{2}$), definer den stokastiske variable X_i ved $X_i = 1$ $i = \{1,2\}$ hvis krone og $X_i = 0$ ellers, vi har nu udfaldene

$$(X_1, X_2) = \begin{cases} (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{cases}$$

hver med sandsynligheden $P(X_1 = l, X_2 = k) = \frac{1}{4}$, $l,k = \{0,1\}$. Da $P(X_1 = k) = P(X_2 = l) = \frac{1}{2}$ har vi at

$$P(X_1 = l, X_2 = k) = P(X_1 = l)P(X_2 = k) = \frac{1}{4}; \quad l,k = \{0,1\}$$

og dermed er X_1 og X_2 uafhængige.

Eksempel 5 forts:

Definer X_1 og X_2 som ovenfor, men vi observere nu kun X_1 og $Z = X_1 + X_2$ vi har nu udfaldene

$$(X_1, Z) = \begin{cases} (0,0) \\ (0,1) \\ (1,1) \\ (1,2) \end{cases}$$

hver med sandsynligheden $\frac{1}{4}$. Dermed har vi også $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$, $P(Z = 0) = P(Z = 2) = \frac{1}{4}$ og $P(Z = 1) = \frac{1}{2}$, vi får nu eksempelvis

$$P(X_1 = 1, Z = 2) = \frac{1}{4} \neq P(X_1 = 1)P(Z = 2) = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$$

og dermed er X_1 og Z ikke uafhængige (hvilket passer godt med vores intuition).

Eksempel 5 forts.: Middelværdi, varians og kovarians

Vi har

$$E[X] = \frac{1}{2}; \quad E[Z] = 1$$
$$V[X] = \frac{1}{4}; \quad V[Z] = \frac{1}{2}$$

Kovariansen fås til

$$\begin{aligned} Cov[X,Z] &= \sum_{x=0}^{1} \sum_{z=x}^{x+1} \left(x - \frac{1}{2}\right)(z-1) P(X=x, Z=z) \\ &= -\frac{1}{2}(0-1)\cdot\frac{1}{4} + \left(-\frac{1}{2}\right)(1-1)\cdot\frac{1}{4} + \left(\frac{1}{2}\right)0\cdot\frac{1}{4} + \left(\frac{1}{2}\right)(2-1)\cdot\frac{1}{4} \\ &= \frac{1}{8} + 0 + 0 + \frac{1}{8} = \frac{1}{4} \end{aligned}$$

Middelværdi vektor og Varians-Kovarians matricen

Lad $\boldsymbol{X} = [X_1, ..., X_n]$ være en stokastisk vektor, da er middelværdivektor og varians-kovarians matricen defineret ved

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_n^2 \end{bmatrix}$$

med $\sigma_{ij} = Cov(X_i, X_j)$ og $\sigma_i^2 = V[X_i]$.

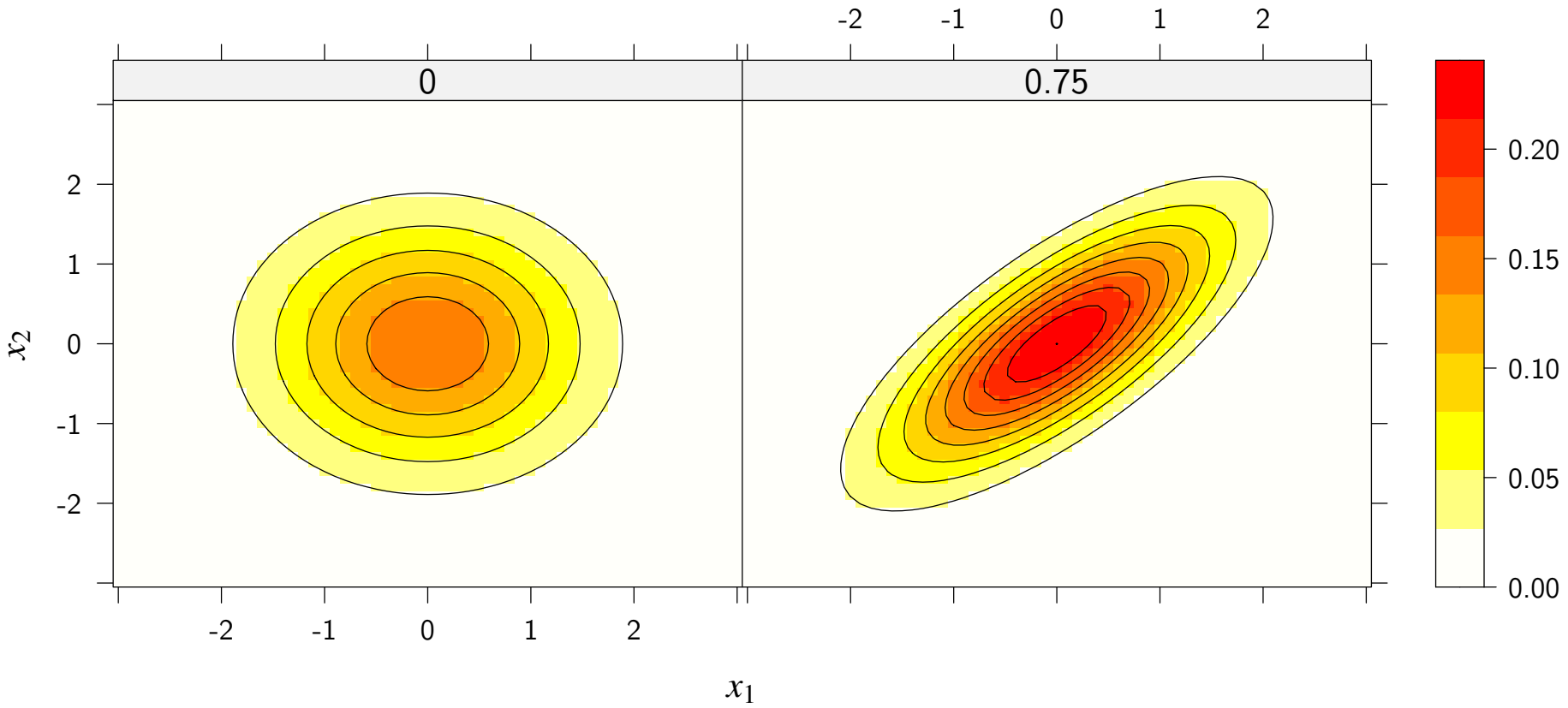
Multivariat normalfordeling

En stokastisk vektor $\boldsymbol{X} = [X_1, ..., X_n]$ siges at følge en multivariat normalfordeling hvis tæthedsfunktionen er givet ved

$$f(x_1, .., x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)}$$

og vi skriver $\boldsymbol{X} \sim N(\mu, \Sigma)$ (Σ er positiv definit).

Eksempel: Bivariat Normal fordeling



Konstruktion af en multivariat normalfordeling

Hvis $\boldsymbol{Z} = [Z_1, ..., Z_n]$ følger uafhængige standard normal fordelinger, og

- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ er en $m \times n$ matrix
- $\boldsymbol{b} \in \mathbb{R}^m$

så følger $\boldsymbol{X} = \boldsymbol{b} + \boldsymbol{A}\boldsymbol{Z}$ en multivariat normal fordeling med middelværdi $\mu = \boldsymbol{b}$, og varians-kovariansmatrix $\Sigma = \boldsymbol{A}\boldsymbol{A}^T$.

Eksempel 6, forts.

Hvis $X \sim N(2,3)$ og $Y \sim N(0,1)$ (X og Y uafhængige), find ved matrix opskrivning $\mathrm{E}(Z)$, $\mathrm{V}(Z)$ og $\mathrm{Cov}(X,Z)$ når $Z = X + 2Y$.

Eksempel 7:

Lad $X_i \sim N(\mu, \sigma^2)$ (i.i.d.) find $\mathrm{Cov}(\overline{X}, X_i - \overline{X})$, efterprøv med simulation.

Oversigt

I morgen

I morgen fortsætter vi i kapitel 2, hvor vi snakker om samplingsfordelinger (t og χ^2 - fordelinger)