Introduktion til Statistik

Forelæsning 3: Kontinuerte fordelinger

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2021

Kapitel 2: Kontinuerte fordelinger

Grundlæggende koncepter:

- Tæthedsfunktion: $f(x)$ *(pdf)*
- Fordelingsfunktion: $F(x) = P(X \leq x)$ *(cdf)*
- Middelværdi (μ) og varians (σ^2)
- Regneregler for stokastiske variabler (lineære funktioner)

Specifikke fordelinger:

- Uniform
- Normal
- Log-Normal
- Eksponential

Funktioner af normalfordeling (afsn. 2.10) (introduceres først i de næste uger):

- t-fordelingen, χ^2-fordelingen *(Chi-i-anden)* og F-fordelingen

Chapter 2: Continuous Distributions

General concepts:

- Density function: $f(x)$ *(pdf)*
- Distribution: $F(x) = P(X \leq x)$ *(cdf)*
- Mean (μ) and variance (σ^2)
- Calculation rules for random variables (linear functions)

Specific distributions:

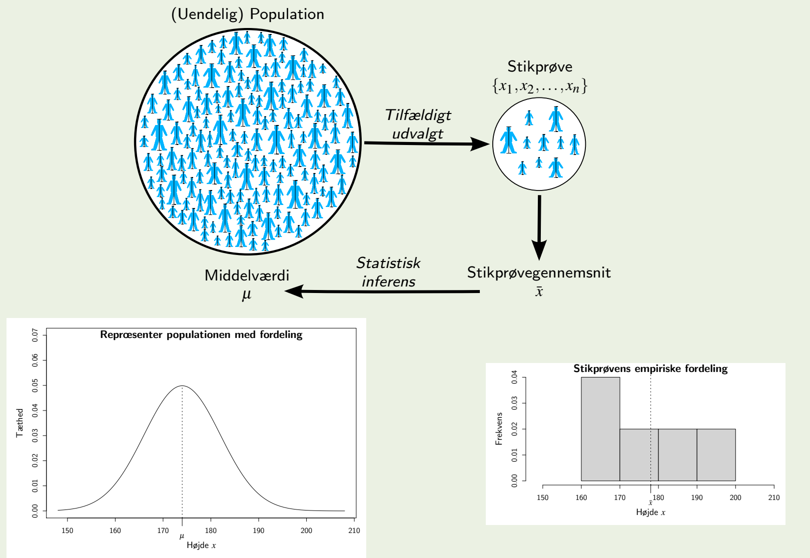
- Uniform
- Normal
- Log-Normal
- Exponential

Funktions of normaldist. (Sec. 2.10) (introduced in the coming weeks):

- t-distribution, χ^2-distribution *(Chi-square)* og F-distribution

Oversigt

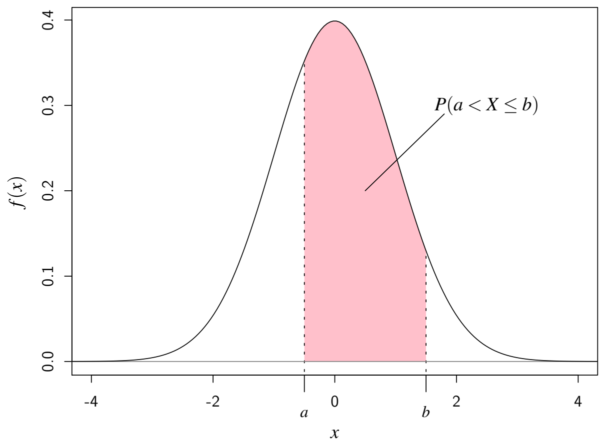
Eksempel: Population og fordeling



Tæthedsfunktion (probability density function (pdf))

- **Tæthedsfunktionen** for en stokastisk variabel betegnes ved $f(x)$
- 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



Tæthedsfunktion for en kontinuert variabel

- Der gælder:
  - Ingen negative værdier

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

  - Areal under kurven er een

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

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

- **Fordelingsfunktion** for en kontinuert stokastisk variabel betegnes ved

$$F(x)$$

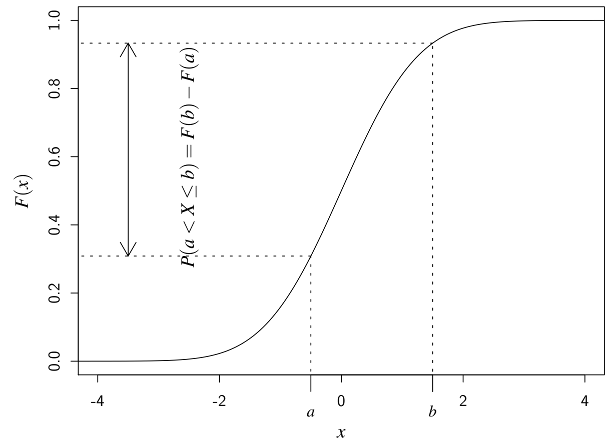
- **Fordelingsfunktionen** svarer til den kumulerede tæthedsfunktion ved

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

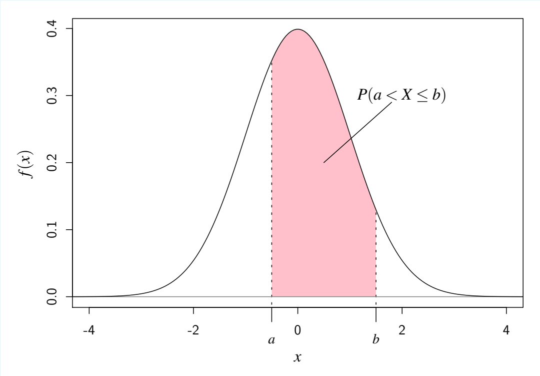
$$F(x) = \int_{-\infty}^{x} f(u)du$$

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

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



Spørgsmål om sandsynligheder (socrative.com, room: PBAC)

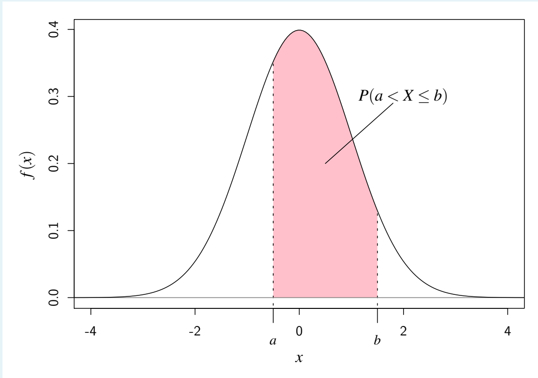


Hvilket udtryk giver den markerede sandsynlighed? (arealet)

A: $\int_{-\infty}^{b} f(x)dx$ B: $1 - \int_{a}^{b} f(x)dx$ C: $\int_{a}^{b} f(x)dx$ D: $1 - \int_{a}^{\infty} f(x)dx$

Svar C: $\int_{a}^{b} f(x)dx$

Spørgsmål om sandsynligheder (socrative.com, room: PBAC)



Hvordan kan vi nemmest udregne den markerede sandsynlighed?

A: $\int_{a}^{b} f(x)dx$ B: $\int_{a}^{b} F(x)dx$ C: $f(b) - f(a)$ D: $F(b) - F(a)$

Svar D: $F(b) - F(a)$ *(vi gør det i R med (normalfordelt):* `pnorm(b) - pnorm(a)`*)*

Middelværdi (mean) af en kontinuert stokastisk variabel

Middelværdien af en kontinuert stokastisk variabel

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

Sammenlign med den diskrete definition: $\mu = \sum_{\text{alle } x} x \cdot f(x)$

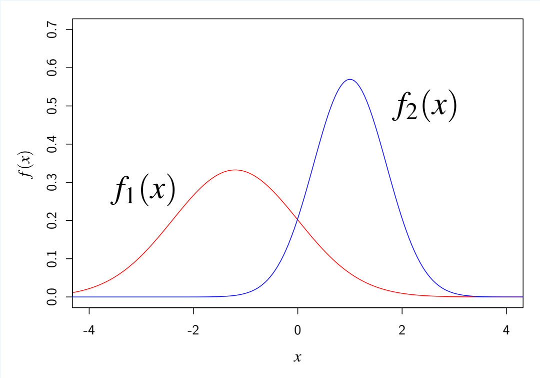
Varians af en kontinuert stokastisk variabel

Variansen af en kontinuert stokastisk variabel:

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

Sammenlign med den diskrete definition: $\sigma^2 = \sum_{\text{alle } x} (x - \mu)^2 \cdot f(x)$

Spørgsmål om middelværdi (socrative.com, room: PBAC)

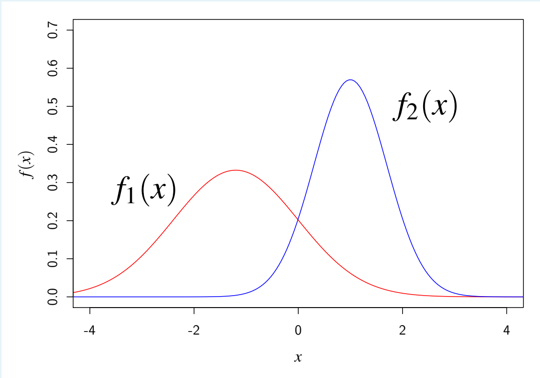


Hvilken pdf har størst middelværdi (begge er symmetriske)?

A: $\mu_1 < \mu_2$ B: $\mu_1 > \mu_2$ C: $\mu_1 = \mu_2$ D: Kan ikke afgøres

Svar A: $\mu_1 < \mu_2$.

Spørgsmål om spredning (socrative.com, room: PBAC)



Hvilken pdf har størst standard afvigelse (begge er symmetriske)?

A: $\sigma_1 < \sigma_2$ B: $\sigma_1 > \sigma_2$ C: $\sigma_1 = \sigma_2$ D: Kan ikke afgøres

Svar B: $\sigma_1 > \sigma_2$ (umiddelbart). Svar D, også fint, da man ikke kan se hvad der er udenfor plottet.

Konkrete statistiske fordelinger

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

- Følgende kontinuerte fordelinger:
  - Uniform fordeling
  - Normalfordelingen
  - Log-normalfordelingen
  - Eksponentialfordelingen

Kontinuerte fordelinger i R

R	Betegnelse
`norm`	Normalfordelingen
`unif`	Uniform fordeling
`lnorm`	Log-normalfordelingen
`exp`	Eksponentialfordelingen

`d` Tæthedsfunktion $f(x)$ (probability density function).

`p` Fordelingsfunktion $F(x)$ (cumulative distribution function).

`q` Fraktil (quantile) i fordeling.

`r` Tilfældige tal fra fordelingen.

Uniform fordeling

Skrivemåde:

$X \sim U(\alpha, \beta)$ (Læses: X følger en uniform fordeling med parametre α og β)

Tæthedsfunktion:

$f(x) = \frac{1}{\beta - \alpha}$

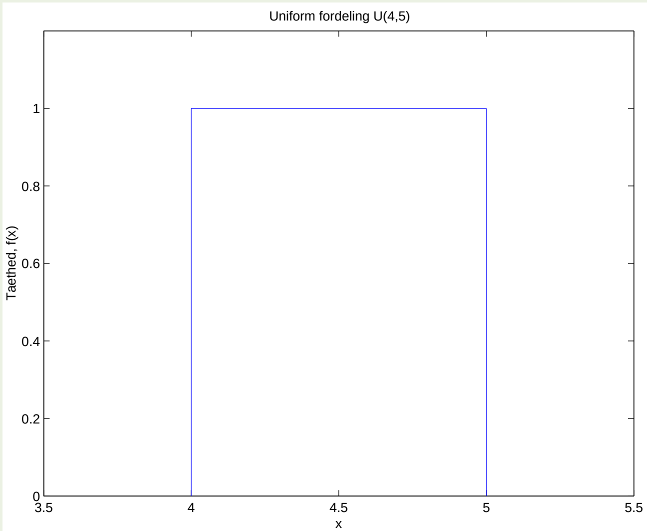
Middelværdi:

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

Varians:

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

Eksempel: Uniform fordeling



Spørgsmål: Uniform fordeling (socrative.com, room: PBAC)

Medarbejdere på en arbejdsplads ankommer mellem klokken 8:00 og 8:30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

Hvad er sandsynligheden for at en tilfældig udvalgt medarbejder ankommer mellem 8:20 og 8:30?

A: 1/2 B: 1/6 C: 1/3 D: 0

Svar C: 10/30=1/3

```
punif(30,0,30) - punif(20,0,30)
```

[1] 0.33

Spørgsmål: Uniform fordeling (socrative.com, room: PBAC)

Medarbejdere på en arbejdsplads ankommer mellem klokken 8:00 og 8:30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

Hvad er sandsynligheden for at en tilfældig udvalgt medarbejder ankommer efter 8:30?

A: 1/2 B: 1/6 C: 1/3 D: 0

Svar: $P(X > 30) = 0$

```
1 - punif(30,0,30)
```

[1] 0

Normalfordelingen

Skrivemåde:

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

Tæthedsfunktion:

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

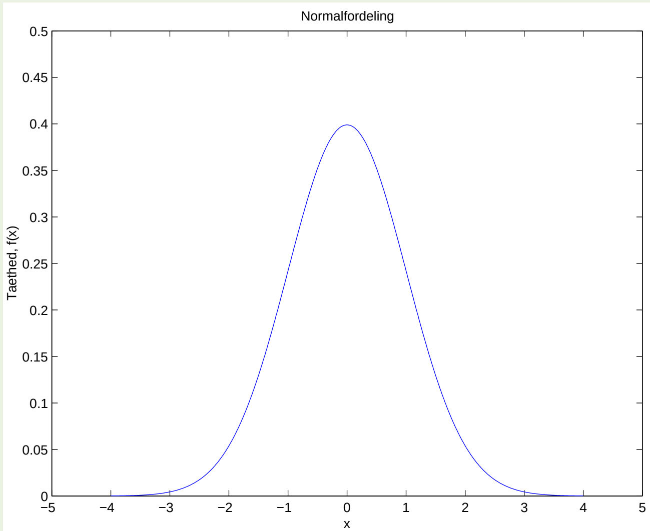
Middelværdi:

$\mu = \mu$

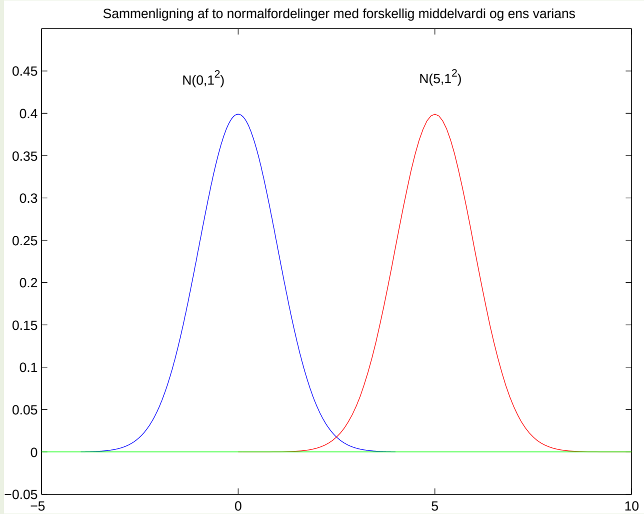
Varians:

$\sigma^2 = \sigma^2$

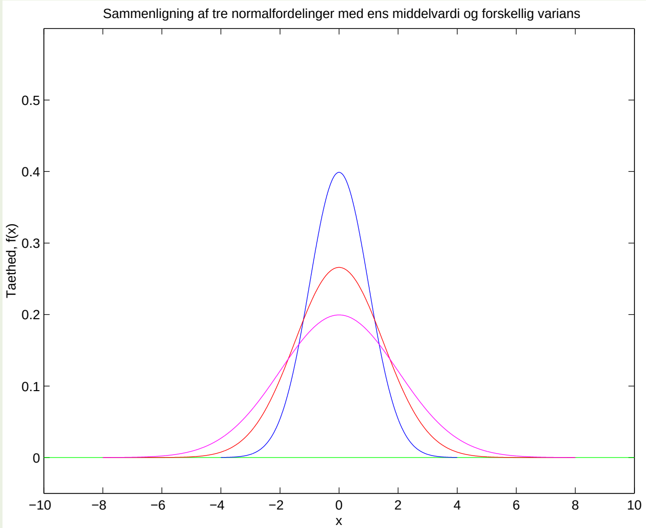
Eksempel: Normalfordelingen



Eksempel: Normalfordelingen



Eksempel: Normalfordelingen



Eksempel: Normalfordeling, sandsynligheder

Fordeling af vægt af rugbrød:

Antag at vægten af et rugbrød fra en produktionslinie kan beskrives med en normalfordeling

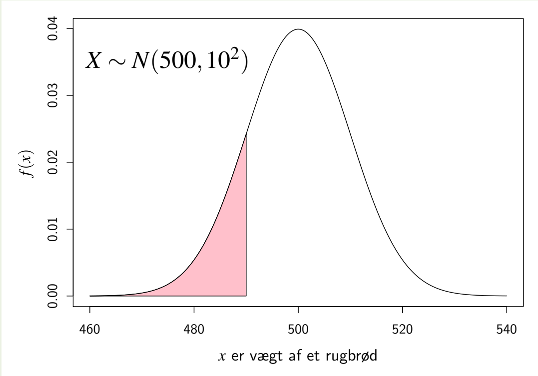
$$X \sim N(500, 10^2)$$

dvs. middelværdi $\mu = 500$ gram og standardafvigelse $\sigma = 10$ gram.
Vi vil måle vægten af ét tilfældigt udvalgt brød.

Spørgsmål:

1: Hvad er sandsynligheden for at brødet vejer under 490 g?
2: Hvad er sandsynligheden for at brødet vejer mere en 20 g forskelligt fra 500 g?

Eksempel: Normalfordeling, spørgsmål 1

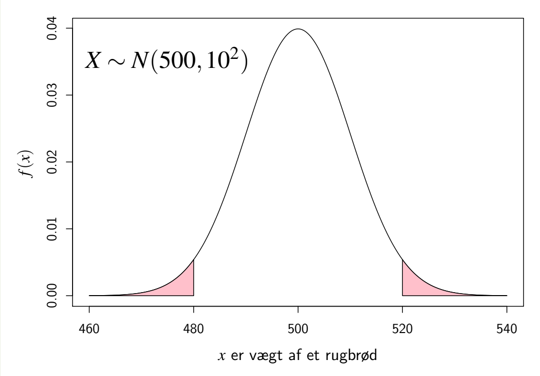


1: Hvad er sandsynligheden for at brødet vejer under 490 g?

Svar: $P(X \leq 490) = F(490) = 0.16$

```
pnorm(490, mean=500, sd=10)
```

Eksempel: Normalfordeling, spørgsmål 2

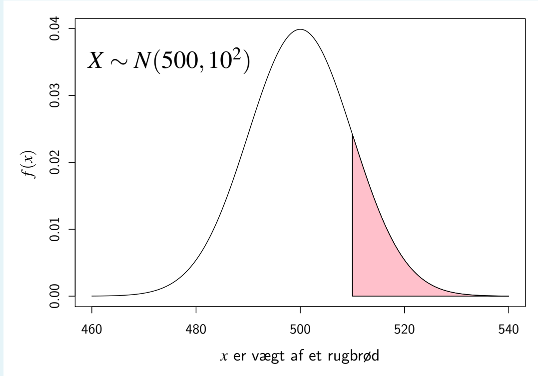


1: *Hvad er sandsynligheden for at brødet vejer mere end 20 g forskelligt fra 500 g?*

Svar: $P(X \leq 480 \vee X > 520) = 2 \cdot P(X \leq 480) = 2 \cdot F(480) = 0.046$

```
2 * pnorm(480, mean=500, sd=10)
```


Spørgsmål: Sandsynlighed i normalfordeling

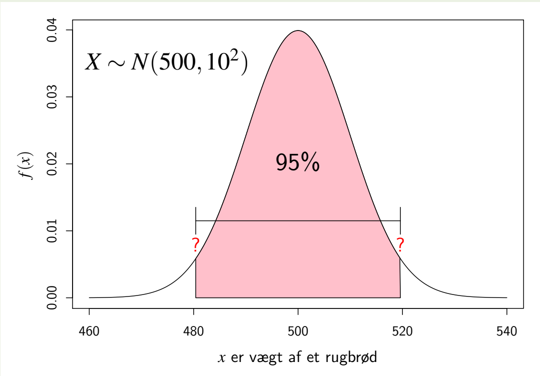


Hvad er sandsynligheden for at rugbrødet vejer over 510 g?

A: $F(510)$ B: $1 - F(490)$ C: $1 - F(520)$ D: $1 - F(510)$

Svar: $P(X > 510) = 1 - P(X \leq 510) = 0.16$

Eksempel: Normalfordeling fraktiler



"Omvendt spørgsmål": *Hvilket interval, symmetrisk om midten, dækker 95% af rugbrødene?*

```
qnorm(c(0.025,0.975), mean=500, sd=10)
```

[1] 480.4 519.6

Standard normalfordelingen

En standard normalfordeling

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

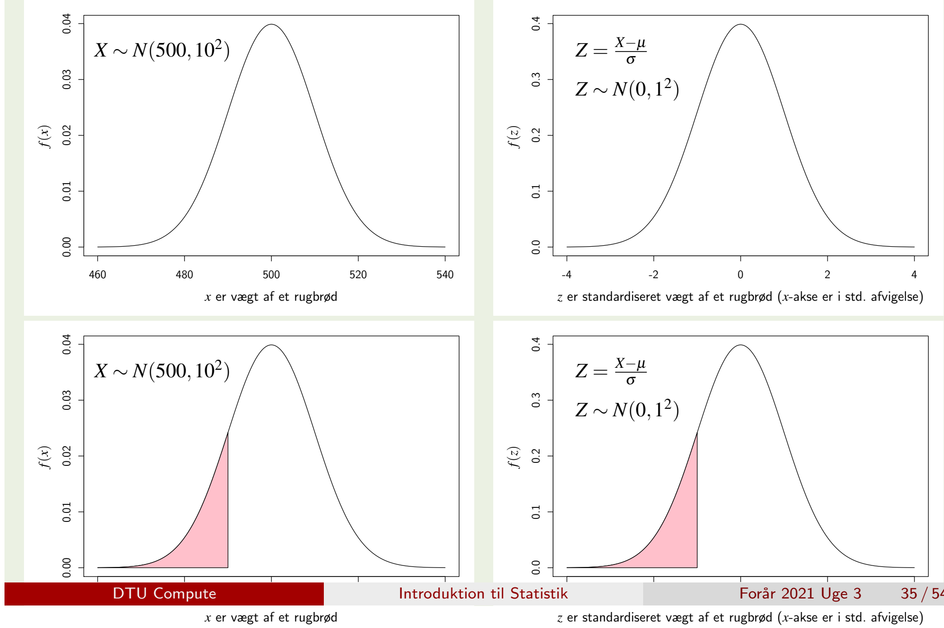
En normalfordeling med middelværdi 0 og varians 1.

Standardisering

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

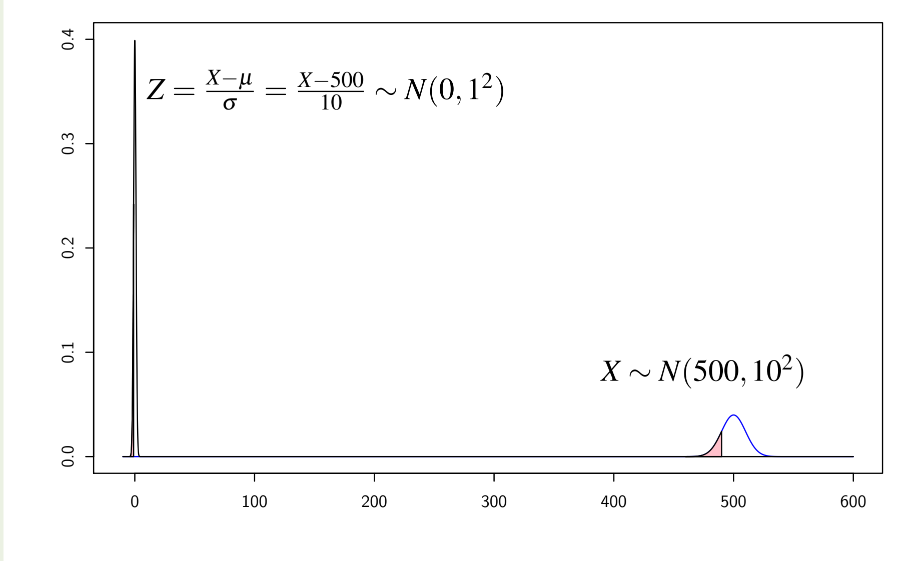
$$Z = \frac{X - \mu}{\sigma}$$

Eksempel: Standard Normalfordeling

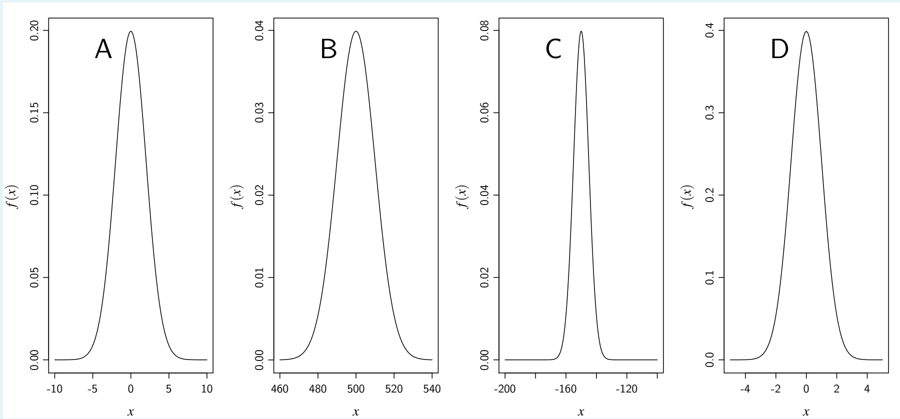


1: *Hvad er sandsynligheden for at brødet vejer under 490 gram?*

Eksempel: Transformation til standard normalfordeling



Eksempel: Transformation til standard normalfordeling



1: *Hvilken af disse er standard normalfordelingens pdf?*

Svar: D, for ca. $\mu \pm 3\sigma$ er $f(x) \approx 0$ for $Z \sim N(0, 1^2)$.

Log-Normalfordelingen

Skrivemåde:

$X \sim LN(\alpha, \beta^2)$ (Hvis X følger log-normal så følger ln(X) normal)

Tæthedsfunktion:

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

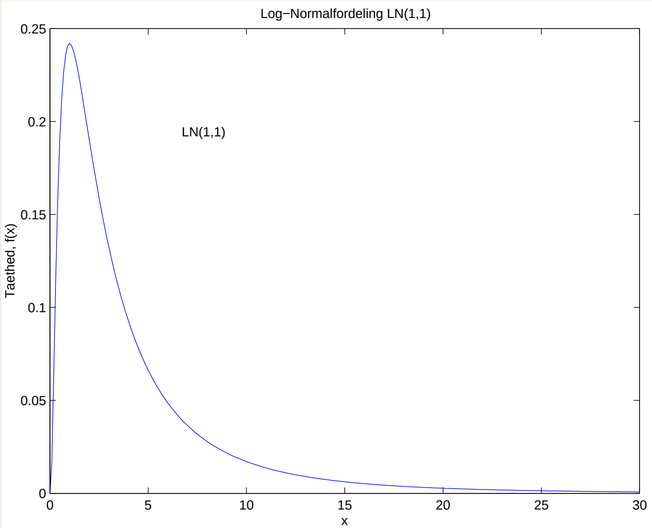
Middelværdi:

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

Varians:

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

Eksempel: Log-normalfordelingen



Log-normalfordelingen

Lognormal og Normalfordelingen:

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

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

dvs.

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

Eksponentialfordelingen

Skrivemåde:

$X \sim Exp(\lambda)$

Tæthedsfunktionen

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

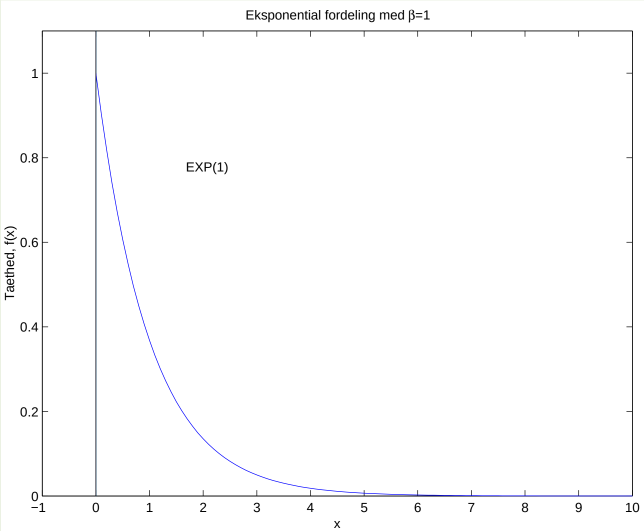
Middelværdi

$\mu = \frac{1}{\lambda}$

Varians

$\sigma^2 = \frac{1}{\lambda^2}$

Eksempel: Eksponentialfordelingen



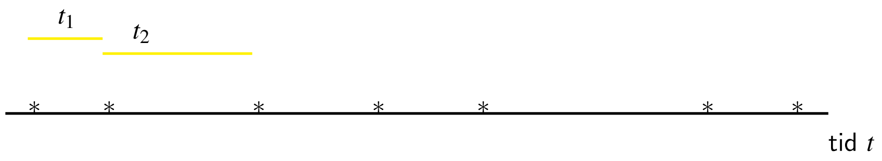
Eksponentialfordelingen

- Eksponentialfordelingen er et special tilfælde af Gammafordelingen
- Eksponentialfordelingen anvendes f.eks. til at beskrive levetider og ventetider
- Eksponentialfordelingen kan bruges til at beskrive (vente)tiden mellem hændelser i poissonproces

Sammenhæng mellem eksponential- og poissonfordelingen

Poisson: Diskrete hændelser pr. enhed

Eksponential: Kontinuert afstand mellem hændelser



Eksempel: Eksponentielfordeling

Kø-model - poissonproces

Tiden mellem kundeankomster på et posthus er eksponentialfordelt med middelværdi $\mu = 2$ minutter, dvs. $\lambda = \frac{1}{\mu} = \frac{1}{2}\frac{1}{\text{min}}$ (skaleret $\lambda_{\text{2min}} = 1\frac{1}{\text{2min}}$).

Spørgsmål:

En kunde er netop ankommet. Beregn sandsynligheden for at der ikke kommer flere kunder indefor en periode på 2 minutter vha. poissonfordelingen

Svar:

Med Poissonfordelingen (periodelængden skal svare til spørgsmålet, brug λ_{2min}):

```
dpois(x=0, lambda=1)
```

Brug Eksponentialfordeling ($X_{\text{exp}} \sim Exp(\lambda)$ med $\lambda = \frac{1}{2}\frac{1}{\text{min}}$, find $P(X_{\text{exp}} > 2)$):

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

Giver 0.37

Regneregler for lineær funktion af et X

Hvis:

- X er en stokastisk variabel
- Vi antager at a og b er konstanter

Da gælder (gælder BÅDE kontinuert og diskret):

Middelværdi-regel:

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$$

Varians-regel:

$$\mathrm{V}(aX + b) = a^2\,\mathrm{V}(X)$$

Eksempel: Regneregler for lineær funktion af et X

X er en stokastisk variabel

En stokastisk variabel X har middelværdi 4 og varians 6.

Spørgsmål:

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

Svar:

$$\mathrm{E}(Y) = \mathrm{E}(-3X+2) = -3\,\mathrm{E}(X) + 2 = -3 \cdot 4 + 2 = -10$$

$$\mathrm{V}(Y) = \mathrm{V}(-3X+2) = (-3)^2\,\mathrm{V}(X) = 9 \cdot 6 = 54$$

Regneregler for lineær funktion af flere Xer

Hvis:

- $X_1, \ldots, X_n$ er stokastiske variable

Da gælder (når de er uafhængige) (gælder BÅDE kontinuert og diskret):

Middelværdi-regel:

$$\mathrm{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$$

Varians-regel:

$$\mathrm{V}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2\,\mathrm{V}(X_1) + a_2^2\,\mathrm{V}(X_2) + \ldots + a_n^2\,\mathrm{V}(X_n)$$

Eksempel: Regneregler for lineær funktion af flere Xer

Flypassager-planlægning

Vægten af een passager på fly på en strækning 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.

Hvad er *den samlede passagervægt Y* på en afgang?

A: $Y = 55 \cdot X$ B: $Y = \sum_{i=1}^{55} X_i$ C: $Y = 55 + X$ D: Ej A,B eller C

Svar B: $Y = \sum_{i=1}^{55} X_i$, *det er summen af 55 forskellige passagerer.*

Eksempel: Regneregler 3

Hvad er den samlede passagervægt Y på en afgang?

$Y = \sum_{i=1}^{55} X_i$, hvor $X_i \sim N(70, 10^2)$

Middelværdi og varians for Y:

$$\mathrm{E}(Y) = \sum_{i=1}^{55} \mathrm{E}(X_i) = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\mathrm{V}(Y) = \sum_{i=1}^{55} \mathrm{V}(X_i) = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Bruger normalfordeling for Y:

```
1-pnorm(4000, mean = 3850, sd = sqrt(5500))
```

```
[1] 0.022
```

Eksempel: Regneregler 3 - FORKERT ANALYSE

Hvad er Y?

I hvert fald IKKE: $Y = 55 \cdot X$!!!!!!

Middelværdi og varians for Y:

$$\mathrm{E}(Y) = 55 \cdot 70 = 3850$$
$$\mathrm{V}(Y) = 55^2\,\mathrm{V}(X) = 55^2 \cdot 100 = 550^2 = 302500$$

Bruger normalfordeling for Y:

```
1-pnorm(4000, mean = 3850, sd = 550)
```

[1] 0.39

Konsekvens af forkert beregning:

MANGE spildte penge for flyselskabet!!!

Lineær kombination af normalfordelte stokastiske variabler er også normalfordelt

- Lineær kombination af normalfordelte stokastiske variabler er også normalfordelt
- Theorem 2.40: Let $X_1, \ldots, X_n$ be independent normal random variables, then any linear combination of $X_1, \ldots, X_n$ will follow a normal distribution, with mean and variance given in Theorem 2.56.