02402 Notes

Daniel Brasholt s214675

E22

Contents

Uge 01: Introduktion til statistik og R

Date: Tuesday 30.08.22

Praktisk

Eksamen d. 17 december. 4 timers multiple choice.

2 projekter skal bestås for at kunne gå til eksamen. For hvert projekt vælges mellem 4 emner.

Forelæsning tirsdag 8-10 i 303A/42-43, opgaver i 324 og 303.

Kursushjemmeside

Definitioner

Spredning & varians

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

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

Kovarians & Korrelation

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}) = \frac{s_{xy}}{s_x \cdot s_y}$$

Uge 02: Stokastiske variable og diskrete fordelinger

Date: Tuesday 06.09.22

Lecture slides

Stokastiske variable og tæthedsfunktioner

Repræsenterer værdien af et udfald *før* et eksperiment finder sted. Deles op i **diskrete** og **kontinuerte**:

- Diskret: huller i intervallet
- Kontinuert: kontinuert over et interval; ingen mellemrum

Før eksperimentet: stokastisk variabel X eller $(X_1, \dots, X_n)$.

Efter eksperimentet: x eller $(x_1, \dots, x_n)$.

Kan simuleres i fx R. Eksempler på terningkast på slides.

Tæthedsfunktionen

pdf.

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

For diskret variabel opfylder den to betingelser:

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

Fordelingsfunktioner

cdf.

Cumulative distribution.

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

Find sandsynligheden for at få et udfald mindre end 3:

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

Figure 1: Eksempel: Fair terning

Binomialfordelingen

To udfald: succes eller ikke succes. Eksperimentet gentages så nogle gange. X er så antallet af succeser efter n kast. Formel for fordelingsfunktion på lecture slide 24.

```r
## 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)
```

Callcenter-eksempel:

$$P(x = 6) = \binom{6}{6} 0.7^6(1 - 0.7)^0 = 11.7\%$$

Hypergeometrisk fordeling

X er igen antallet af succeser, men denne gang er det *uden* tilbagelægning. Formel på slide 33.

Eneste forskel på binomial og Hypergeometrisk: binomial er *med* tilbagelægning, Hypergeometrisk er *uden*.

Binomial bliver brugt oftere end Hypergeometrisk.

Poissonfordelingen

Anvendes ofte som en fordeling for tælletal, hvor der ikke er nogen naturlig øvre grænse. Defineres normalt ved en *intensitet*, som har formen "antal/enhed", ofte benævnt λ. Det er typisk *hændelser per tidsinterval*.

Observationer er tilfældige og uafhængige.

Tæthedsfunktion:

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

Fordelingsfunktion:

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

Fordelinger i R

```
sample(1:6, size=1)
```

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 $F(x)$)

Figure 2: Functions in R

Middelværdi og varians

Middelværdi

Det *sande gennemsnit* af X (i modsætning til stikprøvegennemsnit).

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

Udtrykker hvor *midten* af X er. E engelsk for *expectations*.

Kan simuleres:

```
# 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)

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

Flere prøver giver bedre mål for gennemsnit.

Varians

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

Konkrete fordelinger

Binomial

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

Hypergeometrisk

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

Poissonfordelingen

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

Uge 03: Stokastiske variable og kontinuerte fordelinger

Date: Tuesday 13.09.22

Lecture slides

Kontinuerte fordelinger

Tæthedsfunktionen

For kontinuerte fordelinger: $P(X = x) = 0$ for alle x. Tæthedsfunktionen $f(x)$ hørende til fordelingen af en kontinuert stokastisk variabel opfylder dog stadig, at

$$f(x) \geq 0 \quad \text{og} \quad \int_{-\infty}^{\infty} f(x)dx = 1$$

Tæthedsfunktionen er arealet under et område af grafen;

$$P(x < X < x + \Delta x) \approx \Delta x \cdot f(x)$$

Fordelingsfunktionen

Fordelingsfunktionen er defineret ved

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Gælder så også, at $f(x) = F'(x)$.

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

Da $f(a) = 0 = f(b)$, er det ligegyldigt, om det er skarpt eller svagt større end / mindre end oppe i formlen.

Empirisk Fordelingsfunktion

```r
# Empirical cdf for sample of height data from Chapter 1
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
plot(ecdf(x), verticals = TRUE, main = "")

# 'True cdf' for normal distribution (with sample mean and variance)
xp <- seq(0.9*min(x), 1.1*max(x), length = 100)
lines(xp, pnorm(xp, mean(x), sd(x)), col = 2)
```

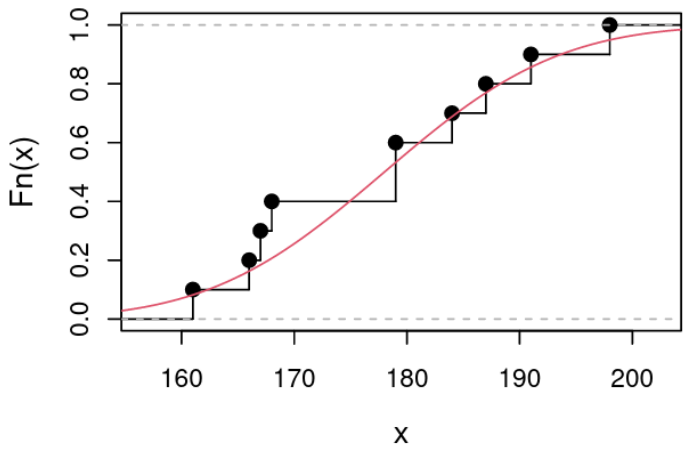

Figure 3: Eksempel på ECDF

Middelværdi

$$\mu = \int_{-\infty}^{\infty} = xf(x)dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Kovarians

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Hvis to stokastiske variable er uafhængige, er deres kovarians 0. Det modsatte er ikke nødvendigvis tilfældet.

Vigtige kontinuerte fordelinger

I R:

R	Distribution
`norm`	Normalfordelingen
`unif`	Uniform fordeling
`lnorm`	Log-normalfordelingen
`exp`	Eksponentialfordelingen

`d` tæthedsfunktion, $f(x)$ (probability density function).

`p` Fordelingsfunktion, $F(x)$ (cumulative distribution function).

`q` Fraktiler i fordeling (quantile).

`r` Tilfældige tal fra fordelingen (random).

Figure 4: Kontinuerte fordelinger i R

Den uniforme fordeling

Tæthed for den uniforme ser "firkantet" ud (se slide 15). Syntaks:

X~U(a, b)

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta$$

$$\mu = \frac{\alpha - \beta}{2}$$

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

Fx studerende kommer mellem 8 og 8:30; hvad er sandsynligheden for, at en studerende kommer mellem 8:20 og 8:30?:

$$\frac{10}{30} = \frac{1}{3}$$

```
punif(30, 0, 30) - punif(20, 0, 30)
```

Fordelingsfunktionen for en uniform fordeling vil være en ret linje.

Normalfordelingen

Normalfordelingen er en klokkeformet fordeling centreret om middelværdien, som er lig medianen.

$X \sim N(\mu, \sigma^2)$

Tæthedsfunktionen står på lecture slide 20. Middelværdi og varians er nemme at finde, da de er parametre i fordelingen.

I R tæller `norm`-funktionen uden andre parametre som standard-normalfordelingen ($\mu = 0$ og $\sigma^2 = 1$). $Z \sim N(0, 1^2)$.

En vilkårlig normalfordelt variabel $X \sim N(\mu, \sigma^2)$ kan standardiseres ved

$$Z = \frac{X - \mu}{\sigma}$$

Log-normalfordelingen

$X \sim LN(\alpha, \beta^2)$ hvor $\beta > 0$

Tæthedsfunktionen, middelværdi og varians står på slide 33.

I princippet er fordelingen ikke vigtig, man kan bare ændre den til en rigtig normalfordeling.

Eksponentialfordelingen

Specialtilfælde af *gammafordelingen*.

Poisson: diskrete hændelser pr. enhed.

Eksponentialfordelingen: kontinuert afstand mellem hændelser.

Regneregler for stokastiske variable

Gælder både for kontinuerte og diskrete stokastiske variable.

X stokastisk variabel, a og b konstanter.

Middelværdiregel

Gælder altid.

$$E(aX + b) = aE(X) + b$$

Variansregel

Skal være uafhængige for at reglen gælder.

$$Var(aX + b) = a^2 Var(X)$$

Uge 04: Konfidensinterval for middelværdi

Date: Tuesday 20.09.22

t-fordelingen

Fordeling for gennemsnittet af normalfordelinger

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

Middelværdi og varians regneregler

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Følger af sætning, at $\bar{X}$ er normalfordelt med middelværdi μ og varians σ^2/n.

Fordelingen af den fejl, vi begår $\bar{X} - \mu$

Spredning af $\bar{X}$:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Spredningen af $(\bar{X} - \mu)$:

$$\sigma_{(\bar{X}-\mu)} = \frac{\sigma}{\sqrt{n}}$$

Det standardiserede stikprøvegennemsnit af Z følger en standard-normalfordeling.

Konfidensintervallet for μ

One-sample konfidensinterval for μ

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

s er her stikprøvespredning i modsætning til σ, som er populationsspredning.

Her forsøger vi at finde $100(1-\alpha)\%$-fraktilen, oftest 95%, så

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

Højde-eksempel

På slides er der 10 målinger med gennemsnit 178. 95%-konfidensintervallet for middelværdien er så:

```
## t-quantiles for n=10
qt(0.975,9)

[1] 2.262
```

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

som er:

$$178 \pm 8.74 = [169.3; 186.7]$$

R-funktion one-sample t test

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
t.test(x, conf.level=0.99)
```

Giver et længere output.

Statistisk sprogbrug og formel ramme

Den formelle ramme for statistisk inferens

- μ og σ er *parametre*, som beskriver populationen
- $\bar{x}$ er *estimatet* for μ - konkret udfald
- $\bar{X}$ er *estimatoren* for μ (nu set som stokastisk variabel)
- Begrebet *statistics* er fællesbetegnelse for begge

Tilfældig stikprøveudtagning (random sampling)

En tilfældig stikprøve fra en (uendelig) population: observationerne $X_1, X_2, \ldots, X_n$ udgør en tilfældig stikprøve af størrelse n fra den uendelige population $f(x)$ hvis:

- Hvert X_i er en stokastisk variabel med fordeling $f(x)$
- De n stokastiske variable er uafhængige

Altså

- Alle observationer skla komme fra den samme population
- De må ikke dele information med hinanden (fx hvis man havde udtaget hele familier i stedet for enkeltindivider)

Ikke-normale data, *den centrale grænseværdisætning*

Den centrale grænseværdisætning (CLT)

Gennemsnittet af en tilfældig stikprøve følger altid en normalfordeling, hvis n er stor nok.

Dvs. hvis n er stor nok, kan vi tilnærmelsesvist antage:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

Forsøg med det i R på slides.

Konsekvens af CLT

Vi kan bruge konfidensinterval baseret på *t*-fordelingen i stort set alle situationer, blot n er "stor nok".

- Tommelfingerregel $n \geq 30$
- Selv for mindre n kan formlen være (næsten) gyldig for ikke-normale data

Formel fortolkning af konfidensintervallet

$$P(\bar{X} - t_{0.975}\frac{S}{\sqrt{n}} < \mu < \bar{X} - t_{0.975}\frac{S}{\sqrt{n}}) = 0.95$$

Konfidensinterval for varians og spredning

Variansestimatet opfører sig som en X^2-fordeling:

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

Er en stokastisk variabel, som er X^2-fordelt med $v = n - 1$ frihedsgrader.

Et $100(1-\alpha)\%$ konfidensinterval for stikprøvevariansen σ^2 er:

$$\left[\frac{(n-1)s^2}{X^2_{1-\alpha/2}}; \frac{(n-1)s^2}{X^2_{1-\alpha/2}}\right]$$

For spredning: det samme, bare med kvadratrod på.

I R

95%-konfidensinterval for variansen - vi skal bruge X^2-fraktilerne:

```
qchisq(c(0.025, 0.975), df = 19)

[1] 8.907 32.852
```

Så for tilfældig stikprøve med $n = 20$ og:

$$\mu = \bar{x} = 1.01,\ \sigma^2 = s^2 = 0.07^2$$

bliver konfidensintervallet for variansen:

$$\left[\frac{19 \cdot 0.7^2}{32.85}; \frac{19 \cdot 0.7^2}{8.907}\right] = [0.002834; 0.01045]$$

Uge 05: Hypotesetests

Date: Tuesday 27.09.22

Eksempel: sovemedicin

Nulhypotese for forsøget opstilles (der er ingen forskel på de to sovemidler).

Spredning og middelværdi udregnes.

Hvis H_0 er sand, følger middelværdien en t-fordeling. Derfor beregner man bare t-værdien for det stikprøvegennemsnit, man har fundet. Tilstrækkeligt usandsynligt at få den middelværdi = nulhypotese må forkastes = der er forskel på de to sovemidler.

$$t_{obs} = \frac{\bar{x} - \mu 0}{s/\sqrt{n}}$$

One-sample t-test og p-værdi

$$\text{p} - \text{value} = 2 \cdot P(T > |t_{obs}|)$$

Man plejer at regne med 5%-konfidensinterval.

```r
x <- c(1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4)
n <- length(x)
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))
pvalue <- 2 * (1-pt(abs(tobs), df=n-1))
pvalue

[1] 0.001156876

## Gør det samme
t.test(x)
```

Kritiske værdier: $t_{\alpha/2}$ og $t_{1-\alpha/2}$

Hvis $t_{obs} >$ kritisk, kan nulhypotese forkastes - så er prøven for ekstrem i forhold til nulhypotese.

Hypotese-test med alternativer

Standard: nulhypotese at $\mu = \mu 0$, så ellers må vi acceptere $\mu \neq \mu 0$.

Det meste af tiden bliver den to-sidede brugt. Man skal dog, før man laver testen, bestemme sig for, hvilken test og hvilket signifikans-niveau.

Fejlslutninger

- Type I: Afvisning af H_0, når H_0 er sand
- Type II: Ikke afvisning/godkendelse af H_0, når H_1 er sand
- $P(\text{TypeIfejl}) = \alpha$
- $P(\text{TypeIIfejl}) = \beta$

Accepting af null hypothesis is not a statistical proof of the null hypothesis being true!

Test af normalfordeling

Først: man kan lave et histogram.

Ellers: QQ-plot

```r
qqnorm(x)
qqline(x)
```

Man kan evt lave en logaritme-transformation og derefter lave QQ-plot og t-test.

Uge 06: Analysis of Two Samples

Date: Tuesday 04.10.22

Eksempel med hospitaler

Forskel i energiforbrug. Giver en p-værdi på 0.0083, derfor statistisk usandsynligt, at H_0 er sand, så der er statistisk forskel på de to hospitaler.

Hypotesetest og signifikans

Starter med nulhypotese. Hvis p-værdi er $< \alpha$, hvor α er på forhånd valgt signifikansniveau, er nulhypotese forkastet.

Nulhypotese er oftest, at der *ikke* er forskel.

Two-sample t-test

$$\delta = \mu_2 - \mu_1$$

$$H_0 : \delta = \delta_0$$

Typisk kigger vi på $\delta = 0$, så at der ikke er nogen forskel.

Teststørrelsen er en Welch two-sample t-teststørrelse.

$$t_{obs} = \frac{(\bar{x_1} - \bar{x_2}) - \delta_0}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$$

$\bar{x}$ er gennemsnit, s^2 er stikprøvevarians. n er antallet af obs.

Teststørrelsen er tilnærmelsesvis t-fordelt med v frihedsgrader, som cirka er givet ved $n_1 + n_2 - 2$, dog ikke helt sådan. Står på slide 13.

I hospitalseksemplet er $t_{obs} = -3.0$, $v = 15.99$, $p = 0.008$. Der er altså stærk evidens *mod* nulhypotesen, som gik på, at der ikke er forskel på energiforbrug på de to hospitaler. Husk at gange med 2 jævnfør slide 14!

Konfidensinterval for $\mu_1 - \mu_2$

Se formel på slide 18. Det er de to fratrukket hinanden plus en given værdi.

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Kan også gøres i R ved:

```
xA <- c(...)
xB <- c(...)

t.test(xB, xA)
```

Giver det hele for de to variable.

Overlappende Konfidensintervaller

Man laver konfidensintervaller for hver gruppe. Ofte barplot med error bars (slide 23). Man skal være varsom, da man ikke bruger den rigtige standard error til at vurdere forskellen:

$$\sigma_{\bar{X}_A - \bar{X}_B} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$Var_{\bar{X}_A - \bar{X}_B} = Var_{\bar{X}_A} + Var_{\bar{X}_B}$$

$$\sigma_{\bar{X}_A - \bar{X}_B} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

Hvis to konfidensintervaller *IKKE* overlapper: der er signifikant forskel.

Hvis de overlapper: vi kan ikke med sikkerhed konkludere.

Det parrede setup

Parret setup og analyse = one-sample analyse.

Med eksemplet med sovemedicin, kan man både lave dif og t-test over det, eller man kan sige

```
t.test(x2, x1, paired = TRUE)
```

Giver det samme resultat.

Parret ift uafhængigt

- Fuldstændigt tilfældigt (uafhængige stikprøver)
  - 20 patienter, som tilfældigt fordeles på to grupper (normalt lige mange i hver gruppe). Dvs. at der er forskellige (uafhængige) patienter i de to grupper.
- Parrede observationer (afhængige stikprøver)
  - Vi har 10 patienter, som alle får begge behandlinger (typisk noget tid imellem og tilfældig rækkefølge af behandlingerne). Dvs. de samme patienter i de to grupper.
  - Træk værdierne i de to grupper fra hinanden og udfør one-sample test.

Vi kan *IKKE* (kun) ud fra data afgøre, om det er et parret eller et uparret setup - vi skal også vide noget om forsøget.

Tjekke normalfordelingsantagelserne

Man kan lave et QQ-plot *for hver af stikprøverne*. Man kan ikke lave det på residualerne, differenser eller samlet.

Styrke og stikprøvestørrelse - forsøgsdesign

Styrke = power

Når vi laver konfidens: det stykke, der er på hver side, kaldes fejlmargen.

Der findes en one-sample CI sample size formula, som giver den stikprøvestørrelse (hvis man kender spredningen), som giver et bestemt konfidensinterval.

Fx hvis man har lavet højdeforsøg og fået $\bar{x} = 178$ og $s = 12.21$. Med 95%-konfidens, skal vi have antal obs for $3cm$ margin of error:

$$n = \left(\frac{1.96 \cdot 12.21}{3}\right)^2 = 63.64 \approx 64$$

Styrke

- Sandsynligheden for at detektere en (påstået) effekt
- $P(H_0\text{afvises})$ når H_1 er sand
- Sandsyngliheden for en korrekt afvisning af H_0
- MEN: en nulhypotese kan være forkert på mange måder
- I praksis: brug en scenarie-baseret tilgang
  - E.g. "Hvis $\mu = 86$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"
  - E.g. "Hvis $\mu = 84$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"
  - Etc.

Hvis vi kender (eller antager) fire ud af de frem følgende størrelser, kan vi finde den manglende:

- Stikprøvestørrelse, n
- Signifikansniveauet α, som vi tester på
- Forskellen i middelværdi (effekt-størrelsen) $\mu_0 - \mu_1$
- Populationsstandardafvigelsen, σ
- Styrken (power), $1 - \beta$

Stikprøvestørrelse

Metode 3.65

$$n = \left(\sigma\frac{z1 - \beta + z_{1-\alpha/2}}{\mu_0 - \mu_1}\right)^2$$

Eksempel - styrken hvis $n = 40$

```
# Giver styrke
power.t.test(n = 40, delta = 4, sd = 12.21, type = "one.sample")

# Giver n
power.t.test(power = .80, delta = 4, sd = 12.21, type = "one.sample")
```

Two-sample:

```
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)
```

Det sammenvejede t-test

Hvis vi antager, at $\sigma_1^2 = \sigma_2^2$, kan nogle ting gøres. Står på de sidste undervisningsslides. Det er dog nogenlunde idiotsikkert at bruge Welch-versionen.

Uge 07: Simulationsbaseret statistik

Date: Tuesday 25.10.22

Introduktion

Mange relevante beregningsstørrelser har komplicerede fordelinger. Da meget ikke nødvendigvis er normalfordelt, kan det være svært, selv for middelværdien. Man kan håbe på CLTs magi, men vi kan ikke være sikre. Derfor kan problemet simuleres.

```
rbinom
rpois
rhyper
rnorm
rlnorm
rexp
runif
rt
rchisq
rf
```

Fx med areal af plade: længde X beskrevet ved $N(2, 0.01^2)$ og Y $N(3, 0.02^2$. $A = XY$, så fordeling af A?

```
k = 10000
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y

mean(A)
[1] 6

var(A)
[1] 0.002458

mean(abs(A-6) > 0.01)
[1] 0.0439
```

Fejlophobningslove

Method 4.3

$$\sigma^2_{f(X_1,\ldots,X_n)} \approx \sum_{i=1}^{n} \left(\frac{df}{dx_i}\right)^2 \sigma_i^2$$

$$\sigma_1^2 = Var(X) = 0.01^2$$

$$\sigma_2^2 = Var(Y) = 0.02^2$$

$$f(x,y) = xy, \frac{df}{dx} = y, \frac{df}{dy} = x$$

$$Var(A) \approx \left(\frac{df}{dx}\right)^2 \sigma_1^2 + \left(\frac{df}{dy}\right)^2 \sigma_2^2$$

$$= y^2\sigma_1^2 + x^2\sigma_2^2$$

$$= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2$$

$$= 0.0025$$

Metode 4.4

Se også metode 4.4, fejlophobning ved simulation, på slide 14.

3 metoder

- Simulationstilgang
- Teoretisk udledning
- Den analytiske, men approksimative, *error propagation* metode

Simulationstilgang har vigtige fordele:

- Nem måde at beregne andre størrelser end blot standardafvigelsen
- Nem måde at bruge andre fordelinger end normalfordelingen, hvis vi tror at det beskriver virkeligheden mere korrekt
- Afhænger ikke af en lineær approksimation (som error propagation) til den underliggende ikke-lineære funktion

Parametrisk bootstrap

Introduktion

- Parametrisk bootstrap: simulér gentagne samples fra den antagede (og estimerede) fordeling
- Ikke-parametrisk bootstrap: simulér gentagne samples direkte fra data

Hvis vi fx har nogle målinger for ventetider i et callcenter, som vi forventer følger en eksponentialfordeling, kan vi udregne estimeret μ og λ, men vi kan ikke finde konfidensinterval for μ.

```
k <- 100000

# Simulate 10 exponentials with the "right" mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Compute the mean of the 10 simulated observations k times
sim_means <- apply(sim_samples, 2, mean)

# Find the relevant quantiles of the k simulated means
quantile(sim_means, c(0.025, 0.975))

##  2.5% 97.5%
## 12.59 44.63

hist(sim_means, col="blue", nclass=30, main="", prob=TRUE, xlab="Sim means")
```

Samme kan gøres for median:

```
k <- 100000

# Simulate 10 exponentials with the "right" mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Compute the mean of the 10 simulated observations k times
sim_medians <- apply(sim_samples, 2, median)

# Find the relevant quantiles of the k simulated means
quantile(sim_medians, c(0.025, 0.975))

##  2.5% 97.5%
## 7.038 38.465

hist(sim_medians, col="blue", nclass=30, main="", prob=TRUE, xlab="Sim medians"
```

Metode 4.7

Konfidensinterval for en vilkårlig feature θ ved parametrisk bootstrap.

Antag at vi har faktiske observationer $x_1, \ldots, x_n$ og at disse kommer fra en sandsynlighedsfordeling med tæthed f:

- Simulér k stikprøver af n observationer fra den antagede fordeling f, hvor middelværdien er lig $\bar{x}$
- Udregn estimatet $\hat{\theta}$ for hver af de k stikprøver, kald disse $\hat{\theta}^*_1, \dots, \hat{\theta}^*_{k-}$
- Find $100(\alpha/2)\%$ og $100(1-\alpha/2)\%$ fraktilerne i $\hat{\theta}^*_1, \dots, \hat{\theta}^*_k$, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $100(1-\alpha)\%$ konfidensinterval

Andre parametre skal også matche data bedst muligt. Nogle fordelinger har mere end én. Man bør anvende *maximum likelihood*-tilgangen.

Eksempel

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)
Q3 <- function(x){ quantile(x, 0.75) }
k <- 100000
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))
simQ3s <- apply(sim_samples, 2, Q3)
quantile(simQ3s, c(0.005, 0.995))

##  0.5% 99.5%
## 172.8 198.0
```

4.10 two-sample

Se slide 27. Stort set det samme, bare med differensen mellem features i hver af de k stikprøver.

Se kode på slide 29.

Pointe: vi antager en fordeling

Ikke-parametrisk bootstrap

Vi antager **ikke** en fordeling.

Eksempel: cigaretforbrug, 4.15

```
x1 <- c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <- c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)

dif <- x1 - x2
mean(dif)

[1] 5.273

k = 100000
sim_samples = replicate(k, sample(dif, replace=TRUE))
sim_means = apply(sim_samples, 2, mean)
quantile(sim_means, c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 1.364 9.818
```

Metode: samme som før, men simuleringen er ved tilfældig trækning med tilbagelægning.

4.17

Samme som ovenstående, men med differenser.

```
k <- 100000
simx_samples <- replicate(k, sample(x, replace=TRUE))
simy_samples <- replicate(k, sample(y, replace=TRUE))
sim_median_difs <- apply(simx_samples, 2, median)-
                        apply(simy_samples, 2, median)
quantile(sim_median_difs, c(0.005, 0.995))

##  0.5% 99.5%
##    -8     0
```

Overblik

4 ikke så forskellige metoder:

- Med eller uden fordeling
- For one- eller two-sample analyse

Vi kan også udføre hypotesetest ved at kigge på konfidensintervallerne.