Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 4: Stikprøvefordelinger

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

Oversigt

Eksempel: Gennemsnit og varians af normalfordelt stikprøve

Antag at vi planlægger et studie hvor vi udtager 5 observations enheder. Vi antager desunden at middelværdi og varians i population er hhv $\mu = 10$ og $\sigma^2 = 2$, hvad er fordelingen af gennemsnit og empirisk varians under disse antagelser?

Der er (mindst to) måder at svare på spørgsmålet

1: Gennemgå de teoretiske udledninger for at få de præcise fordelingsfunktioner

2: Udfør eksperimentet et stort antal gange (eks. 10.000) på din computer og find den empiriske fordeling

Eksempel: Gennemsnit og varians af normalfordelt stikprøve

Simuleringssvaret findes ved

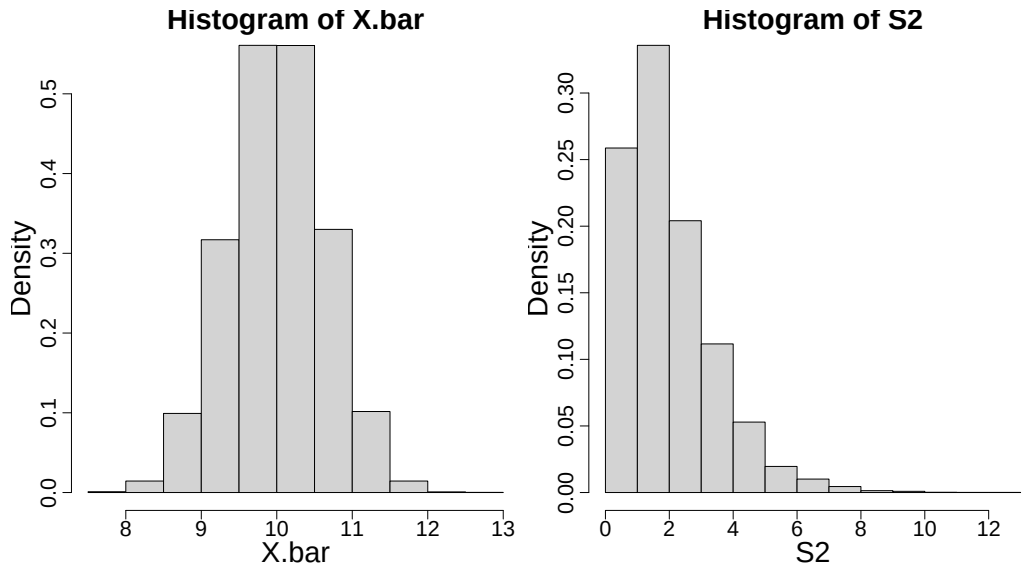
```
>       set.seed(125)
>       n <- 5; k <- 10000
>       ## Simuleringen
>       X <- matrix(rnorm(n * k, mean = 10, sd = sqrt(2)),
+                   ncol = n, nrow = k)
>       head(X)

##          [,1]    [,2]    [,3]    [,4]    [,5]
## [1,] 11.3199  9.4356  6.6461  8.1428  9.5086
## [2,]  9.2575  9.9915 10.2699 10.5822  9.4101
## [3,] 12.5660 10.2551 10.2662  9.7488  9.0320
## [4,] 10.1174 10.5482  7.9771 11.6162 10.2075
## [5,] 10.5596  8.4286 11.9226  9.7385 11.6490
## [6,]  6.8977  8.5435  9.0741  9.9761 10.5987
```

Eksempel: Gennemsnit og varians af normalfordelt stikprøve

Fordelingerne af $\bar{X}$ og S^2 findes ved

```
> X.bar <- apply(X, 1, mean)
> S2 <- apply(X, 1, var)
> hist(X.bar, prob = TRUE); hist(S2, prob = TRUE)
```



Oversigt

Den formelle ramme for *statistisk inferens*

Fra eNote, Chapter 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observationsenhed**)
- The *statistical population* consists of all possible "measurements" on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Sprogbrug og koncepter:

- μ og σ er parametre, som beskriver populationen
- $\bar{x}$ er *estimatet* for μ (konkret udfald)
- $\bar{X}$ og S^2 er *estimatorer* for μ hhv. σ^2(nu set som stokastisk variabel)
- Begrebet *'statistic(s)'* er en fællesbetegnelse for begge

Statistisk inferens

I statistik har vi typisk et antal nøgletal (eks. $\bar{x}$ og s^2) fra en stikprøve.

- Vi ønsker at udtale os om populationens parametre (eks. μ og σ^2)
- Generelt kræver det en fordelingsantagelse for populationen (eksempelvis normalfordelingen)[1]
- For at kvantificere usikkerheder har vi brug for afledte fordelinger, eksempelvis fordelingen for $\bar{X}$ og S^2

Studiet af disse afledte fordelinger under normalfordelingsantagelsen er emnet for dagens forelæsning.

- Vi bruger en blanding af små eller delvise beviser og simulation til at anskueligøre resultaterne

[1] eller mere præcist vi får bedre styrke i konklussioner hvis vi har en fødelings antagelse

Begreber

Central Estimator:
En estimator $\hat{\theta}$ er central (eller ikke-biased), hvis og kun hvis, middelværdien af stikprøvefordelingen for estimatoren er lig θ

Consistent Estimator
En central estimator $\hat{\theta}$, der konvergere i sandsynlighed (du kan tænke på det som $V(\theta_n) \to 0$).

Efficient Estimator
En estimator $\hat{\theta}_1$ er en mere efficient estimator for θ end estimatoren $\hat{\theta}_2$ hvis:

1. $\hat{\theta}_1$ og $\hat{\theta}_2$ begge er centrale estimatorer af θ
2. Variansen af stikprøvefordelingen for $\hat{\theta}_1$ er mindre end for $\hat{\theta}_2$

Estimat
Når vi har udtaget vores stikprøve og udregnet vores nøgle tal har vi et estimat (det er ikke en stokastisk variabel)

Eksempel:

Hvis $X_1, .., X_n$ i.i.d. $N(\mu, \sigma^2)$ stokastiske variable, så er

- $\overline{X} = \hat{\mu}$ en central estimator for μ ($E[\overline{X}] = \mu$).
- $\overline{X}$ er også en konsistent estimator for μ ($V[\overline{X}] = \frac{\sigma^2}{n} \to 0$, $n \to \infty$).
- $\bar{x}$ er et estimat for μ.
- medianen er også en central og konsistent estimator for μ, men medianen er mindre efficient.

Oversigt

Fordeling for gennemsnit af normalfordelinger (Theorem 3.2)

(Stikprøve-) fordelingen/ The (sampling) distribution for $\bar{X}$

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Oversigt

Eksempel: Middelværdi af varians estimator

Lad $X_1 \sim N(\mu, \sigma^2)$ og $X_2 \sim N(\mu, \sigma^2)$ være uafhængige stokastiske variable, hvad er middelværdien af

$$Q = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2$$

Eksempel: Middelværdi af varians estimator

Lad $X_1, \ldots, X_n$ være uafhængige og identisk fordelte stokastiske variable med middelværdi $E[X_i] = \mu$ og varians $V[X_i] = \sigma^2$, lad Q være

$$Q = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Hvad er middelværdien af Q?

Eksempel: Middelværdi af varians estimator

Lad $X_1, \ldots, X_n$ være uafhængige og identisk fordelte stokastiske variable med middelværdi $E[X_i] = \mu$ og varians $V[X_i] = \sigma^2$, lad Q være

$$Q = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Hvad er middelværdien af Q?

$$\begin{aligned} E[Q] &= \sum_{i=1}^{n} E[(X_i - \overline{X})^2] \\ &= \sum_{i=1}^{n} E[(X_i - \mu + \mu - \overline{X})^2] \\ &= \sum_{i=1}^{n} E[(X_i - \mu)^2 + (\mu - \overline{X})^2 + 2(X_i - \mu)(\mu - \overline{X})] \end{aligned}$$

Eksempel forts: Middelværdi af varians estimator

$$
\begin{aligned}
E[Q] =& \sum_{i=1}^{n} E[(X_i - \mu)^2] + E[(\mu - \overline{X})^2] - 2Cov[X_i, \overline{X}] \\
=& n\sigma^2 + \sigma^2 - 2\sum_{i=1}^{n} \frac{1}{n} Cov\left(X_i, \sum_{j=1}^{n} X_j\right) \\
=& (n+1)\sigma^2 - 2\sum_{i=1}^{n} \frac{1}{n} Cov\left(X_i, X_i\right) \\
=& (n+1)\sigma^2 - 2\sigma^2 \\
=& (n-1)\sigma^2
\end{aligned}
$$

Dermed er $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ en central estimator for σ^2.

χ^2-fordelingen

Hvis $X_1, .., X_n$ er iid N(0,1) så følger

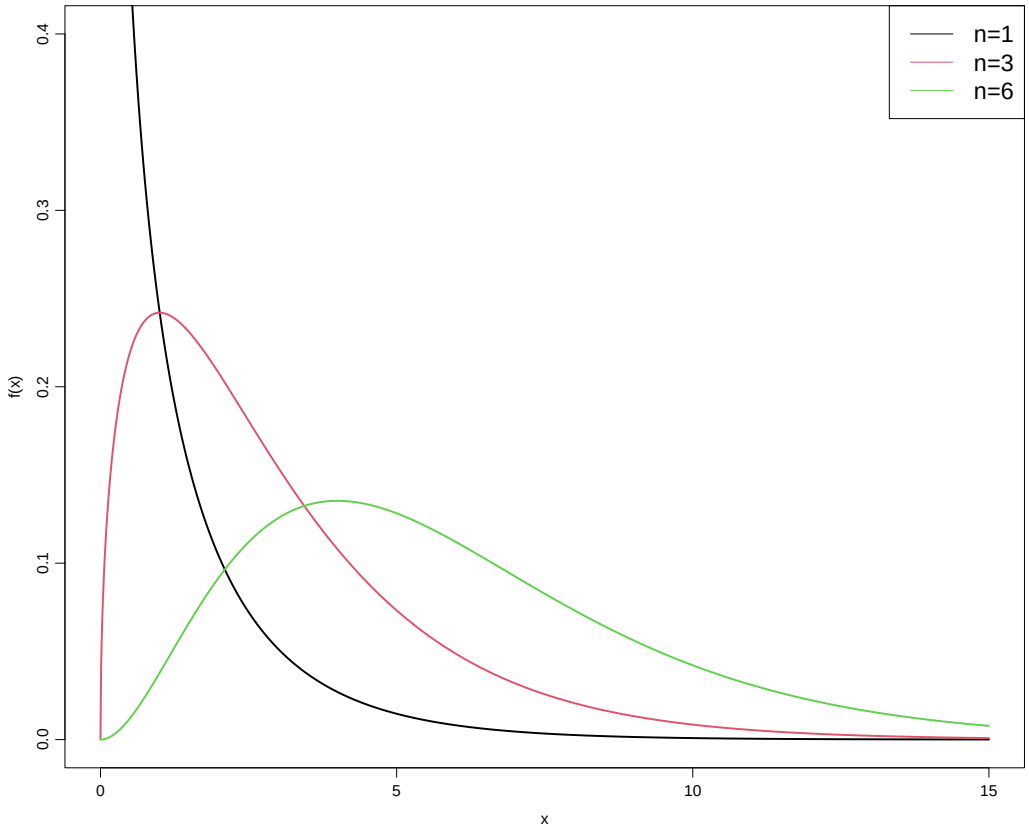
$$Q = \sum_{i=1}^{n} X_i^2$$

en χ^2-fordeling med n-frihedsgrader, vi skriver $Q \sim \chi^2(n)$
Tæthedsfunktionen for en χ^2-fordeling er givet ved

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}; \quad x \geq 0.$$

hvor $\Gamma(\cdot)$ er Gamma funktionen og n er antal frihedsgrader.

χ^2-fordelingen



Egenskaber ved χ^2-fordelingen

Hvis $Q \sim \chi^2(n)$ så er

$$\begin{aligned} E(Q) &= n \\ V(Q) &= 2n \end{aligned}$$

Hvis $Q_1 \sim \chi^2(n_1)$ og $Q_2 \sim \chi^2(n_2)$ er uafhængige så gælder

$$Q = Q_1 + Q_2 \sim \chi^2(n_1 + n_2)$$

Eksempel

Hvis $X_1, \ldots, X_{10}$ er i.i.d. $N(\mu, \sigma^2)$ og

$$Q = \frac{1}{\sigma^2} \sum_{i=1}^{10} (X_i - \mu)^2$$

Hvad er så $P(Q > 10)$?

Fordeling af varians estimator

Hvis $X_1, \ldots, X_n$ er i.i.d. $N(\mu, \sigma^2)$, med $\overline{X}$, S^2 hhv. gennemsnit og empirisk varians. Så gælder at

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

og det følger at

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Bevis (Skitse)

1: $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$ og $\frac{(\bar{X}-\mu)^2}{\sigma^2/n} \sim \chi^2(1)$

2: $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$

3: $Cov(\bar{X}, \bar{X}-X_i) = 0 \Rightarrow S^2$ og $(\bar{X}-\mu)^2$ uafhængige

4: Hvis $Q_1 \sim \chi^2(n_1)$ og $Q_2 \sim \chi^2(n_2)$ uafhængige, så gælder $Q_1 + Q_2 \sim \chi^2(n_1+n_2)$

Eksempel

Find $E(S^2)$ og $V(S^2)$.

Eksempel

Find $E(S^2)$ og $V(S^2)$. Svar:

$$E[S^2] = \frac{\sigma^2}{n-1} E\left[\frac{n-1}{\sigma^2} S^2\right]$$

$$V[S^2] = \left(\frac{\sigma^2}{n-1}\right)^2 V\left[\frac{n-1}{\sigma^2} S^2\right]$$

Da $\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$ følger det at

$$E[S^2] = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

$$V[S^2] = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1}$$

Det betyder at S^2 er en central og konsistent estimator for σ^2

Eksempel: Sammenvægtet varians

Lad $X_1, ..., X_{n_1}$ og $Y_1, ..., Y_{n_2}$ (i.i.d. stokastiske variable), med $X_i \sim N(\mu_1, \sigma^2)$, og $Y_i \sim N(\mu_2, \sigma^2)$. Med $a \in [0, 1]$ find a så $V[S_p^2]$ minimeres når

$$S_P^2 = aS_1^2 + (1 - a)S_2^2$$

hvor S_1^2 og S_2^2 er stikprøve variansen for hhv. X og Y. Opskriv S_P^2 og find $V[S_P^2]$.

Oversigt

t-fordelingen

Hvis $X \sim N(0,1)$ og $Q \sim \chi^2(n)$ og X og Q er uafhængige så følger

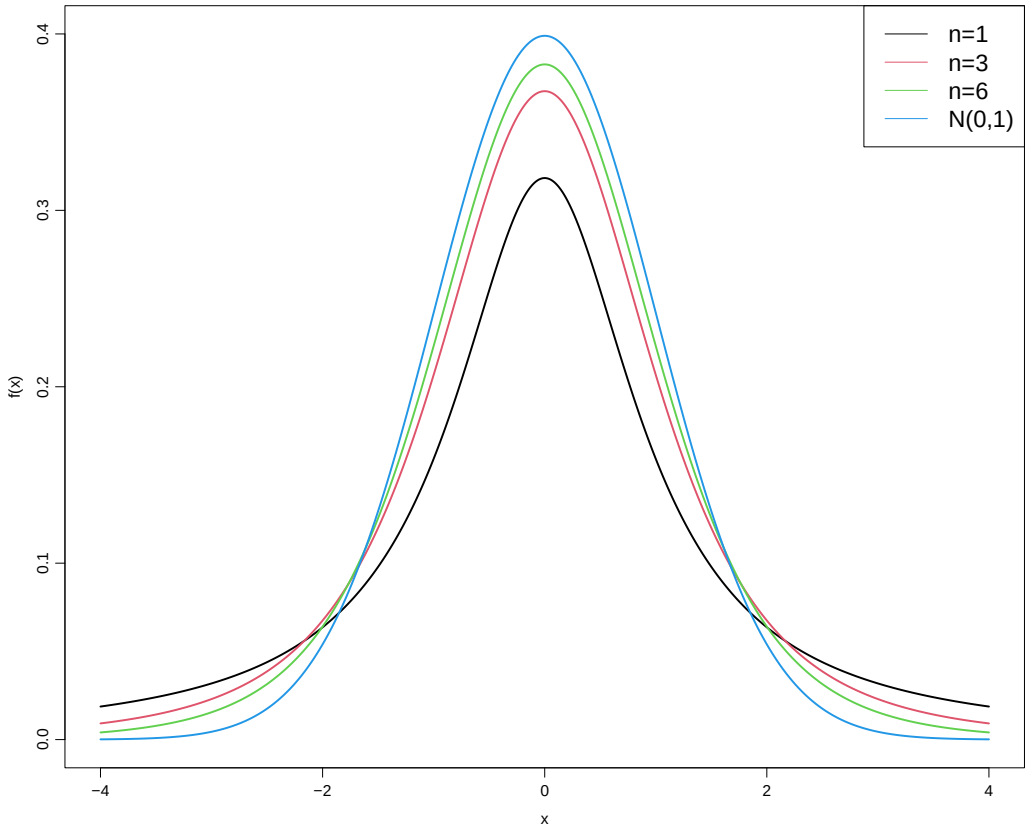
$$T = \frac{X}{\sqrt{Q/n}} \tag{1}$$

en t-fordeling med n-frihedsgrader.
Tæthedsfunktionen for en t-fordeling er givet ved

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad ; t \in \mathbb{R} \tag{2}$$

hvor n er antallet af frihedsgrader og $\Gamma()$ er Gamma funktionen.

t-fordelingen



t-fordelingen som stikprøvefordeling

Lad $X_1, \ldots, X_n$ være i.i.d. $\sim N(\mu, \sigma^2)$ så følger

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{3}$$

en t-fordeling med $n - 1$ frihedsgrader.

t-fordelingen som stikprøvefordeling - Bevis

Vi skal vise at T kan skrives som en standard normal fordeling dividet med en χ^2-fordeling med $n-1$ frihedsgrader (og at tæller og nævner er uafhængige).

1: Vi har vist at $\bar{X}$ og S^2 er uafhængige

2: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ og $Q = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$

3:

$$T = \frac{\frac{1}{\sigma/\sqrt{n}}(\overline{X}-\mu)}{\sqrt{\frac{1}{\sigma^2/n}\frac{n-1}{n-1}S^2/n}} = \frac{Z}{\sqrt{Q/(n-1)}}$$

hvor $Z \sim N(0,1)$ og dermed følger T en t-fordeling med $n-1$ frihedsgrader.

Eksempel: Konfidensinterval

Lad $X_1, \ldots, X_n$ være i.i.d. $\sim N(\mu, \sigma^2)$, find d så $(0 < \alpha < 0.5)$

$$1 - \alpha = P(\mu - d \cdot S < \overline{X} < \mu + d \cdot S)$$

Eksempel: Konfidensinterval

Lad $X_1, \ldots, X_n$ være i.i.d. $\sim N(\mu, \sigma^2)$, find d så $(0 < \alpha < 0.5)$

$$1 - \alpha = P(\mu - d \cdot S < \overline{X} < \mu + d \cdot S)$$

Svar:

$$\begin{aligned} P(\mu - d \cdot S < \overline{X} < \mu + d \cdot S) &= P\left(-d < \frac{\overline{X} - \mu}{S} < d\right) \\ &= P\left(-d\sqrt{n} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < d\sqrt{n}\right) \\ &= F_T(d\sqrt{n}) - F_T(-d\sqrt{n}) = 2F_T(d\sqrt{n}) - 1 \end{aligned}$$

ved at sætte ovenstående lig med $1 - \alpha$ og løse for d fås

$$d = \frac{1}{\sqrt{n}} F_T^{-1}\left(1 - \frac{\alpha}{2}\right) = \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

hvor $t_{1-\frac{\alpha}{2}}$ er $1 - \frac{\alpha}{2}$-fraktilen i en t-fordeling med $n - 1$ frihedsgrader.

Eksempel: Konfidensinterval

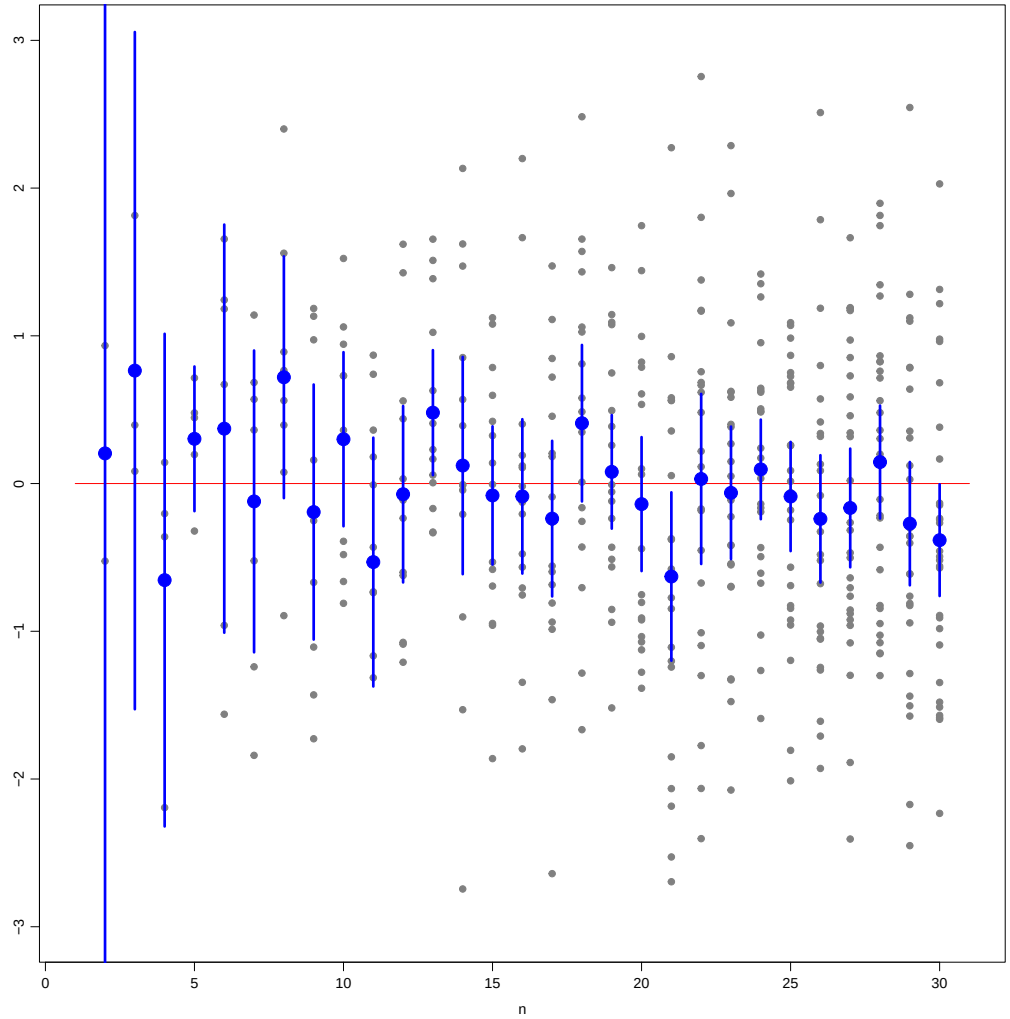
Vi har altså

$$1-\alpha = P\left(\mu - \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S < \overline{X} < \mu + \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S\right)$$
$$= P\left(\overline{X} - \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S < \mu < \overline{X} + \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S\right)$$

Læg mærke til at i den sidste omskrivning er intervallet stokastiske, mens μ er en fast størrelse.
I statistik vil vi typisk udtale os om en ukendt størrelse (eksempelvis μ) på basis af realisationer af gennemsnit ($\bar{x}$) og (empirisk) varians (s^2), og vi kan eksempelvis skrive at vi er 95% sikre på at μ ligger i intervallet

$$\bar{x} \pm t_{0.975} \cdot s/\sqrt{n}$$

Konfidensintervaller for stigende stikprøvestørrelse



Eksempel

Brug $\alpha = 0.05$

- Udregn d for afstandsdata udtaget på dag 1
- opskriv intervallet $\bar{x} \pm d \cdot s$
- giv en fortolkning af dette interval

Oversigt

I morgen

I morgen tager vi hul på eNote 3 og starter dermed med mere specifikke statistiske metoder.