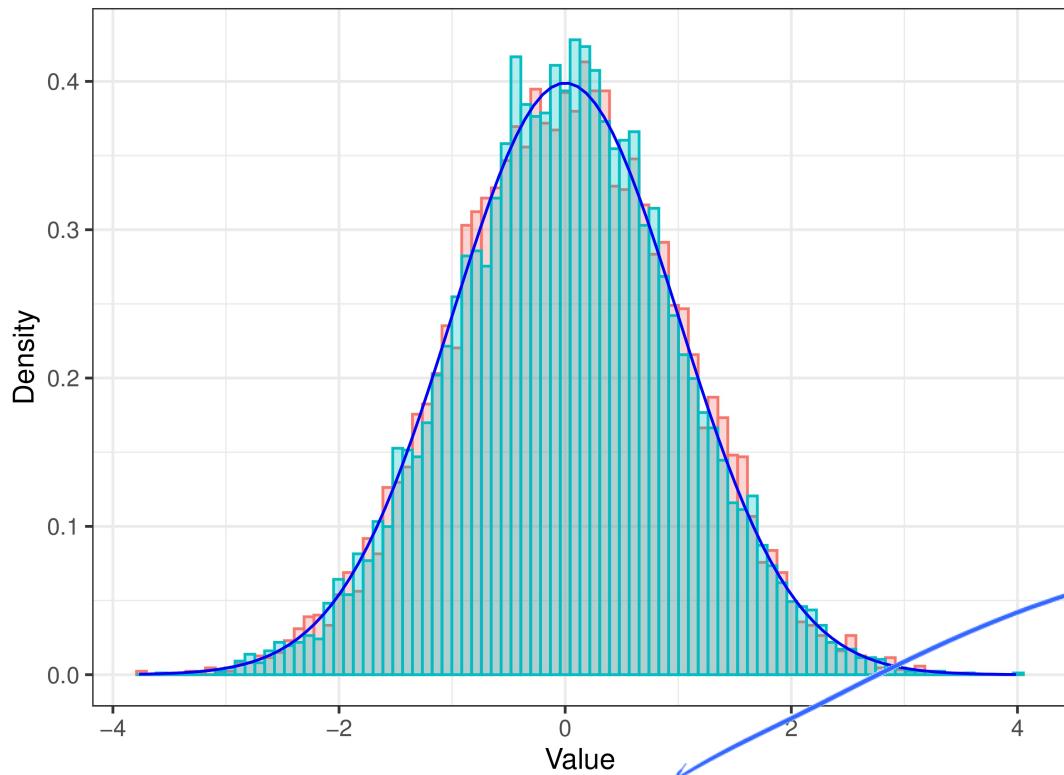


## Box–Muller Transformationen



### Teori

Hvad er en pop., hvad er en stikprøv.?

#### Middelværdi og standardafvigelse

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 3 I BOGEN

I en population findes der to værdier,  $\mu$  og  $\sigma$ , som er interessante at kigge nærmere på. Gennemsnittet af en population kaldes  $\mu$  (middelværdi), mens at populationens standardafvigelse kaldes  $\sigma$ .

Populations standardafvigelse kan udregnes ud fra populationens varians. Nedenstående ligning viser, hvordan variansen for en stikprøve kan udregnes:

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

hvordan udreges  
 $\mu, \sigma, \bar{x}, s^2$ ?

Herunder er der:

- $x_i$ , som svarer til den enkelte observation i stikprøven.
- $\bar{x}$ , som svarer til gennemsnittet af stikprøven.

Sammenhængen mellem en populations standardafvigelse og populationens varians kan ses i nedenstående ligning.:

$$\sigma = \sqrt{\text{var}(x)} = \sqrt{\sigma^2}$$

hvordan ikke  $\pm$ ?

Populationen bliver her betegnet som  $X$

hvordan fortolkes  
 $\mu, \sigma$ ?

## Fordelinger

Der findes en lang række sandsynlighedsfordelinger, og i det følgende afsnit fokuseres der på de fordelinger, som har mest relevans for projektet.

### Standardnormalfordeling

Standardnormalfordeling: <https://mse.redwoods.edu/darnold/math15/UsingRInStatistics/StandardNormal.php>, ? what?

En standardnormalfordeling er kendtegnet ved at  $\mu = 0$ , og at  $\sigma = 1$ . Desuden har en standardnormalfordeling et klokkeformet udseende. Nedenstående er der plottet en standardnormalfordeling:

```
qdist("norm", p=1, mean = 0, sd = 1, xlim = c(-4,4))
```

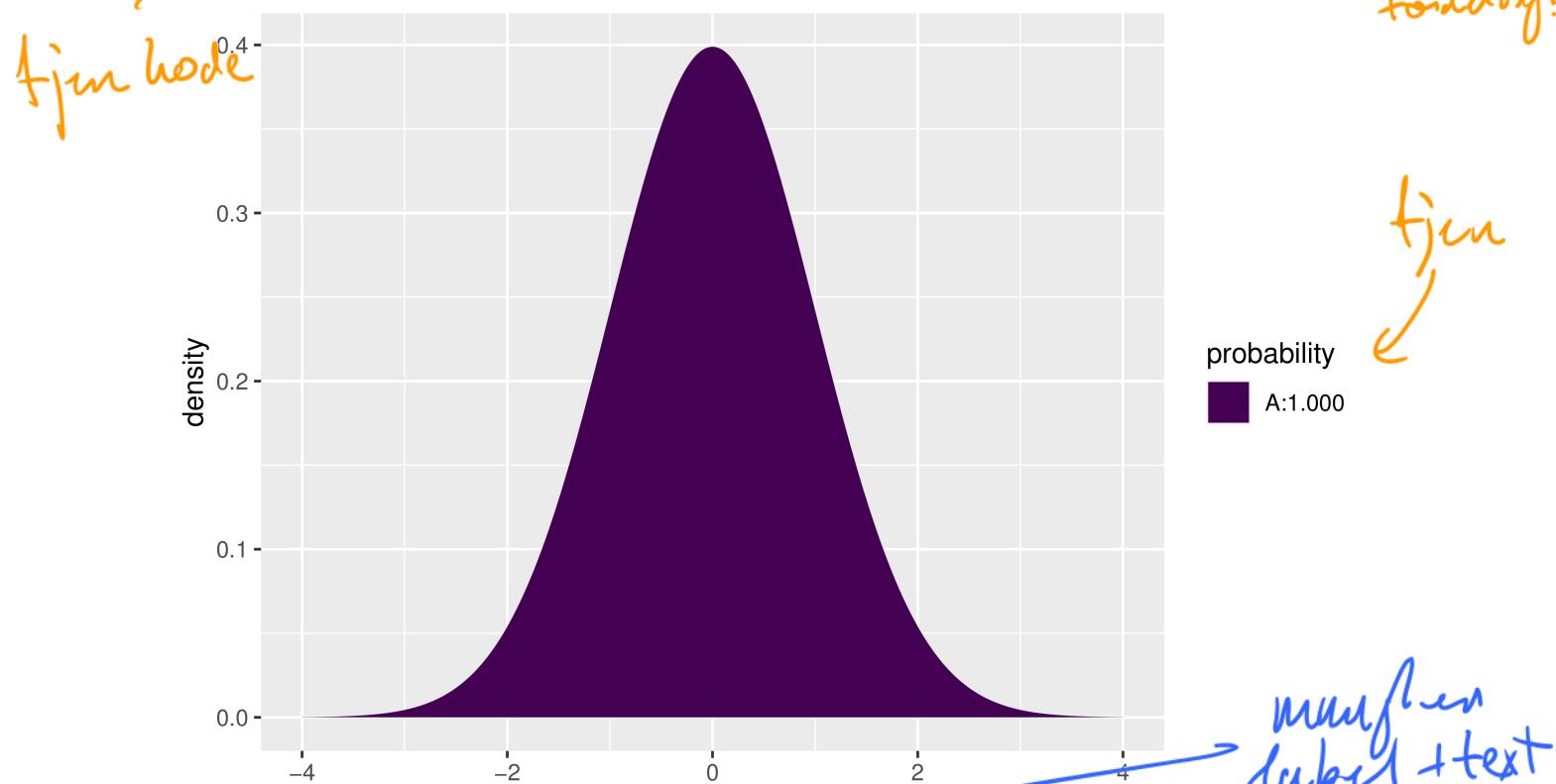


Fig. 1: ... tætst ...

hvordan relateres fig. 1 til  
sandsyn?

Der er en høj sandsynlighed for, at få en værdi tæt på 0. Jo længere man afviger fra  $\mu$ , jo mindre er densiteten, som medfører at sandsynligheden for de givne værdier mindskes. Ergo vil sandsynligheden for at få 2, være mindre end sandsynligheden for at få 0. En af anvendeserne for en standardnormalfordeling er, at den kan benyttes til en Z-test, som benyttes senere i projektet.

Z-fordeling DETTE AFSNIT ER SKREVET UD FRA KAPITEL 4 I BOGEN

ikke den korrekte for tolkning

findes ikke.

hvad er sammenhængen  
med uddannelsesfænrete og stud.  
norm.?

gælder dette kun for  
normalfordelte variabler

For at finde z-scoren, skal  $\mu, \sigma$  være kendt. z-scoren kan udregnes for normalfordelinger ved formlen:

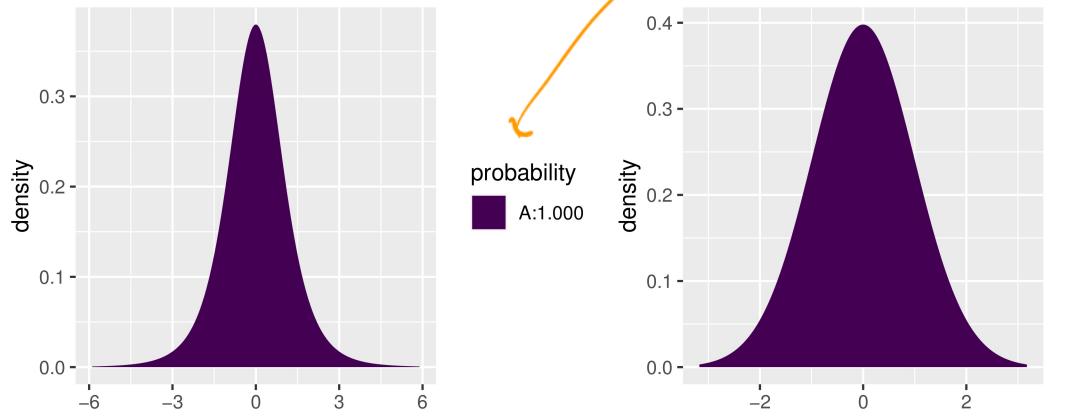
$$z = \frac{y - \mu}{\sigma}$$

hvor  $y$  er en værdi, som er  $z$  standardafvigelse fra  $\mu$ . Desuden, kan det udledes, at hvis  $z > 0$  er  $y$ -værdien på højre side af  $\mu$ , og omvendt, hvis  $z < 0$ , er  $y$ -værdien på venstre side af  $\mu$ .

### T-fordeling DETTE AFSNIT ER SKREVET UD FRA KAPITEL 5 I BOGEN

En T-fordeling laves ud fra en standardnormalfordeling og dens udseende minder også om denne fordeling. Forskellen på en T-fordeling og en standardnormalfordeling, ligger i, at en T-fordeling laves ud fra frihedsgrader. For en stikprøve vil antallet af frihedsgrader være antallet af observationer minus 1, altså  $n - 1$ . Jo flere frihedsgrader der er, jo mere ligner en T-fordeling en standardnormalfordeling. Nedenstående er der plottet 2 T-fordelinger med henholdsvis 5 og 100 frihedsgrader:

```
p1 <- qdist("t", df = 5, p = 1, return = "plot")
p2 <- qdist("t", df = 100, p = 1, return = "plot")
#p3 <- qdist(mean = 0, sd = 1, p = 0, return="plot")
grid.arrange(p1, p2, ncol=2)
```



### Binomialfordeling

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 6 I BOGEN

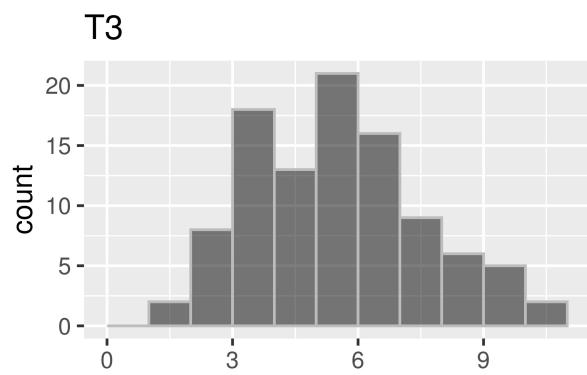
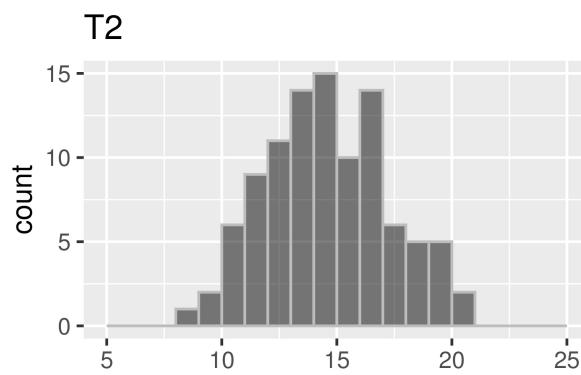
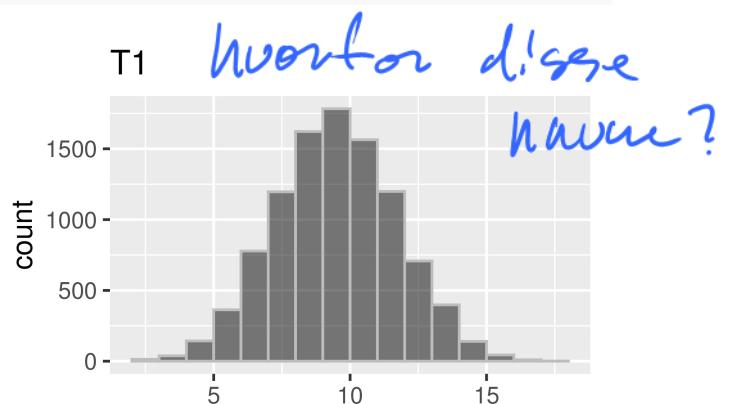
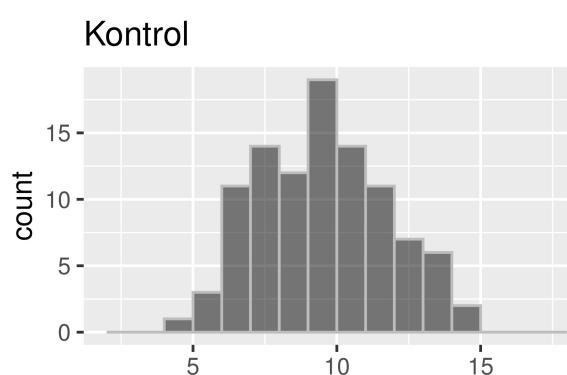
Binomialfordelingen benyttes i sammenhænge med kvalitative data. Specifikt er der kun to mulige udfald. Hvis der er lige stor chance for hvert af udfaldene, altså  $\pi = 0.50$ , og en stor stikprøve, vil en binomialfordeling vise sig at være symmetrisk. Nedenstående er et eksempel på 4 binomialfordelinger.

```
set.seed(987)
rbino1 <- c(rbinom(100, 20, 0.5))
p1 <- gf_histogram(~ rbino1, fill="black", col="grey", breaks = seq(2,18, by=1), title = "Kontrol", xla
#Test 2: Antallet af test forøges
set.seed(987)
rbino2 <- c(rbinom(10000,20,0.5))
p2 <- gf_histogram(~ rbino2, fill = "black", col = "grey", breaks = seq(2,18, by = 1), title = "T1", xla
```

```

#Test 3: Antallet af kroner forøges
set.seed(987)
rbino3 <- c(rbinom(100,30,0.5))
p3 <- gf_histogram(~ rbino3, fill = "black", col = "grey", breaks = seq(5,25, by =1), title = "T2", xla
Fjern  
kode
#Test 4: Sandsynligheden ændres fra 50%
set.seed(987)
rbino4 <- c(rbinom(100,20,0.3))
p4 <- gf_histogram(~ rbino4, fill = "black", col = "grey", breaks = seq(0,11, by =1), title = "T3", xla
grid.arrange(p1, p2, p3, p4, ncol=2, nrow = 2)

```



I ovenstående figur er der altså fire grafer, som er lavet ud fra binomialfordelingen. Der er en kontrolgraf, hvor 100 observationer hver bliver testet 20 gange med en 50% chance for success. Yderligere er der T1, T2 og T3, hvor der ved hver graf er ændret en af parameterene fra kontrol grafen. T1 har flere observationer, som gør at grafen får et udseende af en normalfordeling, her kan det aflæses at  $\bar{x} = 10$ . Ved T2 bliver hver observation testet 30 gange i stedet for 20 gange, og dette resulterer i at middelværdien (hvis antallet af observationer er stort nok) i stedet vil være  $\bar{x} = 15$ . I T3 er sandsynligheden for success ændret fra 50% til 30%. Dette ændrer igen udseendet på grafen, hvor middelværdien bliver lavere, da hver test har en mindre sandsynlighed for at være en success.

Poisson Fordeling

<https://www.statology.org/plot-poisson-distribution-r/>

En poisson fordeling beskriver chancen for at en begivenhed sker et givet antal gange over et kendt tidsinterval. Dette kunne være hvor mange gange der har været stormvejr indenfor det sidste år. Fordelingen laves

? *Har I tænkt jer ut  
bruge Poisson?*

holdes normalt  
 inting i tet

ud fra følgende formel:  

$$P(x; \lambda) = \frac{(e^{-\lambda} \cdot \lambda^x)}{x!}$$

Hvor:  $e$  er eulers tal.  $\lambda$  er begivenhedsraten, altså det forventede antal gange begivenheden vil ske.

Hvis  $\lambda$  er 5, og der undersøges hvad chancen for at begivenheden sker 7 gange, så vil  $x = 7$ . Dette eksempel kan ses forneden:

```
(exp(1)^-5*5^7)/factorial(7)
## [1] 0.1044449
```

fjern

Der er altså en 10.4% chance for at en begivenhed sker 7 gange, hvis begivenhedsraten er 5. Nedenstående er resultaterne af denne poissonfordeling plottet:

```
success <- 0:20
dpois(7, 5)
```

```
## [1] 0.1044449
```

```
plot(success, dpois(0:20, 5), type = "h", xlab = "", ylab = "")
```

{ fjern

Eles?  
Samme  
Sfg. seen

{ fjern

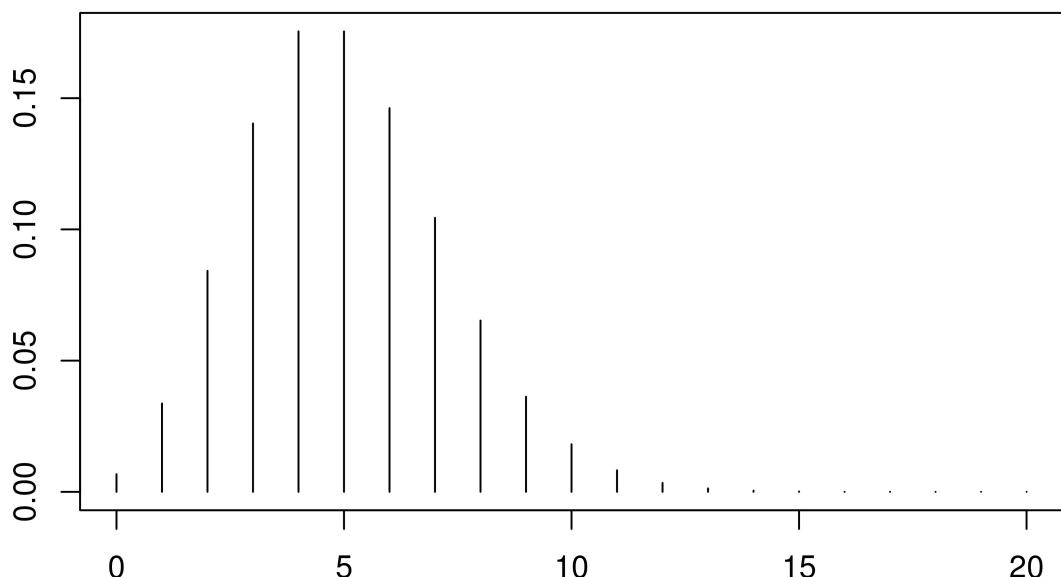


Fig X. ... test...

ihle un  
god bestudere

## Uniform fordeling

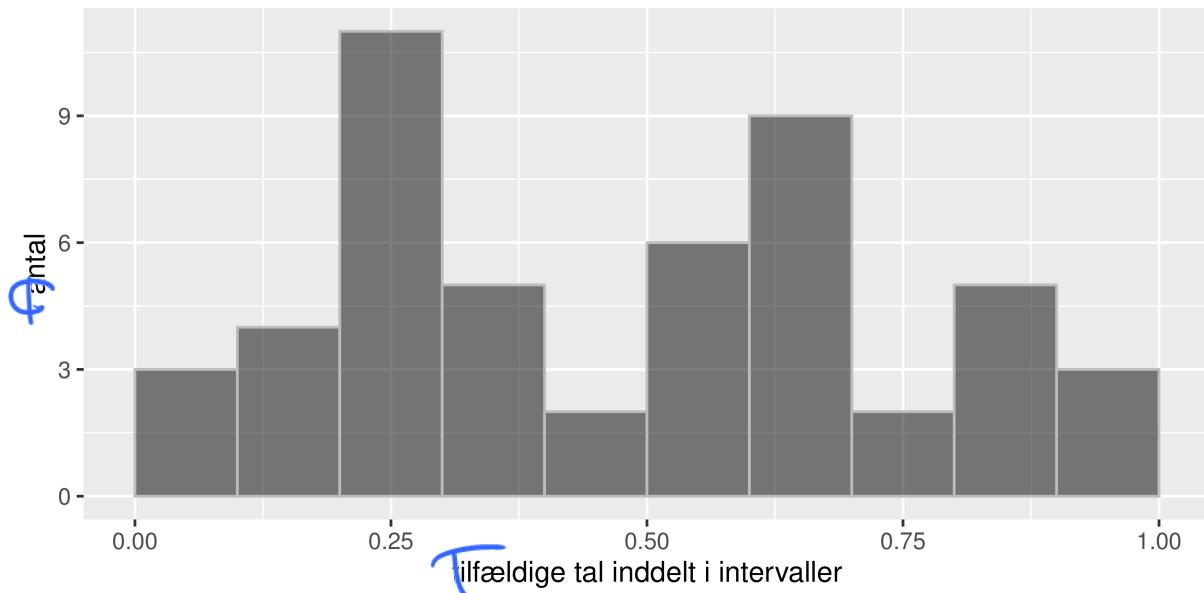
En uniform fordeling, er hvor dataet bliver inddelt i lige store intervaller, og alle intervaller, har lige antal observationer. Det er vigtigt at notere at hvis antallet af observationer er lavt, vil fordelingen ikke tilsvarelende være uniform. Dette vises ved at lave to uniformfordelinger på det samme seed, hvor den eneste forskel er antallet af observationer. Dette kan ses i de nedenstående grafer.

```
set.seed(1234)
unifx <- runif(50, min = 0, max = 1)
set.seed(1234)
unifx1 <- runif(5000, min = 0, max = 1)

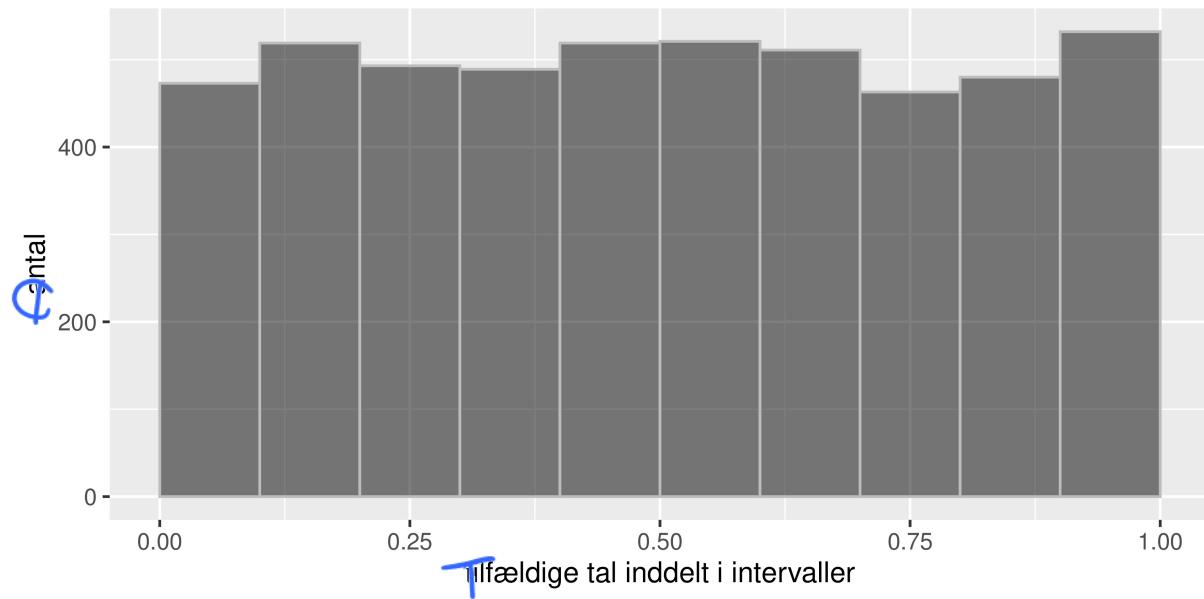
p1 <- gf_histogram(~unifx, breaks = seq(0,1,by=0.1), fill="black", col="grey", xlab = "tilfældige tal inde
p2 <- gf_histogram(~unifx1, breaks = seq(0,1,by=0.1), fill="black", col="grey", xlab = "tilfældige tal inde
grid.arrange(p1, p2, nrow=2)
```

Afjørn

### Uniform fordeling ved 50 observationer



### Uniform fordeling ved 5000 observationer



### Statistisk inferens

Indenfor statistisk analyse er der to kategorier; den første, deskriptiv statistik har i fokus at beskrive data, hvor den anden statistisk inferens har i fokus at lave forudsigelser om et element på baggrund af data og tendenser dertil. I rapporten anvendes statistisk inferens, det gøres i form af estimation, der benyttes både punktestimater og intervallestimate. Hvilket vil sige, at der først findes et punktestimat, altså et gæt, eksemplificeret ved middelværdien fra en stikprøve. Selvom der så er meget meget lille sandsynlighed for, at det også er middelværdien i hele populationen, kan estimatet anvendes, til at lave et intervallestimat, eller et konfidensinterval, hvori hele populationens middelværdi med ret stor sikkerhed hører til.

ihke  
nælt  
nigtigt

meget meget lille er  
et punkt  
int. est.?  
17  
en underdrivelse  
ihke helt  
nærligt

## Konfidensinterval

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 5 I BOGEN

estimeres har ikke  
bruges på denne  
måde...

Når der indenfor statistik estimeres på en populations parameter ud fra en stikprøve, vil resultatet aldrig være perfekt. Dette skyldes at der er en fejlmargin, som der skal tages højde fra. En måde hvorpå man kan tage højde for denne fejlmargin er et konfidensinterval.

hvor kommer den fra?

Et konfidensinterval for en givet parameter, er et interval mellem to tal, hvori det estimeres at parameteren ligger. Sandsynligheden for at producere et interval, som indeholder parameteren kaldes for et konfiden-

sniveau. Denne værdi er valgt til et tal tæt på 1, som regel enten 0.95 eller 0.99.

En konfidensinterval skabes på baggrund, af et punktestimat og fejlmarginen. Måden dette gøres på, kan ses forneden:

$$KI = \text{Punktestimat} \pm \text{fejlmargin}$$

Måden at finde konfidensintervallet varierer baseret på, hvilken fordeling man bruger. I de følgende afsnit vil der forklares, hvordan konfidensintervallet for kvalitative og kvantitative variabler findes, og opstilles ved hjælp af eksempler.

hvor kommer dette fra?

**Kvalitative variabler** Hvis man spurte en population, hvorvidt de kunne lide deres job eller ej, ville en andel svare "ja", og en andel svare "nej". Er man interesseret i hvor mange procent svarede "ja" til spørgsmålet, ville man dividere antallet der svarede "ja" med antallet af observation. Se følgende:

$$\hat{\pi}_{ja} = \frac{n_{ja}}{n_{total}}$$

Før selve konfidensintervallet kan opstilles, skal standardfejlen ( $sf$ ) findes. Følgende formel bruges for at finde standardfejlen for kvalitative variabler:

$$sf(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

hvor kommer dette fra?

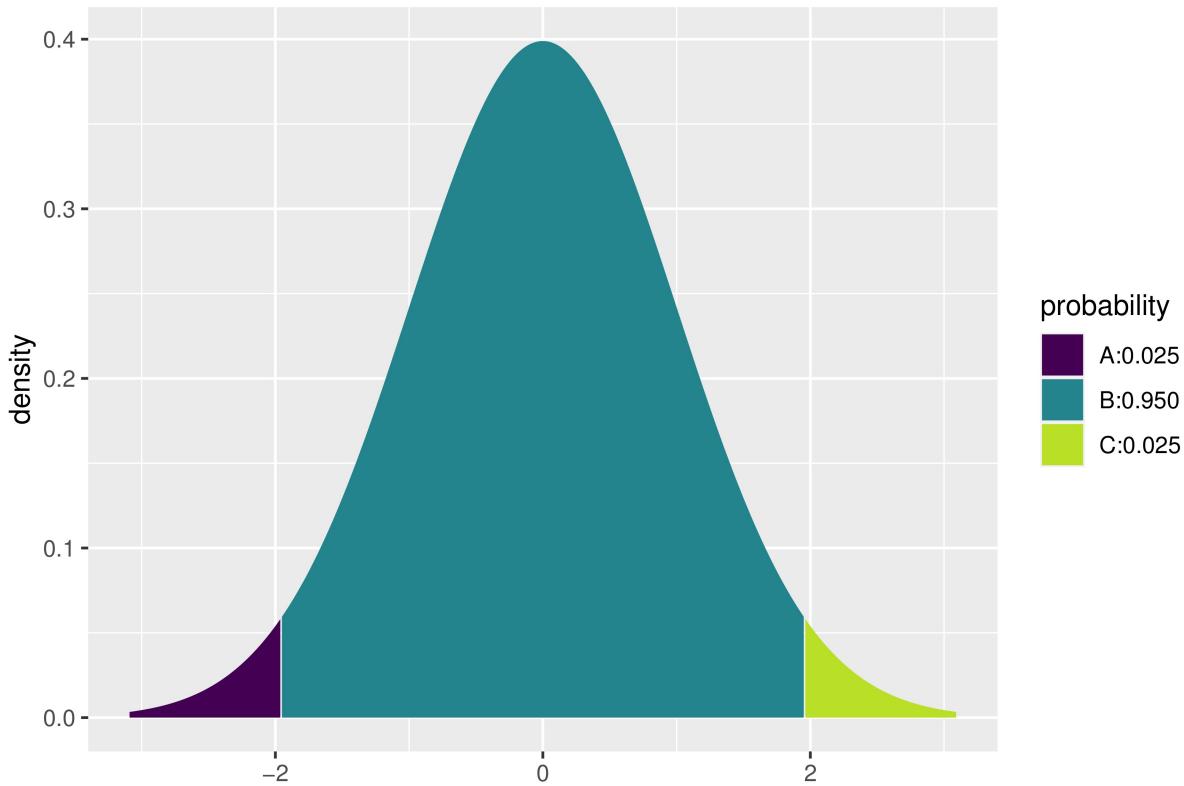
Når standardfejlen er fundet, kan denne ganges med en værdi som afhænger af signifikansniveauet. Denne værdi kaldes  $z_{crit}$ . For at finde ud af hvilken værdi  $z_{crit}$  skal være, benyttes en z-test. En z-test er i realiteten bare en standardnormalfordeling. For at finde  $z_{crit}$  for signifikansniveauet 95%, findes x-værdien ved 97.5% og 2.5%.

hvorom z

hvor er en  
z-test?

```
qdist("norm", mean = 0, sd = 1, p = c(0.025, 0.975))
```

{  
ffern hode



`## [1] -1.959964 1.959964` de er præcis 1,96

Denne værdi aflæses til at være  $\sim 1.96$ . Nu kan konfidensintervallet opstilles via følgende ligning:

$$KI = \hat{\pi} \pm 1.96 \cdot sf(\hat{\pi})$$

Da dette blev lavet udfra signifikansniveauet 95%, kan der altså med 95%'s sikkerhed siges at  $\hat{\pi}$  ligger indenfor intervallet:

$$(\hat{\pi} - 1.96 \cdot sf(\hat{\pi}), \hat{\pi} + 1.96 \cdot sf(\hat{\pi}))$$

hender vi  
i lu denne?

**Kvantitative variabler** Konfidensintervallet for kvantitative variabler har næsten samme formel som ved kvalitative variabler. Dog er måden hvorpå standardfejlen findes anderledes. Desuden skal der bruges en anden slags test ved hjælp af signifikansniveauet.

$$KI = \bar{y} \pm t_{crit} \cdot sf(\bar{y})$$

Herunder er standardfejlen udregnet ved:

$$sf(\bar{y}) = \frac{s}{\sqrt{n}}$$

hvor  $s$  er standardafvigelsen for stikprøven, og  $n$  er antallet af observationer i stikprøven.

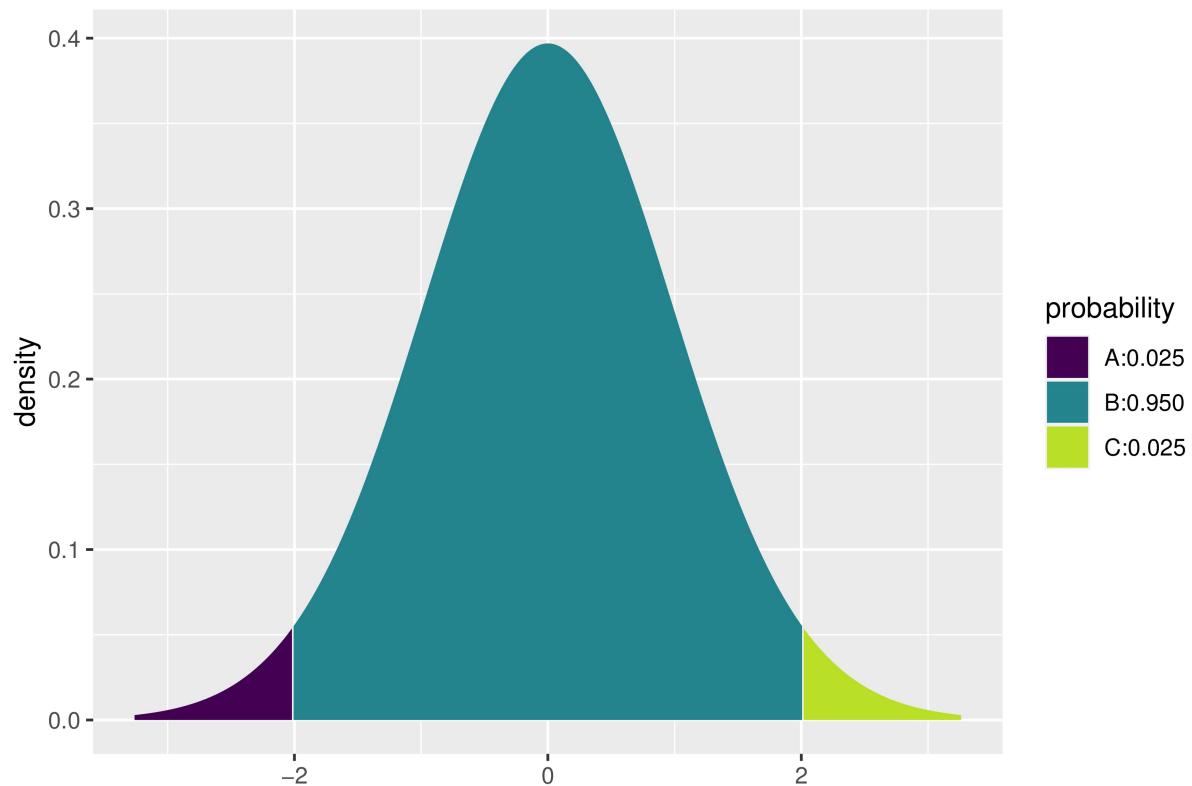
For at finde  $t_{crit}$  benyttes der en t-test. Hvis det ønskede signifikansniveau igen er 95%, hvor der igen aflæses x-værdien ved 2.5% og 97.5%. Nedenstående er der et eksempel, hvor en t-test med 50 frihedsgrader benyttes til, at finde 95% signifikansniveauet:

ihle sandt.

hvorfor  $t_{97.5}$ ?

En den forskel  
mellan standart  
normal forud. og  $t_{97.5}$

```
qdist("t", df = 50, p = c(0.025, 0.975))
```



```
#> [1] -2.008559 2.008559
```

Denne værdi aflæses til at være  $\sim 2.01$ . Nu kan konfidensintervallet igen opstilles:

$$(\bar{x} - 2.01 \cdot sf(\bar{x}), \bar{x} + 2.01 \cdot sf(\bar{x}))$$

## Hypoteser

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 6 I BOGEN

Hver signifikantest har to hypoteser. En hypotese er et udsagn omkring værdien af population parameter. En hypotese forudsiger, hvorvidt en parameter påvirker en numerisk værdi eller falder indenfor en specifik værdimængde. Det gør det så muligt at beregne, om en hypotese er sand eller falsk.

For at sørge for, at man anvender en pålidelig analysemethode på en stikprøve, så opsættes en hypotese omkring sammenhængen af ens data, hvorefter der opstilles en hypotesetest. Det er en metode til at teste en hypotese, (Alternativ hypotese -  $H_a$ ) hvor den sættes op mod en konkurrerende hypotese, nulhypotesen ( $H_0$ ). En stikprøve tages ud af en population, hvor  $H_0$  testes imod  $H_a$  ved brug af et signifikansniveau ( $\alpha$ ) og en p-værdi.

Et klassisk eksempel på dette kunne være at opstille:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \mu \neq \mu_0 \end{aligned}$$

*ihle sandt.*

Hvor  $\mu$  er den ene stikprøves populations middelværdi og  $\mu_0$  er den andens stikprøves populations middelværdi.

For at tjekke hvilken af hypoteserne er sand eller falsk, vælges et signifikansniveau som ofte er  $\alpha = 5\%$  eller  $\alpha = 1\%$ . Ud fra dette signifikansniveau bliver der udregnet en kritisk værdi på en sandsynlighedsfordeling.

*shal  
skriv  
med om:*

$p > \alpha$  så forkastes  $H_0$  ikke.

$p \leq \alpha$  så forkastes  $H_0$ , og  $H_1$  foretrækkes.

*hvor er  $H_1$ ?*

For at finde  $p$ -værdien skal der foretages en hypotesetest. Der findes en række hypotesetest, som vil være relevante i forskellige sammenhænge.

### Hypotesetest

*hvilken test er dette?*

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 6 I BOGEN

For hypotesetester for kvantitative variabler, er der et par nødvendige antagelser. Først og fremmest antages det at populationen er normalfordelt, samt at stikprøven som der undersøges ud fra er taget ud fra populationen tilfældigt.

Det næste der skal gøres, er at opsætte nogle hypoteser, altså en  $H_0$  og en  $H_a$ :

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

Det næste der skal gøres, er at udregne en  $t$ -score, som kan gøres ved hjælp af følgende formel:

$$t = \frac{\bar{y} - \mu_0}{s_f}, \quad s_f = \frac{s}{\sqrt{n}}$$

*hældos en  
test størrelse på denne  
test størrelse på denne*

Det kan ses at  $t$ -scoren afhænger af  $\bar{y}$ , og den estimerede  $\mu_0$  værdi. Yderligere afhænger  $t$ -scoren også af standardfejlen. Standardfejlen udregnes ved at dividere afvigelsen af stikprøven med kvadratrodren af antallet af observationer. Nu kan  $t_{crit}$  udregnes ved hjælp af en T-fordeling med  $n - 1$  frihedsgrader. Yderligere skal der vælges et signifikansniveau, og dette sættes til 95%.

`qdist("t", p = c(0.025, 0.975), df = 99)`

*hvorfor har denne  
nyt acum?*

*ihle et estimat.*

*signifikansniveauet  
er 5%, ihle*

*95%.*

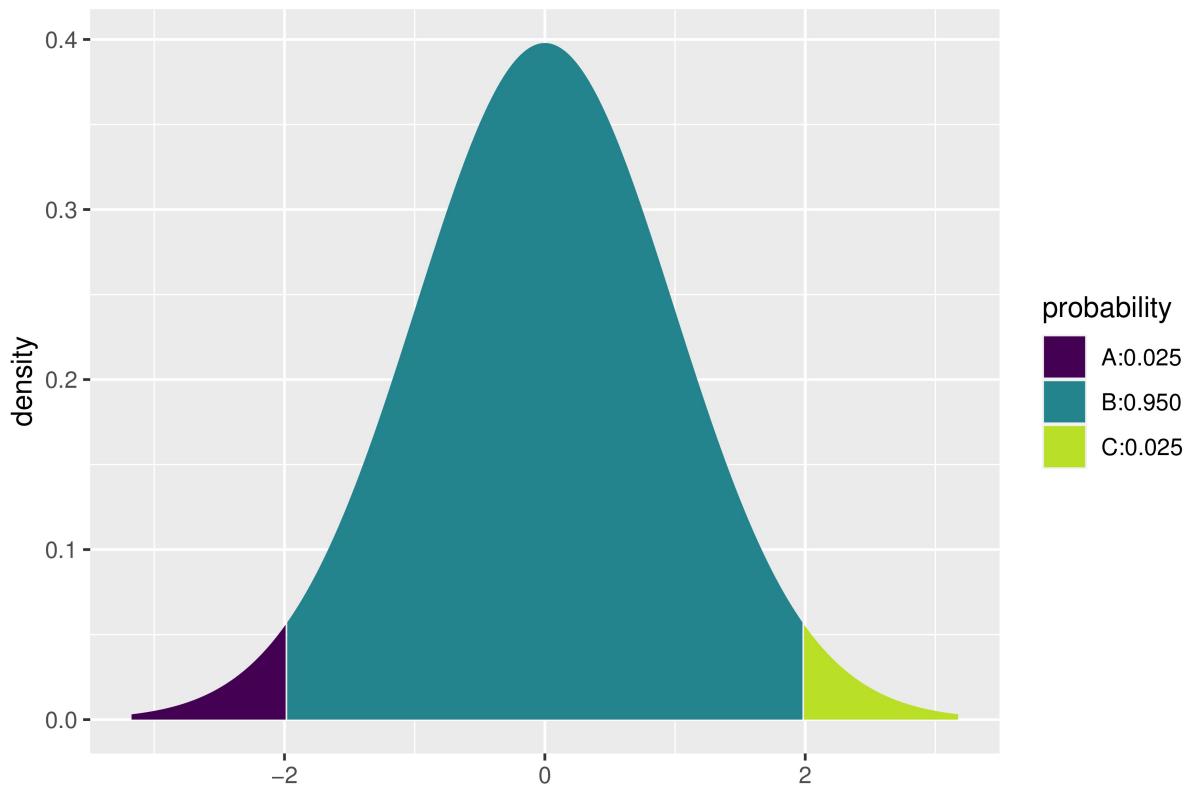


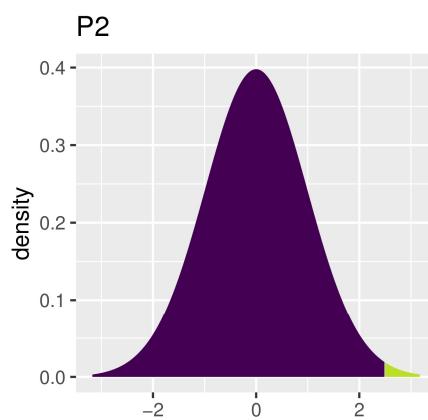
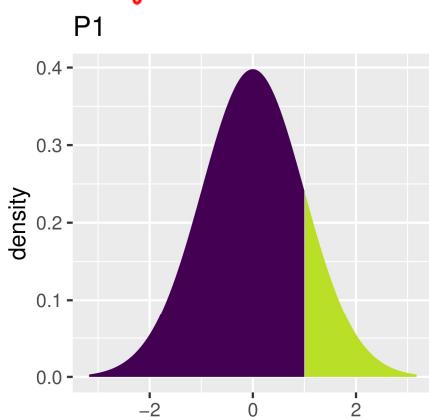
Fig X :

## [1] -1.984217 1.984217

Nu kan  $t_{crit}$  aflæses til at være  $\pm 1.984$ . Hernæst skal den udregnede  $t$ -score benyttes i en T-fordeling med  $n - 1$  frihedsgrader. Dette vil give  $p$ -scoren. Nedenstående er der 2 eksempler hvor  $t$ -scoren henholdsvis er 1 og 2.5:

hvaad er det?

~~p1 <- pdist("t", q = 1, df = 99, return = "plot", title = "P1")  
p2 <- pdist("t", q = 2.5, df = 99, return = "plot", title = "P2")  
grid.arrange(p1, p2, ncol=2)~~



I figur P1 er  $p$ -scoren  $2 \cdot 0.16 = 0.32$ , mens at figur P2 har en  $p$ -score på  $2 \cdot 0.007 = 0.014$ . Da sig-

? hvaad  
er dette  
for et  
eksempel!

nifikansniveauet  $\alpha = 0.05$ , betyder det altså for P1 at  $p \geq \alpha$ . Dette betyder at der ikke er nok evidens til at forkaste  $H_0$ . For P2 betyder det at  $p \leq \alpha$ . Dette betyder at der er nok evidens til at forkaste  $H_0$ , og at  $H_a$  er mere sandsynlig.

*andele*  
 Fremgangsmåden til hypotesetest ved proportioner, er i stort omfang lig med hypotesestesten for kvantitative variabler. En af forskellene er at en antagelse for proportioner er at man forventer mindst 10 observationer. Hypotesen er for test er bygget op på fuldstændig samme måde, dog er forskellen denotationen, hvor der heri benyttes  $\pi$  i stedet for  $\mu$ . Dvs. at Nulhypotesen ser således ud.

$$H_0 : \pi = \pi_0 \quad H_a : \pi \neq \pi_0$$

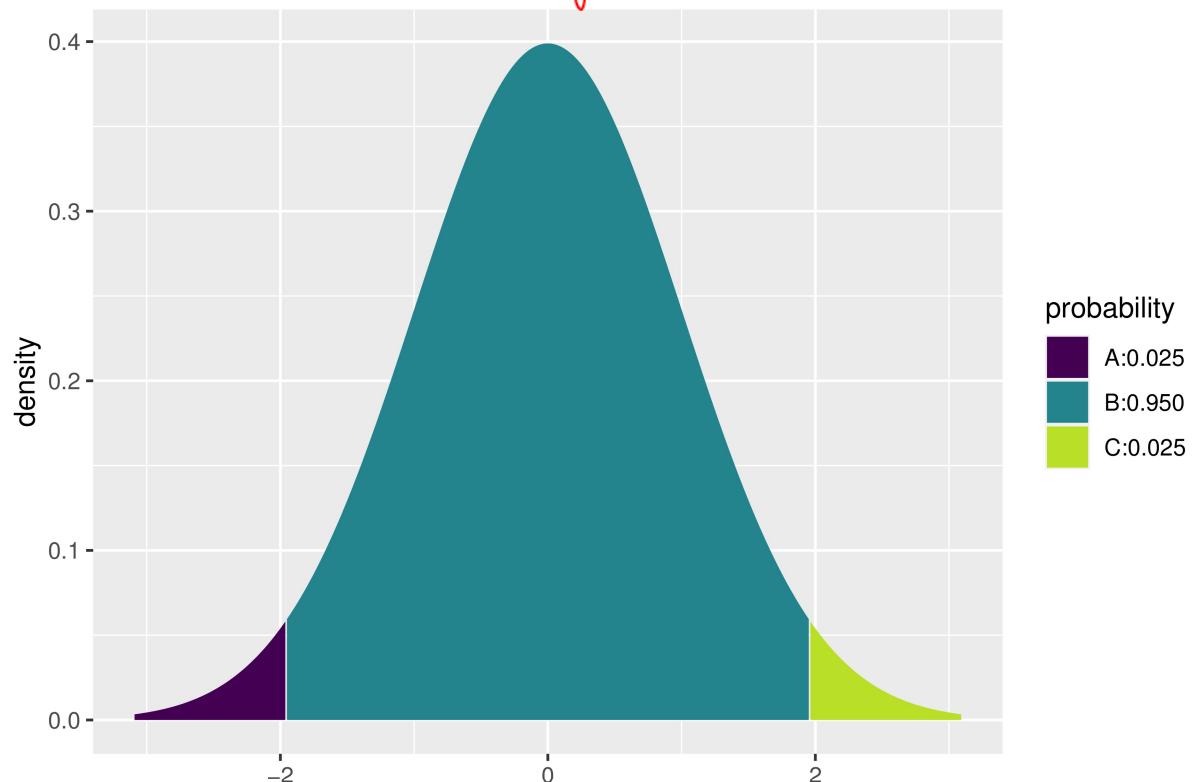
Hvor  $\pi_0$  er den forventede proportion. Desuden benyttes der nu z-score, hvor z-værdien får ud fra følgende ligning:

$$z = \frac{\hat{\pi} - \pi_0}{s_{f_0}}, \quad s_{f_0} = \sqrt{\pi_0(1 - \pi_0) / n}$$

Det kan ses at z-testen indeholder en  $\hat{\pi}$ , som svarer til proportionen for stikprøven. Standardfejlen udregnes på en anden måde end ved t-test.

Man kan finde  $z_{crit}$  ved hjælp af en standardnormalfordeling, hvor man igen skal have et signifikansniveau. I nedenstående graf er 95% brugt:

```
qdist("norm", mean = 0, sd = 1, p = c(0.025, 0.975))
```



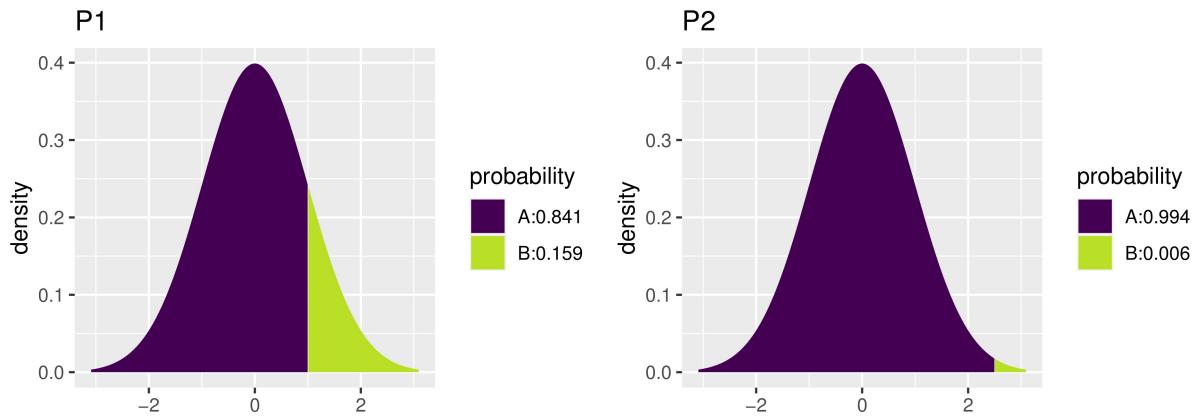
Fj X: . . . .

```
## [1] -1.959904 1.959904
```

Hvorfor bliver vi zrit, hvis vi alligevel udregner p-værdi?

Ergo er  $z_{crit}$  altså  $\pm 1.96$ . Nu kan  $z$ -scoren benyttes til at finde  $p$ -værdien. Nedenstående er der to eksempler, hvor  $z$ -scoren er henholdsvis 1 og 3:

```
p1 <- pdist("norm", mean = 0, sd = 1, q = 1, return = "plot", title = "P1")
p2 <- pdist("norm", mean = 0, sd = 1, q = 2.5, return = "plot", title = "P2")
grid.arrange(p1, p2, ncol=2)
```



I figur P1 er  $p$ -scoren  $2 \cdot 0.159 = 0.318$ , mens at figur P2 har en  $p$ -score på  $2 \cdot 0.006 = 0.012$ . Da signifikansniveauet  $\alpha = 0.05$ , betyder det altså for P1 at  $p \geq \alpha$ . Dette betyder at der ikke er nok evidens til at forkaste  $H_0$ . For P2 betyder det at  $p \leq \alpha$ . Dette betyder at der er nok evidens til at forkaste  $H_0$ , og at  $H_a$  er mere sandsynlig.

## Metoder

Fejl i signifikantest *hyp. - test*

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 6 I BOGEN

Indenfor signifikantest, er der to mulige konklusioner; at nulhypotesen skal forkastes, eller at nulhypotesen ikke kan forkastes. Dette betyder at der er to typer tilfælde, hvor man har ret og to, hvor man tager fejl. De to fejtyper hedder type 1 og type 2. Den førstnævnte, er når nulhypotesen forkastes, selvom den i virkeligheden, var sand. Dermed afhænger type 1 af signifikansniveauet, som sædvanligvis sættes til 5% således at hvis  $p$ -værdien, altså "probability" er under 0,05 forkastes hypotesen. Det har den virkning at der i 5 procent af tilfældene, laves en type 1 fejl. Type 2 fejl er således fejl, hvor nulhypotesen ikke forkastes, på trods af at den i virkeligheden er forkert. Denne type fejl sker oftere jo lavere signifikansniveauet er sat til, hvorfor signifikansniveauet ikke bare kan sættes til 0,00001 eller et andet meget lavet tal.

	Sand	Forkert
Nulhypotesen forkastes ikke	Korrekt beslutning Sandsynlighed: $1 - \alpha$	Type 2 fejl False positive Sandsynlighed: $\beta$
Nulhypotesen forkastes	Type 1 fejl True negative Sandsynlighed: $\alpha$	Korrekt beslutning Sandsynlighed: $1 - \beta$

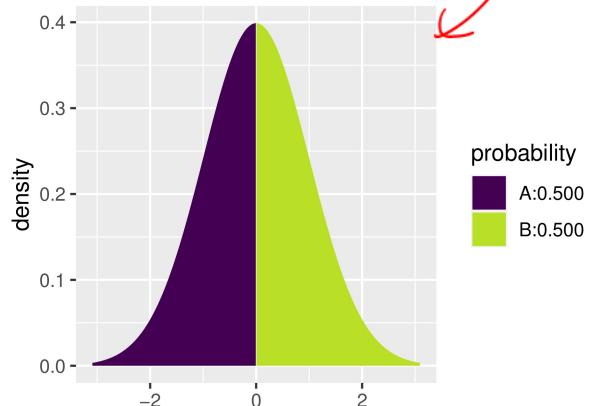
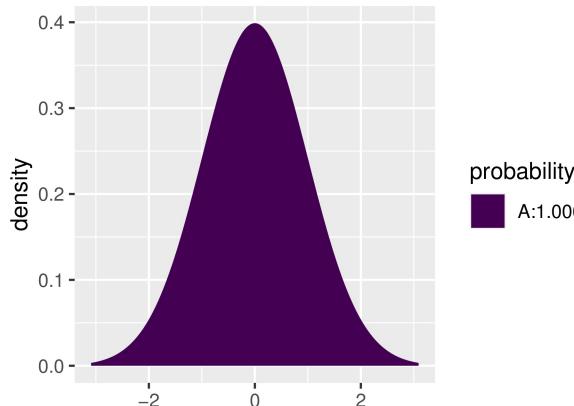
→ ikke nødvendigt  
hjemmel

Tabel / Fig X: ....

```

p1 <- qdist(mean = 0, sd = 1, p = 1, return = "plot")
p2 <- qdist(mean = 0, sd = 1, p = 0.5, return = "plot")
#p3 <- qdist(mean = 0, sd = 1, p = 0, return="plot")
grid.arrange(p1, p2, ncol=2)

```



fjernkode  
hvontor?

## Lineær regression

*DETTE AFSNIT ER SKREVET UD FRA KAPITEL 9 I BOGEN*

Lineær regression er en model, hvori det ønskes at forudsige  $y$  (en responsvariabel), ud fra  $x$  (en forklarende variabel). Ud fra denne undersgelse vil der opnås en graf, med en regressionslinje. En regressionslinje er en ret linje som minimerer den vertikale afstand mellem alle punkterne og linjen. Der vil desuden også være en ligning for denne regressionslinje, se nedenstående:

$$\tilde{E}(y) = \alpha + \beta \cdot x$$

hvor:

•  $E(y)$  er den forventede værdi af  $y$ ,

•  $\alpha$  er skæringen i  $y$ -aksen,

•  $\beta$  bestemmer hældningen af regressionslinjen baseret på  $x$ 's værdi.

*alle nordund, y=*  
*hvor er y lig med?*

Nedenstående er der plottet et eksempel på en lineær model. Desuden er der indskrevet ligningen for regressionslinjen:

```

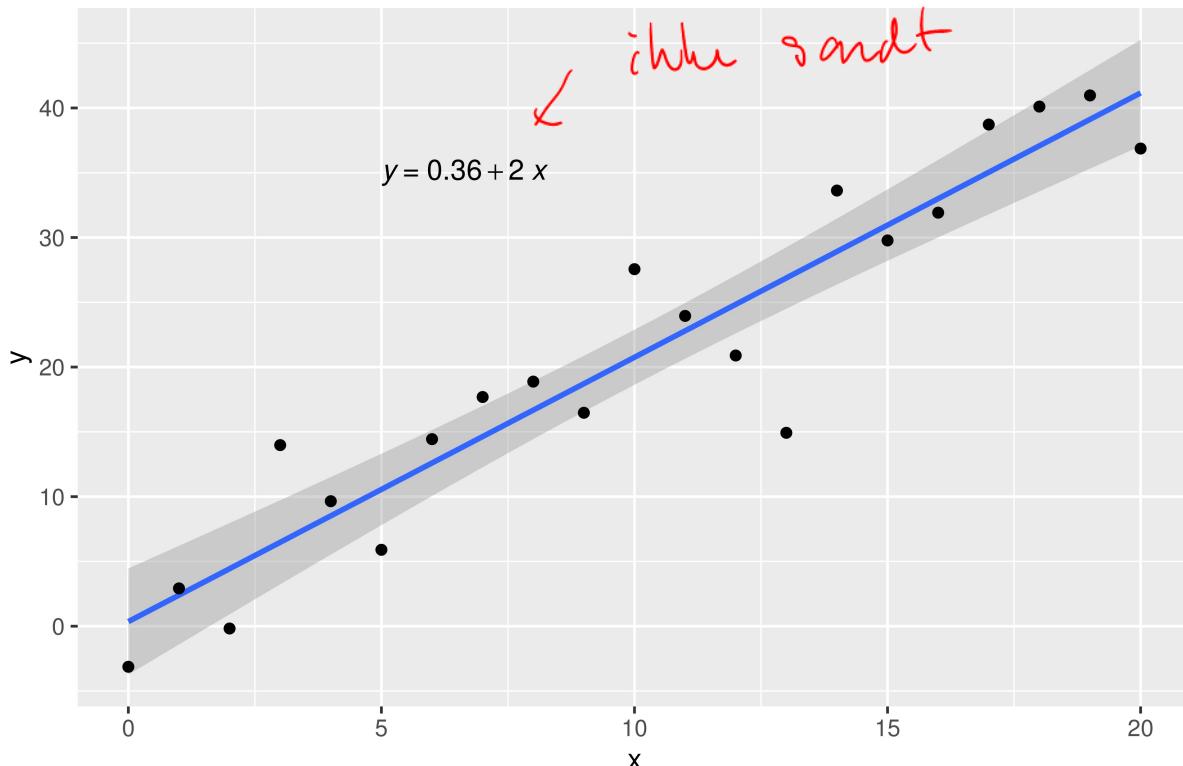
ggplot(data = df, aes(x = x, y = y)) +
  geom_smooth(method = "lm") +
  geom_point() +
  stat_regress_line_equation(label.x = 5, label.y = 35) +
  ggtitle("Eksempel på lineær regression")

```

*geom\_smooth()* using formula 'y ~ x'

fjern kode

## Eksempel på lineær regression



Hvilken  
linje?

Som det kan ses på figuren, er der få punkter som ligger tæt på regressionslinjen. Dette skyldes at denne regressionslinje er den "bedste" rette linje, altså den rette linje, hvor den vertikale afstand mellem punkterne og linjen er mindst.

For at undersøge hvad den vertikale afstand mellem punkterne og regressionslinjen, benyttes en metode kaldes "Sum of Squares Error" (også kaldet  $SSE$ ), som udregnes på følgende måde:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Hvad er  $\hat{y}_i$ ?

Jo lavere  $SSE$  er, jo bedre passer punkterne altså på regressionslinjen, den forklarer yderligere, hvor langt hvert datapunkt er fra de forudsagte datapunkter. Yderligere kan "Total Sum of Squares" (også kaldet  $TSS$ ) udregnes på følgende måde:

$$TSS = \sum (y_i - \bar{y})^2$$

Denne formel forklarer forskellen mellem hvert punkt og  $y$ 's gennemsnit. Endnu en formel man kan udregne kaldes "Sum of Squared Regression" (også kaldet  $SSR$ ) og udregnes på følgende måde:

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Hvad er  $\hat{y}_i$ ?

Denne formel forklarer forskellen mellem de forudsagte datapunkter og  $y$ 's gennemsnit. Der er en sammenhæng mellem disse tre formler, nemlig at:

$$TSS = SSE + SSR$$

Med disse værdier kan det udregnes hvor god  $x$  er til at forudsige  $y$ . Måden hvorpå dette kan udregnes, står forneden:

$$R^2 = \frac{SSR}{TSS} = \frac{TSS - SSE}{TSS}$$

Denne  $R^2$ -værdi vil altid være imellem 0 og 1, hvor 1 betyder at  $y$  kan forudsiges ud fra  $x$  alene, og hvor 0 betyder at  $y$  slet ikke kan forudsiges ud fra  $x$ .

Hvad bruges dette  
til

26

Hvordan fortolkes  
denne værdi?  
Hvad er sammenhæng  
med korrelation?

# Hvad er antagelserne vi gør os når vi laver lineær regression?

## Multipel lineær regression

En multipel lineær regression minder meget om en lineær regression, dog med den forskel at der her er flere forklarende variabler. Dette vil altså sige, at den fornævnte ligning for regressionslinjen, her vil se ud på denne måde:

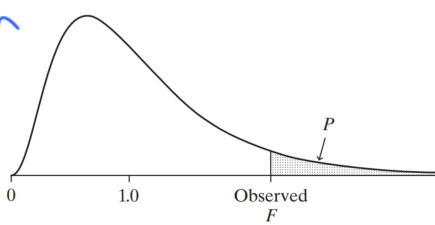
$$E(y) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n$$

Er der ekstra  
antagelser når vi  
laver multipel?

### F-test

DETTE AFSNIT ER SKREVET UD FRA KAPITEL 11 I BOGEN

En F-test er en statistisk test som har en F-fordelingen under nul hypotese. Den er mest ofte brugt når man sammenligner statistiske modeller i et datasæt til at identificere den model som passer bedst til populationen fra dataen. Det er en test som hæder statistikeren, Ronald A. Fisher, som opdagede F-fordelingen i 1922. F-fordelingen kan kun tage ikke-negative værdier og den er noget skævt til højre, ligesom en chi i andenfordelingen. Den nedenstående figur illustrerer dette. I forhold til T-test som har til formål at sammenligne middelværdierne i to populationer, så vil F-test sammenligne spredninger (eller varianser). Principperne for denne test er identiske med principperne for t-testene. Blot beregnes tesetstørrelsen på en anden måde, og der skal bruges F-fordelingen til at beregnes testsandsynligheden.



Denne figur viser F-fordelingen og P-værdien for F-test. Hvor højere F-værdier betyder at der er større evidens for at kunne forkaste  $H_0$ .

For at udregne en f-score skal man opstille en F-brøk. En F-brøk udregnes på følgende måde:

$$F = \frac{MSR}{MSE}$$

$MSR$  og  $MSE$  er varianser fra variationerne  $SSR$  og  $SSE$ . Disse varianser står for  $MSR$  betyder "Mean Square Regression" og  $MSE$  betyder "Mean Square Error". De bliver så udregnet på følgende måde:

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

Varians udregnes ved at dividere variationen med dens frihedsgrad. Dette er bestemt af to frihedsgrader som er noteret som  $df_1$  og  $df_2$

Dette er antallet af variabler i modellen.

$$df_1 = k$$

Dette er  $n$  - antal af parameter i regressionsudregningen.

$$df_2 = n - (k + 1)$$

Den første frihedsgrad,  $df_1 = k$  er tælleren ( $R^2$ ) i F-testen. Den anden frihedsgrad  $df_2 = n - (k + 1)$  er nævneren ( $1 - R^2$ ).

$$\frac{R^2}{(1 - R^2)}$$

Det vil sige hvor højere  $R^2$  er, desto større er ratioen  $\frac{R^2}{1 - R^2}$ , og desto større er F test værdien. Ved en høj F test værdi hvor større evidens er der for at forkaste  $H_0$ .