PSY2014 – Kvantitativ metode

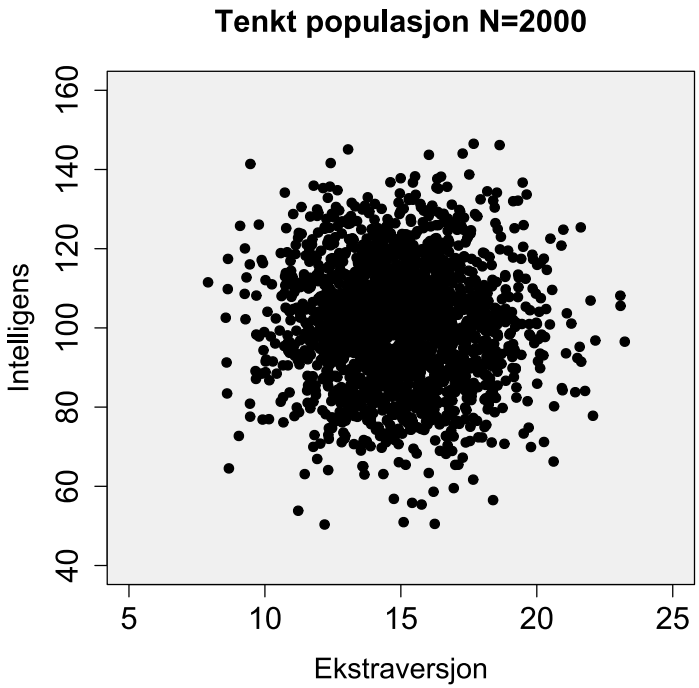
Forelesning 4: Inferens i regresjon

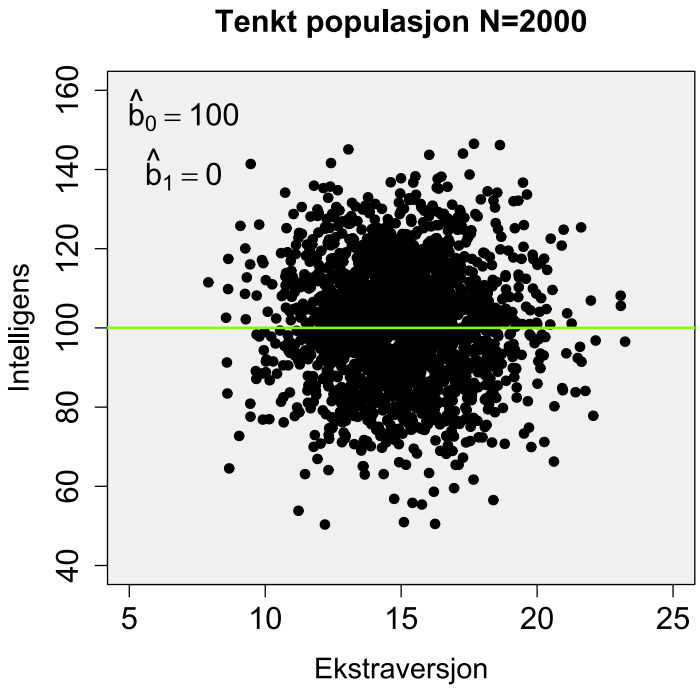
Nikolai Czajkowski

Opptak

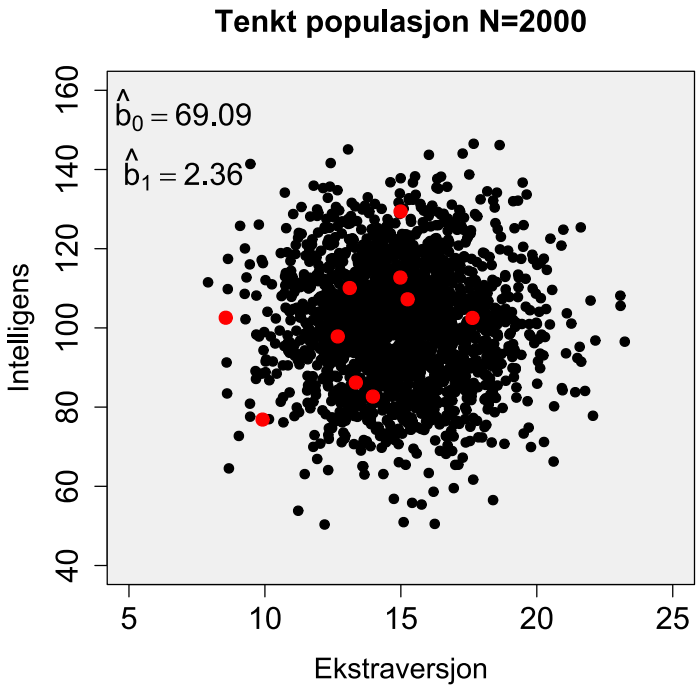
- **Det gjøres opptak av forelesningen**
- Opptaket vil bli lagret på emnesiden til PSY2014 UiO, og en lenke vil tilgjengelig for de som følger kurset.
- Opptaket skal bli slettet etter 2023.

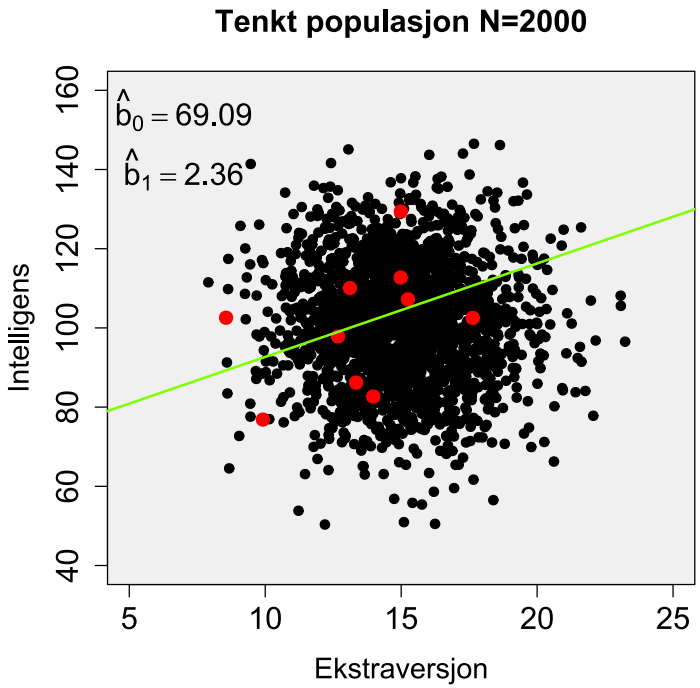
Ekstraversjon og intelligens (1)





Ekstraversjon og intelligens (2)





Hva er slutningsstatistikk?

- I første forelesningen tok vi for oss *deskriptiv statistikk*
  - verktøy for å kvantifisere viktige egenskaper til et datasett (snitt, varians etc).
  - Som navnet antyder er dette beskrivelser av utvalget, og vi trekker ikke slutninger om hva den "sanne" verdien i befolkningen er.

- Kan det hende at du i en regresjonsanalyse får $\hat{b}_1 \neq 0$ selv om variablene egentlig ikke er assosierte?
  - Kan sammenhenger komme til uttrykk ved tilfeldighet i noen utvalg?

- Idag skal vi lære om **slutningsstatistikk (inferens)**.
  - Det å trekke slutninger om populasjoner gjennom egenskaper ved utvalg.
  - F.eks. dersom gjennomsnittlig IQ i et utvalg er 110, kan du være helt sikker på at utvalget er trukket fra en populasjon der snittet er over 100?

Hva dere bør få med dere av inferens?

Repetisjon fra PSY1010:

- Forstå hva vi mener med begrepene **samplingfordeling** og **standardfeil**, og forstå hvordan disse brukes i **statistisk hypotesetesting**.

Gjennom PSY2014 burde etter hvert være i stand til å regne ut og tolke:

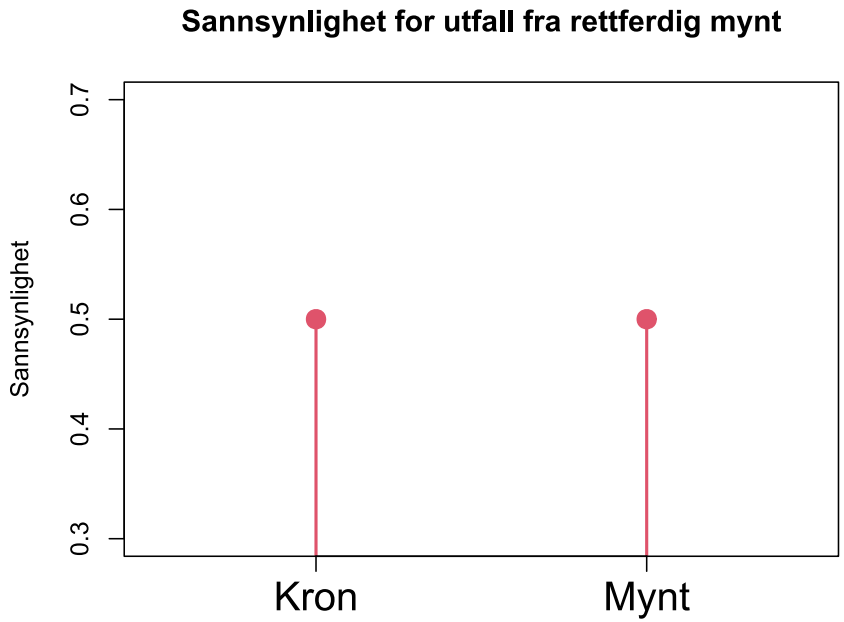
1. **Standardfeilen** til $\hat{b}_1$.
2. **t-statistikken**. som brukes til å teste om individuelle regresjonskoeffisienter et statistisk signifikant.
3. **F-statistikken**, som brukes til å teste den samlede innflytelsen til flere uavhengige variabler.
4. Et **konfidensintervall** for $\hat{b}_1$.

Agresti Kapittel 6 (repetisjon fra PSY1010)

- **Utvalg (sample):** Et sett med observasjoner som er trukket fra en større populasjon.
- **Sampling:** det å trekke ut et subset fra populasjonen.
- **En statistikk:** Enhver egenskap (verdi) regnet ut på et utvalg (altså ikke på hele populasjonen).

- Merk: Når vi trekker et tilfeldig utvalg (random sample) fra populasjonen, vil enhver verdi (statistikk) du regner ut også være en stokastisk variabel, og følgelig variere fra utvalg til utvalg

Sannsynlighet for kron/mynt



Et myntkast har to mulige utfall, og dersom mynten er rettferdig har hvert av dem like stor sannsynlighet for å inntreffe.

- P(Mynt) = P(Kron)=0.5.

Hvordan avgjøre om en mynt er rettferdig?

- **Hvordan kan du finne ut om en mynt er rettferdig?**
  - Forslag: Kaste mynten 100 ganger og telle antall kron.

- **Hvor mange kron ville du forvente å få på 100 kast dersom mynten er rettferdig?**
  - I snitt 50.

- **Dersom du teller 65 kron, kan du da konkludere noe om rettferdigheten av mynten?**
  - Du har ikke nok informasjon. Du vet ikke om 65 kron er et sannsynnlig eller usannsynlig utfall når mynten er rettferdig.

Hypotesetesting (A: 6.1)

- Innen statistikk er en *hypotese* en påstand om en populasjon, altså at en parameter ved populasjonen har en spesifikk verdi.* Typisk har vi to rivaliserende hypoteser:
  - **Nullhypotesen (H_0)** er sier at en parameter tar en verdi som tilsvarer «ingen effekt»
  - **Forskningshypotese (H_1)**

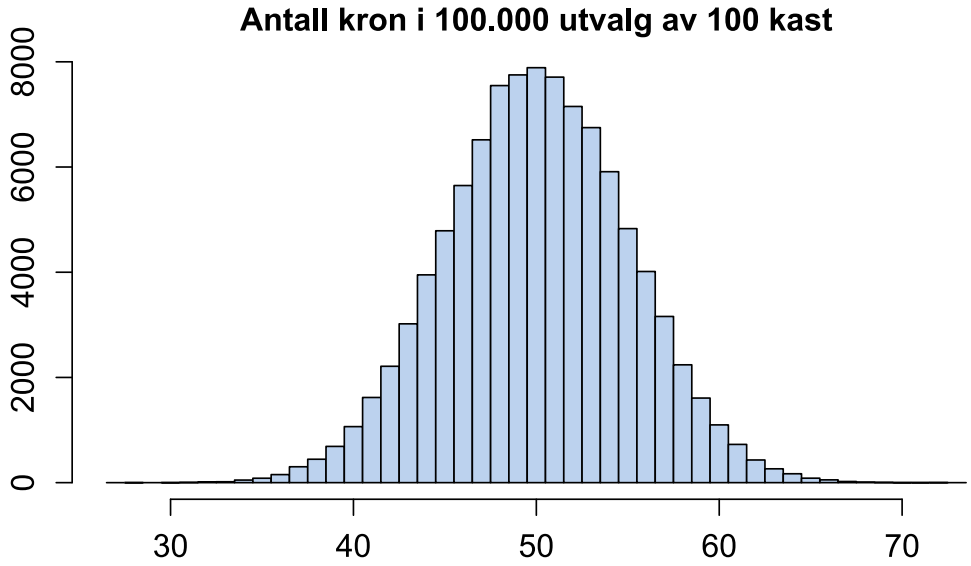
- Hva er H_0 og H_1 in vårt mynteksempel?
  - H_0: Mynten er rettferdig. Både kron og mynt er like sannsynlige.
  - H_1: Sannsynlighet for kron er ikke lik sannsynligheten for mynt.

La oss telle antall kron i 100 kast av en rettferdig mynt.

- **I det første utvalget av 100 kast får jeg 51 kron:**
  - K K K K M K K K M K K M K K K K M K M M K K M M K K M K K K M M K K K M M K K M M M M K M M M M K K M K M M K M K M K K K K M M K K M M M M K M K K K K M M M K M M M K K K M K K M M M M K K M M M K M

- **I det andre utvalget av 100 kast får jeg 46 kron:**
  - M M M M M K M M M K M K K K K K K M K K M M K K K M M K K M K M M K K M M M K M K M K M K K K M M M M K M M K K K M K M K M K M K M M K M M K M M M K K M M

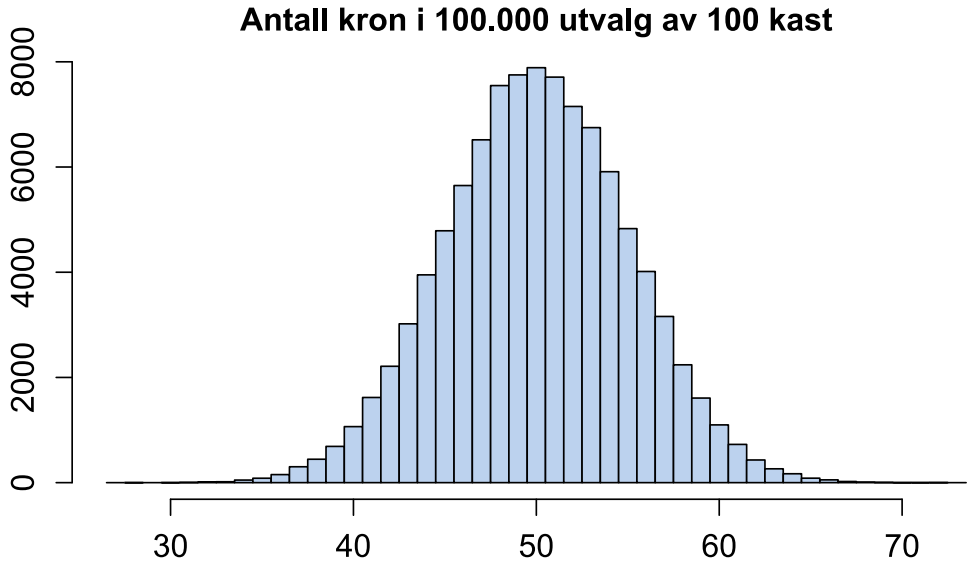
Antallet kron varierer mellom utvalgene. Dersom vi tok 100.000 utvalg og lage et histogram over antall kron i hvert, hvordan ville dette histogramet se ut?



Samplingfordelingen til en statistikk er fordelingen av de verdiene statistikken tar på tvers av ulike utvalg (av lik størrelse, og trukket fra samme populasjon).

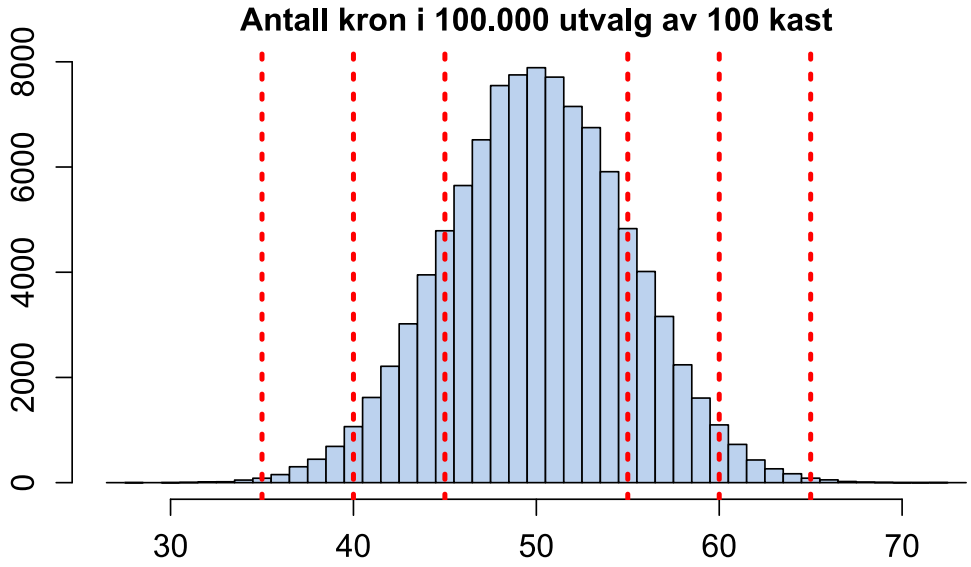
- Standardavviket til en samplingfordeling kalles **standardfeilen (SE)**.
- Fordelingen over kalles også gjerne **nullfordelingen**
  - Viser den relative sannsynligheten til ulike antall kron når nullhypotesen er sann (altså her at mynten er rettferdig).

Bruk av samplingfordelingen



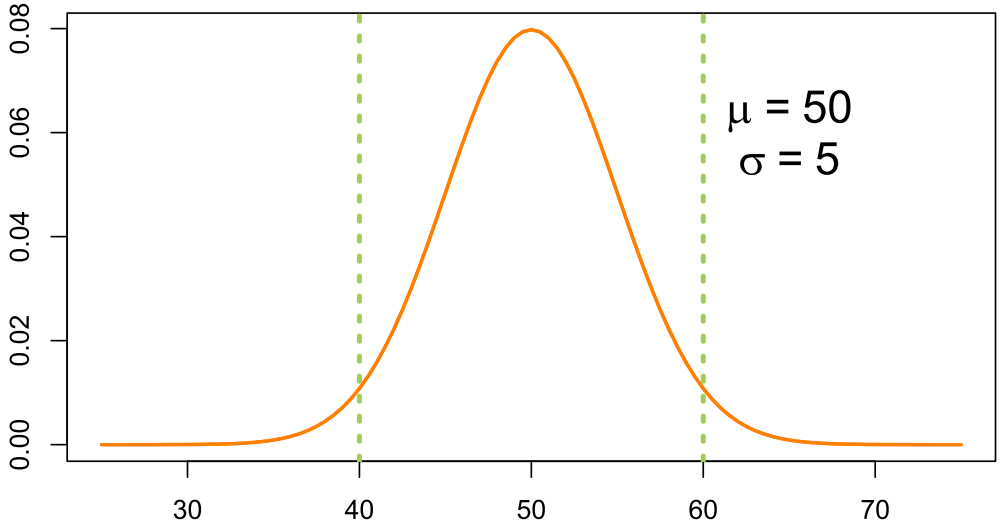
Hvor mange kron må du få for å kunne *forkaste nullhypotesen*?

Bruk av samplingfordelingen (2)



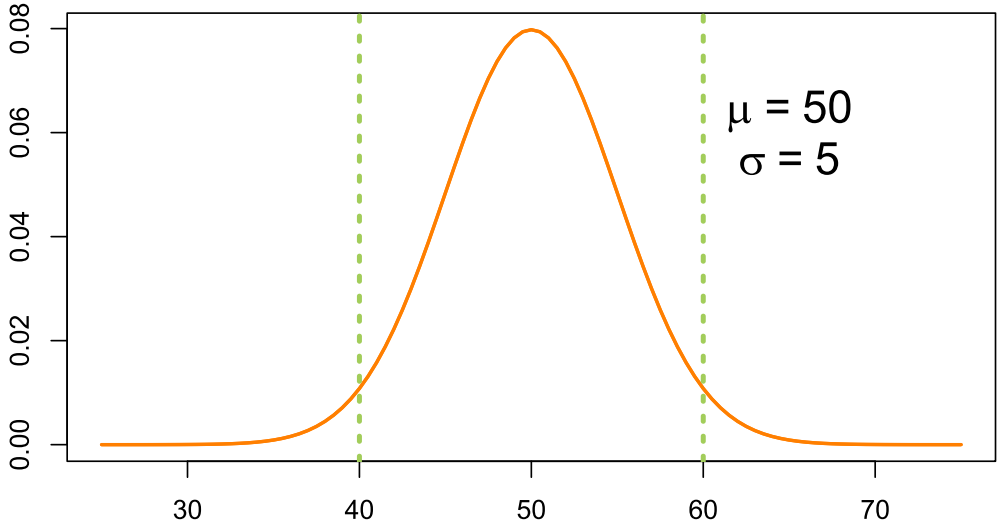
Hvor mange kron må du få for å kunne *forkaste nullhypotesen*?

Bruk av samplingfordelingen (3)



- **Sentralgrenseteoremet (Central limit theorem):** Summen av uavhengige og identisk fordelte variabler er nærmer seg en normalfordelt ettersom utvalgsstørrelsen vokser.
  - Dette er det viktigste teoremet innen statistikk, og basisen for nesten all klassisk slutningsstatistikk.
  - Moralen er: Noen ganger vet vi (gjennom sentralgrenseteoremet) hva samplingfordelingen til vår statistikk er.

Bruk av samplingfordelingen (2)



Dersom jeg fortalte deg at samplingfordelingen til antall kron under H_0 er normalfordelt med snitt 50 og standardavvik 5, N(50,5).

Hvilket intervall rundt snittet dekker 95% av de resultatene du vil se på tvers av ulike utvalg?

Hypotesetesting oppsummert

1. Formuler **nullhypotese (H_0)**, og en alternativ hypotese H_1.

2. Finn en **teststatistikk** med kjent samplingfordeling under H_0.

3. Samplingfordelingen brukes så til å finne **p-verdien**, definert som; *sannsynligheten for å en verdi på teststatistikken som er lik elle større enn den observerte (gitt at H_0 er sann).*

4. Velg **alfa** (α), en terskel for hvor liten p-verdien må være før du forkaster H_0,
   - Den verdien du trenger på teststatistikken din for å kunne forkaste H_0 på et bestemt alfa-nivå kalles den **kritiske verdien**.
   - Typisk 0.05 or 0.01.

Litt om frihetsgrader

Frihetsgrader (Degrees of freedom [df])

- Frihetsgrader (df) til en statistikk antall verdier i statistikken "får lov til" å variere.
- F.eks. Dersom du vet at summen av tre tall er 6, og at de første to har verdiene 2 og 3, hvilken Verdi må det tredje ha?
  - 1
- Dersom snittet av tre tall er 2. Det første tallet er 3 og det andre er 3 hva vet du om det tredje? [(3+3+?)/3=2]
  - 0
- Antall frihetsgrader vil være gitt som; antall observerte verdier – antall estimerte parameter i en statistikk.

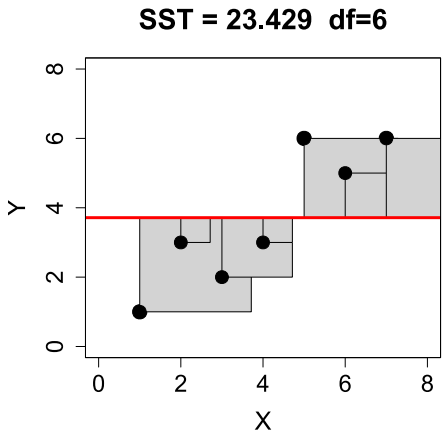
$$Var(X) = s_X^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

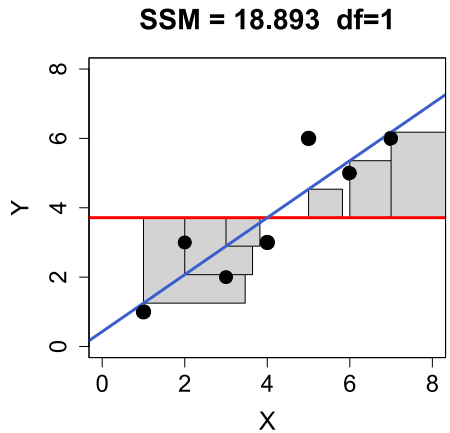
- I varians er det bare (n-1) observasjoner som brukes til utregningen, siden den siste "låses" av den estimerte verdien for gjennomsnittet.

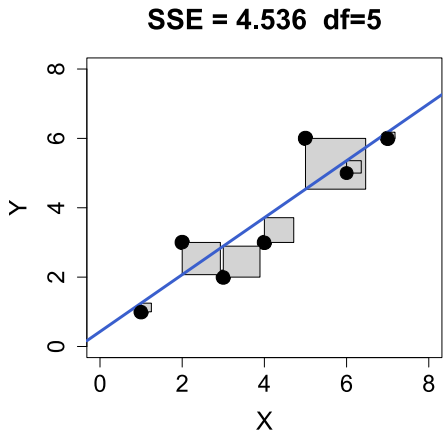
Kvadratsummer med frihetsgrader (p: antall uavhengige variabler)

$$TSS = SSM + SSE$$

$$df_{TSS} = df_{SSM} + df_{SSE}$$







$$TSS = \sum(Y_i - \bar{Y})^2$$

$$df_{TSS} = n - 1$$

$$SSM = \sum(\hat{Y}_i - \bar{Y})^2$$

$$df_{SSM} = p$$

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$

$$df_{SSE} = n - p - 1$$

Hva er utvalgstørrelsen?

```
> anova(lm(Y~X))
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X           1  100.00  100.00  49.019 3.192404e-10 ***
Residuals  98  200.00    2.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Hva er TSS? p? n?

$100 + 200$, 1

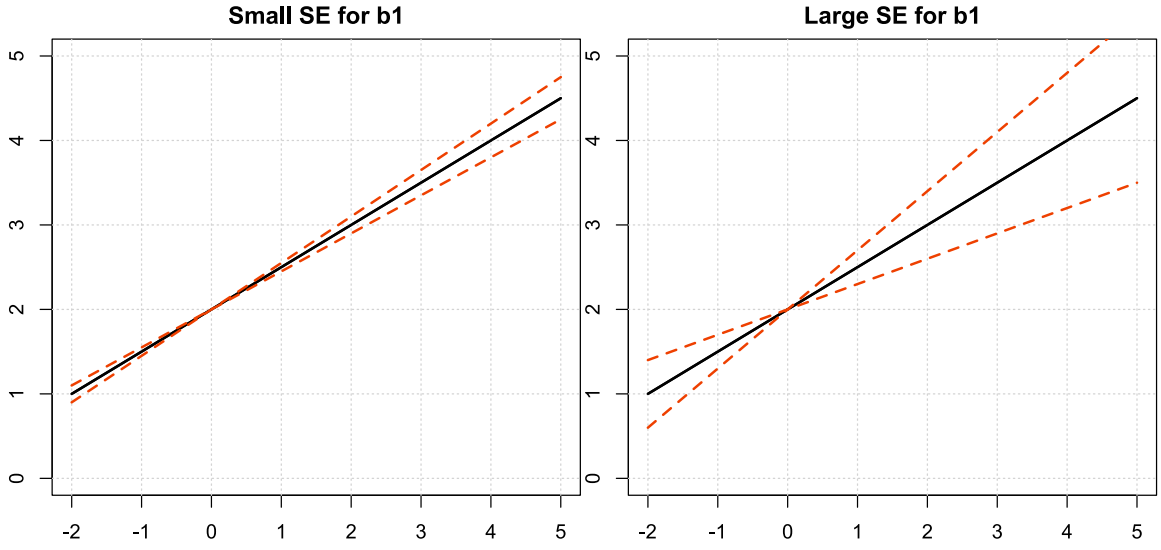
$$df_{TSS} = df_{SSM} + df_{SSE}$$

$$(n-1) = (p) + (n-p-1) = 1 + 98 = 99$$

$$n = 99 + 1 = 100$$

Standardfeilen til b_1

Standardfeilen til b_1



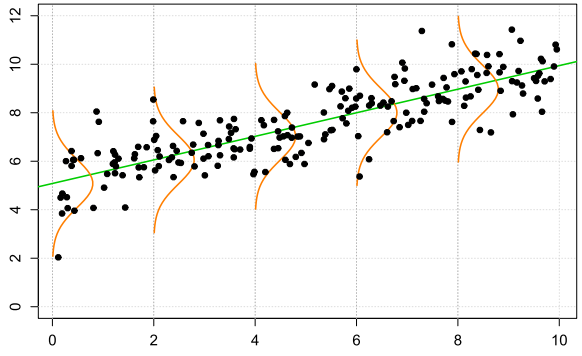
- `(SE(\hat{b}_1))` er et estimat på variabiliteten (standardavviket) i stigningstallet du ville ha sett på tvers av ulike utvalg tatt fra den samme populasjonen.
    - Du kan bruke ditt utvalg til å estimere denne variabiliteten

Residual standard error (A:9.3)

$$Y_i = b_0 + b_1 \cdot X_i + \epsilon_i \, , \epsilon_i \sim N(0, \sigma^2)$$

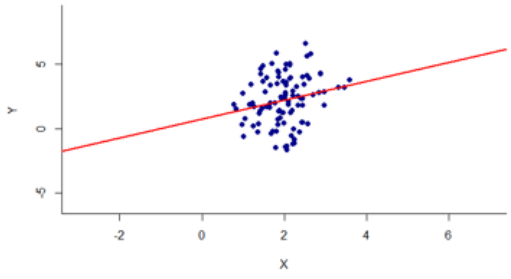
Residual standard error (det Agresti kaller *conditional standard deviation* p.269) er et estimat av σ i uttrykket.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14941    0.07527   1.985   0.0485 *
X            0.46826    0.07504   6.241  2.6e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.054 on 198 degrees of freedom
Multiple R-squared:  0.1644,    Adjusted R-squared:  0.1601
F-statistic: 38.94 on 1 and 198 DF,  p-value: 2.597e-09
```



$$s = \sqrt{\frac{SSE}{df_{SSE}}}$$

Standard error of the slope (A:9.5)

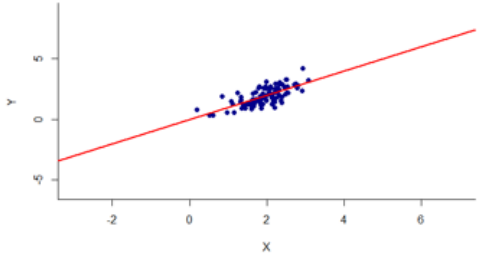


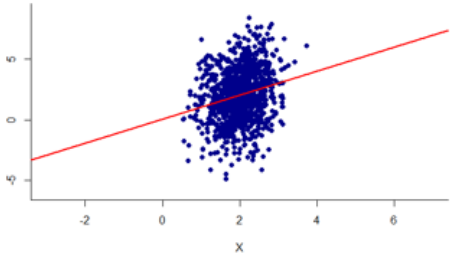
$$SE(\hat{b}_1) = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}}$$

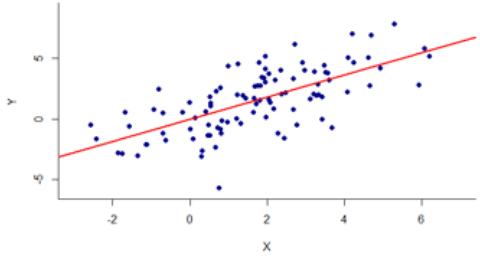
Liten *s, altså lite variasjon rundt regresjonslinjen resulterer i liten SE.*

Mange observasjonen resulterer i mindre SE

Stor spredning i den uavhengige variabelen gir mindre SE.

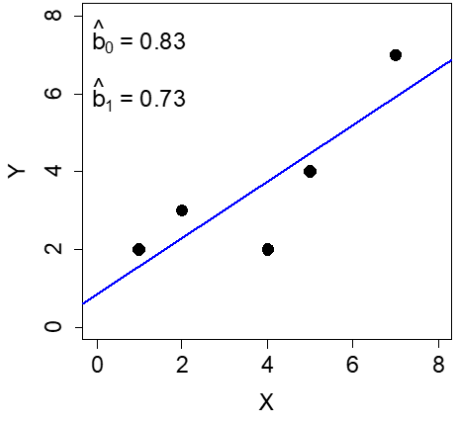






Hvordan beregne SE(B_1)

X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$	$X - \bar{X}$	$(X - \bar{X})^2$
1,00	2,00	1,56	0,44	0,19	-2,80	7,84
2,00	3,00	2,29	0,71	0,50	-1,80	3,24
4,00	2,00	3,75	-1,75	3,06	0,20	0,04
5,00	4,00	4,48	-0,48	0,23	1,20	1,44
7,00	7,00	5,94	1,06	1,12	3,20	10,24
				5.114		**22,80**



$$s = \sqrt{\frac{SSE}{df_{SSE}}} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - p - 1}} = \sqrt{\frac{5.114}{3}} = 1.306$$

$$SE(\hat{b}_1) = \frac{s}{\sqrt{\sum(X - \bar{X})^2}} = \frac{1.305}{\sqrt{22.80}} = 0.275$$

```
lm(formula = Y ~ x)

Residuals:
      1       2       3       4       5
 0.4386  0.7105 -1.7456 -0.4737  1.0702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8333     1.1919   0.699   0.5348
x             0.7281     0.2734   2.663   0.0762 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.306 on 3 degrees of freedom
Multiple R-squared:  0.7027,    Adjusted R-squared:  0.60
F-statistic:  7.09 on 1 and 3 DF,  p-value: 0.07616
```

t-statistikken: Testing av individuelle regresjonskoeffisienter

100.000 TRUKKET OG FØLGENDE UTREGNET I HVER

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2}$$

```
First 10: 0.46 0.38 -0.50 0.16 -0.73 -0.07 -0.33 -0.20 0.31 -0.24...
```

$$\hat{SE}(\hat{b}_1) = \frac{s}{\sqrt{\sum (X_i - X)^2}}$$

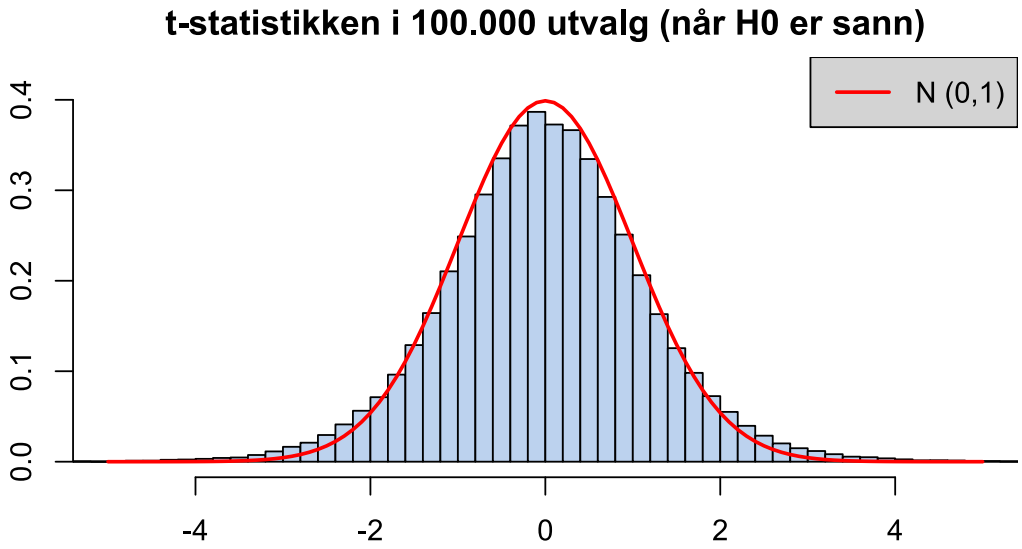
```
First 10: 0.23 0.56 0.16 0.24 0.45 0.30 0.45 0.32 0.27 0.30..
```

Nå introduserer jeg en ny *test-statistikk* som jeg kaller "t":

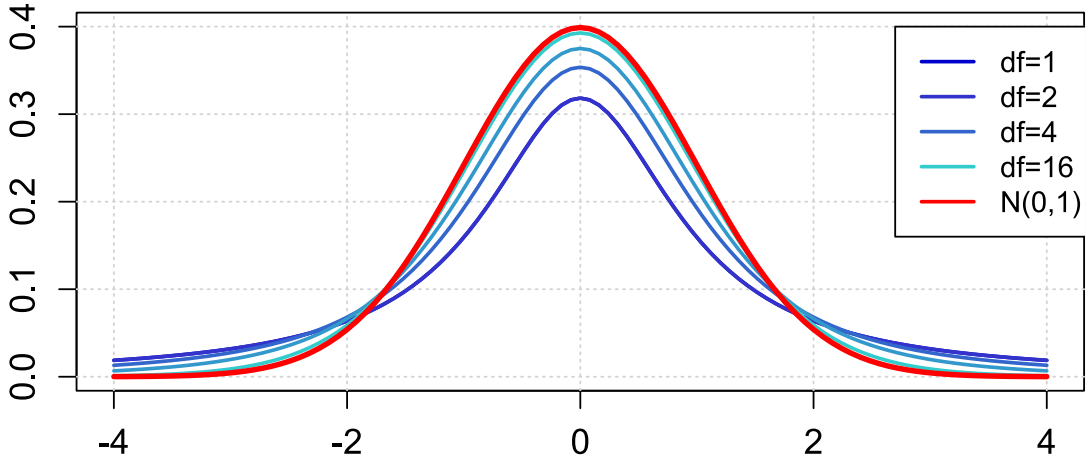
$$t = \frac{\hat{b}_1}{\hat{SE}(\hat{b}_1)}$$

```
First 10: 2.00 0.69 -3.12 0.65 -1.60 -0.25 -0.75 -0.62 1.16 -0.80...
```

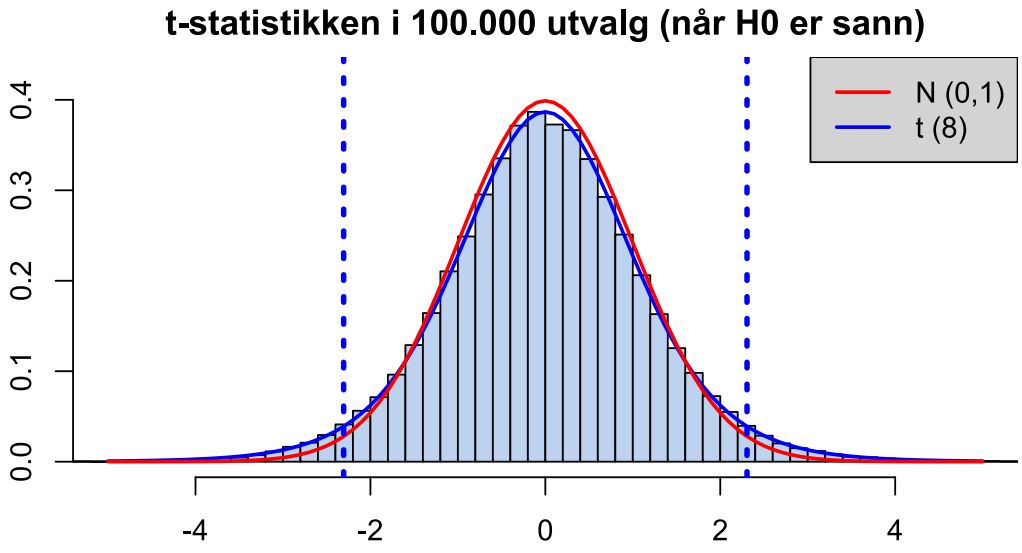
Sampling distribution for $\hat{B}_1$



Her er et histogram over alle 100.000 verdiene, samt en standard normalfordeling i rødt. Passer standard normalfordelingen heeelt til de observerte verdiene?



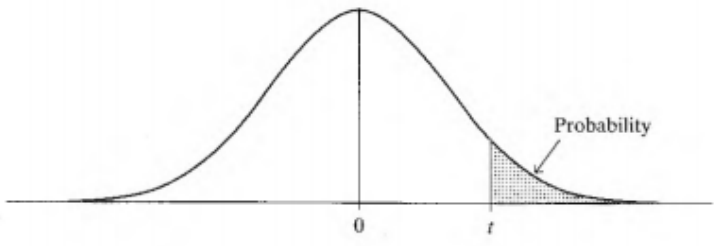
- Formen på en t-fordeling bestemmes av ett parameter, som kalles frihetsgrader (df).
- De blå kurvene over er t-fordelinger med 1 (mørkeblå) til 16 (lyseblå) frihetsgrader.
- Legg merke til at når antall frihetsgrader øker, nærmer en t-fordeling seg en standard normalfordeling, og de er praktisk talt identiske når df > 30.



- Når nullhypotesen er sann (b_1=0), så følger t-statistikken i ulike tilfeldige utvalg på størrelse 10 en t-fordeling med 8 frihetsgrader.
- Jeg kan da bruke en tabell for å finne de *kritiske verdiene* til t-statistikken.
- Under H0 er t-statistikken t-fordelt med (n-p-1) frihetsgrader.

Bruk av t-tabellen for å finne kritiske verdier

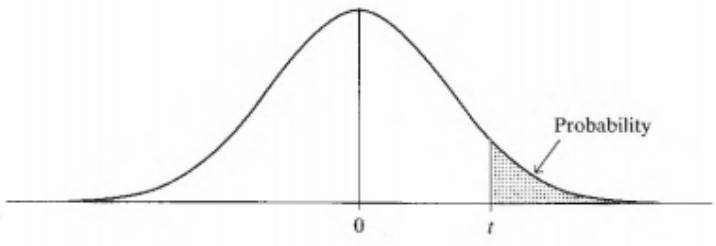
TABLE B: *t* Distribution Critical Values



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552

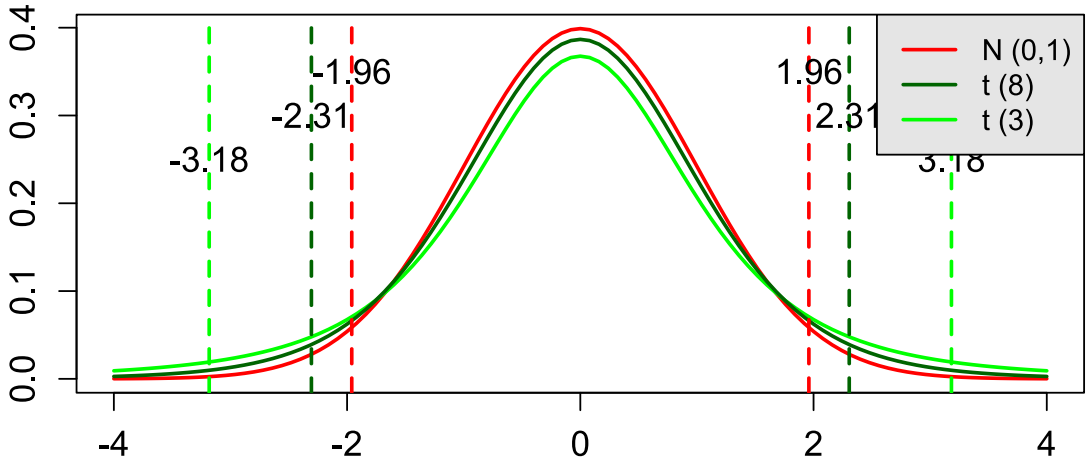
Bruk av t-tabellen for å finne kritiske verdier (2)

TABLE B: *t* Distribution Critical Values



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552

Mer om kritiske verdier

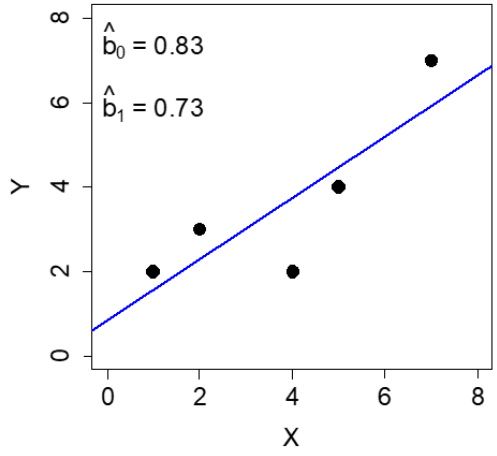


While in a standard normal distribution, 95% of the area is between +-1.96, on a t(8) distribution, 95% is between +-2.31 and for a t(3) distribution, +- 3.18.

Moral: with fewer degrees of freedom there is more uncertainty.

```
lm(formula = Y ~ X)

Residuals:
      1       2       3       4       5
 0.4386  0.7105 -1.7456 -0.4737  1.0702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8333     1.1919   0.699   0.5348
X             0.7281     0.2734   2.663   0.0762 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.306 on 3 degrees of freedom
Multiple R-squared:  0.7027,    Adjusted R-squared:  0.6056
F-statistic:  7.09 on 1 and 3 DF,  p-value: 0.07616
```

```
> anova(M1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 12.086 12.0860  7.0899 0.07616 .
Residuals  3  5.114  1.7047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



$$t = \frac{\hat{b}_1}{\hat{SE}(\hat{b}_1)} = \frac{0.728}{0.273} = 2.663$$

Under H_0 fordelt (t_{n-p-1})

	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173

Hypotesene

$H_0 : b_j = 0$
$H_1 : b_j \neq 0$

Test statistikken

$$t = \frac{\hat{b}_j}{\hat{SE}(\hat{b}_j)}$$

Samplingfordelingen under H_0:

Under H_0 følger t-statistikken en $t(n - p - 1)$ fordeling.

Hvilken uavhengig variabel har lavest p-verdi?

```
Call:
lm(formula = Happiness ~ Age + Gender + Ses)

Residuals:
     Min       1Q   Median       3Q      Max
-2.67936 -0.72015  0.04902  0.67117  2.82479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14941    1.3250   XXXXX   XXXXX
Age          1.00000    4.6920   XXXXX   XXXXX
Gender       1.00000    0.0210   XXXXX   XXXXX
Ses          1.00000    0.4160   XXXXX   XXXXX
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.054 on 198 degrees of freedom
Multiple R-squared:  0.2814,    Adjusted R-squared:  0.2713
F-statistic: 38.94 on 3 and 196 DF,  p-value: 2.597e-09
```

Hvilken uavhengig variabel har lavest p-verdi?

```
Call:
lm(formula = Happiness ~ Age + Gender + Ses)

Residuals:
     Min       1Q   Median       3Q      Max
-2.67936 -0.72015  0.04902  0.67117  2.82479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.3290    11.2530  2.7840   XXXXX
Age          -0.2130     0.0240 -8.8750   XXXXX
Gender        3.4810     4.6920  0.2420   XXXXX
Ses           0.4180     0.1120  3.7320   XXXXX
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.054 on 198 degrees of freedom
Multiple R-squared:  0.2814,    Adjusted R-squared:  0.2713
F-statistic: 38.94 on 3 and 196 DF,  p-value: 2.597e-09
```

Konfidensintervaller

Konfidensintervaller

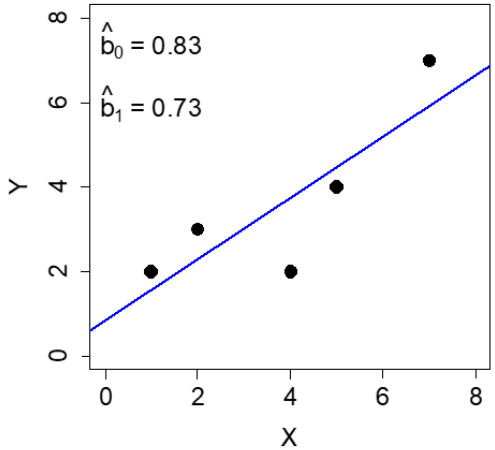
- Konfidensintervaller (CIs) er en annen måte å uttrykke usikkerhet ved estimatene.
- Enhver verdi innad i et 95% konfidensintervall utgjør en mulig hypotese om parameterets Verdi som du ikke kan forkaste på et 0.05 nivå. Enhver verdi utenfor intervallet kan forkastes på et 0.05 nivå.

$$\hat{b}_1 = 10(8, 12)$$

- Konfidensintervall gir mer informasjon enn en p-Verdi.
  - Om 0 er utenfor intervallet betyr det at vi kan forkaste hypotesen om at parameteret har verdien 0.
  - I tillegg kan vi se hvilke andre verdier vi kan forkaste.

Utregning av konfidensintervaller

```
> confint(M1)
                2.5 %    97.5 %
(Intercept) -2.9597439 4.626411
X           -0.1421214 1.598262
```



	Level of Significance for One-Tailed					
	0.25	0.20	0.15	0.10	0.05	0.025
	Level of Significance for Two-Tailed					
df	0.50	0.40	0.30	0.20	0.10	0.05
1	1.000	1.376	1.963	3.078	6.314	12.706
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776

2. For hånd

$$95\%\ Konfidensintervall:\ \hat{b} \pm t_{df,alpha} \times SE(\hat{b})$$

Dette gir den kritiske verdien ved en t-fordeling med df-frihetsgrader og et gitt alfa nivå.

Øvre grense: $\hat{b} + t_{n-p-1} \times SE(\hat{b}) = 0.728 + 3.18 \times 0.273 = 1.598$

Nedre grense: $\hat{b} - t_{n-p-1} \times SE(\hat{b}) = 0.728 - 3.18 \times 0.273 = -0.142$

F-testen: Testing av flere regresjonskoeffsienter samlet

Oppsummering av F-testen

Hypotesene

$H_0 : b_1 = b_2 = b_3 = \cdots = b_p = 0$
H_1 : minst en $b_j \neq 0$

Test statistikken:

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

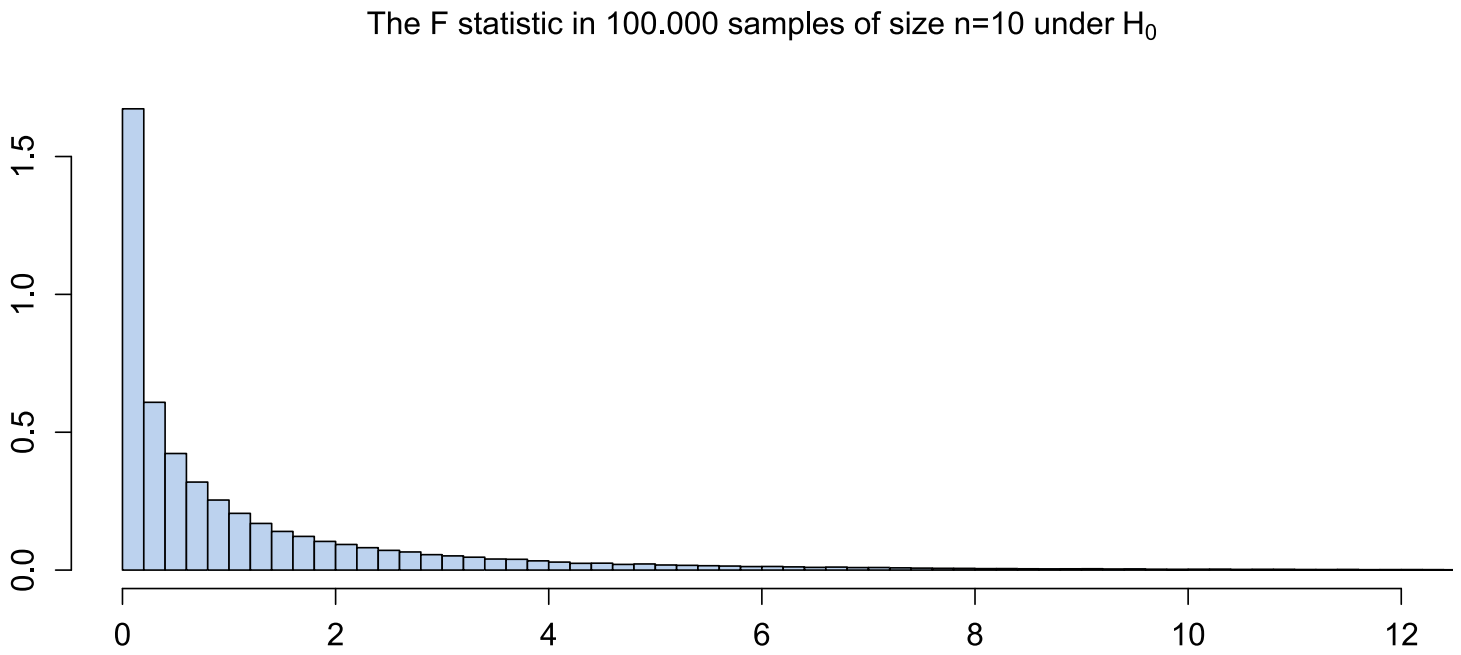
$$MSM = \frac{MSM}{df_{MSM}} = \frac{\sum(Y_i - \bar{Y})^2}{p}$$

$$MSE = \frac{SSE}{df_{SSE}} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p - 1}$$

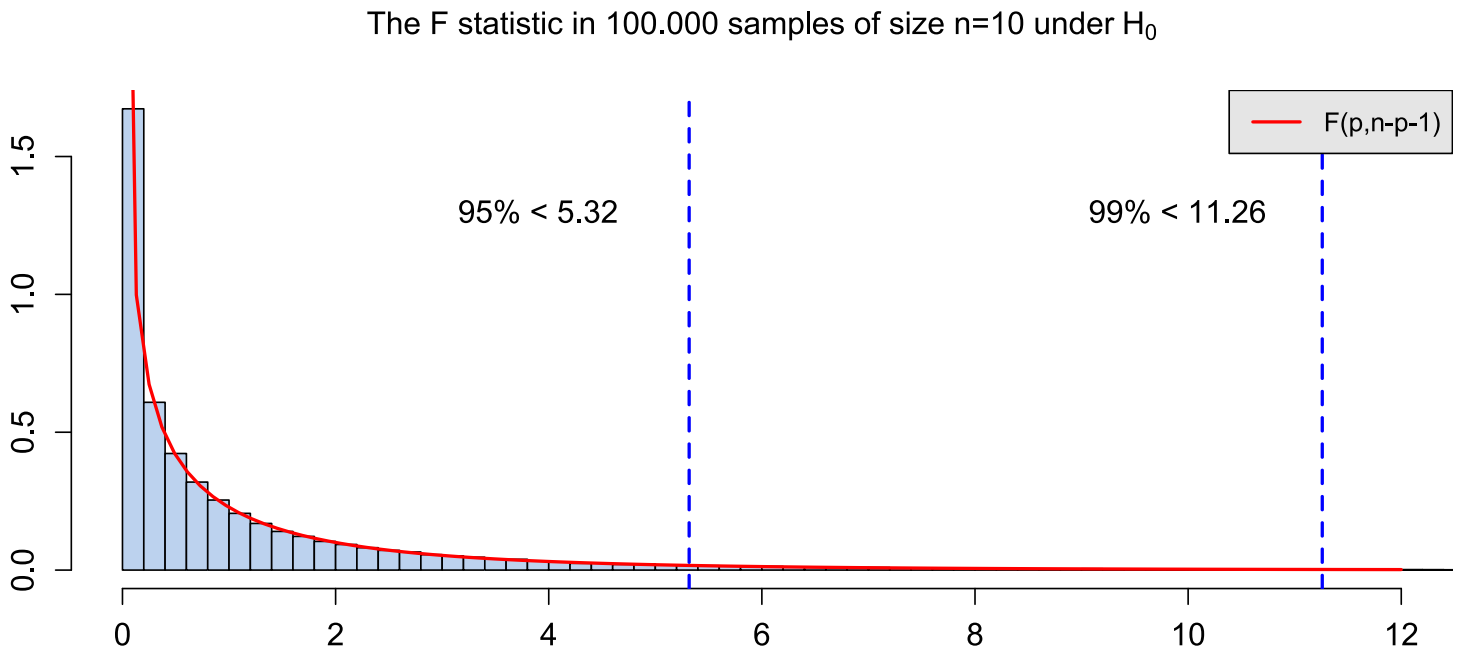
Samplingfordelingen under H0:

Under H_0 følger F-statistic en $F(p, n - p - 1)$ fordeling.

Samplingfordelingen til F-statistikken

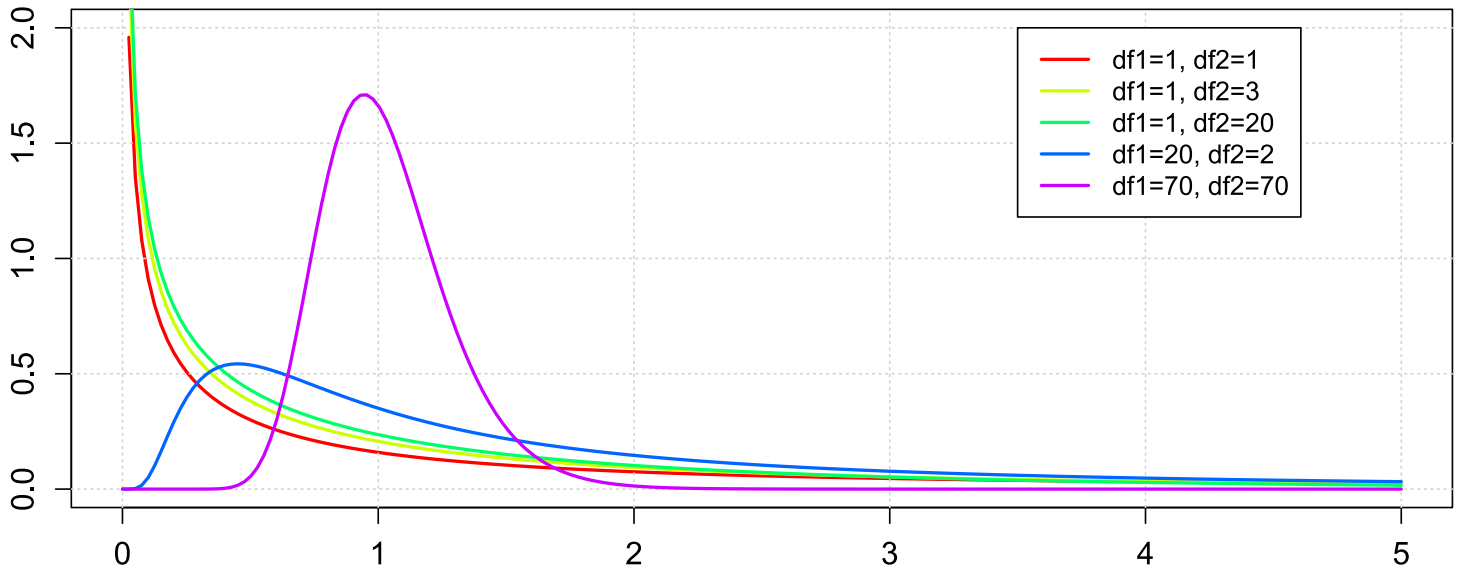


Sampling distribution of F



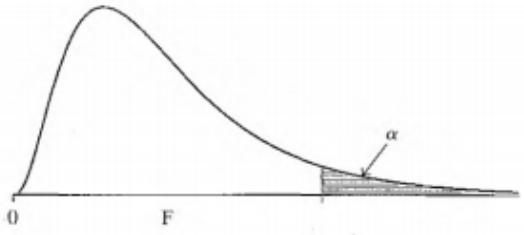
For the F-distribution, only a one tailed test is reasonable.

F-FORDELINGEN



F-TABELLEN (1)

TABLE D: *F* Distribution

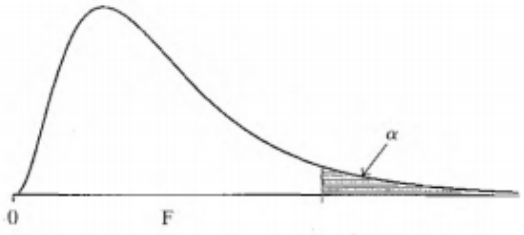


$\alpha = .05$

	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07

F-TABELLEN (2)

TABLE D: *F* Distribution



$\alpha = .05$

	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07

F-UTSKRIFT

```
lm(formula = Y ~ X)

Residuals:
      1       2       3       4       5
 0.4386  0.7105 -1.7456 -0.4737  1.0702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8333     1.1919   0.699   0.5348
X             0.7281     0.2734   2.663   0.0762 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.306 on 3 degrees of freedom
Multiple R-squared:  0.7027,    Adjusted R-squared:  0.6046
F-statistic:  7.09 on 1 and 3 DF,  p-value: 0.07616
```

```
> anova(M1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 12.086 12.0860  7.0899 0.07616 .
Residuals  3  5.114  1.7047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Neste uke

Avsluttende om regresjon:

- Kollinearitet
- Kategoriske prediktorer
- Interaksjon
- Mediering
- Modellvalg