02323 Introduktion til statistik

Uge 2: Stokastiske variable og diskrete fordelinger

Nicolai Siim Larsen
DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

Overview

Opsummering: Uge 1

Vi ønsker at undersøge en population.

Populationen kan beskrives ved bl.a. positionsmål og spredningsmål. Hvis populationen består af N individer, kan populationsgennemsnittet og -variansen beregnes ved

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2.$$

Opsummering: Uge 1

Hvis vi har en repræsentativ stikprøve med n observationer, og vi ønsker at estimere populationsparametrene (lave statistisk inferens), kan man udregne stikprøvegennemsnittet og -variansen ved

$$\bar{x} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Vi bemærker, at der divideres med $n-1$ i beregningen af stikprøvevariansen, da vi benytter det estimerede gennemsnit $\hat{\mu}$ i stedet for μ. Hvis μ kendes, kan denne anvendes i formlen, og man dividerer med n.

Stikprøvevariansen er blot estimatet for populationsvariansen. Det er ikke variansen i stikprøven!

Overview

Eksperimenter og stokastiske variable (random variables)

Setup: Et eksperiment.

Udfaldsrummet S er mængden af alle eksperimentets mulige udfald.

En stokastisk variabel X er en afbildning/funktion

$$X : S \to \mathbb{R}.$$

En stokastisk variabel repræsenterer værdien af udfaldet *før* det tilhørende *eksperiment* finder sted.

Eksempler

Nogle eksempler på stokastiske variable:

- Forelæsningens varighed
- Antallet af seksere i ti terningkast
- Andelen af stemmer til Det Republikanske Parti ved næste præsidentvalg
- En patients blodsukkerniveau
- Årsresultatet i Novo Nordisk
- Antal placeringer DTU er steget på QS University Ranking siden sidste år
- Ventetiden til Danmark vinder VM i fodbold

Hvilke egenskaber karakteriserer stokastiske variable?

Diskret eller kontinuert stokastisk variabel

Vi skelner mellem *diskrete* og *kontinuerte* stokastiske variable.

- Diskret: Værdimængden er tællelig
  - Antal personer, der bruger briller i lokalet
  - Antal passagerer, der letter fra Københavns Lufthavn inden for en time
- Kontinuert: Værdimængden er utællelig
  - Vindmåling
  - Transporttid til DTU
- I dag behandler vi diskrete variable, medens næste uges pensum omhandler kontinuerte variable.

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, og vi har et udfald. Udfaldet giver anledning til en observation (observeret værdi)

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

noteret med små bogstaver.

Simulation: Kast en terning i R

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

```
[1] 2
```


Fordelinger (Distributions)

- Stokastiske variable beskriver udfaldet af et eksperiment før det udføres.
- Hvordan kan vi regne på eksperimentet før det er udført?
- Løsning: *Fordelinger* (Distributions).

En univariat fordeling beskriver, hvordan sandsynlighedsmassen fordeles over de reelle tal.

Klassificering af fordelinger

Man kan klassificere en fordeling på flere måder:

- Fordelingsfunktionen
- Tæthedsfunktionen
- Laplacetransformationen
- Den momentgenererende funktion
- Den karakteristiske funktion

I dette kursus benytter vi kun de to første!

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.

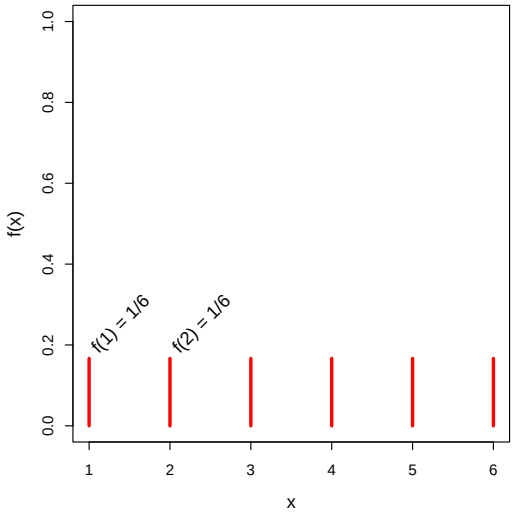
Tæthedsfunktion, diskret stokastisk variabel, Definition 2.6

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: Tæthedsfunktion - Kast med en fair terning



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)

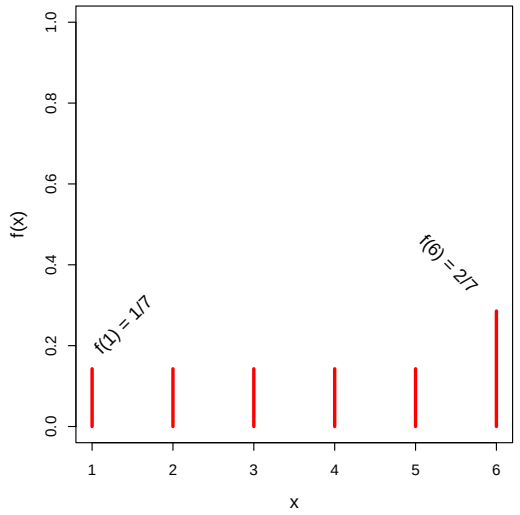
$$\{x_1, x_2, \ldots, x_n\},$$

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)
```


Eksempel: Tæthedsfunktion - Kast med en unfair terning



Eksempel: Simulér n kast med en unfair terning

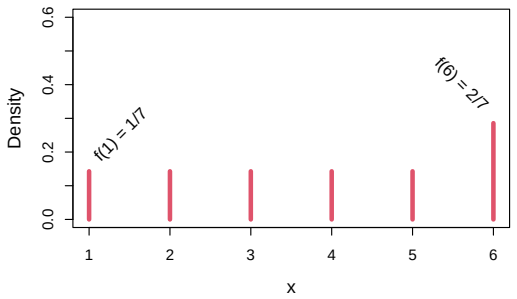
```r
# Number of simulated realizations (sample size)
n <- 30

# 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 det antal øjne, der fås ved 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: Kast med en fair terning

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

Find sandsynligheden for at observere en værdi mindre end 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: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

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

Overview

Konkrete (diskrete) statistiske fordelinger

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

Binomialfordelingen

- Vi betragter et eksperiment med to udfald: "succes" og "fiasko", som gentages et vist antal gange (uafhængige gentagelser).
- Lad X være antallet af succeser efter n gentagelser.
- Så følger X en binomialfordeling m. antalsparameter n og succesparameter p:

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

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

Binomialfordelingens tæthedsfunktion

Sandsynligheden for at observere x antal succeser gives ved

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

hvor binomialkoefficienten kan beregnes som

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

Eksempel - Binomialfordelingen

Antag $X \sim B(4,p)$, dvs. $n = 4$. Find sandsynligheden for at observere 3 succeser.

- Sandsynligheden for 3 succeser er $P(X=3)$.
- De tre succeser kan fremkomme på fire 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
  $$P(X=3) = 4p^3(1-p).$$

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: Kast med en fair terning

```
# 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 prøver man at forbedre kundetilfredsheden. Det er især 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?

- Trin 1) Hvad skal repræsenteres af den stokastiske variabel X?
  Antallet af udbedrede fejl.
- Trin 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?

- Trin 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

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?**

- Trin 1) Hvad skal repræsenteres af den stokastiske variabel X?
  Antallet af harddiske med skramme i stikprøven.
- Trin 2) Hvad er fordelingen af X?
  En hypergeometrisk fordeling med $n=3$, $a=2$ og $N=10$.
- Trin 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 λ.
- Typisk *hændelser per tidsinterval*.

Poissonfordelingen

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

Tæthedsfunktion:

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

Fordelingsfunktion:

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

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?

- Trin 1) Hvad skal repræsenteres af den stokastiske variabel X?
  Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.
- Trin 2) Hvad er fordelingen af X?
  En poissonfordeling med $\lambda = 0.3$.
- Trin 3) Hvilken sandsynlighed skal udregnes?
  $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

Overview

Fordelinger i R

R	Name
`binom`	Binomialfordeling
`hyper`	Hypergeometrisk fordeling
`pois`	Poissonfordeling

- `d` Tæthedsfunktion (density)
- `p` Fordelingsfunktion (probability)
- `r` Tilfældighedsgenerator: Simulerer tilfælige tal (random number)
- `q` Fraktilfunktion ("invers" af fordelingsfunktionen) (quantile)

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"* af X (i modsætning til stikprøvegennemsnittet).

Eksempel: Kast med en fair terning

Lad X repræsentere antallet af øjne ved et kast med en fair terning. Så følger X en diskret uniform fordeling (diskret ligefordeling) på intervallet $[1,6]$ og har middelværdi:

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

Sammenligning med stikprøvegennemsnittet - lær fra simulationer

```r
# 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] 3.733
```


Asymptotisk resultat: Store tals lov

Des flere observationer, 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 rundt om middelværdien.
- Den "rigtige varians" af X (i modsætning til stikprøvevariansen).

Eksempel: Kast med en fair terning

Lad X repræsentere antallet af øjne ved et kast med en fair terning. Så følger X en diskret uniform fordeling (diskret ligefordeling) på intervallet $[1,6]$ og har varians:

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

Sammenligning med stikprøvevariansen - 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] 3.597

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:
  $\sigma^2 = \lambda$

Dagsorden