D - Hypotesetesting

Nødvendige data

Rådata, lest fra den minuttfordelte fila (skal få levert rådata på et tidspunkt)

Trenger forventningsverdier for f.eks. en splitt på t=5 minutter istedenfor 30, for å få et fornuftig antall forsøk.

```
B = textscan(fopen('statistics/minutes.csv'), '%d %d', 'Delimiter',',');
minute = B{1};
minute_bikes = B{2};
totals = zeros(6, 1);
for i = 1:6
    totals(i) = sum(minute_bikes(((i-1)*5+1):(i*5)));
end
totals
```

```
totals = 6×1
    20
    13
     9
    11
    14
    11
```

```
average = sum(totals) / 6
```

```
average = 13
```

```
average_rate = sum(totals ./ 5) / 6
```

```
average_rate = 2.6000
```

Får altså samme gjennomsnittlige estimat for raten λ som tidligere.

Estimatoren for standardavviket vil nok bli noe annerledes siden *t* er forskjellig.

Testobservator

Antall observasjoner, *n*, er 6. $n - 1 = 5$ frihetsgrader.

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{6}}}$$

```
N = 6
```

```
N = 6
```

```
X = average
```

```
X = 13
```

```
S = sqrt(average_rate * 5)
```

```
S = 3.6056
```

```
T = @(mu) (X - mu) / (S / sqrt(N))
```

```
T = function_handle with value:
    @(mu)(X-mu)/(S/sqrt(N))
```

P-verdier

$$z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

P-verdien sier noe om hvor lurt det er å forkaste nullhypotesen - den angir sannsynligheten for forkastningsfeil.

I vårt tilfelle ser det ut til at z_0 kan beregnes med samme funksjon som *T*.

Trenger normalfordeling:

```
normal = makedist("Normal");
G = @(z) cdf(normal, z);
```

Høyresidig test

Med signifikansnivå 0.05 og 5 frihetsgrader:

$t_\alpha = t_{0.05} = 2.015$

Hypotesene:

$H_0 : \mu \leq 12, H_1 : \mu > 12$

Testobservator:

```
t = T(12)
```

```
t = 0.6794
```

Signifikans:

```
kvantil = 2.015
```

```
kvantil = 2.0150
```

Da er $T < t_\alpha$ og den originale hypotesen bør *ikke* forkastes.

P-verdi

$P(Z > z_0) = 1 - P(Z < z_0)$

```
pz = G(t)
```

```
pz = 0.7515
```

```
p = 1-pz
```

```
p = 0.2485
```

Altså en betydelig sannsynlighet for forkastningsfeil. $> \alpha$ så vi bør ikke forkaste.

Venstresidig test

Med signifikansnivå 0.01 og 5 frihetsgrader:

$t_\alpha = t_{0.01} = 3.365$

Hypotesene:

$H_0 : \mu \geq 18, H_1 : \mu < 18$

Testobservator:

```
t = T(18)
```

```
t = -3.3968
```

Signifikans:

```
kvantil = 3.365
```

```
kvantil = 3.3650
```

Da er $T < -t_\alpha$ og den originale hypotesen *kan forkastes*.

P-verdi

$P(Z < z_0)$

```
p = G(t)
```

```
p = 3.4086e-04
```

Når signifikanssannsynligheten er såpass lav er det lite tenkelig at vi begår type I-feil ved å forkaste H_0.

Tosidig test

Med signifikansnivå 0.05 og 5 frihetsgrader:

$t_{\alpha/2} = t_{0.025} = 2.571$

Hypotesene:

$H_0 : \mu = 12, H_1 : \mu \neq 12$

Testobservator:

```
t = T(12)
```

```
t = 0.6794
```

Signifikans:

```
kvantil = 2.571
```

```
kvantil = 2.5710
```

Da er $|T| < t_{\alpha/2}$ og den originale hypotesen kan *ikke* forkastes.

P-verdi:

$$
\begin{aligned}
&P(Z < -|z_0| \cup Z > |z_0|)\\
&= P(Z < -|z_0|) + P(Z > |z_0|)\\
&= P(Z > -(-|z_0|)) + P(Z > |z_0|)\\
&= 2P(Z > |z_0|)\\
&= 2(1 - P(Z < |z_0|))
\end{aligned}
$$

```
pz = G(t)
```

```
pz = 0.7515
```

```
p = 2*(1-pz)
```

```
p = 0.4969
```

Fra tabellen i boka får jeg at $P(Z < 0.68) = 0.7517$ som stemmer godt med MATLABs verdi.

Det er nesten 50% sannsynlighet for at vi gjør en forkastningsfeil hvis vi forkaster denne H_0 i H_1s favør.

Det gir spesielt mye mening når testobservatøren er såpass langt innenfor kvantilet.