

## Produkt og marked - matematiske og statistiske metoder

Rasmus Waagepetersen  
Institut for Matematiske Fag Aalborg Universitet

February 12, 2020

### Kursusindhold:

- ▶ Sandsynlighedsregning og lagerstyring
- ▶ Normalfordelingen og Monte Carlo-metoder

1 / 28

2 / 28

### Monte Carlo

Antag vi gerne vil beregne sandsynligheden for at et kast med 5 terninger giver mindst 15 i alt.

Er vi ikke i stand til at beregne denne sandsynlighed teoretisk, kan vi i princippet kaste 5 terninger et stort antal gange og udregne andelen af kastene, hvor summen af terningernes øjne giver 15 eller derover. Denne andel vil være et estimat af den ønskede sandsynlighed.

I praksis lader vi en computer foretage kastene vha. computer-genererede tilfældige tal.

Estimat af  $P(X \geq m)$ : Lad  $I_i = 1[X_i \geq m]$

$$P(X \geq m) \approx \frac{1}{n} \sum_{i=1}^n I_i = \bar{I}$$

Eksempel: hvis vi bruger  $n = 1000$  simulationer af kast med 5 terninger fås estimat 0.775. Men kun estimat - hvis vi tager 1000 nye kast fås andet resultat - f.eks. 0.795.

Dvs. der er en vis usikkerhed på Monte Carlo estimatet.

Usikkerheden bliver mindre jo større  $n$ , der benyttes.

3 / 28

4 / 28

## Basal statistik - empirisk middelværdi og varians

Observationer  $X_1, X_2, \dots, X_n$  (kan være "syntetiske" - genereret på computer) - alle samme fordeling som  $X$ .

Empirisk middelværdi og empirisk varians

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

('empirisk middelværdi af kvadrerede afvigelser')

Empirisk spredning:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

5 / 28

## Udregning af varians

Antag  $X$  kan antage værdierne  $m = 0, 1, 2, 3, \dots, M$  med sandsynligheder  $p(m) = P(X = m)$ .

$$\begin{aligned} \text{Var}X &= \sum_{m=0}^M p(m)(m - \mu)^2 = \\ & p(0)(0 - \mu)^2 + p(1)(1 - \mu)^2 + \dots + p(M)(M - \mu)^2 \end{aligned}$$

Eksempel: varians af binomialfordeling  $S \sim b(1, p)$  med  $\mu = p$ .

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = (1 - p)(0 - p)^2 + p(1 - p)^2 = p(1 - p)$$

7 / 28

Når  $n$  meget stor vil  $\bar{X}$  blive nøjagtig tilnærmelse af  $\mu = \mathbb{E}X$  og  $s^2$  vil tilnærme sig *variansen* af  $X$ :

$$\sigma^2 = \text{Var}X = \mathbb{E}(X - \mu)^2$$

Varians: forventede værdi af den kvadrerede afvigelse fra middelværdien.

Spredning:

$$\sigma = \sqrt{\text{Var}X}$$

(samme enhed som  $X$ )

6 / 28

Regneregler for varians:

$$\text{Var}aX = a^2 \text{Var}X$$

Hvis  $X$  og  $Y$  uafhængige:

$$\text{Var}[X + Y] = \text{Var}X + \text{Var}Y$$

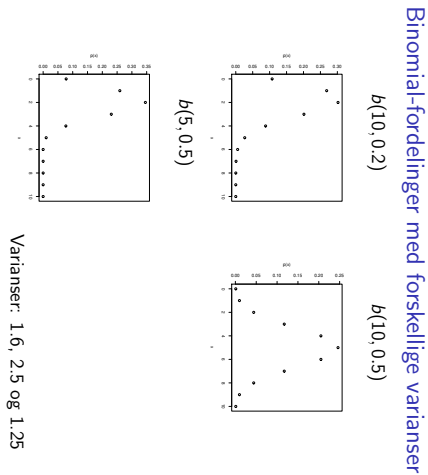
Spredning af  $aX$ :

$$\sqrt{\text{Var}aX} = a\sqrt{\text{Var}X}$$

Eksempel: varians af binomialfordeling  $X = \sum_{i=1}^n S_i \sim b(n, p)$  ( $S_i \sim b(1, p)$  og uafh.)

$$\text{Var}X = \sum_{i=1}^n \text{Var}S_i = np(1 - p)$$

8 / 28



## Kontinuerte stokastiske variable

Hidtil har vi set på *diskrete* stokastiske variable, som antog heltallige værdier.

En *kontinueret* stokastisk variabel kan antage alle reelle værdier.

For en kontinuert stokastisk variabel angives sandsynligheder vha. en funktion  $f$  defineret på de reelle tal. Stor/lille sandsynlighed for at  $X$  falder i intervaller hvor  $f$  er stor/lille.

Eksempel: uniform (ensartet) fordeling på  $[0, 1]$ :

$$f(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{ellers} \end{cases}$$

## Varians for Poisson

Poisson fordeling med middelværdi  $\mu$  fremkommer som grænseværdi af  $b(n, \mu/n)$ .

Varians for  $X \sim b(n, \mu/n)$ :

$$\text{Var}X = n \frac{\mu}{n} \left(1 - \frac{\mu}{n}\right) = \mu \left(1 - \frac{\mu}{n}\right)$$

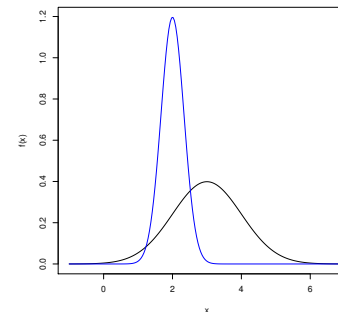
Dvs. variansen for  $X$  går mod  $\mu$  når  $n$  går mod uendelig.

Konklusion: variansen for  $Poisson(\mu)$  er  $\text{Var}X = \mu = \mathbb{E}X$  !

## Normal fordeling

Normalfordeling med middelværdi  $\mu$  og varians  $\sigma^2$ :  $f$  klokkeformet

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



Middelværdi 3 eller 2 og spredning 1 eller 1/3.

## Normalfordeling - 2 slags intervaller

For en normalfordeling  $N(\mu, \sigma^2)$  har vi altid

$$P(\mu - \sigma 1.96 \leq X \leq \mu + \sigma 1.96) = 0.95$$

Sagt i ord: med 95% sandsynlighed afviger en normalfordelt stokastisk variabel ikke mere end to standardafvigelser ( $1.96\sigma \approx 2\sigma$ ) fra middelværdien.

Dette er ækvivalent med, at med 95% sandsynlighed ligger  $\mu$  i det tilfældige interval  $[X - 1.96\sigma; X + 1.96\sigma]$ .

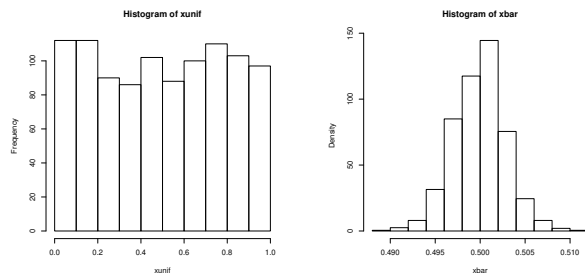
Sagt på en tredje måde: forskellen  $|X - \mu|$  er mindre end  $1.96\sigma$  med 95% sandsynlighed.

$2\sigma$ : 95,4%  $3\sigma$ : 99,7%  $4\sigma$ : 99,99%

13 / 28

## Tilnærmelsesvis normal fordeling af gennemsnit af uniformt fordelte variable

Histogram af uniformt fordelte variable samt histogram af  $\bar{X}$  (gennemsnit af 1000 uniformt fordelte)



15 / 28

## Centrale grænseværdi-sætning

Middelværdi for  $\bar{X}$  er  $\mu$ . Variansen er  $\sigma^2/n$

CLT: når  $n$  stor vil  $\bar{X}$  tilnærmelsesvist have en *normal*-fordeling (med middelværdi  $\mu$  og varians  $\sigma^2/n$ ).

**NB:** ligegyldigt hvilken fordeling  $X_i$ 'erne har fås samme grænse-fordeling af  $\bar{X}$  når  $n \rightarrow \infty$  (dog skal  $X_i$ 'erne være ens fordelt og uafhængige).

14 / 28

## Intervaller for Monte Carlo estimater - vurdering af usikkerhed

Estimat  $\bar{X}$  tilnærmelsesvist  $N(\mu, \sigma^2/n)$ .

Dvs. med 95% sandsynlighed er estimations-fejlen

$$|\mu - \bar{X}|$$

mindre end  $1.96\sigma/\sqrt{n}$ .

Dette er ækvivalent med, at med 95% sandsynlighed ligger  $\mu$  i det tilfældige interval  $[\bar{X} - \sigma 1.96/\sqrt{n}, \bar{X} + \sigma 1.96/\sqrt{n}]$ .

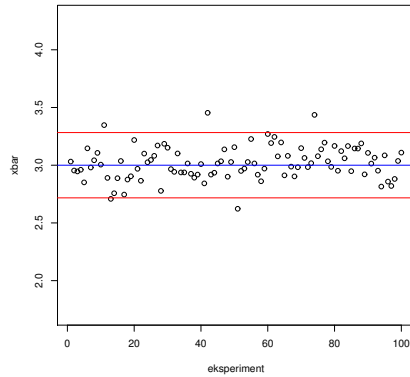
Intervaller giver et bud på usikkerheden af estimationen af  $\mu$  - kaldes *konfidens*-intervallet.

I praksis erstattes ukendte  $\sigma$  af  $s = \sqrt{s^2}$  - den empiriske spredning.

16 / 28

## Simulerede $\bar{X}$

(hver  $\bar{X}$  gennemsnit af 100 normalfordelte variable med middelværdi 3 og varians 2)

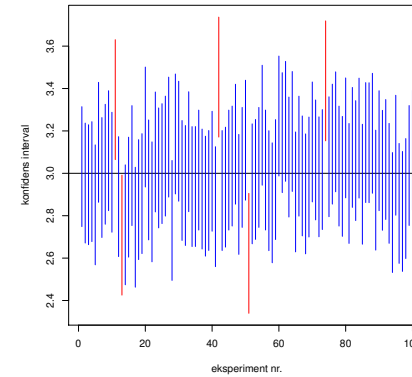


Ca. 5% af de simulerede  $\bar{X}$  ligger udenfor 95%-intervallet.

17 / 28

## Simulerede konfidensintervaller

(hver baseret på  $\bar{X}$  gennemsnit af 100 normalfordelte variable med middelværdi 3 og varians 2)



Konfidensintervallerne indeholder sande middelværdi  $\mu = 3$  i ca. 95% af tilfældene.

18 / 28

## Praktisk brug af konfidensinterval

Hvis vi for et givet eksperiment/datasæt/Monte Carlo beregning hævder, at den ukendte middelværdi ligger i det beregnede interval, så tager vi kun fejl i 5% af tilfældene, hvis der er tale om et 95% interval.

Hvis vi vil have større sikkerhed kan vi i beregningen af konfidensintervallet erstatte  $1.96\sigma$  med  $3\sigma$  eller  $4\sigma$  - giver 99.7% eller 99.99% intervaller.

Da tager vi kun fejl i 0.3% eller 0.01% af tilfældene.

19 / 28

## Tilbage til Monte Carlo

Med tusind (computer) kast af 5 terninger får vi estimeret forventet total antal øjne 17.49 og estimeret varians 13.91.

Estimererne for  $\mu$  og  $\sigma^2/n$  er 17.49 og  $13.91/1000 = 0.0139$ . Den estimerede spredning for  $\bar{X}$  er  $\sqrt{0.0139} = 0.11$ .

NB: hvad er den sande forventede værdi ?

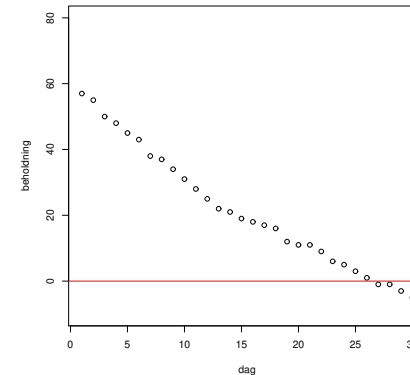
Bud på usikkerheden: trolige værdier af  $\mu$  er  $[17.49 - 2 \times 0.11, 17.49 + 2 \times 0.11]$ .

Estimation af sandsynlighed for mere end eller lig 15: her benytter vi variable  $I_i = 1[\text{total sum}_i \geq 15]$ . Empirisk middelværdi og varians 0.778 og 0.17. Dvs. estimeret spredning på  $\bar{I}$  er  $\sqrt{0.17/1000} = 0.013$ . Dvs. vurdering af usikkerhed: sandsynlighed mellem  $0.778-0.026$  og  $0.778+0.026$ .

20 / 28

Lager:  $V$  ventetid til lager tømmes. Empirisk middelværdi 28.33 og varians 4.42. Antal simulationer=1000. Dvs. sande forventede ventetid  $28.33 \pm 2 \times 0.066$ .

## En simulation af lager



I dette tilfælde gik der 26 dage før lageret blev tømt.

21 / 28

22 / 28

## Øvelser

1. Beregn varianserne af de stokastiske variable fra opgave 1, 2, 6 og 7 i første sæt af opgaver.
2. Udgangspunktet er følgende 10 observationer (uafhængige, samme middelværdi og varians):  
3.10 1.84 1.66 1.44 1.26 2.63 2.93 3.61 0.36 2.45
  - 2.1 Beregn empirisk middelværdi  $\bar{x}$  og varians  $s^2$ .
  - 2.2 Beregn et 95% konfidensinterval for den ukendte (teoretiske) middelværdi  $\mu$ .
  - 2.3 Hvor stor skulle stikprøven være, hvis intervallets bredde skulle være mindre end en halv ?
3. Antag  $X$  er normalfordelt med middelværdi 3 og varians 2.
  - 3.1 Hvad er sandsynligheden for, at  $X$  er mindre end 3 ?
  - 3.2 Angiv et interval, som  $X$  tilhører med 95% sandsynlighed.
4. Antag  $X_1, \dots, X_{50}$  er uafhængige og alle har middelværdi 3 og varians 2.
  - 4.1 Angiv et interval, som  $\bar{X}$  tilhører med sandsynlighed 95%.
  - 4.2 Angiv et (tilfældigt) interval, som med sandsynlighed 95% vil indeholde den sande middelværdi 3.
5. Antag  $X = (X_1 + \dots + X_{100}) \sim b(100, 0.25)$  og lad  $\bar{X} = X/100$ .
  - 5.1 Hvordan kan vi udregne en tilnærmet værdi for  $P(\bar{X} \leq 0.1)$  vha. den centrale grænseværdi-sætning ?
  - 5.2 Antag, at  $p$  er ukendt men vi har observeret  $\bar{X} = 0.65$ . Angiv da et 95% konfidensinterval for  $p$  (vink: husk variansen for  $b(n, p)$  er  $np(1-p)$ ).

23 / 28

24 / 28

6. Antag, at en fabriks produktion af bolte skal leve op til følgende specifikationer: middelværdien af boltens længde skal være 10 mm og der skal være mindst 99.7% sandsynlighed for, at en bolts længde er i intervallet 9.85 til 10.15 mm. En ingeniør måler nu 100 bolte og observerer, at den empiriske middelværdi  $\bar{X}$  og varians  $s^2$  for de målte bolte er henholdsvis 9.91 og 0.000289.
- 6.1 Taler ingeniørens data imod, at den sande middelværdi er 10 ? (vink: beregn konfidensinterval)
- 6.2 Antag, at boltens længder er normalfordelte. Giver data anledning til at tro, at de producerede bolte med 99.7% sandsynlighed overholder specifikationens tolerance-interval ?
7. En produktion af elektriske komponenter skal overholde, at i snit er højst 0.5% af komponenterne defekte. I en stikprøve af 1000 komponenter er 7 komponenter defekte. Det antages, at komponenternes fejlstatus er uafhængige. Giver de observerede data anledning til at mene, at kravet til produktionen ikke er overholdt ? (vink: udregn 95% konfidensinterval for den ukendte andel defekte)

25 / 28

8. (Monte Carlo beregning af sandsynlighed). Brug computer til at estimere sandsynligheden for, at det maksimale antal øjne i et kast med 5 terninger er mindre end eller lig 5. Beregningens nøjagtighed skal opfylde, at 99.7% konfidensintervallet for sandsynligheden har bredde højst 0.02.

Vink: modificer min kode for terningkast. Beregning af max-værdi i en vektor x:

```
> xmax=max(x)
```

Kan I udregne sandsynligheden eksakt ?

26 / 28

## Facit

1: 0.96, 3, 0.5625, 3.6, 2.1: 2.13 0.98, 2.2: [1.52;2.74], 2.3: 60, 3.1: 0.5, 3.2:  $3 \pm 2.77$ , 4.1:  $3 \pm 0.392$ , 4.2:  $\bar{X} \pm 0.392$ , 5.1 :  $\bar{X}$  er approksimativt  $N(0.25, 0.001875)$ . 5.2: Varians  $\sigma^2$  for  $X_i$  estimeres til  $0.65(1 - 0.65) = 0.2275$ . Konfidensinterval:  $0.65 \pm 1.96\sqrt{0.2275/100} = 0.65 \pm 0.093$ . 6.1: ja, 99.7% konfidensintervallet er [9.9049, 9.9151] 6.2: ja, på baggrund af  $\bar{X}$  og  $s^2$  er det estimerede 99.7% interval [9.859, 9.961]. 7: nej, 95% konfidensintervallet er [0.0017, 0.0123]. 8: med 100000 simulationer beregner jeg sandsynligheden til 0.3999 med et 99.7% konfidensinterval [0.3952, 0.04045], som har bredde 0.0092 (husk, disse beregninger er behæftet med Monte Carlo fejl, da de er beregnet på baggrund af simulationer).

27 / 28