# 02402: Introduktion til Statistik

## Uge 2: Stokastiske variable og diskrete fordelinger

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby



## Agenda







## Overview







## Stokastiske variable (Random variables)

En stokatisk variabel repræsenterer værdien af et udfald *før* det tilhørende *eksperiment* finder sted.

- Et terningekast
- Antallet af seksere i ti terningekast
- Hvor stor en andel, der svarer "ja"til et spørgsmål
- En bils benzinforbrug.
- Måling af sukkerniveau i blodprøve
- ...

## Diskret eller kontinuert stokastisk variabel

- Vi skelner mellem *diskrete* og *kontinuerte* stokastiske variable.
- Diskret:
  - Hvor mange i dette rum der bruger briller.
  - Antal der letter fra Københavns Lufthavn inden for en time.
- Kontinuert:
  - Vindmåling.
  - Transporttid til DTU.
- I dag er det diskret, i næste uge er det kontinuert.





## Stokastisk variabel

Før eksperimentet udføres har vi en stokastisk variabel

$$X \text{ (eller } X_1, \ldots, X_n)$$

noteret med store bogstaver. Så udføres eksperimentet, vi har da et *udfald* (realisation) eller en *observation*

$$x \text{ (eller } x_1, \ldots, x_n)$$

noteret med små bogstaver.





## Simulation: kast en terning i R

```
# One random draw from (1,2,3,4,5,6)
# with equal probability for each outcome
sample(1:6, size = 1)
```

[1] 1





## Diskrete fordelinger

- Stokastiske variable beskriver udfaldet af et eksperiment, før det udføres.
- Hvordan kan vi regne på eksperimentet før det er udført?
- Løsning: Brug *tæthedsfunktionen* (density function).

## Tæthedsfunktion, diskret stokastisk variabel, Definition 2.6

*Tæthedsfunktionen* (density function / probability density function, forkortelse: pdf) for en diskret stokastisk variabel:

**Definition**

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

Sandsynligheden for at $X$ antager værdien $x$ når eksperimentet udføres.





## Tæthedsfunktion, diskret stokastisk variabel, Definition 2.6

Tæthedsfunktionen for en diskret stokastisk variabel opfylder følgende to betingelser:

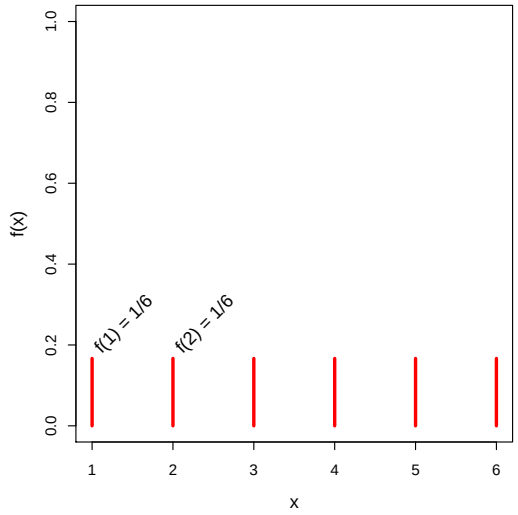
**Definition**

$$f(x) \geq 0 \text{ for alle } x \quad \text{og} \quad \sum_{\text{alle } x} f(x) = 1$$





## Eksempel: En fair ternings tæthedsfunktion








## Stikprøve

Hvad nu hvis vi kun har én observation, kan vi da se fordelingen? Nej!

Men hvis vi har $n$ observationer, så har vi en *stikprøve* (sample).

$$\{x_1, x_2, \ldots, x_n\}$$

og da kan vi begynde at 'se' fordelingen.

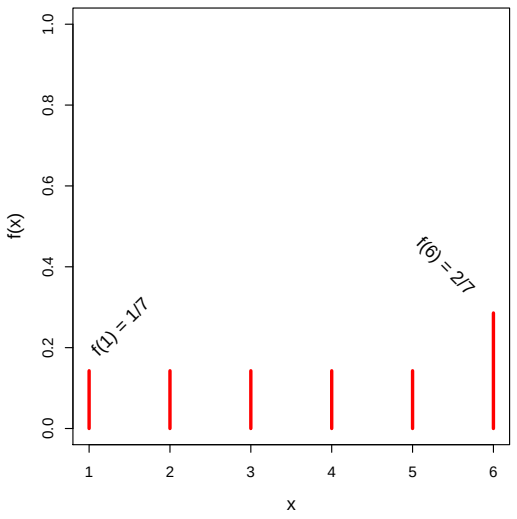
# Eksempel: Simulér $n$ kast med en fair terning

```r
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability of each outcome
xFair <- sample(1:6, size = n, replace = TRUE)
xFair

# Count number of each outcome using the 'table' function
table(xFair)

# Plot the empirical pdf
plot(table(xFair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
     ylab = "Density f(x)")
# Add the true pdf to the plot
lines(rep(1/6,6), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```





# Tæthedsfunktionen for en "unfair" terning








# Eksempel: Simulér $n$ kast med en unfair terning

```r
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with higher probability of getting a six
xUnfair <- sample(1:6, size = n, replace = TRUE, prob = c(rep(1/7,5),2/7))
xUnfair

# Plot the empirical pdf
plot(table(xUnfair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
     ylab = "Density f(x)")
# Add the true pdf to the plot
lines(c(rep(1/7,5),2/7), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

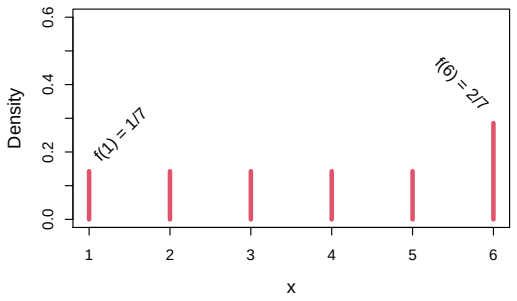




# Spørgsmål

Lad $X$ beskrive et kast med den *unfair* terning Hvad er:

- Sandsynligheden for at få 4?
- Sandsynligheden for at få 5 eller 6?
- Sandsynligheden for at få mindre end 3?



## Overview







## Fordelingsfunktion for en diskret stokastisk variabel: Definition 2.9

*Fordelingsfunktionen* (cumulative distribution function, cdf) for en diskret stokastisk variabel:

### Definition

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

Der gælder for en fordelingsfunktion (cdf):

- Det er en 'ikke-aftagende' funktion
- Den nærmer sig (konvergerer mod) 1 når $x \to \infty$.





## Eksempel: Fair terning

Lad $X$ repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at få et udfald mindre end 3:

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





## Eksempel: Fair terning

Find sandsynligheden for at få et udfald større end eller lig 3:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - F(2) \text{ fordelingsfunktionen} \\ &= 1 - \frac{1}{3} = \frac{2}{3} \end{aligned}$$

## Overview







## Konkrete statistiske fordelinger

- Der findes en række statistiske fordelinger, som kan bruges til at beskrive og analysere forskellige problemstillinger med.
- I dag er det diskrete fordelinger:
  - Binomialfordelingen
  - Den hypergeometriske fordeling
  - Poissonfordelingen





## Binomialfordelingen

- Vi betragter et eksperiment med to udfald, "succes" og "ikke-succes", som gentages et vist antal gange (uafhængige gentagelser).
- $X$ være antallet af succeser efter $n$ gentagelser
- $X$ følger en binomialfordeling:

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

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





## Binomialfordelingens tæthedsfunktion

Sandsynligheden for $x$ antal succeser:

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

hvor

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

## Eksempel

Antag $X \sim B(4,p)$, i.e. $n = 4$. Find sandsynligheden for 3 successer.

- Sandsynligheden for 3 successer: $P(X = 3)$.
- De tre successer kan "fås" på fire forskellige måder: SSSF, SSFS, SFSS, FSSS.
- Altså:

$$\binom{n}{x} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4,$$

og

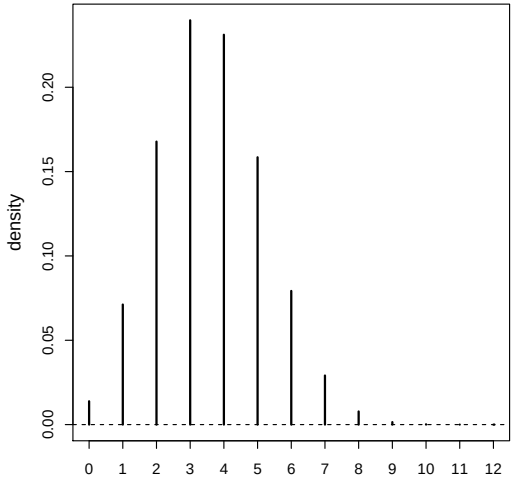
$$P(X = 3) = 4p^3(1-p).$$





## Binomialfordelingen

Eksempel, $B(12, 0.3)$:








## Simulation med binomialfordeling

```
## Probability of success
p <- 0.1

## Number of repetitions
nRepeat <- 30

## Simulate Bernoulli experiment 'nRepeat' times
tmp <- sample(c(0,1), size = nRepeat, prob = c(1-p,p), replace = TRUE)

# Compute 'x'
sum(tmp)

## Or: Use the binomial distribution simulation function
rbinom(1, size = 30, prob = p)
```





## Eksempel: Fair terninger

```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability for each outcome
xFair <- sample(1:6, size = n, replace = TRUE)

# Count the number of six'es
sum(xFair == 6)

## Do the same using 'rbinom()' instead
rbinom(n = 1, size = 30, prob = 1/6)
```

## Eksempel 1

I et kundecenter i et telefonselskab søger man at forbedre kundetilfredsheden. Især er det vigtigt at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.
Antag at sandsynligheden for at en fejl bliver udbedret i løbet af samme dag er 70%.

*I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?*

- Step 1) Hvad skal repræsenteres af den stokastiske variabel $X$?
  Antallet af udbedrede fejl.
- Step 2) Hvad er fordelingen af $X$?
  En binomialfordeling med $n = 6$ og $p = 0.7$.





## Eksempel 1

I et kundecenter i et telefonselskab søger man at forbedre kundetilfredsheden. Især er det vigtigt at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.
Antag at sandsynligheden for at en fejl bliver udbedret i løbet af samme dag er 70%.

*I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?*

- Step 3) Hvilken sandsynlighed skal udregnes
  $P(X = 6) = f(6; 6, 0.7)$





## Overview







## Den hypergeometriske fordeling

- $X$ er igen antallet succeser, men nu *uden* tilbagelægning ved trækningen.
- $X$ følger da den hypergeometriske fordeling

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

  - $n$ er antallet af trækninger (gentagelser)
  - $a$ er antallet af succeser i populationen
  - $N$ er antallet af elementer i (hele) populationen

## Den hypergeometriske fordeling

- Sandsynligheden for at få $x$ succeser er

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

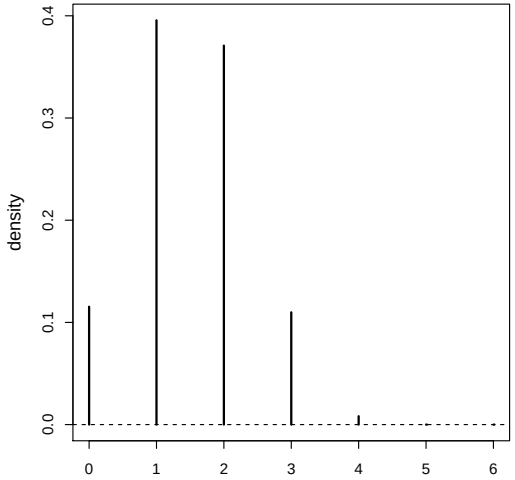
  - $n$ er antallet af trækninger (gentagelser)
  - $a$ er antallet af succeser i populationen
  - $N$ er antallet af elementer i (hele) populationen





## Den hypergeometriske fordeling

Eksempel, $H(6,4,12)$:








## Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- Step 1) Hvad skal repræsenteres af den stokastiske variabel $X$?
  Antallet af harddiske med skramme i stiprøvenr.
- Step 2) Hvad er fordelingen af $X$?
  En hypergeometrisk fordeling med $n=3$, $a=2$, $N=10$.
- Step 3) Hvilken sandsynlighed skal udregnes?
  $P(X \geq 1) = 1 - P(X=0) = 1 - f(0;3,2,10)$





## Binomial vs. hypergeometrisk

- Binomialfordelingen bruges til at analysere stikprøver med tilbagelægning.
- Den hypergeometriske fordeling bruges til at analysere stikprøver uden tilbagelægning.

## Overview







## Poissonfordelingen

- Poissonfordelingen anvendes ofte som en fordeling (model) for tælletal, hvor der ikke er nogen naturlig øvre grænse.
- Poissonfordelingen karakteriseres/defineres normalt ved en *intensitet*, som har formen "antal/enhed", ofte benævnt $\lambda$.
- Typisk *hændelser per tidsinterval*





## Poissonfordelingen

$$X \sim Po(\lambda)$$

Tæthedsfunktion:

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

Fordelingsfunkion:

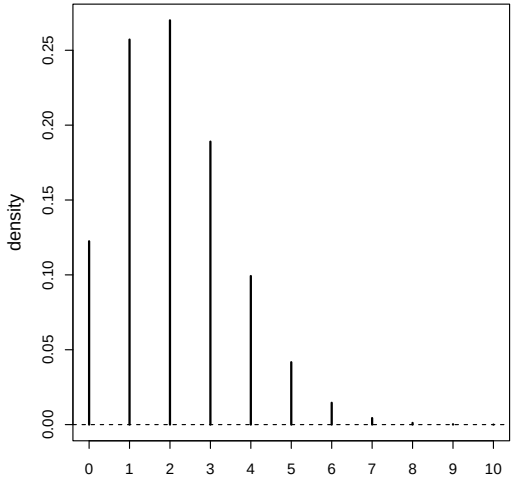
$$F(x) = P(X \leq x)$$





## Poissonfordelingen

Eksempel, $Po(2.1)$:



# Eksempel 3

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 af den stokastiske variabel $X$?
  Antal patienter på en given dag.
- Step 2) Hvad er fordelingen af $X$?
  En poissonfordeling med $\lambda = 0.3$.
- Step 3) Hvilken sandsynlighed skal udregnes?
  $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$





# Overview







| R | Name |
|---|---|
| `binom` | Binomialfordelig |
| `hyper` | Hypergeometrisk fordeling |
| `pois` | Poissonfordeling |

`d` $f(x)$, tæthedsfunktion

`p` $F(x)$, fordelingsfunktion

`r` trækker tilfælige tal fra fordelingen (simulation)

`q` fraktiler fra fordelingen ("invers" of $F(x)$)

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

```
pbinom(q = 5, size = 10, prob = 0.6)
```

[1] 0.3669

```
# Get help with:
?pbinom
```





# Overview

## Middelværdi (expectation, expected value)

Middelværdien af en diskret stokastisk variabel, definition 2.13:

### Definition

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

- Det *"sande gennemsnit"* of X (i modsætning til stikprøvegennemsnittet).
- Udtykker hvor "midten" af $X$ er.





## Eksempel: Middelværdi af et kast med en fair terning

$$\begin{aligned}\mu &= \mathrm{E}(X) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5\end{aligned}$$





## Sammenligning med stikprøvegennemsnittet - lær fra simulationer

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample mean
mean(xFair)
```

```
[1] 3.3
```





## Asymptotik: når man øger stikprøvestørrelsen (sample size)

Des flere observationer (sample size), des tættere kommer vi på den sande middelværdi:

$$\lim_{n \to \infty} \hat{\mu} = \mu$$

- Kaldes *store tals lov* (law of large numbers)

## Varians

Variansen af en diskret stokastisk variabel, Definition 2.16:

### Definition

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

- Måler den gennemsnitlige spredning.
- Den "rigtige varians" af $X$ (i modsætning til stikprøvevariansen).





## Eksempel: Varians af et kast med en fair terning

$$\begin{aligned}\sigma^2 &= \mathrm{E}[(X-\mu)^2] \\ &= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4-3.5)^2 \cdot \frac{1}{6} + (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} \\ &\approx 2.92\end{aligned}$$





## Sammenligning med stikprøvevarians – lær fra simulationer

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample variance
var(xFair)
```

```
[1] 2.437
```





## Middelværdi og varians for konkrete fordelinger

Binomialfordelingen:

- Middelværdi:
  $\mu = n \cdot p$
- Varians:
  $\sigma^2 = n \cdot p \cdot (1-p)$

## Middelværdi og varians for konkrete fordelinger

### Den hypergeometriske fordeling

- Middelværdi:
  $\mu = n \cdot \frac{a}{N}$
- Varians:
  $\sigma^2 = \frac{n \cdot a \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$





## Middelværdi og varians for konkrete fordelinger

### Poissonfordelingen

- Middelværdi:
  $\mu = \lambda$
- Varians:
  $\sigma^2 = \lambda$





## Agenda