Introduksjon til statistikk og data analyse

Kap 1

Statistikk og sannsynlighet handler om tall (data) og modeller for variasjon/ struktur i tall (data).

Hollywood-filmer fra ett år

- 135 filmer
- Samla budsjett: $ 7 166 500 000
- Samla billettsalg: $ 20 199 000 000

Hvordan oppsummere tall?

- Steg 1: **gjennomsnittet (mean)**.

Angir senteret som observasjonene (tallene) er spredt rundt. 10 data:

$$\mathrm{m}= \frac{1}{10}\sum_{i=1}^{10} x(i)$$

Gjennomsnitt, Hollywood-filmer

Sjanger	Antall	Gjennomsnitt	
		Budsjett (millioner $)	Billettsalg (millioner $)
Action	32	89,63	249,05
Animation	12	114,92	286,58
Comedy	27	38,50	107,53
Drama	21	25,33	44,63
Fantasy	2	62,60	664,72
Horror	17	25,79	73,23
Romance	11	38,40	135,95
Thriller	13	30,79	86,91
Alle	135	53,48	150,74

- **Median = den midterste observasjonen**
- En annen måte å angi senteret som observasjonene er spredt rundt

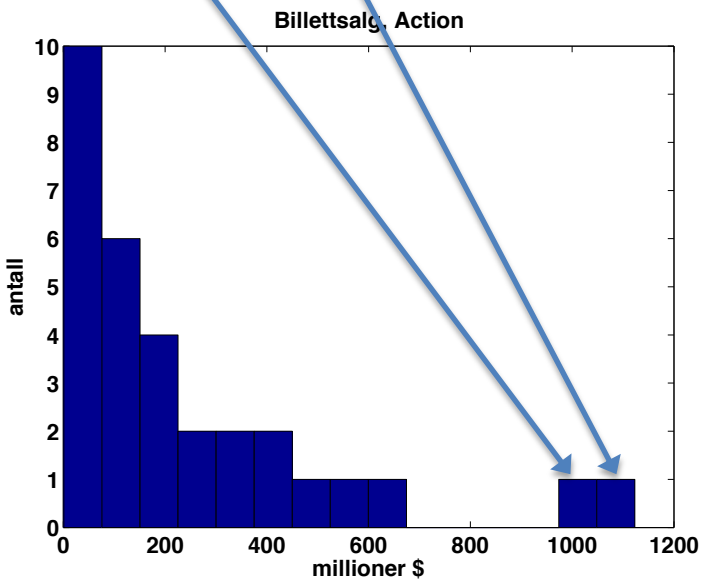
F.eks. en arbeidsplass med 10 ansatte

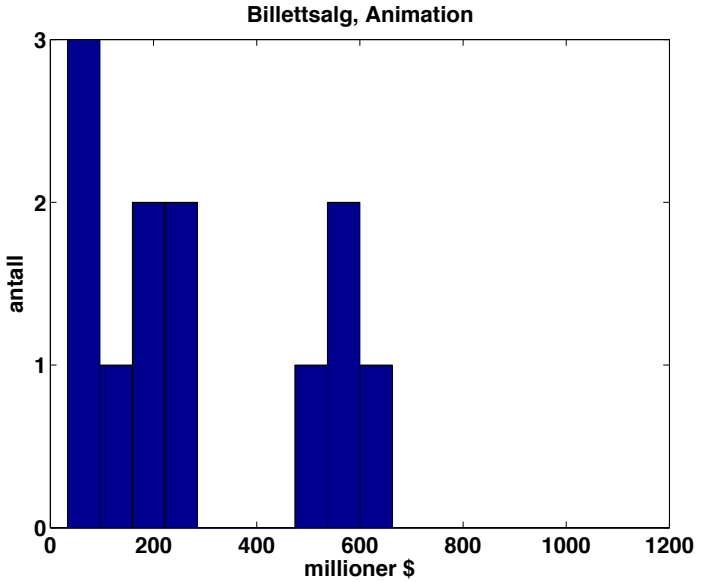
Alder:	20	21	21	22	22	22	23	23	66	67
Gjennomsnitt:	31									
Median:	22									

Gjennomsnitt vs. median

Billettsalg (millioner $)

Sjanger	**Antall**	Median	Gjennomsnitt
Action	32	132,15	249,05
Animation	12	219,56	286,58





Empirisk varians

- Den **empiriske variansen** er et mål på hvor mye spredning det er rundt gjennomsnittet i et datasett

For observasjonene $x_1, x_2, ..., x_{10}$ så er den empiriske variansen:

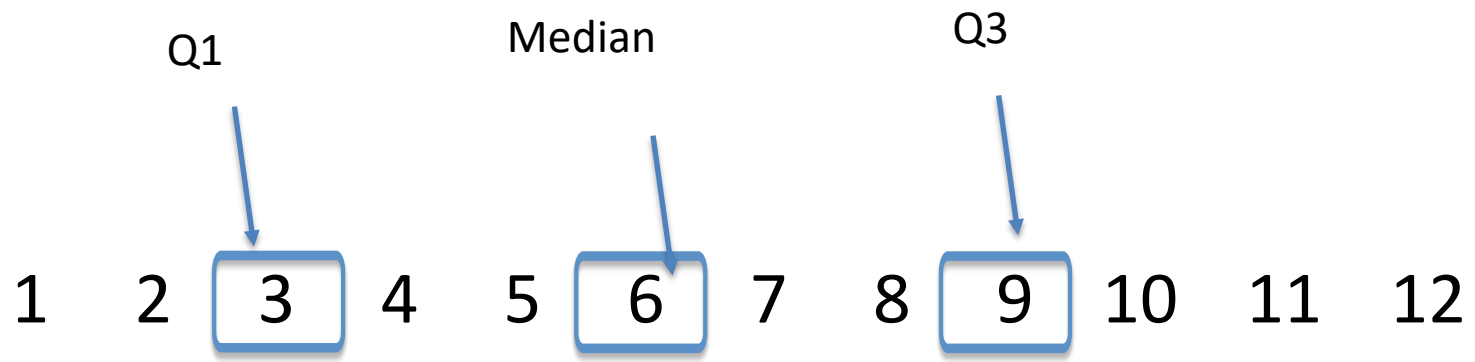
$$s^2 = \frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{x})^2$$

s kalles standardavvik.

Varians, Hollywood-filmer

		Empirisk varians	
Sjanger	**Antall**	Budsjett	Billettsalg
Action	32	3 787	78 560
Animation	12	2 613	49 473
Comedy	27	542	14 541
Drama	21	166	2492
Fantasy	2	7 788	880 190
Horror	17	278	5 400
Romance	11	1 104	31 800
Thriller	13	325	3 475
Alle	135	2 418	46 233

- På samme måte som median er et alternativ til snitt for å se på senteret til et datasett, gir **kvartilene** en alternativ måte å se på spredninga



$$\mathrm{IQR} = Q_3 - Q_1 = 9 - 3 = 6$$

Levetiden til pattedyr

Table 2.14 *Longevity of mammals*

Species	Longevity	Species	Longevity	Species	Longevity
Baboon	20	Elephant	40	Mouse	3
Black bear	18	Elk	15	Opossum	1
Grizzly bear	25	Fox	7	Pig	10
Polar bear	20	Giraffe	10	Puma	12
Beaver	5	Goat	8	Rabbit	5
Buffalo	15	Gorilla	20	Rhinoceros	15
Camel	12	Guinea pig	4	Sea lion	12
Cat	12	Hippopotamus	25	Sheep	12
Chimpanzee	20	Horse	20	Squirrel	10
Chipmunk	6	Kangaroo	7	Tiger	16
Cow	15	Leopard	12	Wolf	5
Deer	8	Lion	15	Zebra	15
Dog	12	Monkey	15		
Donkey	12	Moose	12		

Table 2.14

Levetiden til pattedyr

© Britta Kasholm-Tengve/iStockphoto

Page 61

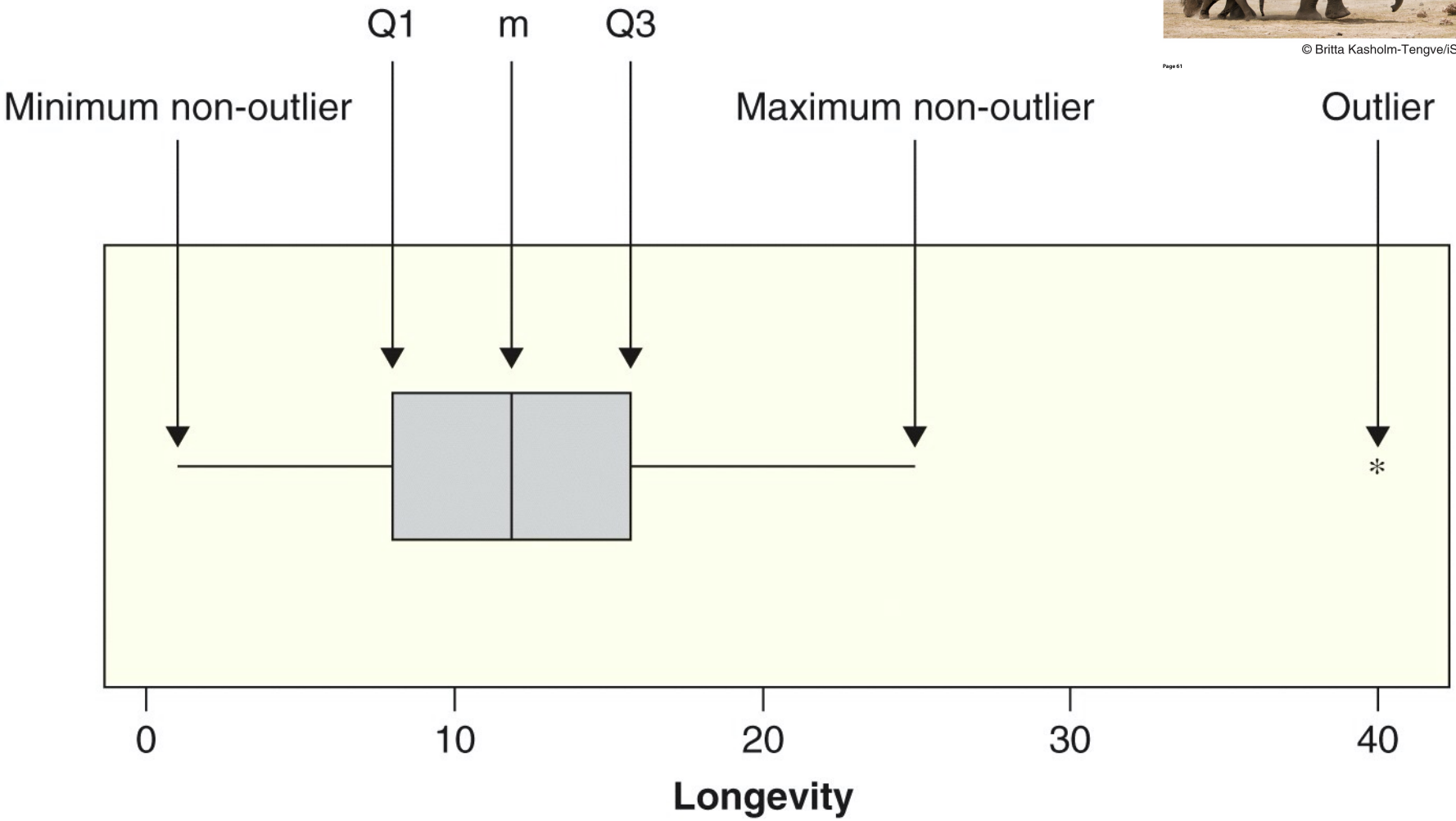
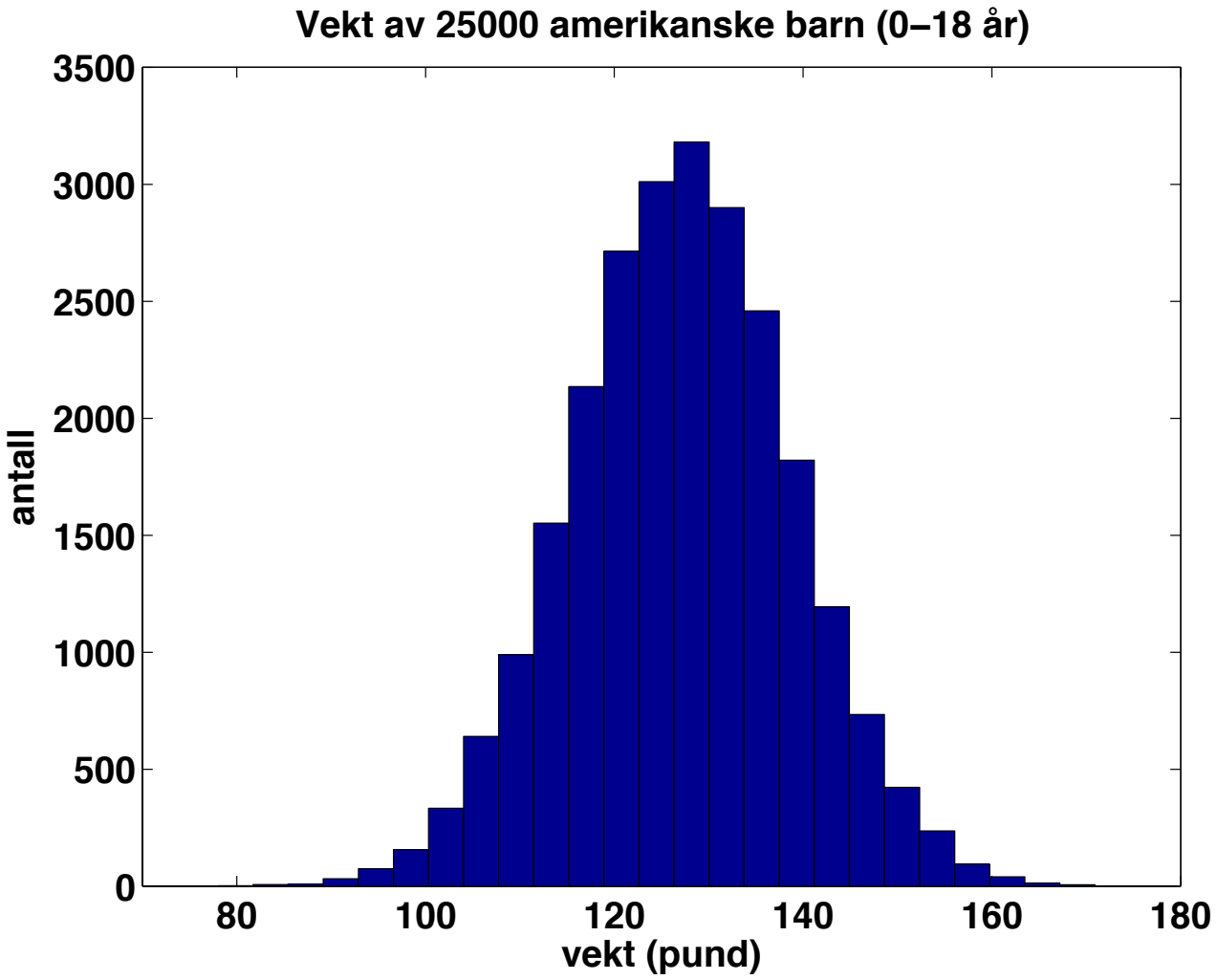


Figure 2.31

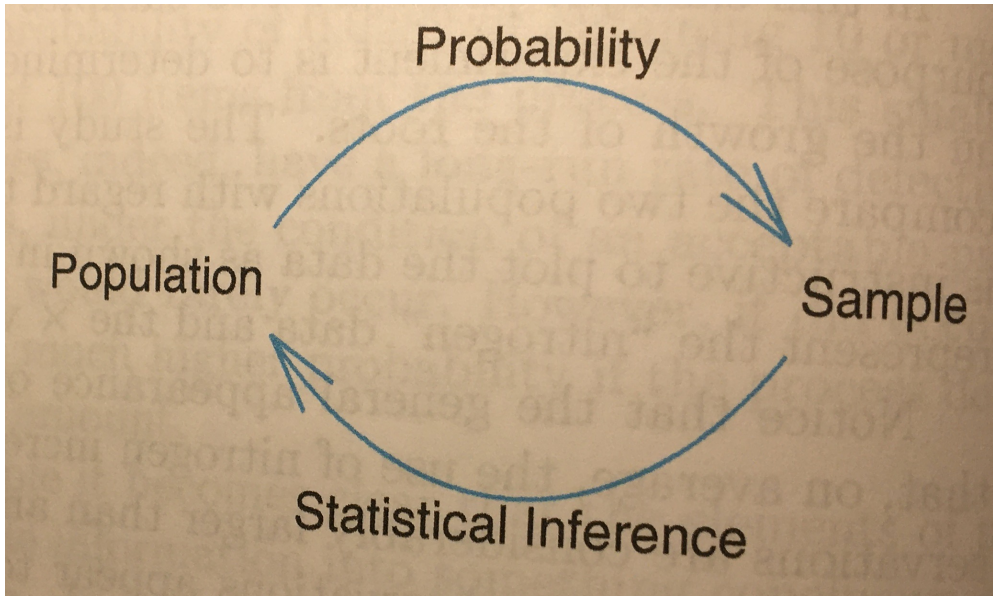
- **Histogram** viser data sortert i bins.




```
>> mean(weights)

ans =

  127.0775

>> var(weights)

ans =

  136.1753
```

Sannsynlighet og statistikk

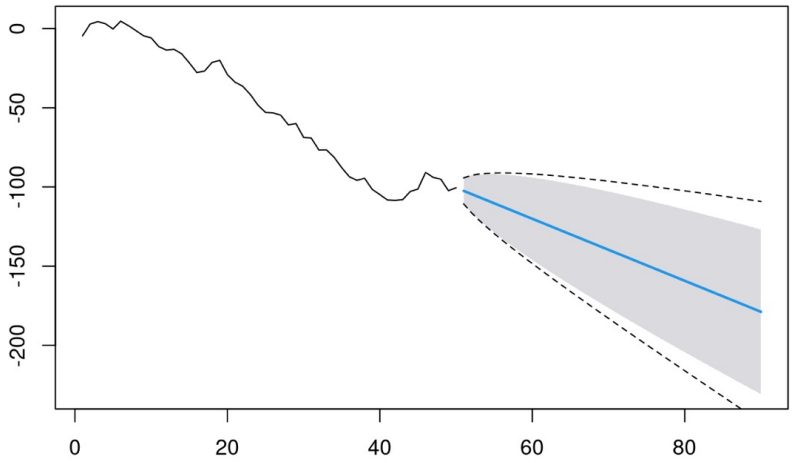
Kap 1, Walpole, Myers, Myers og Ye



Sannsynlighet og statistikk

Koblingen muliggjør

- Usikkerhetskvantifisering
- Ta beslutninger under usikkerhet
- Risikovurdering
- Hensyn til metoden for datainnsamling
- Egenskaper til prediksjoner (mer enn bare ett tall)



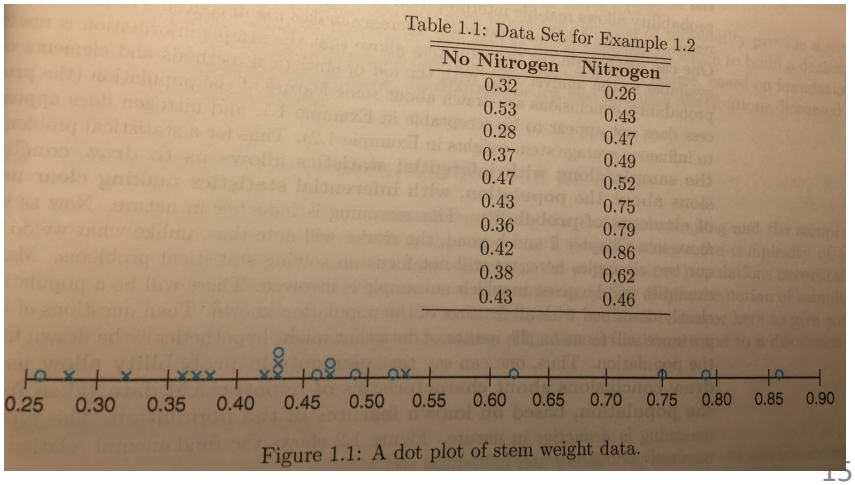


Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

Figure 1.1: A dot plot of stem weight data.

Sannsynlighet og statistikk

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

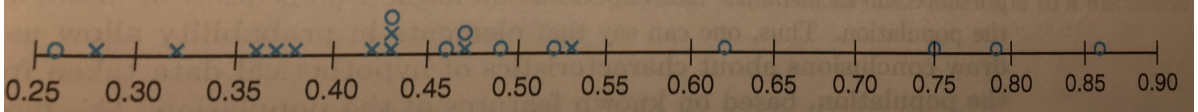


Figure 1.1: A dot plot of stem weight data.

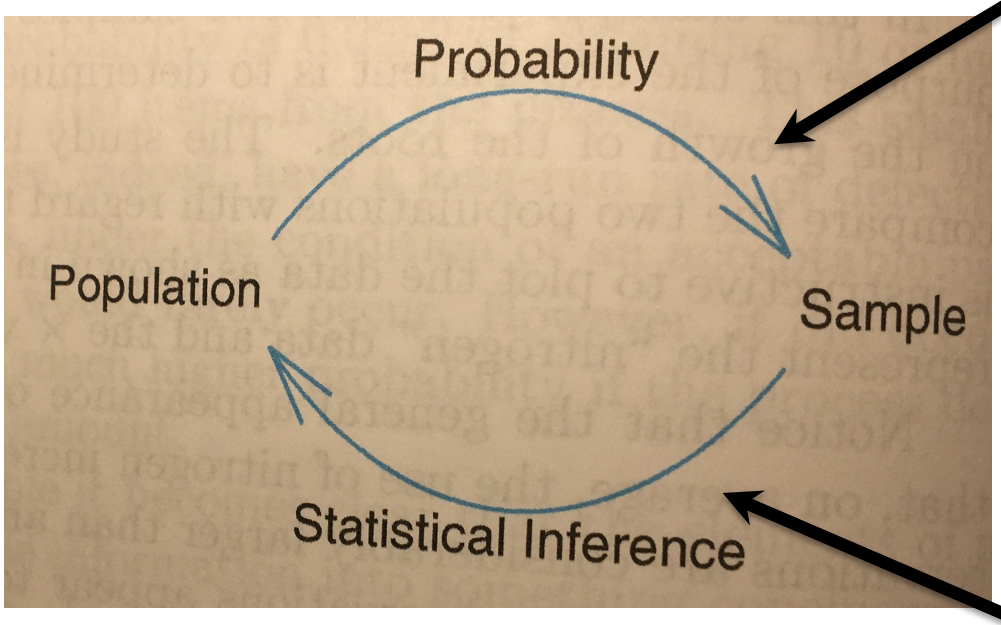
Er den ene gruppen 'bedre' enn den andre?

Det kan se slik ut, men det kan også være et utslag av tilfeldig variasjon.

Mean (No nitgogen) = 0.399
Mean (Nitrogen) = 0.565

Stdev (No nitrogen) = 0.073
Stdev (Nitrogen) =0.187

Sannsynlighet og statistikk



Første del
(6 uker).

Siste del
(6 uker).