Introduktion til Statistik

Forelæsning 13: Eksamensspørgsmål

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Spring 2023

Overview

1. Basale metoder
2. Distribution
3. Korrelation
4. Regneregler
5. t-test
6. Simulering
7. Regression
8. Andele

Hvad er medianen af følgende tal?

$$12, 15, 17, 22, 24, 27, 29, 32$$

A 15

B 16

C 22

D 23

E 24

Svar: D

Find pn hvor $p = 0.50$, så $pn = 4$ og tag da $\frac{x_{(pn)} + x_{(pn+1)}}{2}$.

Hvad er 1. kvartil af følgende tal?

$$12, 15, 17, 22, 24, 27, 29, 32$$

A 15

B 16

C 22

D 23

E 24

Svar: B

Find pn hvor $p = 0.25$, så $pn = 2$ og tag da $\frac{x_{(pn)} + x_{(pn+1)}}{2}$, se Def. 1.7.

```
quantile(c(12,15,17,22,24,27,29,32), type=2)
```

Hvilken af følgende påstande om tæthedsfunktionen (pdf) for normalfordelingen $N(1, 2^2)$ er falsk?

A Det totale areal under kurven er 1.0 sand

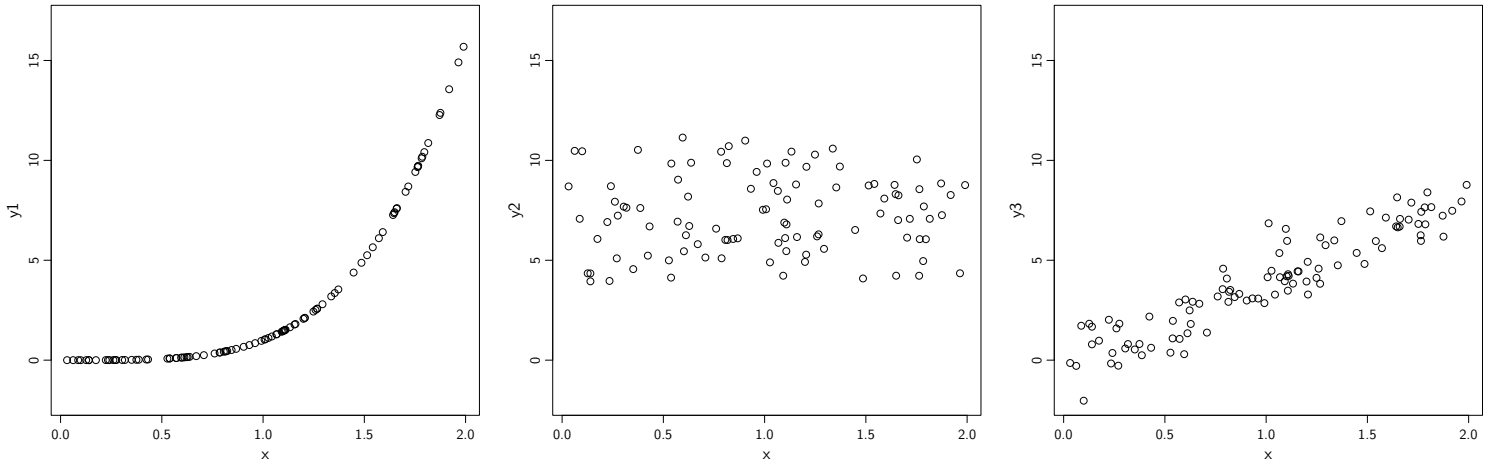
B Middelværdien er lig med 1^2 sand

C Variansen er lig med 2 falsk

D Kurven er symmetrisk omkring middelværdien sand

E De to haler af kurven fortsætter uendeligt i begge retninger sand

Givet følgende 3 plots af sammenhørende værdier af x og y for stikprøver fra 3 forskellige populationer:



Hvilket af følgende er det eneste passende udsagn om korrelationerne af de populationer som stikprøverne er taget fra?

A $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = 0.33$ nope

B $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = -0.89$ nope

C $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0.61$ og $\rho_{XY_3} = 0.91$ nope

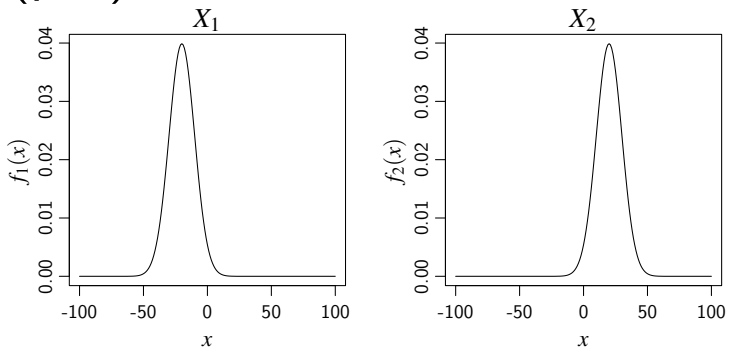
D $\rho_{XY_1} = 0.87$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = 0.92$ Yeah

E $\rho_{XY_1} = 0.22$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = -0.34$ nope

Lad følgende to uafhængige stokastiske variabler være givet ved

$$X_1 \sim N(-20, 10^2) \quad \text{og} \quad X_2 \sim N(20, 10^2).$$

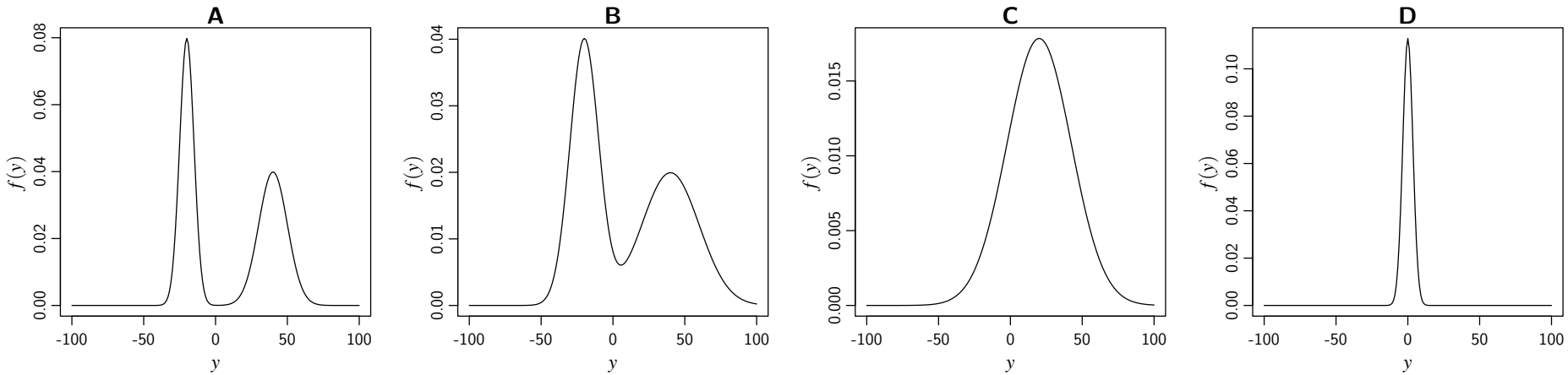
Deres tæthedsfunktioner (pdf) er da:



Nu defineres en ny stokastisk variabel ved

$$Y = X_1 + 2X_2$$

Hvilket af følgende plots er af tæthedsfunktionen (pdf) for Y? SVAR C



Uddybende slide: Varians af ikke-lineære funktioner

Simuler resultat af statistik for ikke-lineære frunktioner.

Varians af ikke-lineær funktion

Find variansen af

$$\mathrm{V}(X^2 + \exp(Y))$$

hvor $X \sim N(\mu_X = 2, \sigma_X^2 = 3)$ og $Y \sim N(\mu_y = 4, \sigma_Y^2 = 2)$

Simuler værdier og beregn statistikken på de simulerede værdier

```
k <- 100000
x <- rnorm(k, mean=2, sd=sqrt(3))
y <- rnorm(k, mean=4, sd=sqrt(2))
simvals <- x^2 + exp(y)
var(simvals)
```

eller brug lineær approximations metode 4.3.

For at sammenligne to undervisningsmetoder i færdighedsregning indgik 18 elever i et mindre eksperiment. Af disse var 8 udvalgt tilfældigt og undervist efter metode 1, og de resterende 10 efter metode 2. Efter undervisningsperioden blev alle eleverne testet på tempoet. Følgende observationer af regnetider (i minutter) blev opnået for den samme test (hurtighed (små tider) anses for at være godt):

```
t.test(x=x1, y=x2) # x1 tider for metode 1 og x2 for metode 2

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -4.13, df = 15.7, p-value = 0.00081
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.3486 -3.0014
## sample estimates:
## mean of x mean of y
##    12.125    18.300
```

Find det *falske* udsagn:

A Der er signifikant forskel på metoderne på niveau $\alpha = 0.01$ sand

B Metode 1 virker signifikant bedre end metode 2 på niveau $\alpha = 0.05$ sand

C Metode 1 virker signifikant bedre end metode 2 på niveau $\alpha = 0.01$ sand

D Metode 1 virker signifikant bedre end metode 2 på niveau $\alpha = 0.0005$ falsk

Der planlægges en ny undersøgelse for at bestemme den gennemsnitlige regnetid for testen mere præcist. Hvis man antager, at standardafvigelsen i regnetider er 3 minutter, og vil finde hvor mange børn skal man så teste for at opnå en styrke (power) på 80% på signifikansniveau 5% for at påvise en forskel på 1 minut. Hvilket af følgende R kald beregner dette korrekt?

A
```
power.t.test(delta=1, sd=3, sig.level=0.05, power=0.8,
             type="two.sample", alternative="two.sided")
```

B
```
power.t.test(n=40, sd=3, sig.level=0.05, power=0.8,
             type="two.sample", alternative="two.sided")
```

C
```
power.t.test(delta=0.5, sd=3^2, sig.level=0.05, power=0.8,
             type="two.sample", alternative="two.sided")
```

D
```
power.t.test(n=400, delta=1, sd=3, sig.level=0.05,
             type="two.sample", alternative="two.sided")
```

E Ved ikke

Svar: A

Hvis en ny test planlægges og der ønskes et konfidensinterval med middelbredde på 1 minut, hvad er da ME (Margin of Error)?

A $ME = 0.25$

B $ME = 0.5$

C $ME = 1$

D $ME = 2$

E Ved ikke

Svar: B. Husk, ME er:

- halvdelen af konfidensintervallets forventede bredde
- (minimum) effektstørrelsen ved hypotesetest, som skal "opdages" med sandsynlighed $1 - \beta$ (styrken)
- `delta` i R funktionen `power.t.test()` (ved `"alternative=two-sided"` som er det eneste vi bruger)

En chef ønsker leverandør A og B sammenlignet uden brug af nogen antagelse om fordeling, og får lavet følgende kørsel i R:

```
xA=c(17,25,22,21,16,22,23,20,17)
xB=c(21,25,20,19,24,19,21,21,17)
k = 10000

## OPTION A:
Asamples = replicate(k, rnorm(9, mean(xA), sd(xA)))
Bsamples = replicate(k, rnorm(xB, mean(xB), sd(xB)))
mymeandifs = apply(Asamples, 2, mean) - apply(Bsamples, 2, mean)

## OPTION B:
Asamples = replicate(k, rlnorm(9, mean(xA), sd(xA)))
Bsamples = replicate(k, rlnorm(xB, mean(xB), sd(xB)))
mymeandifs = apply(Asamples, 2, mean) - apply(Bsamples, 2, mean)

## OPTION C:
Asamples = replicate(k, sample(xA, replace = TRUE))
Bsamples = replicate(k, sample(xB, replace = TRUE))
mymeandifs = apply(Asamples, 2, mean) - apply(Bsamples, 2, mean)
```

Hvilken af de tre options giver de korrekte bootstrap samples? Svar: C

Vi har følgende observationer af x_1, x_2 og y fra 15 personer:

	x1	x2	y
1	7.90	16.70	59.00
2	4.60	13.80	44.00
3	5.10	20.20	59.00
...			
14	5.80	14.60	47.00
15	4.20	20.50	57.00

og den følgende R kode er kørt:

```
lm(y ~ x1 + x2)
```

Hvilken analyse er lavet her?

- A: En two-sample t-test
- B: En en-vejs ANOVA
- C: En simpel lineær regressionsanalyse
- D: En multipel lineær regressionsanalyse, MLR
- E: Ved ikke

I følgende tabel ses antallet af såkaldte "challenges" i en tennisturnering, der bruger det elektroniske "instant replay" system Hawk-Eye, opgjort efter køn af tennisspilleren og om den var berettiget eller ej: (Forklaring: En "challenge" af spilleren giver mulighed for øjeblikkeligt at se, om dommernes vurdering af en bold (inde/ude) er korrekt eller ej)

	Berettiget	
	Kvinder	Mænd
Ja	135	209
Nej	252	295

Den χ^2-fordelte test statistik i en relevant test for om der er forskel på succes-sandsynligheden for kvinder og mænd er givet ved:

A $\frac{(209-194.6)^2}{194.6} + \frac{(295-309.4)^2}{309.4}$

B $(135-149.4)^2 + (252-237.6)^2$

C $(135-209)^2 + (252-295)^2$

D $\frac{(135-149.4)^2}{149.4} + \frac{(209-194.6)^2}{194.6} + \frac{(252-237.6)^2}{237.6} + \frac{(295-309.4)^2}{309.4}$ KORREKT

E Ved ikke

Uddybet svar på forrige slide

Man skal udregne en χ^2-fordelt teststørrelse, se eksempel 7.21, under H_0. Man siger altså andelen af ja er den samme

$$H_0 : p_{\text{kvinder}} = p_{\text{maend}} = p_{\text{ja}}$$

dvs. at man forventer at forholdet er det samme for kvinder og mænd

$$H_0 : \frac{x_{\text{kvinder}}}{n_{\text{kvinder}}} = \frac{x_{\text{maend}}}{n_{\text{maend}}} = \hat{p}_{\text{ja}} = \frac{x_{\text{ja}}}{n_{\text{total}}}$$

Man finder række- og kolonnetotalerne

	Berettiget		
	Kvinder	Mænd	Total
Ja	135	209	x_{ja}
Nej	252	295	x_{nej}
Total	n_{kvinder}	n_{maend}	n_{total}

hvilket man så bruger til at finde de forventede (under H_0) i hver celle

$$e_{11} = \hat{p}_{\text{ja}} n_{\text{kvinder}} = \frac{x_{\text{ja}}}{n_{\text{total}}} n_{\text{kvinder}} = \frac{135+209}{135+209+252+295}(135+252) = 149.4$$

$$e_{12} = \hat{p}_{\text{ja}} n_{\text{maend}} = \frac{x_{\text{ja}}}{n_{\text{total}}} n_{\text{maend}} = \frac{135+209}{135+209+252+295}(209+295) = 194.6$$

Tilsvarende findes de forventede for andelen af nej i hver gruppe ($\hat{p}_{\text{nej}} = 1 - \hat{p}_{\text{ja}} = \frac{x_{\text{nej}}}{n_{\text{total}}}$)

$$e_{21} = \hat{p}_{\text{nej}} n_{\text{kvinder}} = \frac{x_{\text{nej}}}{n_{\text{total}}} n_{\text{kvinder}} = \frac{252+295}{135+209+252+295}(135+252) = 237.4$$

$$e_{22} = \hat{p}_{\text{nej}} n_{\text{maend}} = \frac{x_{\text{nej}}}{n_{\text{total}}} n_{\text{maend}} = \frac{252+295}{135+209+252+295}(209+295) = 309.4$$

Det de værdier som bruges i formlen for χ^2 test statistikken, se metode 7.20.