# Bivariate analyser

Analyse av sammenhengen mellom to variabler

**$H_0$ : Ingen sammenheng**
**$H_1$ : Sammenheng**

Hvis den ene variabelen er kategorisk er en slik analyse det samme som å sammenligne grupper.

**Ulike grupper vil alltid være litt forskjellige.**

**En sammenligning bør omfatte:**

- Beskrivelse av gruppene:
  Figurer, tabeller, oppsummeringstall

- Hypotesetest (statistisk analyse):
  Er de observerte forskjellene statistisk signifikante?
  $H_0$ : Gruppene er like
  $H_1$ : Gruppene er ikke like
  Effektmål og konfidensintervall

- En klinisk* vurdering av observasjonene:
  Er de observerte forskjellene klinisk interessante?
  (* Når forskningsfeltet ikke er medisin, må observasjonene sees i sammenheng med det fagfeltet du forsker i)

# Sammenligning av to grupper med uavhengige, kontinuerlige data

**$H_0$:** Gruppene er like
**$H_1$:** Gruppene er ulike

Det er da to* mulige analysemetoder:

# To-utvalgs t-test

Brukes hvis det er normalfordelte data i hver gruppe.
OBS: Hvis det er mange nok observasjoner i hver gruppe (n er stor), eller fordelingen til dataene i hver gruppe ikke er for langt fra en normalfordeling, kan man allikevel bruke t-test.

# Wilcoxon Rank Sum test (Mann-Whitney U)

Brukes hvis det er ikke-normalfordelte data i hver gruppe.
(Men lik form på fordelingen til dataene i de to gruppene; at de er skjeve til samme side.)
OBS: Det er altså *ikke* en regel at vi må bruke MW hvis vi har få observasjoner. Det er *fordelingen* til dataene som avgjør.

* I STK1000 lærer vi om forskjellige testobservatorer (z og t), og nyansene mellom dem. Hvilken vi skal bruke avhenger av hvor stor n er, om data er normalfordelte, og om vi kjenner σ. Det er viktig for eksamen i faget å forstå dette.
I en praktisk forskningshverdag vil vi så godt som aldri kjenne σ for de to gruppene på forhånd. Da må σ erstattes med sd. Hvis data er normalfordelte, vil vi alltid ende opp med en t-observator og en t-test når vi er interessert i forskjellen på forventningsverdier.
Hvis data ikke er normalfordelte, men n er stor, skulle vi i følge teorien og CLT brukt en z-observator og en z-test. Men med en stor n (hvor stor *stor* er, avhenger av hvor skjeve data er), vil t-observatoren være så lik z-observatoren at vi kan ofte tillate oss å bruke t-observatoren allikevel. Statistikkprogrammet SPSS har for eksempel bare t-test som valgalternativ hvis du ønsker å teste forskjellen på to forventningsverdier, og til og med i R må du gjøre beregningene på egen hånd hvis du absolutt skal gjøre en to-utvalgs z-test. Her er en nettside som viser deg hvordan: https://www.r-bloggers.com/two-sample-z-test/
Men: i STK1000 **er** det allikevel pensum å vite forskjellen på en z-test og en t-test, fordi dette viser bakgrunnen for testene og hvordan vi tenker når vi formulerer en signifikanstest.

# Testens idé: T-test

Fordelingene i de to gruppene er *like* (normalfordeling) bortsett fra senteret (muligens forskjøvet).

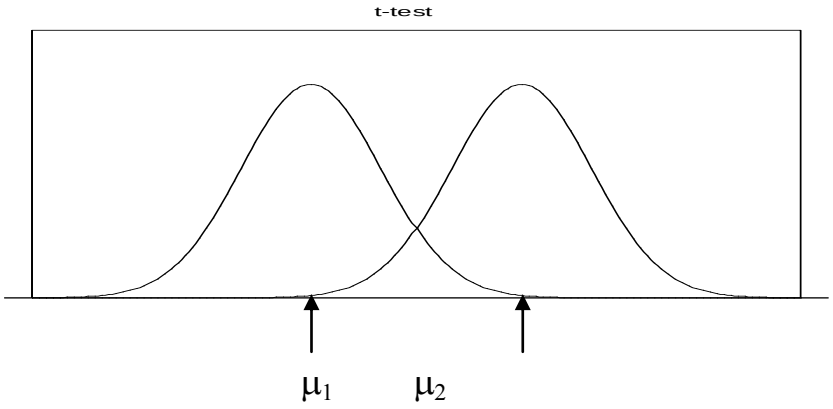
Senteret i fordelingene er gitt ved parameterne $\mu_1$ og $\mu_2$.
t-test kalles en **parametrisk test**

Undersøker om gruppene er like, altså om $\mu_1 = \mu_2$

ved å sammenligne $\bar{x}_1$ og $\bar{x}_2$

Hvis $H_0$ er sann: $\mu_1 = \mu_2 \Rightarrow \bar{x}_1 \approx \bar{x}_2$

Hvis $H_1$ er sann: $\mu_1 \neq \mu_2 \Rightarrow \bar{x}_1 \neq \bar{x}_2$




Hvis gruppene er like, forventes også ganske like gjennomsnitt i de to gruppene.

# Testens idé, WRS-test:

Den ene gruppa har systematisk høyere verdier enn den andre gruppa.

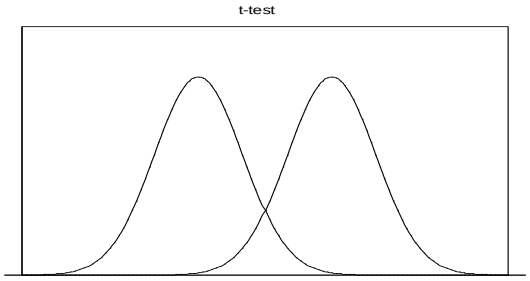
Antar ingen spesiell fordeling, og kan derfor ikke relatere forskjellen på gruppene til en spesifikk parameter (m.a.o. **ikke-parametrisk test**).

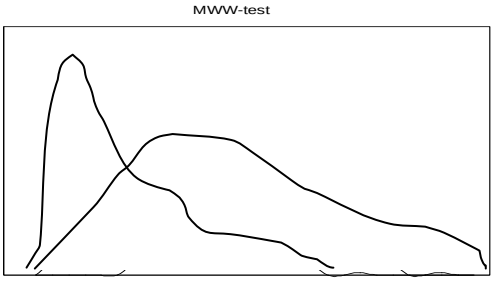
Undersøker om gruppene er like ved å sammenligne rangeringene av observasjonene.

Sorter samtlige observasjoner i stigende rekkefølge og gir dem rangeringer. Beregn "**rangsummene"** i hver gruppe.
(Alt dette gjøres heldigvis av dataprogrammet)

Hvis $H_0$ er sann: Ganske like rangsummer i hver gruppe

Hvis $H_1$ er sann: Ganske forskjellige rangsummer i hver gruppe.




Parametrisk test:
$H_0$ : gruppene er like
$\Rightarrow \bar{x}_1 \approx \bar{x}_2$

Ikke-parametrisk test:
$H_0$ : gruppene er like
$\Rightarrow$ rangsum$_1 \approx$ rangsum$_2$

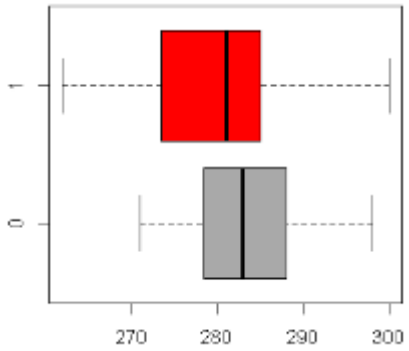
# Eksempel 1: Med Ericsson til Barcelona?

Problemstilling: Har førstegangsfødende en annen svangerskapslengde enn de som har født før?

Gruppe 1: Førstegangsfødende (ferskmor). Vi har $n_1$=19
Gruppe 0: Har født før Vi har $n_1$=16

$H_0$ : Like lange svangerskap i gruppene, $\mu_0 - \mu_1 = 0$
$H_1$ : Ikke like lange svangerskap i gruppene, $\mu_0 - \mu_1 \neq 0$

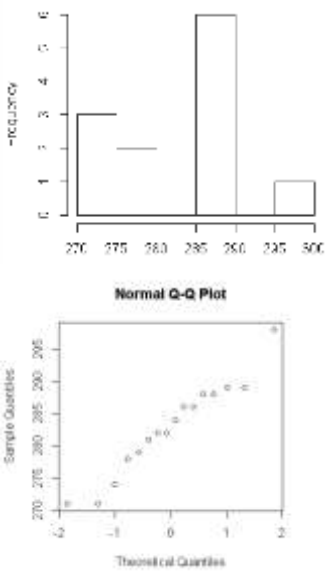
```
boxplot(dager ~ gruppe,horizontal=TRUE, col=c("dark grey", "red"))
```



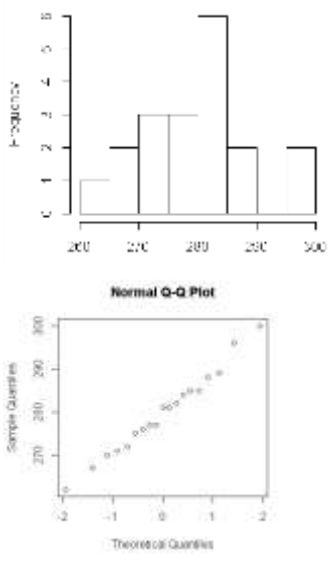

```
summary(dager[gruppe==1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  262.0   273.5   281.0   279.9   285.0   300.0
sd(dager[gruppe==1])
[1] 9.6

summary(dager[gruppe==0])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  271.0   278.8   283.0   282.9   288.0   298.0
sd(dager[gruppe==0])
[1] 7.2
```

```
hist(dager[gruppe ==0])
qqnorm(dager[V1==0])
```




```
hist(dager[gruppe ==1])
qqnorm(dager[gruppe ==1])
```




Kommentar:

Begge gruppene ser rimelig normalfordelte ut. Derfor gir det mening å formulere hypotesene i form av parametere for forventningsverdien. Dette kan vi ikke alltid vite før vi har undersøkt fordelingene i gruppene.

Betingelsene for å kjøre t-test er oppfylt $\Rightarrow$ kjør t-test!

```
t.test(dager ~ gruppe)

      Welch Two Sample t-test

data:  dager by gruppe
t = 1.0446, df = 32.611, p-value = 0.3039
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.826878  8.787404
sample estimates:
mean in group 0 mean in group 1
       282.8750        279.8947
```

## Testobservator:

$t = \frac{\bar{x}_0 - \bar{x}_1}{\sqrt{\frac{sd_0^2}{n_0} + \frac{sd_1^2}{n_1}}}$, som er t-fordelt.

## Frihetsgrader, df:

I følge boka bruker vi df = $n_1 - 1$ eller df = $n_0 - 1$ når σ-ene er ukjente og ulike. I dette eksemplet tilsvarer det enten at df=19-1=18, eller df=16-1=15.

Men hvis vi kan anta at σ-ene er like (selv om de er ukjente), kan vi bruke df = $n_1+n_0-2$. I dette eksemplet tilsvarer det at df = 19+16-2 = 33. Vi ser at $sd_1$=9.6, og $sd_0$=7.2, som er litt forskjellig, men ikke veldig forskjellig. I R-utskriften står det **`df = 32.611`**, og det betyr at R har beregnet df så det blir riktigst mulig ut i fra hvor like sd-ene er.

## Effektmål:

Effektmålet er et tall som oppsummerer forskjellen på gruppene («effekten» av å ha født før). I dette eksemplet er verdiene i hver gruppe relativt normalfordelt, og gjennomsnittene er derfor gode oppsummeringstall for hver gruppe. Dermed er forskjellen på gjennomsnittet en god oppsummering av forskjellen på gruppene. Et naturlig effektmål er derfor $\bar{x}_0 - \bar{x}_1 = 3$. Ser vi på gruppestatistikken, tolkes dette som at vi estimerer at de som har født før har gjennomsnittlig 3 dager lenger svangerskap enn førstegangsfødende.

## 95% konfidensintervall for $\mu_0 - \mu_1$, altså den sanne forskjellen i svangerskapslengde for førstegangs- og flergangsfødende:

Fra R: 95% KI er (-2.8, 8.8). Det betyr at selv om vi estimerer forskjellen til å være 3, kan vi ikke (med 95% sikkerhet) utelukke at den er så stor som 8.8 dager. Vi kan heller ikke (med 95% sikkerhet) utelukke at det faktisk er førstegangsfødende som

har lengst svangerskap, med 2.8 dager. Men vi tror at dette intervallet dekker den sanne forskjellen.

**p-verdien og selve testen:**

p-verdien =

P(å observere minst like ekstreme observasjoner som det vi har observert, gitt $H_0$) =

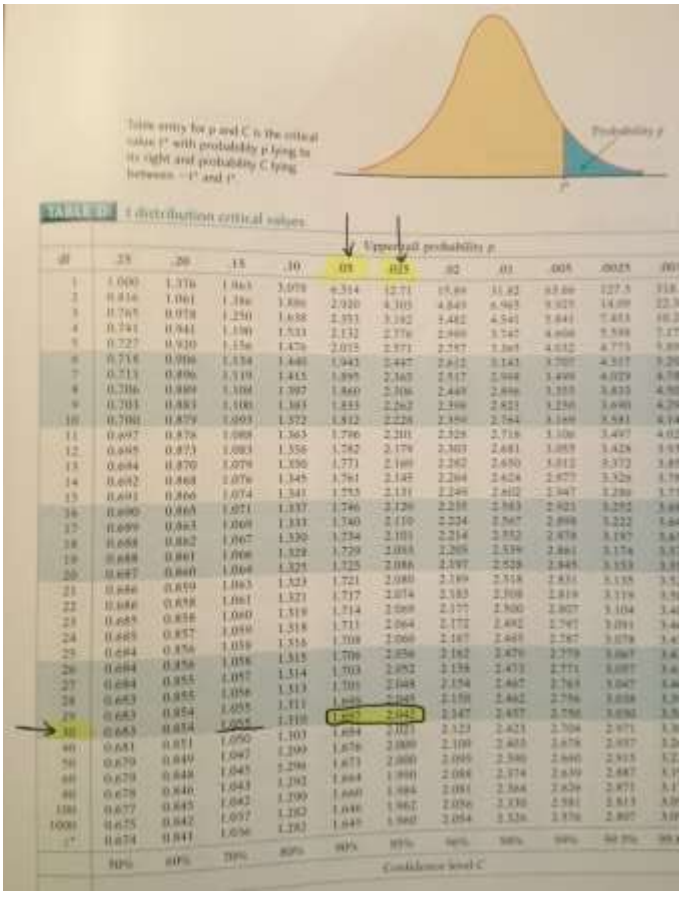
$$P\left( t = \frac{\bar{x}_0 - \bar{x}_1}{\sqrt{\frac{sd_0^2}{n_0} + \frac{sd_1^2}{n_1}}} \geq \frac{3}{\sqrt{\frac{7.2^2}{19} + \frac{9.6^2}{16}}} \mid t \sim T_{df} \right) = P( t \geq 1.03 ) = 0.30.$$

Forskjellen på 1.03 og R-utskriften (t = 1.04) skyldes avrunding av sd-ene.

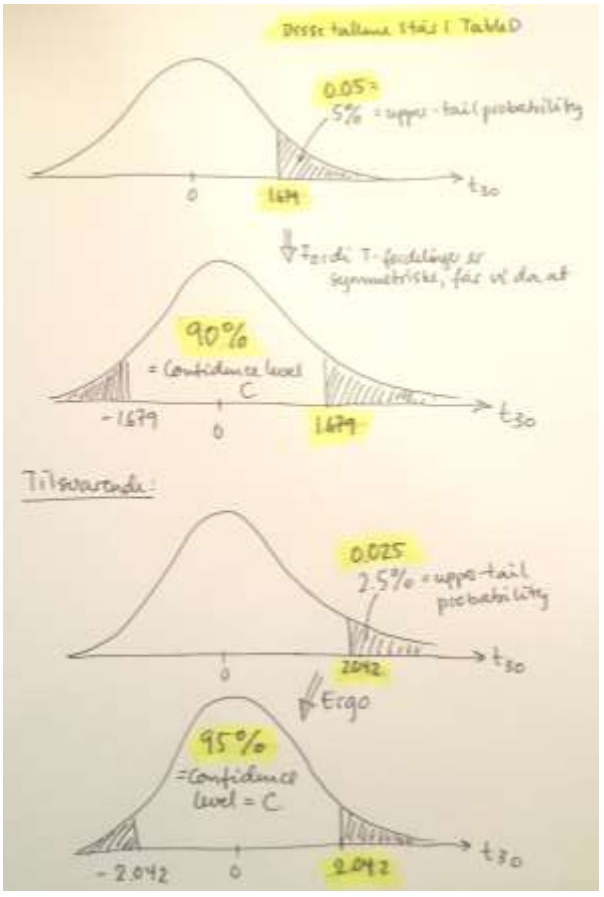
**Tabell-lesing:**

Mange har lurt på hvordan t-fordelingstabellen brukes.

Her er illustrasjoner, basert på tallene over. Tolkningene kommer på neste side.



Figur 1




Figur 2

Hvis $H_0$ er sann, forventer vi at 95% av t-verdiene skal være mellom -2.042 og 2.042, når det er 30 frihetsgrader (se de nederste to fordelingene i Figur 2, og det gule tallet til høyre inne i tabellen i Figur 1), og at en t-verdi på 1.03 tilsvarer en ensidig p-verdi på omtrent 0.15: t-verdien 1.055 er den nærmeste vi finner i tabellen (finn tallet med strek under i Figur 1), og den tilsvarer en ensidig p-verdi på 0.15 (finn tallet i øverste linje i Figur 1). Dermed blir den tosidige p-verdien det dobbelte, altså 0.30, og konfidensgraden C blir 70% (finn tallet i den nederste linja i Figur 1).

Når vi bruker R, er vi ikke bundet av begrensningene i de trykte t-fordelingstabellene, og derfor får vi et mer nøyaktig resultat i R enn det vi klarer å regne frem til for hånd og ved å bruke tabeller.

**Konklusjon:**
En så høy p-verdi (som 0.30) betyr at det er ganske sannsynlig («vanlig») å observere så store forskjeller (som 3 dager) på grupper av denne størrelsen og med disse standardavvikene. Disse observasjonene er altså ikke i konflikt med $H_0$, og vi beholder derfor $H_0$.

**Vurdering av klinisk signifikans:**
Hvis forskjellen på gruppene (i populasjonen) virkelig **er** så stor som 3 dager (nesten en halv uke), ville det vært klinisk interessant.
Her har vi to valg: Hvis støtten for dette i fysiologiske teorier er svak, slår vi oss til ro med at observasjonene våre bare var tilfeldige forskjeller. Hvis det derimot er sterk støtte for en slik forskjell i fysiologiske teorier, kan vi ta utgangspunkt i det, og bruke observasjonene i denne studien til å gjøre en utvalgsberegning (ofte kalt en styrkeberegning) for en større studie som vil gi oss sikrere svar på problemstillingen.

OBS: En større studie vil ikke garantere at vi finner en stor forskjell, men den vil gjøre konklusjonene våre sikrere, enten vi konkluderer med at det er en forskjell, eller at det ikke er en forskjell.

Tilsvarende vil et nytt konfidensintervall basert på et større utvalg bli smalere, selv om det ikke er sikkert at det nye intervallet har 3 i midten.

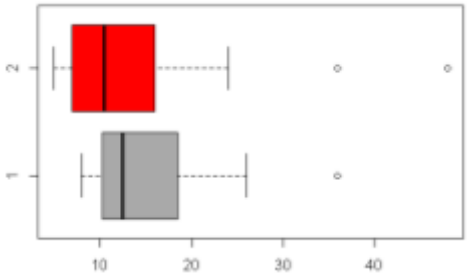
**Eksempel 2: Sammenligner to metoder for å operere bort visdomstenner.**

Registrerer rekonvalesenstid i timer (antall timer etter operasjonen før pasienten er smertefri) for de to metodene. Vi lurer på om rekonvalesentidene for den ene operasjonsmetoden er systematisk høyere eller lavere enn dem for den andre metoden. Opererer 16 med metode 1 og 18 med metode 2.

$H_0$ : Metodene er like gode.
Ingen sammenheng mellom operasjonsmetode og smertevarighet

$H_1$ : Metodene er ikke like gode.
Smh operasjonsmetode og smertevarighet

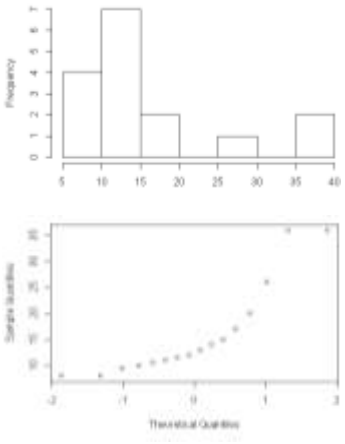
```
boxplot(rekonv~metode,horizontal=TRUE, col=c("dark grey", "red"))))
```



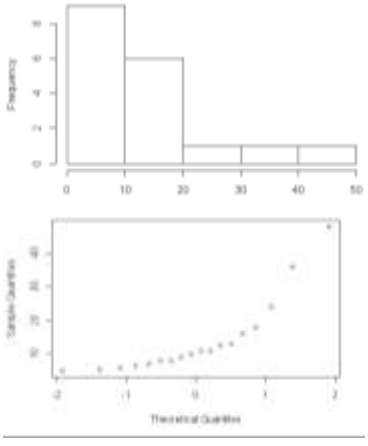

```
summary(rekonv[metode==1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   10.38   12.50   16.09   17.75   36.00
sd(rekonv[metode==1])
[1] 9.05

summary(rekonv[metode==2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00    7.25   10.50   14.14   15.25   48.00
sd(rekonv[metode==2])
[1] 11.42
```

```
hist(rekonv[metode==1])
qqnorm(rekonv[metode==1])
```



```
hist(rekonv[metode==2])
qqnorm(rekonv[metode==2])
```



Kommentar:
Begge gruppene ser rimelig skjeve og ikke normalfordelte ut. Derfor gir det ikke mening å formulere hypotesene i form av parametere for forventningsverdien. Dette kan vi ikke vite før vi har undersøkt fordelingene i gruppene.

Gruppene er små og begge fordelingene er skjeve ⇒ kjør WRS-test!

```
wilcox.test(rekonv~metode)

      Wilcoxon rank sum test with continuity correction

data:  rekonv by metode
W = 186, p-value = 0.1515
alternative hypothesis: true location shift is not equal to 0
```

**Testobservator:**
W, som i følge læreboka er rangsummen i den ene gruppa. R beregner W litt annerledes enn boka gjør, og derfor vil du få et annet tall enn R hvis du regner det ut for hånd (som vi nesten aldri gjør i praksis). Les her om hvordan R beregner W: https://www.r-bloggers.com/wilcoxon-mann-whitney-rank.../
Hvis $H_0$ er sann, har vi statistisk teori som gir oss fordelingen til W. (For eksempel vil en standardisert W være N(0,1)-fordelt når n er stor.)

**Effektmål:**
Effektmålet er et tall som oppsummerer forskjellen på gruppene («effekten» av å bruke en annen operasjonsmetode). I dette eksemplet med skjevfordelte data, er , og gjennomsnittene er ikke gode oppsummeringstall. Vi kan bruke medianene og forskjellen på medianene i stedet. Jeg velger altså effektmålet median2-median1=2. Ser vi på gruppestatistikken, tolkes dette som at vi estimerer at de som opereres med metode 1 blir smertefrie 2 timer tidligere enn de som opereres med metode 2.

**95% konfidensintervall for populasjonsmedian2-populasjonsmedian1, altså den sanne forskjellen i rekonvalesenstid i populasjonen:**
Det er ikke pensum i kurset å beregne konfidensintervaller for medianer eller differanser av medianer, for da gjelder ikke CLT. Den som vil finne ut av det på egen hånd, må lese om bootstrapping.

**p-verdien og selve testen:**
p-verdien =
P(å observere minst like ekstreme observasjoner som det vi har observert, gitt $H_0$) = $P( W \geq 186 \mid H_0 ) = 0.15$.

**Konklusjon:**
En så høy p-verdi (som 0.15) betyr at det er ganske sannsynlig («vanlig») å observere så store forskjeller (som 3 dager) på grupper av denne størrelsen og med disse standardavvikene. Disse observasjonene er altså ikke i konflikt med $H_0$, og vi beholder derfor $H_0$.
OBS: For å forkaste $H_0$ er det vanlig å velge signifikansnivået $\alpha=0.05$, altså at vi forkaster $H_0$ hvis p-verdien er mindre enn 0.05. Men andre valg av $\alpha$ er fullt mulig. Hvis det har alvorlige negative konsekvenser å gjøre type I-feil, altså forkaste $H_0$ på feil grunnlag, velges kanskje $\alpha = 0.01$ eller enda mindre.

**Vurdering av klinisk signifikans:**
Hvis forskjellen på gruppene (i populasjonen) virkelig **er** så stor som 2 timer, ville det muligens være klinisk interessant, fordi det kan bety en ekstra dose smertestillende.
Her har vi to valg: Hvis støtten for dette i fysiologiske teorier er svak, slår vi oss til ro med at observasjonene våre bare var tilfeldige forskjeller, og at $H_0$ er sann. Hvis det derimot er sterk støtte for en slik forskjell i fysiologiske teorier, kan vi ta utgangspunkt i det, og bruke observasjonene i denne studien til å gjøre en utvalgsberegning (ofte kalt en styrkeberegning) for en større studie som vil gi oss sikrere svar på problemstillingen.

OBS: En større studie vil ikke garantere at vi finner en stor forskjell, men den vil gjøre konklusjonene våre sikrere, enten vi konkluderer med at det er en forskjell, eller at det ikke er en forskjell.

Ekstrakommentar: t-testen er ganske robust, det vil si at dersom fordelingene ikke er for skjeve (altså at de ikke har for ekstreme ekstremverdier), vil p-verdien fra t-testen også kunne brukes for å sammenligne grupper. La oss sjekke her, siden data er skjevfordelte, men ikke ekstremt skjevfordelte:

```
t.test(rekonv~metode)

      Welch Two Sample t-test

data:  rekonv by metode
t = 0.55599, df = 31.621, p-value = 0.5821
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.210395  9.120117
sample estimates:
mean in group 1 mean in group 2
       16.09375        14.13889
```

Da er det ikke aktuelt å rapportere gjennomsnitt, forskjell på gjennomsnitt eller konfidensintervall, men vi ser at p-verdien på 0.58 gir samme konklusjon som WRS-testen, altså Behold $H_0$. Hvis dere velger denne løsningen i praktiske situasjoner, er dette med robusthet vs skjevhet noe som må kommenteres.

# Generelle betraktninger om de to testene

Testene er basert på ulike forutsetninger/betingelser/antakelser, og vil ofte gi ulike (og også uriktige) svar hvis forutsetningene ikke er oppfylt.

Hvis tvil: Ja takk begge deler, håp på samme konklusjon!

Tilleggsinformasjon:

- t-tester gir samtidig deskriptiv statistikk for gruppene ( $\bar{x}$ og SD), samt estimater og KI for effektmål.
- Å velge WRS-test innebærer at median og kvartiler bør brukes som deskriptive mål for gruppene.
  Dette må beregnes i tillegg til MWW-testen.