Course 02403 Introduktion til Matamatisk Statistik

Forelæsning 2: Stokastiske variabel og fordelinger

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 305, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

DTU Compute
Institut for Matematik og Computer Science

Oversigt

Stokastisk variabel

Før eksperimentet er udført stokastisk variabel haves

$$X_i$$

noteret med stort bogstav.
Så udføres eksperimentet, og vi har da en *realisation* eller *observation*

$$x_i$$

noteret med småt bogstav.

Tæthedsfunktion

En stokastisk variabel har en *tæthedsfunktion* (probability density function (pdf))

Definition

$$f(x) = P(X = x)$$

Sandsynligheden for at X bliver udfaldet x når eksperimentet udføres.
Der gælder at:

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

$$\sum_x f(x) = 1$$

Fordelingsfunktion (distribution function eller cumulative density function (cdf))

Definition

Fordelingsfunktionen (cdf) er tæthedsfunktionen akkumuleret

$$F(x) = P(X \leq x) = \sum_{j \text{ hvor } x_j \leq x} f(x_j)$$

Fair terning eksempel

Lad X repræsentere et kast med en fair terning
Udregn sandsynligheden for at få udfald under 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ } \textit{fordelingsfunktionen} \\ &= P(X = 1) + P(X = 2) \\ &= f(1) + f(2) \text{ } \textit{tæthedsfunktioner} \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \end{aligned}$$

Oversigt

DTU Compute
Institut for Matematik og Computer Science

Konkrete Statistiske fordelinger

- Der findes en række statistiske fordelinger, som kan bruges til at beskrive og analysere forskellige problemstillinger med
- Vi starter med diskrete fordelinger:
  - Binomial fordelingen
  - Den hypergeometriske fordeling
  - Poisson fordelingen

Binomial fordelingen

- Et eksperiment med to udfald (succes eller ikke-succes) gentages
- X er antal succeser efter n gentagelser
- Så følger X en binomial fordelingen

$$X \sim B(n, p)$$

- Binomial fordelingens tæthedsfunktion giver sandsynligheden for x antal succeser

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

- n antal gentagelser
- p sandsynligheden for succes i hver gentagelse

Binomial fordelingen: Eksempel

Hvad er sandsynligheden for at slå 2 seksere i 4 slag med en ærlig terning?

Binomial fordelingen: Eksempel

Hvad er sandsynligheden for at slå 2 seksere i 4 slag med en ærlig terning?

```
6 * (1 / 6)^2 * (5 / 6)^2
```

[1] 0.11574

```
dbinom(2,4,1/6)
```

[1] 0.11574

Hypergeometrisk fordeling

- X er igen antal succeser, men nu er det *uden tilbgelægning ved gentagelsen*
- X følger da den hypergeometriske fordeling

$$X \sim H(n, a, N)$$

- Sandsynligheden for at få x succeser er

$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

hvor

- n er antallet af trækninger
- a er antallet af succeser i populationen
- N elementer i population

Hypergeometrisk fordelingen: Eksempel

En skål indeholder 10 sorte og 2 hvide kugler, hvis der trækkes 4 kugler hvad er da sandsylingheden for at 2 er hvide?

```
choose(2,2) * choose(10,2) / choose(12,4)
```

[1] 0.090909

```
dhyper(2,2,10,4)
```

[1] 0.090909

Poissonfordelingen

- Poisson fordelingen anvendes ofte som en fordeling (model) for tælletal, hvor der ikke er nogen naturlig øvre grænse
- Poisson fordelingen karakteriseres ved en intensitet, dvs. på formen antal/enhed
- Parameteren λ angiver intensiteten
- Typisk hændelser per tidsinterval
- Intervallerne mellem hændelserne er uafhængige, dvs. processen er hukommelsesløs

Poissonfordelingen

X følger Poisson fordelingen

- $X \sim P(\lambda)$
- Tæthedsfunktion:
  $f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$

Eksempel

Det antages, at der i gennemsnit bliver indlagt 0.3 patienter pr. dag på københavnske hospitaler som følge af luftforurening.

Hvad er sandsynligheden for at der på en vilkårlig dag bliver indlagt højst 2 patienter som følge af luftforurening?

- Step 1) Hvad skal repræsenteres: X er antal patienter pr. dag
- Step 2) Hvilken fordeling: X følger Poisson fordelingen
- Step 3) Hvilken sandsynlighed:$P(X \leq 2)$
- Step 4) Hvad er raten: $\lambda = 0.3$ patienter per dag

```
ppois(2,lambda=0.3)
```

```
[1] 0.9964
```

Binomial, hypergeometrisk, eller Poisson

- Binomial fordelingen anvendes også for at analysere stikprøver med tilbagelægning (Tænk på en terningekast)
- Når man vil analysere stikprøver uden tilbagelægning anvendes den hypergeometriske fordeling (Tænk på træk fra en hat).
- Poisson anvendes når der ikke er et naturligt loft over antal observationer.

R og diskrete fordelinger

R	Betegnelse
`binom`	Binomial
`hyper`	hypergeometrisk
`pois`	Poisson

- `d` Tæthedsfunktion $f(x)$ (probability density function).
- `p` Fordelingsfunktion $F(x)$ (cumulative distribution function).
- `r` Tilfældige tal fra den anførte fordeling. (Forelæsning 10)
- `q` Fraktil (quantile) i fordeling.

Husk at hjælp til funktion mm. fåes ved at sætte '?' foran navnet.

Eksempel binomial fordelt: $P(X \leq 5) = F(5; 10, 0.6)$

```
pbinom(q=5, size=10, prob=0.6)
## Få hjælpen med
?pbinom
```

Middelværdi (mean) og forventningsværdi (expectation)

Stokastisk variabels middelværdi

$$\mu = E(X) = \sum_{\text{alle x}} x f(x)$$

- Det "rigtige gennemsnit"
- Fortæller hvor "midten" af X er

Varians

Definition

$$\sigma^2 = \mathsf{Var}(X) = \sum_{\text{alle x}} (x-\mu)^2 f(x)$$

- Et mål for spredningen
- Den "rigtige spredning" af X (modsat empirisk varians (sample variance))

Diskrete fordelinger: oversigt

Fordeling	Ω	pdf	μ	σ^2
Generel		$f(x)$	$\sum x_i f(x_i)$	$\sum(\mu - x_i)^2 f(x_i)$
Binomial	$0, \ldots, n$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
Hypergeo.	$\max(0, n+a-N),$ $\ldots, \min(n, a)$	$\frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$	$\frac{na}{N}$	$n\frac{a}{N}\frac{(N-a)}{N}\frac{N-n}{N-1}$
Poisson	$0, 1, \ldots$	$\frac{\lambda^x}{x!} e^{-\lambda}$	λ	λ

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
- For kontinuerte variable svarer tætheden ikke til sandsynligheden, dvs. $f(x) \neq P(X = x)$
- Et godt plot af $f(x)$ er et histogram (kontinuert)

Tæthedsfunktion for en kontinuert variabel

For en kontinuert stokastisk variabel skrives tæthedsfunktionen som:

$$f(x)$$

Der gælder:

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

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

Fordelingsfunktion (distribution function eller cumulative density function (cdf))

- Fordelingsfunktion for en kontinuert stokastisk variabel betegnes ved $F(x)$.
  Fordelingsfunktionen svarer til den kumulerede tæthedsfunktion:

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

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

$$f(x) = F'(x)$$

- Et godt plot for fordelingsfunktionen er den kumulative fordeling

Middelværdi (mean) og varians 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

DTU Compute
Institut for Matematik og Computer Science

Konkrete statistiske fordelinger

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

Vi betragter nu kontinuerte fordelinger

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

Uniform fordelingen

Skrivemåde:

$$X \sim U(\alpha, \beta)$$

Tæthedsfunktion:

$$f(x) = \frac{1}{\beta - \alpha}; \quad x \in [\alpha, \beta]$$

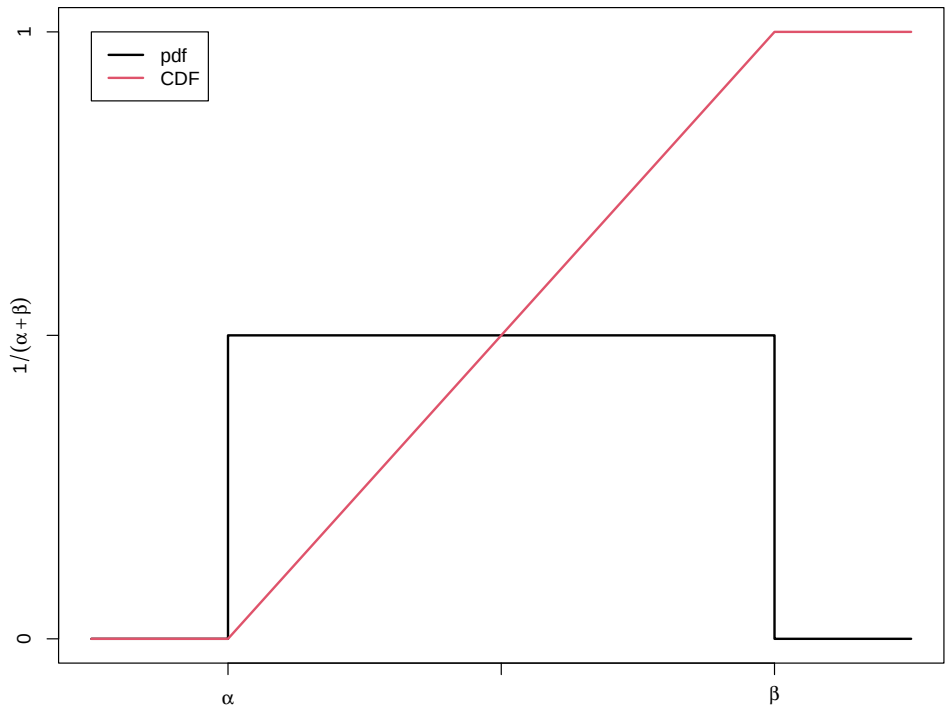
Middelværdi:

$$\mu = \frac{\alpha + \beta}{2}$$

Varians:

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

Uniform fordelingen



Eksponentialfordelingen

- Eksponential fordelingen er et special tilfælde af Gamma fordelingen
- Eksponential fordelingen anvendes f.eks. til at beskrive levetider og ventetider
- Eksponential fordelingen kan bruges til at beskrive (vente)tiden mellem hændelser i poisson fordelingen
- Tæthedsfunktion

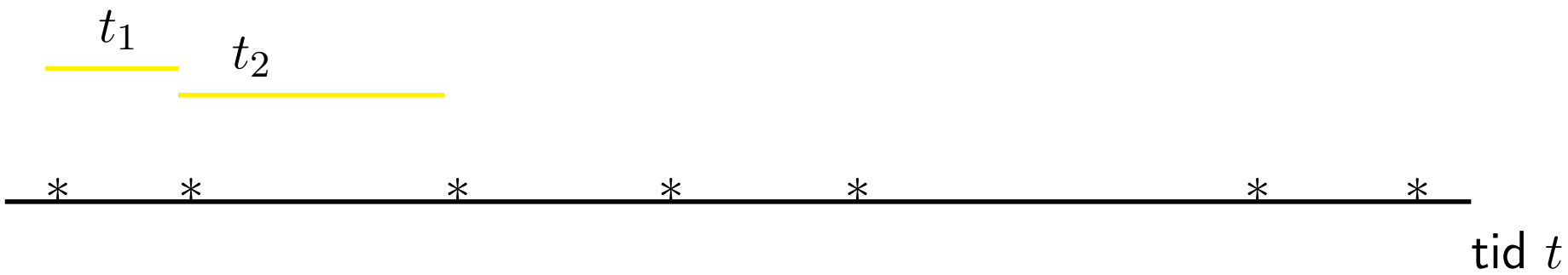
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0,\ \lambda > 0 \\ 0 & \text{ellers} \end{cases}$$

- Middelværdi $\mu = \frac{1}{\lambda}$
- Varians $\sigma^2 = \frac{1}{\lambda^2}$

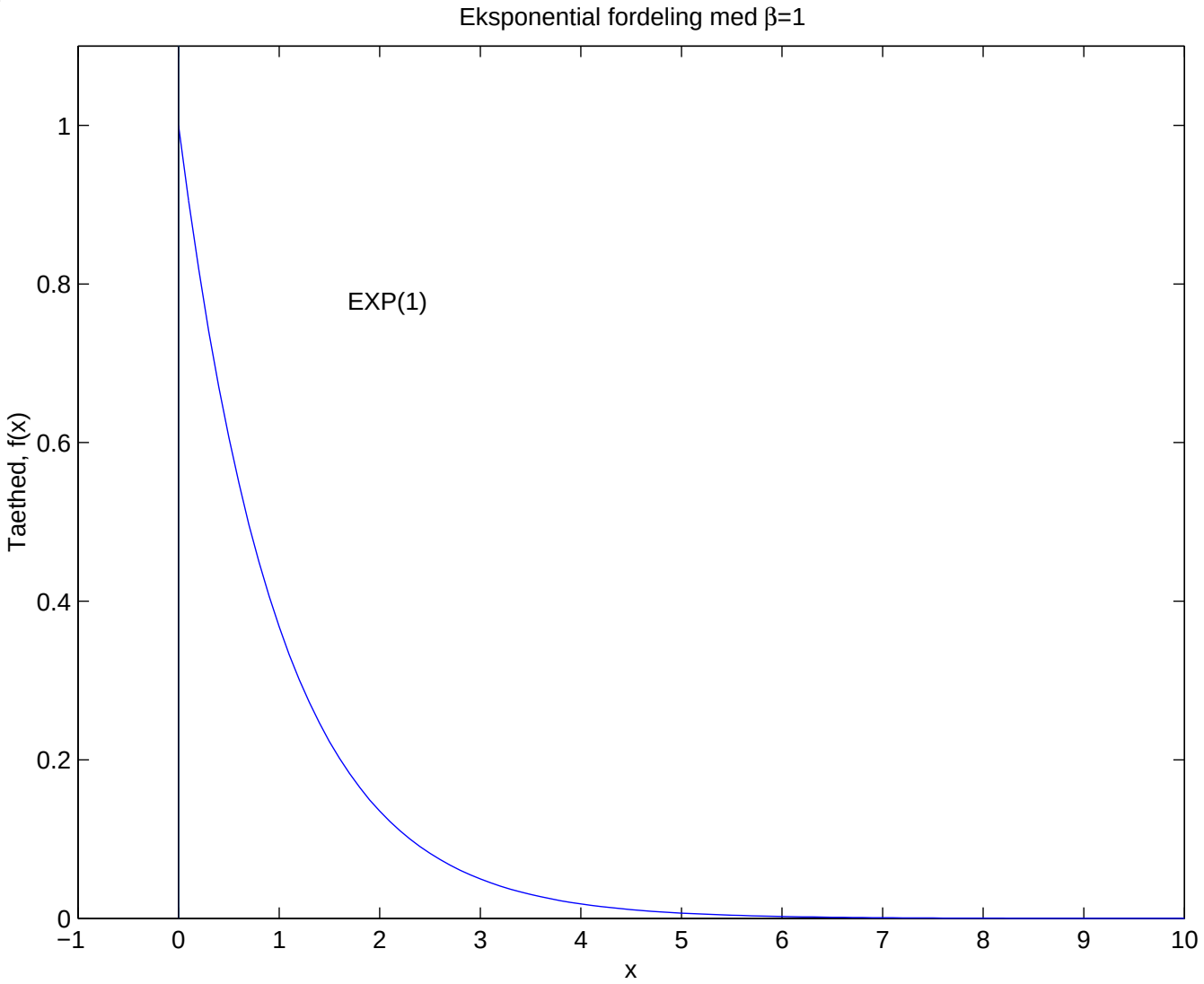
Sammenhæng mellem Eksponential og Poisson fordelingen

Poisson: Diskrete hændelser pr. enhed

Eksponential: Kontinuert afstand mellem hændelser



Eksponential fordelingen



DTU Compute
Institut for Matematik og Computer Science

Eksempel

Kø-model - poisson proces

Tiden mellem kundeankomster på et posthus er eksponential fordelt med middelværdi $\mu = 2$ minutter.

Spørgsmål:

En kunde er netop ankommet. Hvad er sandsynligheden for at der ikke kommer flere kunder indenfor en periode på 2 minutter?

```
1-pexp(2,rate=1/2)
```

```
[1] 0.36788
```

Normal fordelingen

Skrivemåde:

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

Tæthedsfunktion:

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

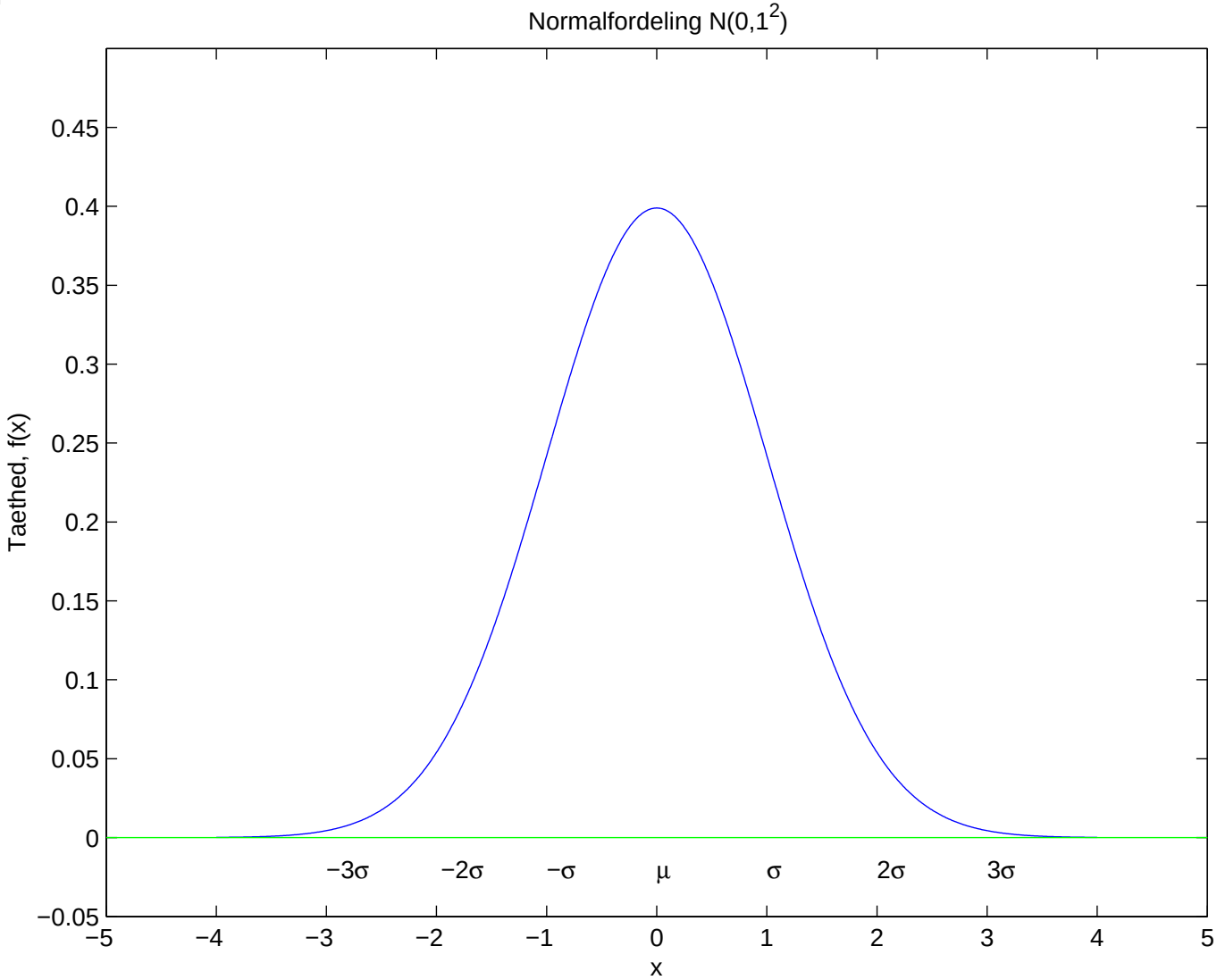
Middelværdi:

$$\mu = \mu$$

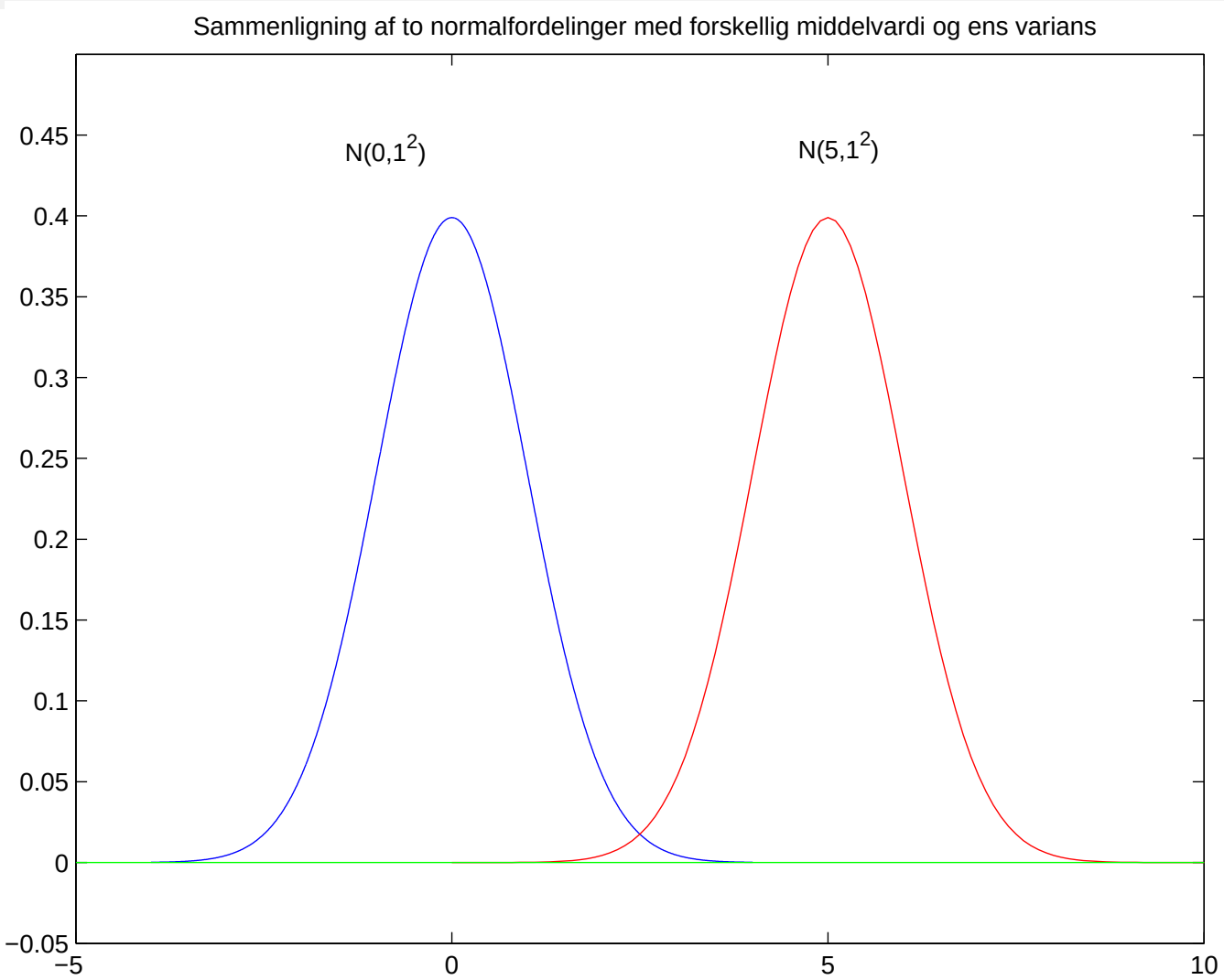
Varians:

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

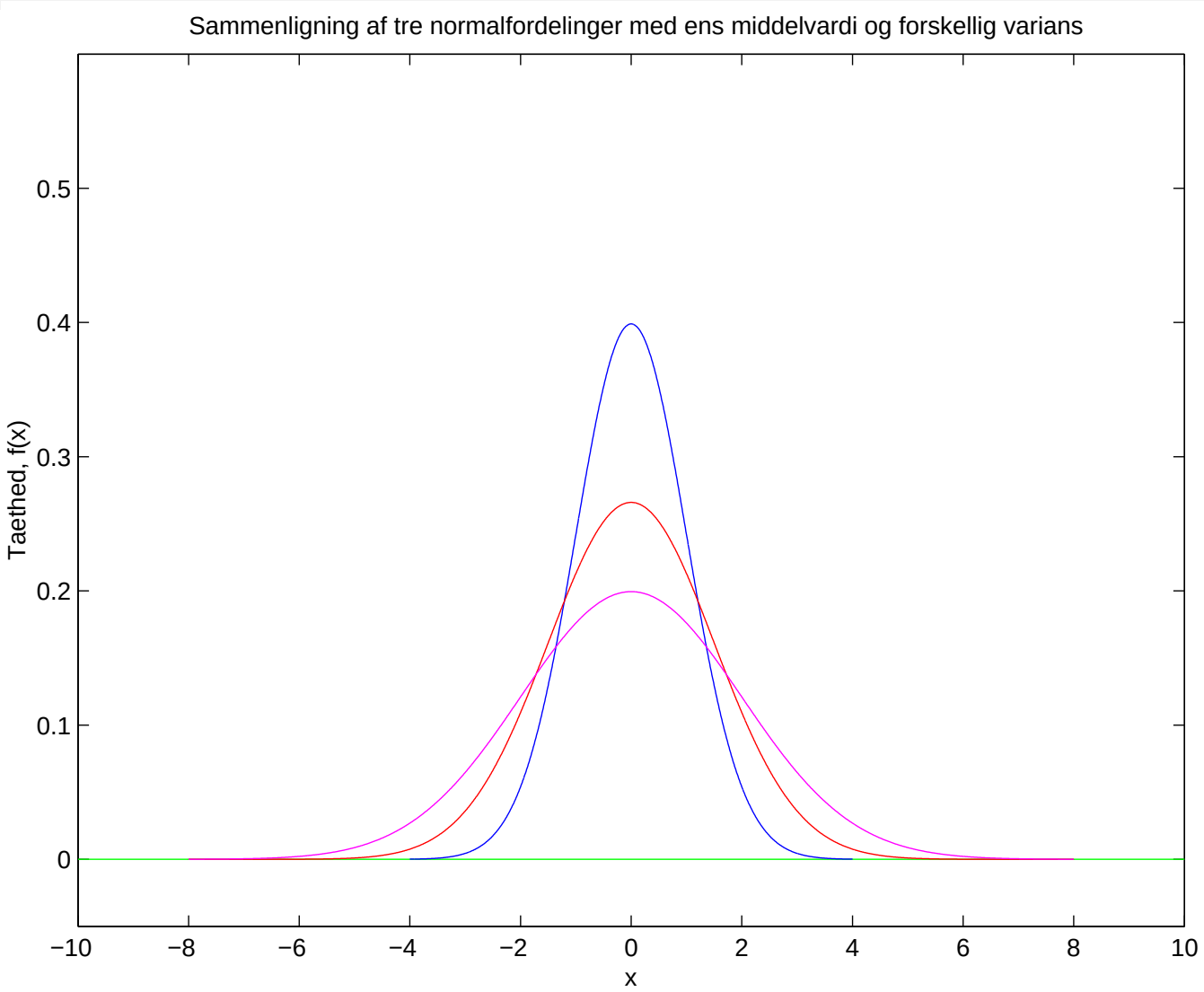
Normalfordelingen



Normalfordelingen



Normalfordelingen



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

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

Eksempel

Målefejl:

En vægt har en målefejl, Z, der kan beskrives ved en standard normalfordeling, dvs

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

dvs. middelværdi $\mu = 0$ og spredning $\sigma = 1$ gram.
Vi måler nu vægten af ét emne

Spørgsmål a):

Hvad er sandsynligheden for at vægten måler mindst 2 gram for lidt?

Eksempel

Spørgsmål b):

Hvad er sandsynligheden for at vægten måler mindst 2 gram for meget?

Spørgsmål c):

Find d så $P(-d < Z < d) = 0.95$

Spørgsmål d):

Hvis $X \sim N(\mu, \sigma^2)$, find d så $P(\mu - d < X < \mu + d) = 0.95$

Log-Normal fordelingen

Lad $X \sim N(\alpha, \beta^2)$ så følger $Y = e^X$ en log-normal fordeling

Skrivemåde:

$Y \sim LN(\alpha, \beta^2)$

Tæthedsfunktion:

$$f(x) = \begin{cases} \frac{1}{x\beta\sqrt{2\pi}} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0,\ \beta > 0 \\ 0 & \text{ellers} \end{cases}$$

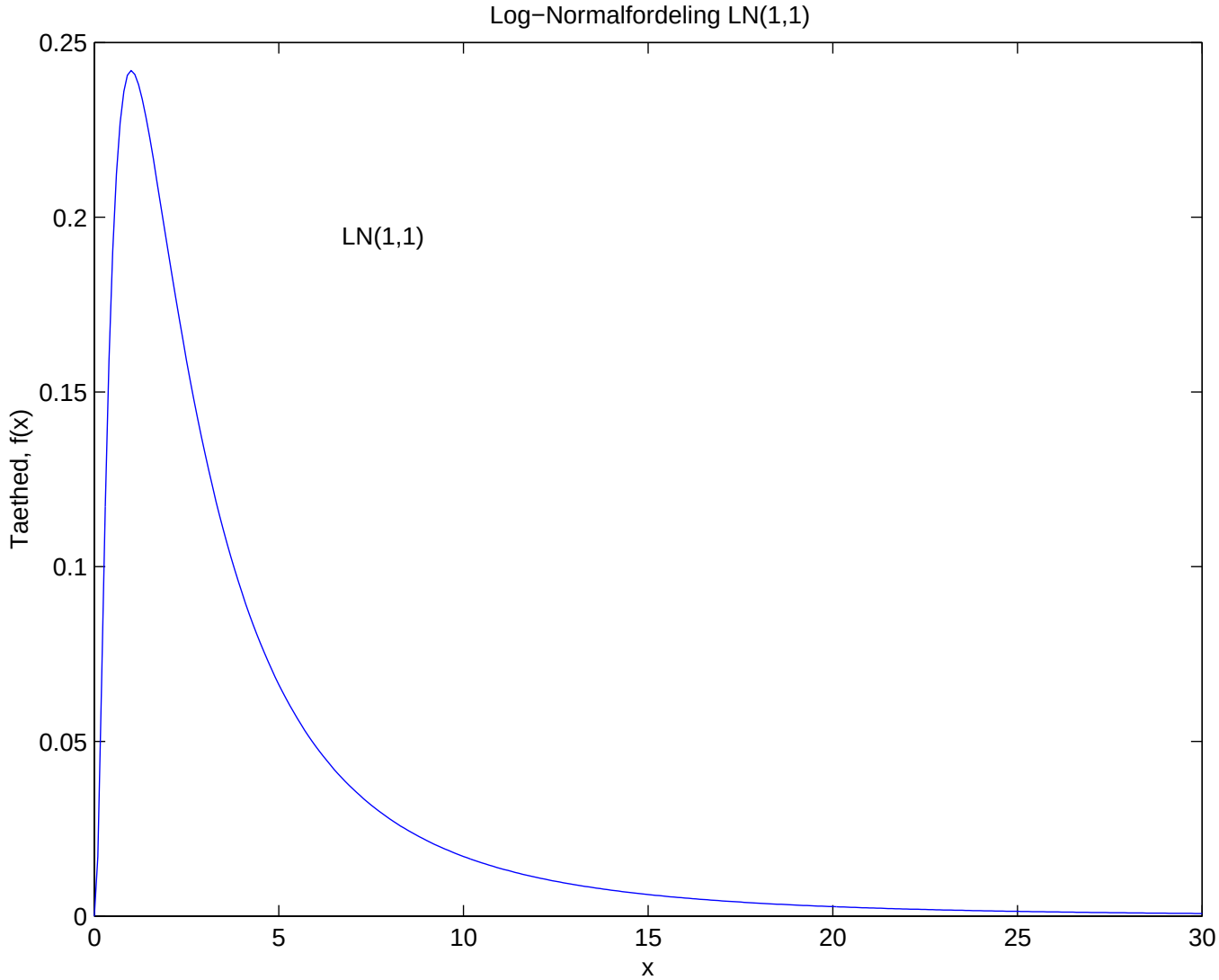
Middelværdi:

$\mu = e^{\alpha+\beta^2/2}$

Varians:

$\sigma^2 = e^{2\alpha+\beta^2}(e^{\beta^2} - 1)$

Log-Normal fordelingen



DTU Compute
Institut for Matematik og Computer Science

Log-Normal fordelingen

Lognormal og Normalfordelingen:

En log-normal fordelt variabel $Y \sim LN(\alpha, \beta^2)$, kan transformeres til en standard normal fordelt variabel X ved

$$X = \frac{\ln(Y) - \alpha}{\beta}$$

dvs.

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

Ved antagelse om log-normalfordelte data foretages analysen sædvanligvis på log-transformerede data.

Kontinuerte fordelinger i R

Sammelign klorofyll data fra Skive fjord med en relevant fordelinger.
Sammelign log-klorofyll data fra Skive fjord med en relevant fordelinger.

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.

Kontinuerte fordelinger: oversigt

Fordeling	Ω	pdf	μ	σ^2
Generel		$f(x)$	$\int xf(x)dx$	$\int(\mu-x)^2 f(x)dx$
Uniform	$[\alpha,\beta]$	$\frac{1}{\beta-\alpha}$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$
Exponential	$[0,\infty)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
log-Normal	$(0,\infty)$	$\frac{1}{x\beta\sqrt{2\pi}}e^{-(\ln(x)-\alpha)^2/2\beta^2}$	$e^{\alpha+\beta^2/2}$	$\mu^2(e^{\beta^2}-1)$

Oversigt

Kontinuerte og diskrete fordelinger: some facts

Diskret	Kontinuert
$f(x) = P(X = x)$	$f(x) \neq P(X = x) = 0$
$F(x) = \sum_{-\infty}^{x} f(x)$	$F(x) = \int_{-\infty}^{x} f(x)dx$
$P(X \leq x) \neq P(X < x)$	$P(X \leq x) = P(X < x)$
$P(x_1 < X \leq x_2) = \sum_{x_1+1}^{x_2} f(x)$	$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$
$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$	$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$
$E[X] = \sum_{-\infty}^{\infty} xf(x)$	$E[X] = \int_{-\infty}^{\infty} xf(x)$
$V[X] = \sum_{-\infty}^{\infty} (E[X] - x)^2 f(x)$	$V[X] = \int_{-\infty}^{\infty} (E[X] - x)^2 f(x)dx$

DTU Compute
Institut for Matematik og Computer Science

Oversigt

Forelæsning 3

I forelæsning 3 fortsætter vi med kapitel 2, samt første del af kapitel 4

- Regneregler for middelværdi og varians
- Simulation som generelt værktøj
- Kovarians og uafhængighed