02402: Introduktion til Statistik

Uge 3: stokastiske variable og kontinuerte fordelinger

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

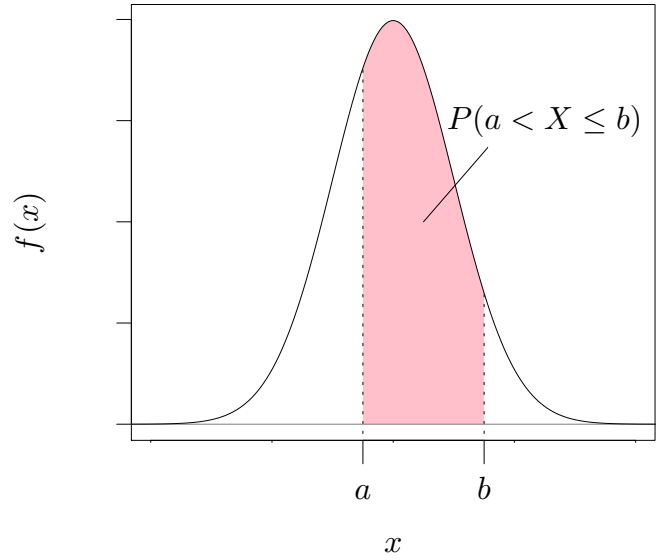
Overview

Tæthedsfunktionen, Definition 2.32

- Tæthedsfunktionen (density function/probability density function, pdf) for en stokastisk variabel betegnes med $f(x)$.
- Tæthedsfunktionen $f(x)$ siger noget om hyppigheden af udfaldet x for den stokastiske variabel X.
- Tæthedsfunktionen for en kontinuert stokastisk variabel svarer *ikke* til sandsynligheden. Der gælder faktisk, $P(X = x) = 0$ for alle x.
- Tæthedsfunktionen $f(x)$ hørende til fordelingen af en kontinuert stokastisk variabel opfylder at:

$$f(x) \geq 0 \text{ for alle } x \quad \text{og} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

Tæthedsfunktionen



Fordelingsfunktionen, Definition 2.33

- Fordelingsfunktionen (distribution function/cumulative density function, cdf) hørende til en kontinuert stokastisk variabel benæves med $F(x)$.
- Fordelingsfunktionen er defineret ved

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

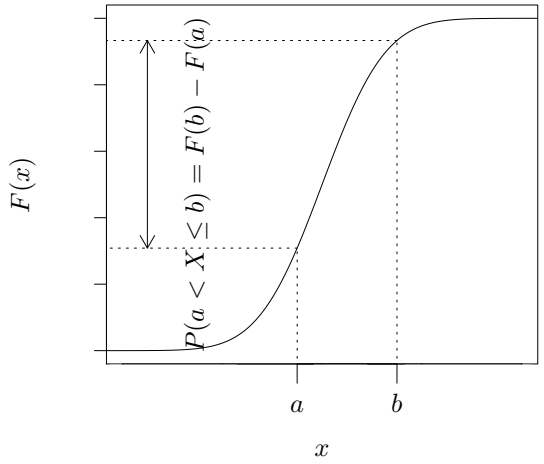
- Som følge af denne definition gælder

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

- Det er rigtigt smart at bemærke at:

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)\,dx.$$

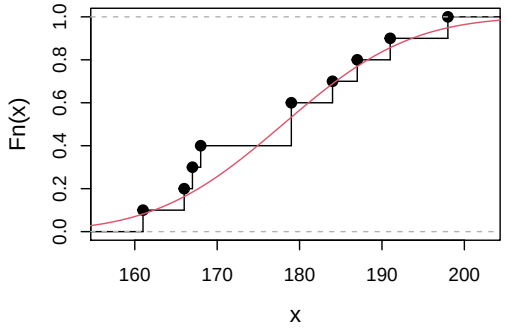
Fordelingsfunktionen



Empirisk fordelingsfunktion (ecdf)

```
# Empirical cdf for sample of height data from Chapter 1
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
plot(ecdf(x), verticals = TRUE, main = "")

# 'True cdf' for normal distribution (with sample mean and variance)
xp <- seq(0.9*min(x), 1.1*max(x), length = 100)
lines(xp, pnorm(xp, mean(x), sd(x)), col = 2)
```



Middelværdi (mean) af en kontinuert stokastisk variabel , Definition 2.34

Middelværdien af en kontinuert stokastisk variabel:

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

Sammenlign med den diskrete definition:

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

Varians af en kontinuert stokastisk variabel, Definition 2.34

Variansen af en kontinuert stokastisk variabel:

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

Sammenlign med den diskrete definition:

$$\sigma^2 = \sum_{\text{all } x} (x-\mu)^2 f(x)$$

Kovarians, Definition 2.58

Kovariansen af to stokastisk variable:

Lad X og Y være to stokastiske variable. Kovariansen mellem X and Y er defineret ved

$$\text{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$$

Kovariansen:

Hvis to stokasiske variable er *uafhængige*, så er kovariansen 0. *Det modsatte er ikke nødvendigvis tilfældet!*

Overview

Vigtige kontinuerte fordelinger

Der findes en række statistiske fordelinger (både kontinuerte og diskrete), som kan bruges til at beskrive og analysere forskellige problemstillinger med

I dag ser vi nærmere på følgende kontinuerte fordelinger:

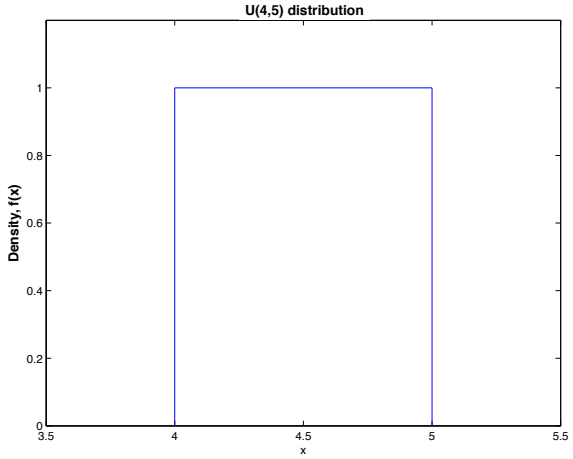
- Den uniforme fordeling
- Normalfordelingen
- Log-normalfordelingen
- Eksponentialfordelingen

Kontinuerte fordelinger in R

R	Distribution
`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` Fraktiler i fordeling (quantile).
- `r` Tilfældige tal fra fordelingen (random).

Tæthed for en uniform fordeling (eksempel)



Den uniforme fordeling, Def. 2.35 & Theo. 2.36

Syntaks:

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

Tæthedsfunktion:

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

Middelværdi:

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

Varians:

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

Eksempel 1

Studerende på et statistikkursus til forelæsning mellem 8.00 og 8.30. Det antages at ankomsttiden kan beskrives ved en uniform fordeling.

Spørgsmål:

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

Svar:

$10/30 = 1/3$

Lad $X \sim U(0,30)$ repræsentere ankomsttid:
$P(20 \leq X \leq 30) = P(X \leq 30) - P(X \leq 20) = 1 - 2/3 = 1/3$

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

[1] 0.33

Eksempel 1 (fortsat)

Spørgsmål:

Hvad er sandsynligheden for at en tilfældigt udvalgt studerende ankommer efter 8:30?

Svar:

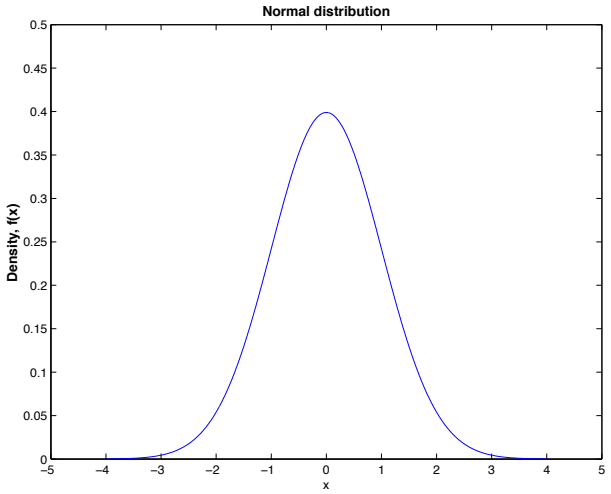
0

Lad $X \sim U(0,30)$ repræsentere ankomsttid:
$P(X > 30) = 1 - P(X \leq 30) = 1 - 1 = 0$

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

[1] 0

Tætheden for en normalfordeling (eksempel)



Normalfordelingen, Def. 2.37 & Theo. 2.38

Skrivemåde:

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

Tæthedsfunktion:

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

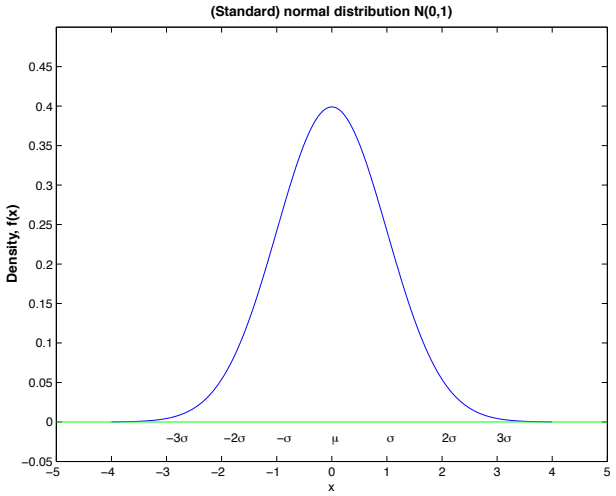
Middelværdi:

$\mu = \mu$

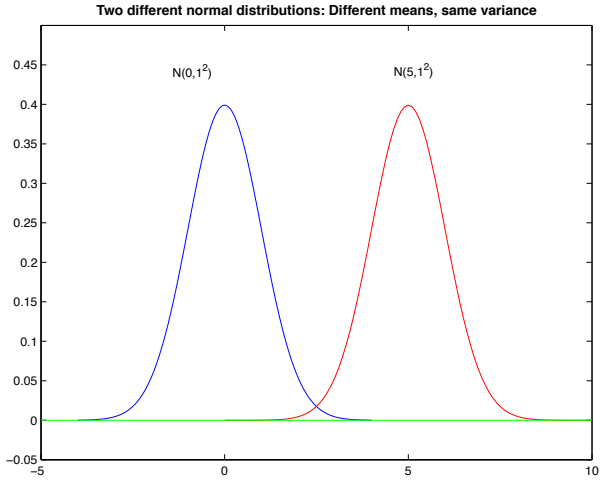
Varians:

$\sigma^2 = \sigma^2$

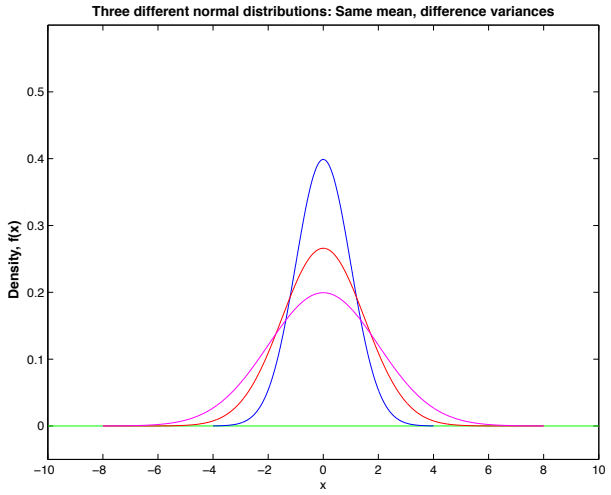
Tætheden for en standard-normalfordeling



Tætheden for to normalfordelinger (eksempel)



Tætheden for tre normalfordelinger (eksempel)



Standard-normalfordelingen

Standard-normalfordelingen:

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

Normalfordelingen med middelværdi 0 og varians 1.

Standardisering:

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

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

Eksempel 2

Målefejl:

En given vægt har en målefejl, Z, som kan beskrives standard normalfordeling,

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

Dvs. at den gennemsnitlige målefejl er $\mu = 0$ med standardafgivelse $\sigma = 1$ gram. Antag at vægten bruges til at veje et produkt.

Spøgsmål a):

Hvad er sandsynligheden for at vægten giver et resultat, som er mindst 2 gram mindre end den sande vægt af produktet?

Svar:

$P(Z \leq -2) = 0.02275$

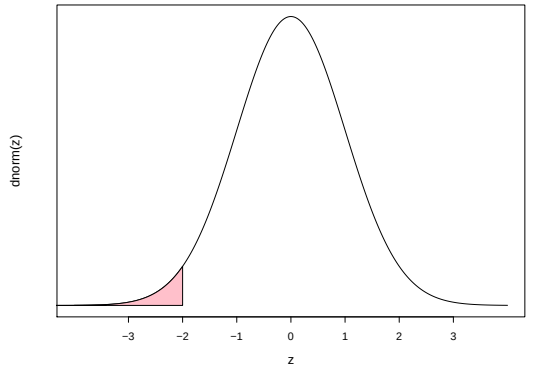
```
pnorm(-2)
```


Eksempel 2

Svar:

```
pnorm(-2)
```

```
[1] 0.023
```



Eksempel 2

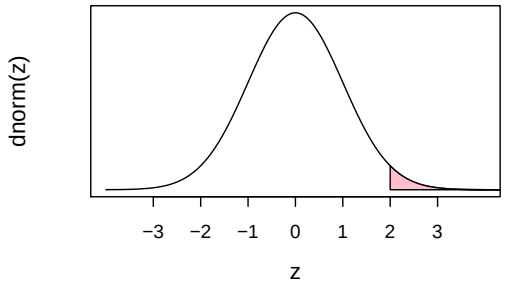
Spørgsmål b):

Hvad er sandsynligheden for at vægten giver et resultat, som er mindst 2 gram større end den sande vægt af produktet?

Svar:

$P(Z \geq 2) = 0.02275$

```
1 - pnorm(2)
```



Eksempel 2

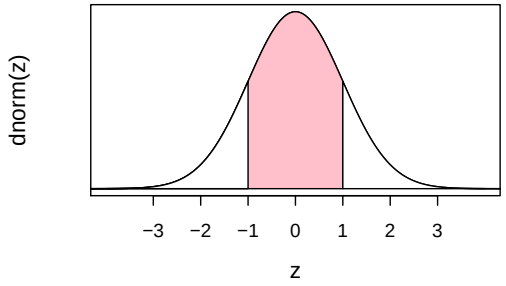
Spørgsmål c):

Hvad er sandsynligheden for at vægten har en afvigelse på højst et ± 1 gram?

Svar:

$P(|Z| \leq 1) = P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = 0.683$

```
pnorm(1) - pnorm(-1)
```



Eksempel 3

Indkomstfordeling:

Det antages at folkeskolelæreres løn kan beskrives med en normalfordeling med middleværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

Spørgsmål a):

Hvad er sandsynligheden for at en tilfældigt udvalgt lærer tjener mere end 300.000 kr?

Eksempel 3

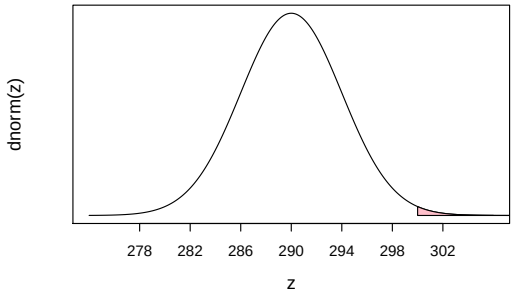
Spørgsmål a):

Hvad er sandsynligheden for at en tilfældigt udvalgt lærer tjener mere end 300.000 kr?

Svar:

```
1 - pnorm(300, m = 290, s = 4)
```

[1] 0.0062



Eksempel 4

(Samme indkomstfordeling):

Det antages at folkeskolelæreres løn kan beskrives med en normalfordeling med middleværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

"Modsat" spørgsmål:

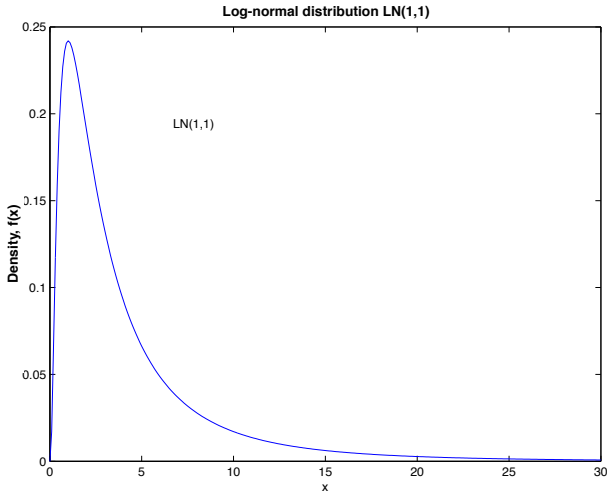
Specificér et løninterval (som er symmetrisk omkring middelværdien), som dækker 95% af lærernes lønninger.

Svar:

```
qnorm(c(0.025, 0.975), m = 290, s = 4)
```

[1] 282 298

Log-normalfordelingen



Log-normalfordelingen, Def. 2.46 & Theo. 2.47

Skrivemåde:

$X \sim LN(\alpha, \beta^2)$ (hvor $\beta > 0$)

Tæthedsfunktion:

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

Middelværdi:

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

Varians:

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

Log-normalfordelingen

Log-normal- og normalfordelingen:

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

$$X = \ln(Y),$$

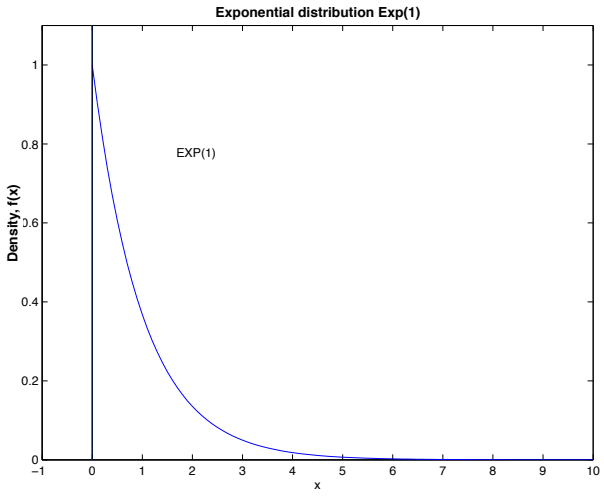
som er normalfordelt med middelværdi α og varians β^2, ie. $X \sim N(\alpha, \beta^2)$.

Variablen Z,

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

er standard-normalfordelt, ie. $Z \sim N(0, 1)$.

Eksponentialfordelingen



Eksponentialfordelingen, Def. 2.48 & Theo. 2.49

Skrivemåde:

$$X \sim \text{Exp}(\lambda)$$

hvor $\lambda > 0$.

Tæthedsfunktion:

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

Middelværdi:

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

Varians:

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

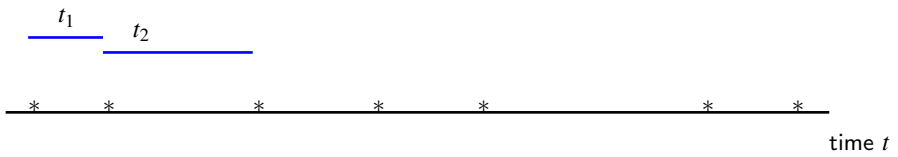
Eksponentialfordelingen

- Eksponentialfordelingen er et specialtilfælde af *gammafordelingen*.
- Eksponentialfordelingen anvendes f.eks. til at beskrive levetider og ventetider.
- Eksponentialfordelingen kan bruges til at beskrive (vente)tiden mellem hændelser i en poissonproces.

Sammenhæng mellem eksponential- og poissonfordelingen

Poisson: Diskrete hændelser pr. enhed

Eksponential: Kontinuert afstand mellem hændelser



Eksempel 5

Kø-model – poissonproces

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

Spørgsmål:

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

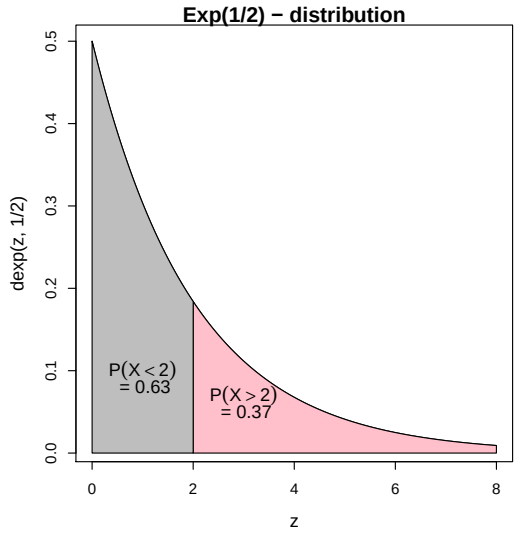
Svar:

$X \sim \text{Exp}(1/2)$ repræsenterer ventetiden indtil en ny kunde.
$P(X > 2) = 1 - P(X \leq 2)$

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

[1] 0.37

Eksempel 5



Eksempel 6

Spørgsmål:

En kunde er netop ankommet.
Brug poissonfordelingen til at beregne sandsynligheden for at der ikke kommer flere kunder inden for de næste to minutter.

Svar:

$\lambda_{2min} = 1$, $P(X = 0) = \frac{e^{-1}}{1!} 1^0 = e^{-1}$

```
dpois(0,1)
```

[1] 0.37

```
exp(-1)
```

[1] 0.37

Overview

Regneregler for stokastiske variable

Disse regneregler gælder både for kontinuerte og diskrete stokastiske variable!

Lad X være en stokatisk variabel, a og b være konstanter.

Middelværdi-regel:

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

Varians-regel:

$$\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X)$$

Eksempel 7

Lad X være en stokatisk variabel med middelværdi 4 og varians 6.

Spørgsmål:

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

Svar:

$$\mathsf{E}(Y) = -3\mathsf{E}(X) + 2 = -3 \cdot 4 + 2 = -10$$
$$\mathsf{Var}(Y) = (-3)^2\mathsf{Var}(X) = 9 \cdot 6 = 54$$

Regneregler for stokastiske variable

Lad $X_1, \ldots, X_n$ være *uafhængige* stokastiske variable.

Middelværdi-regel:

$$\begin{aligned} &\mathsf{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1\mathsf{E}(X_1) + a_2\mathsf{E}(X_2) + \cdots + a_n\mathsf{E}(X_n) \end{aligned}$$

Varians-regel:

$$\begin{aligned} &\mathsf{Var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1^2\mathsf{Var}(X_1) + \cdots + a_n^2\mathsf{Var}(X_n) \end{aligned}$$

Eksempel 8

Flypassager-planlægning

Vægten af én passagerer på et flytur antages at være normalfordelt $X \sim N(70, 10^2)$.

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

Spørgsmål:

Beregn sandsynligheden for at flyet bliver overlastet.

Hvad er den samlede passagervægt Y på en afgang?

Hvad er Y?

IKKE: $Y = 55 \cdot X$

Eksempel 8

Hvad er den samlede passagervægt Y

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

Middelværdi og varians af Y:

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

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

Y er normalfordelt, så vi kan finde $P(Y > 4000)$ ved:

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

[1] 0.022

Eksempel 8 - FORKERT analyse

Hvad er Y?

IKKE: $Y = 55 \cdot X$

Middelværdi og varians af FORKERT Y:

$$\mathsf{E}(Y) = 55 \cdot 70 = 3850$$

$$\mathsf{Var}(Y) = 55^2\mathsf{Var}(X) = 55^2 \cdot 100 = 550^2$$

Det FORKERTE Y er også normalfordelt. Her finder vi $P(Y > 4000)$ med FORKERT Y:

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

[1] 0.39

Konsekvens af forkert beregning:

MANGE spildte penge for flyselskabet!!!

Overview

1. Kontinuerte fordelinger
   - Tætheds- og fordelingsfunktioner
   - Middelværdi, varians og kovarians
2. Vigtige kontinuerte fordelinger
   - Den uniforme fordeling
   - Normalfordelingen
   - Log-normalfordelingen
   - Eksponentialfordelingen
3. Regneregler for stokastiske variable