Introduktion til Statistik

Forelæsning 2: Stokastisk variabel og diskrete fordelinger

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Spring 2017

Kap. 2: Diskrete fordelinger

Grundlæggende koncepter:

- Stokastisk variabel *(værdi afhængig af udfald af endnu ikke udført eksperiment)*
- Tæthedsfunktion: $f(x) = P(X = x)$ *(pdf)*
- Fordelingsfunktion: $F(x) = P(X \leq x)$ *(cdf)*
- Middelværdi: $\mu = E(X)$
- Standard afvigelse: σ
- Varians: σ^2

Specifikke distributioner:

- Binomial *(terningekast)*
- Hypergeometrisk *(trækning uden tilbagelægning)*
- Poisson *(antal hændelser i interval)*

Kap. 2: Discrete Distributions

General concepts:

- Random variable *(value is outcome of yet not carried out experiment)*
- Density function: $f(x) = P(X = x)$ *(pdf)*
- Distribution function: $F(x) = P(X \leq x)$ *(cdf)*
- Mean: $\mu = E(X)$
- Standard deviation: σ
- Variance: σ^2

Specific distributions:

- The binomial distribution *(dice roll)*
- The hypergeometric distribution *(draw without replacement)*
- The Poisson distribution *(number of events in interval)*

Oversigt

Praktisk information

SE PRAKTISK INFORMATION PÅ HJEMMESIDEN OG I STARTEN AF SLIDES FRA UGE 1

Stokastisk variabel

En **stokastisk variabel** (random variable) tildeler en værdi til udfaldet af et eksperiment *der endnu ikke er udført*, f.eks.:

- Et terningekast
- Antallet af seksere i 10 terningekast
- Hvor stor en andel svarer ja til et spørgsmål
- km/l for en bil
- Måling af sukkerniveau i blodprøve
- ...

Diskret eller kontinuert

- Vi skelner mellem diskret og kontinuert
- **Diskret** (kan ofte tælles):
  - Hvor mange der bruger briller herinde
  - Antal mange flyvere letter den næste time
  - ...
- **Kontinuert**:
  - Vindmåling
  - Tiden det tog at komme til DTU
  - ...
- Der er en gråzone, f.eks. for observationer målt med lav opløsning.
- I dag er det diskret og i næste uge er det kontinuert.

Stokastisk variabel

- Før eksperimentet udføres: en **stokastisk variabel**

  $$X_1$$

  noteret med stort bogstav

- Så udføres eksperimentet: vi har da en *realisation* eller *observation*

  $$x_1$$

  noteret med småt bogstav.

Stokastisk variabel og stikprøve

- Før eksperimentet udføres: **stikprøven** som n stokastiske variable

$$X_1, X_2, \ldots, X_n$$

  noteret med stort bogstav

- Så udføres eksperimentet: vi har da n *realisationer* (*observationer*)

$$x_1, x_2, \ldots, x_n$$

  noteret med småt bogstav

- Dvs. vi udfører eksperimentet n gange for at lave stikprøven

Eksempel: Simuler et terningekast

- Vælg et tal fra $(1, 2, 3, 4, 5, 6)$ med lige stor sandsynlighed for hvert udfald
- Simuler i R

```r
## Simuler et terningekast

## Vælg et tal fra (1,2,3,4,5,6) med lige sandsynlighed for hvert udfald
sample(1:6, size=1)

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
sample(1:6, size=n, replace=TRUE)
```

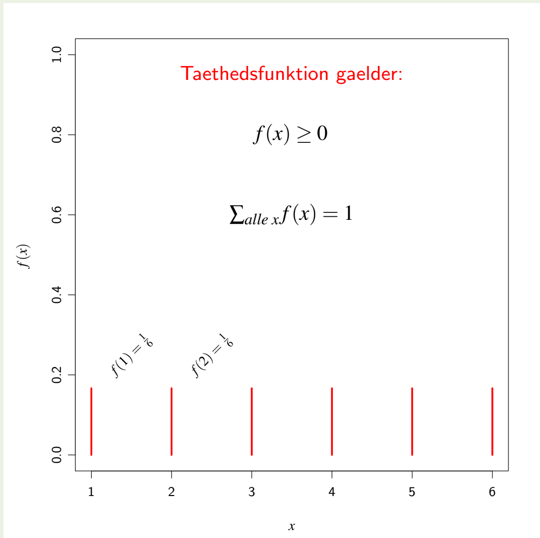

Tæthedsfunktion (probability density function (pdf))

En stokastisk variabel har en **tæthedsfunktion**

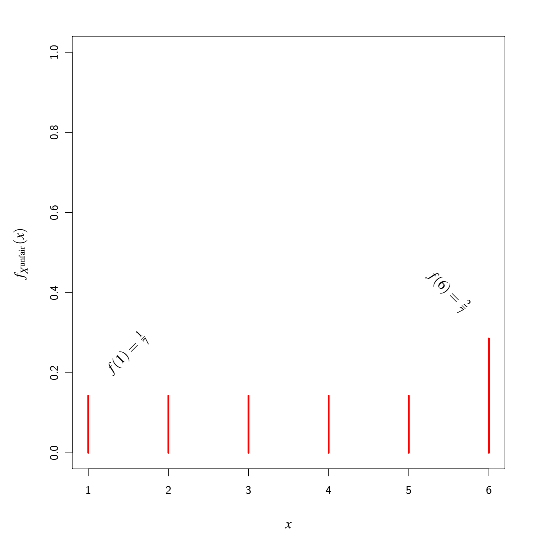
$$f(x) = P(X = x)$$

Sandsynligheden for at X antager værdien x når eksperimentet udføres

Eksempel: En fair ternings tæthedsfunktion



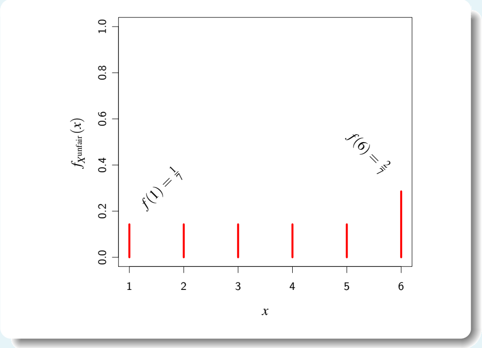
Eksempel: En unfair ternings tæthedsfunktion



Spørgsmål om unfair terning (socrative.com, room: PBAC)

Find nogle sandsynligheder for X^{unFair}:

- Sandsynligheden for at få en fire? Svar: E
- Sandsynligheden for at få en femmer eller en sekser? Svar: A
- Sandsynligheden for at få mindre end tre? Svar: D



Svarmuligheder:

A: $\frac{3}{7}$

B: $\frac{1}{6}$

C: $\frac{4}{7}$

D: $\frac{2}{7}$

E: $\frac{1}{7}$

Stikprøve

- Vi har en terning og vil nu undersøge om en terningen er fair.
- Hvis vi kun har en 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.

Simuler n kast med en fair terning:

```
## Simuler en fair terning

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)
## Tæl antallet af hvert udfald
table(xFair)
## Plot den empiriske tæthedsfunktion (pdf), altså et density histogram
plot(table(xFair)/n, ylim=c(0,1), lwd=10, xlab="x", ylab="f(x)")
## Tilføj den rigtige tæthedsfunktion til plottet
lines(rep(1/6,6), type="h", lwd=3, col="red")
## legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```

Simuler n kast med en ikke-fair terning:

```
## Simuler en ikke-fair terning

## Antal simulerede realiseringer
n <- 30
## Træk uafhændigt fra mængden (1,2,3,4,5,6) med højere sandsynlighed for en sekser
xUnfair <- sample(1:6, size=n, replace=TRUE, prob=c(rep(1/7,5),2/7))
## Tæl antallet af hvert udfald
table(xUnfair)
## Plot den empiriske tæthedsfunktion
plot(table(xUnfair)/n, lwd=10, ylim=c(0,1), xlab="x", ylab="Density")
## Tilføj den rigtige tæthedsfunktion
lines(c(rep(1/7,5),2/7), lwd=4, type="h", col=2)
## En legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```


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

Fordelingsfunktionen (cdf) er tæthedsfunktionen akkumuleret

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

Der gælder for en fordelingsfunktion (cdf):

- Den er en 'ikke-aftagende' funktion
- Den akkumuleres (assymtotisk) til 1 når $x \to \infty$

Eksempel: Fair terning

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

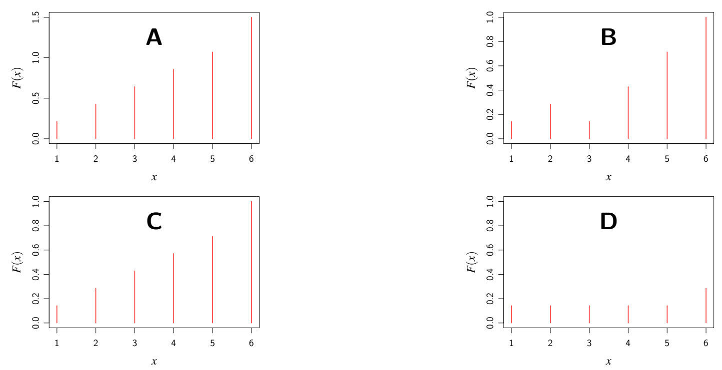
$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ \textit{fordelingsfunktionen}} \\ &= P(X = 1) + P(X = 2) \\ &= f(1) + f(2) \text{ \textit{tæthedsfunktionen}} \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \end{aligned}$$

Eksempel: Fair terning

- Udregn sandsynligheden for at få udfald over eller lig 3:

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

Spørgsmål: Fordelingsfunktion (cdf) (socrative.com, room: PBAC)



Hvilket et af ovenstående plots kan være en fordelingsfunktion (akkumuleret tæthedsfunktion, cdf)?

A, B, C eller D? Svar: C

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

- Lad X repræsentere antal succeser efter n gentagelser af handling (eksperiment) med to udfald (succes eller ikke-succes)
- X følger **binomialfordelingen**

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

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

- Tæthedsfunktion: Sandsynlighed for x antal succeser

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

Eksempel: Binomialfordelingen

Eksempel: Sandsynlighed for 2 plat ved 5 plat-eller-krone kast med mønt

$$f(2;5,0.5) = P(X=2) = \binom{5}{2} 0.5^2 (1-0.5)^{5-2} = 0.3125$$

```
## Sandsynlighed for 2 plat (success) i 5 kast med mønt

## Slå op med binomial tæthedsfunktion
dbinom(x=2, size=5, prob=0.5)
```

Binomialfordeling simuleringseksempel i R:

```
## Simuler en binomialfordeling

## Sandsynlighed for success
p <- 0.1
## Antal gentagelser af succes og ikke-succes eksperimentet
nRepeat <- 30
## Simuler Bernoulli eksperiment nRepeat gange
tmp <- sample(c(0,1), size=nRepeat, prob=c(1-p,p), replace=TRUE)
## x er nu
sum(tmp)

## Lav tilsvarende med funktion til simulering af binomialfordeling
rbinom(1, size=30, prob=p)
```

Binomialfordeling med terning i R:

```
## Fair terning eksempel

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)
## Tæl sammen hvor mange seksere
sum(xFair == 6)

## Lav tilsvarende med rbinom()
rbinom(n=1, size=30, prob=1/6)
```


Hypergeometrisk fordeling

- X er igen antal succeser, men nu er det *uden tilbagelægning ved gentagelsen*
- X følger en **hypergeometrisk fordeling**

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

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

- Tæthedsfunktion: Sandsynlighed for at få x succeser

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

Binomial vs. hypergeometrisk

- Binomialfordelingen 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)

PAUSE

R navn	Betegnelse
binom	binomial
hyper	hypergeometrisk

- 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.

Eksempel binomialfordelt:

Find

$$P(X \leq 5) = F(5; 10, 1/6)$$

```
## Binomial fordelingsfunktion (cdf)

## Sandsynlighed for at få 5 eller færre succeser i 10 kast med terning
pbinom(q=5, size=10, prob=1/6)
## Få hjælpen med
?pbinom
```

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

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 altid er 0.7.

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

- Step 1) Hvad skal repræsenteres: X er antal udbedrede fejl
- Step 2) Fordeling: X følger A: binomial, B: hypergeometrisk? binomialfordelingen
- Step 3) Hvilken sandsynlighed: $P(X\,?\,?)$ $P(X = 6) = f(6; n, p)$
- Step 4)
  - Hvad er antal trækninger? $n = 6$
  - Hvad er succes-sandsynligheden? $p = 0.7$

    Udregn i R

Eksempel 1

Hvad er sandsynligheden for at 2 eller færre fejl bliver udbedret samme dag?

- Step 1) Hvad skal repræsenteres: X er antal udbedrede fejl
- Step 2) Hvilken fordeling: X følger binomialfordelingen
- Step 3) Hvilken sandsynlighed:

A: $P(X < 2)$ B: $P(X \leq 1)$ C: $P(X < 3)$ D: $1 - P(X \geq 2)$

$P(X < 3) = P(X \leq 2) = F(2; n, p)$

- Step 4)
  - Hvad er antal trækninger? $n = 6$
  - Hvad er succes-sandsynligheden? $p = 0.7$

    Udregn i R

Eksempel 2

I bankoklubben trækkes der lod om 5 flasker Gammel Dansk. Der er udstedt 50 lodder og du har 3 lodder (der er altså 50 sedler med numrene fra 1 til 50 og der er 3 af de numre som du vinder på).

Hvad er sandsynligheden for at du vinder præcis en flaske Gammel Dansk?

- Step 1) Hvad skal repræsenteres: X er antal flasker du vinder
- Step 2) Hvilken fordeling: X følger den hypergeometriske fordeling
- Step 3) Hvilken sandsynlighed: $P(X\,?\,?)$ $P(X=1)=f(1;n,a,N)$
- Step 4)
  - Hvad er antal trækninger? $n=5$
  - Hvor mange succeser er der? $a=3$
  - Hvor mange er der i alt? $N=50$

  Udregn i R

Poissonfordelingen

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

Poissonfordelingen

- X følger **Poissonfordelingen**

$$X \sim P(\lambda)$$

  - Parameteren λ angiver intensiteten

- Tæthedsfunktion: Sandsynligheden for x antal i intervallet

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

Eksempel 3.1: Poissonfordelingen

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 Poissonfordelingen
- Step 3) Hvilken sandsynlighed: $P(X\,?\,?)$ $P(X \leq 2)$
- Step 4) Hvad er raten: $\lambda = 0.3$ patienter per dag

  Udregn i R

Eksempel 3.2: Poissonfordelingen

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 præcis 2 patienter?

- Step 3) Hvilken sandsynlighed: $P(X\,?\,?)$ $\quad P(X=2)$
  Udregn i R

Eksempel 3.3: Poissonfordelingen

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 mindst 2 patienter?

- Step 3) Hvilken sandsynlighed: $P(X\,?\,?)$ $\quad P(X \geq 2) \quad = 1 - P(X \leq 1)$
  Udregn i R

Eksempel 3.4: Skalering af intensiteten i Poissonfordeling

Hvad er sandsynligheden for at der i en periode på 3 dage bliver indlagt præcis 1 patient?

- Step 1) Hvad skal repræsenteres:
  - Fra X som er *patienter per dag*
  - Til X^{3dage} som er *patienter per 3 dage*
- Step 2) Hvilken fordeling følger X^{3dage}: Poissonfordelingen
- Step 3) Hvilken sandsynlighed: $P(X^{\text{3dage}}\,?\,?)$ $\quad P(X^{\text{3dage}} = 1)$
- Step 4) Skaler raten
  - Fra $\lambda_{\text{dag}} = 0.3$ patient/dag til $\lambda_{\text{3dage}} = 0.9$ patient/3 dag
    Udregn i R

Middelværdi (mean) og forventningsværdi (expectation)

Definition: Middelværdi af stokastisk variabel

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

- Populationsgennemsnittet (det "rigtige gennemsnit")
- Fortæller hvor "midten" af tæthedsfunktion for X er

Eksempel: Middelværdi

Middelværdi af et terningekast

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

Eksempel: Simuler terningekast og beregn gennemsnit

```
## Simuler stikprøve af en fair terning og beregn gennemsnit

## Antal simulerede realiseringer (stikprøve på n elementer)
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)

## Udregn stikprøvegennemsnit (sample mean)
mean(xFair)
```


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

Hvad sker der generelt med gennemsnittet af en stikprøve *når man får flere observationer*?

A: Det er uafhængigt af antal observationer

B: Det kommer generelt længere væk fra middelværdien

C: Det kommer generelt tættere på middelværdien

Svar C: Des flere observationer, des tættere kommer man generelt på middelværdien.

Prøv at lege med det i ved simulering i R

Varians (variance)

Definition: Varians af stokastisk variabel

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

- Et mål for spredningen
- Populationsvariansen
- Den “rigtige spredning“ af X tæthedsfunktion

Eksempel: Varians

Varians af terningekast

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

Eksempel: Varians

```
## Simuler stikprøve med udfald af en fair terning og beregn stikprøvevarians

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)

## Udregn empirisk varians (sample variance, læg mærke til
## at i R hedder funktionen 'var')
var(xFair)
```


Middelværdi og varians for de diskrete fordelinger

Fordeling	**Middelværdi**	**Varians**
Binomialfordelingen	$\mu = n \cdot p$	$\sigma^2 = n \cdot p \cdot (1-p)$
Hypergeometrisk	$\mu = n \cdot \frac{a}{N}$	$\sigma^2 = \frac{na \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$
Poissonfordelingen	$\mu = \lambda$	$\sigma^2 = \lambda$

Eksempel: Forskel på stikprøvegennemsnit (sample mean) og middelværdi (mean, dvs. populationsgennemsnittet)

Se stikprøvegennemsnittet i forhold til middelværdien:

```
## Simuler en binomialfordeling, terninge eksempel

## Gentag 10 gange: Tæl sammen for mange seksere på 30 slag
antalSeksere <- rbinom(n=10, size=30, prob=1/6)

## Endelig kan vi se på stikprøvegennemsnittet (sample mean)
mean(rbinom(n=10, size=30, prob=1/6))
## versus Middelværdien (mean)
n * 1/6
```

Oversigt