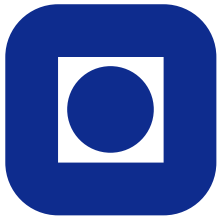


NTNU
Norwegian University of
Science and Technology

Introduksjon til dataanalyse

Deskriptiv statistikk

Kapittel 1

Denne timen og delvis forrige time er inspirert av Kapittel 1, men vi kommer ikke til å gå igjennom alt fra dette kapittelet i forelesning. Siden Kapittel 1 er pensum er det viktig at dere leser resten selv.

Denne forelesningen vil bruke slides, men forelesningen i morgen og de etterfølgende forelesningene vil for det meste være tavleforelesninger.

Datasett

Vi skal se på to datasett fra
`http://www.math.hope.edu/swanson/statlabs/data.html`:

1. Høyde til 50 mannlige og 50 kvinnlige studenter gitt i meter
2. Kroppstemperatur for 65 menn og 65 kvinner gitt i grader Celsius

Datasett

Vi skal se på to datasett fra
`http://www.math.hope.edu/swanson/statlabs/data.html`:

1. Høyde til 50 mannlige og 50 kvinnlige studenter gitt i meter
2. Kroppstemperatur for 65 menn og 65 kvinner gitt i grader Celsius

Vi ønsker å finne ut om det er en forskjell mellom høyden og kroppstemperaturen til kvinner og menn

Hvor begynner man?

Datasettene er for store til å bare skrive ut tallene for menn og for kvinner og sammenligne dem, men vi kan

- Regne ut tall som beskriver sentrene til de observerte verdiene for menn og kvinner
- Regne ut tall som beskriver spredningene til de observerte verdiene for menn og kvinner
- Lage plott som viser hvordan de observerte verdiene er spredt i forhold til hverandre

Hvor begynner man?

Datasettene er for store til å bare skrive ut tallene for menn og for kvinner og sammenligne dem, men vi kan

- Regne ut tall som beskriver sentrene til de observerte verdiene for menn og kvinner
- Regne ut tall som beskriver spredningene til de observerte verdiene for menn og kvinner
- Lage plott som viser hvordan de observerte verdiene er spredt i forhold til hverandre

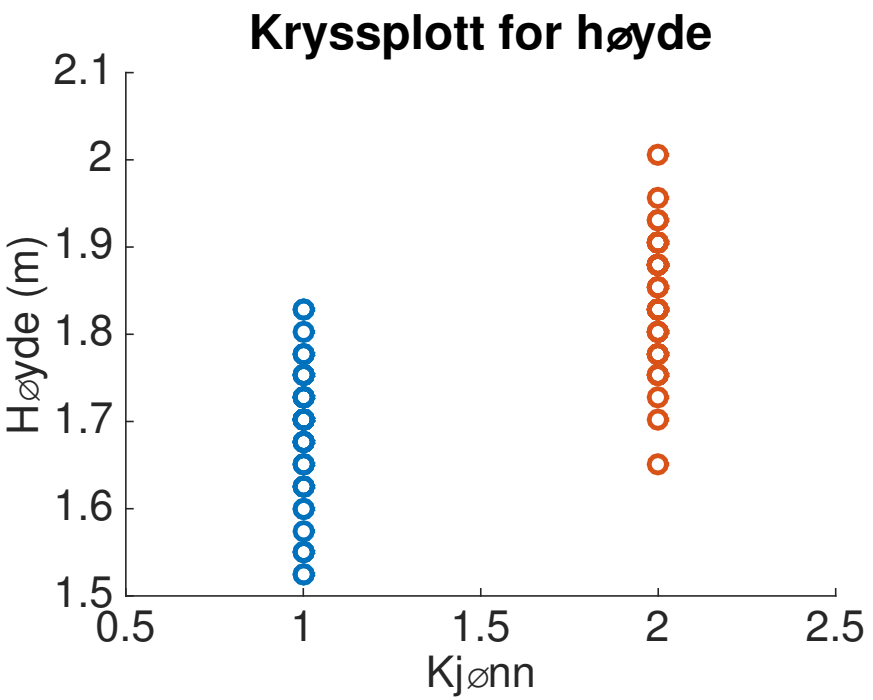
Dette er metoder for å beskrive de observerte verdiene og bruk av disse kalles **deskriptiv statistikk**

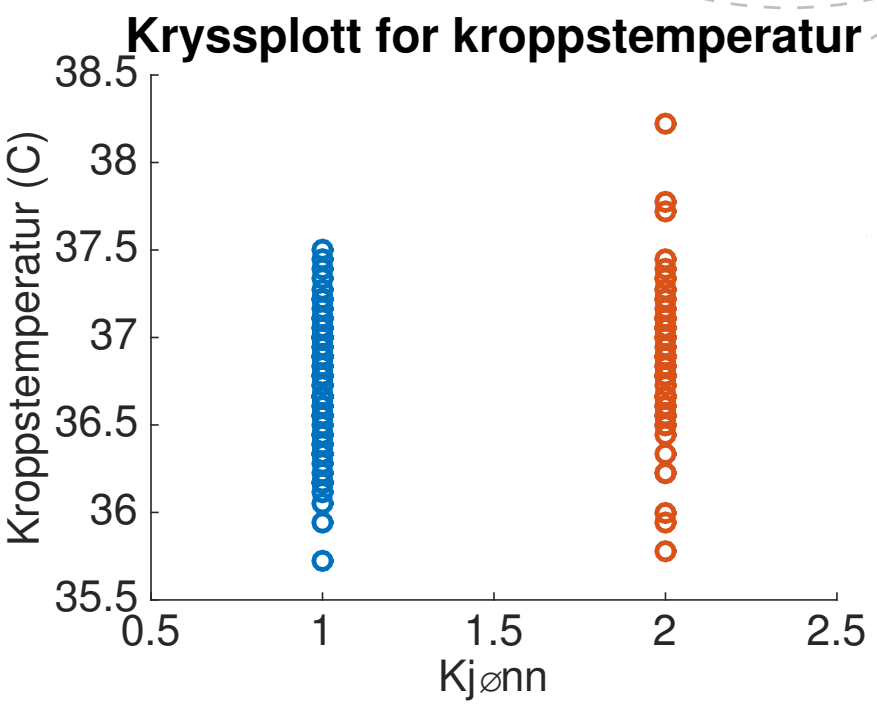
Deskriptiv statistikk

Deskriptiv statistikk er metoder for å beskrive eller oppsummere de observerte verdiene på en slik måte at man kan se mønstre. Typisk gjennom enkle grafiske presentasjoner eller ved utregning av tall som oppsummerer de observerte verdiene slik som middelverdi og spredning.

Kryssplott

Et enkelt startpunkt er å plotte verdiene for hvert av kjønnene og sammenligne





Sentralmål

Den vanligste måten å oppgi senteret til de observerte verdiene er å regne ut den empiriske middelverdien

Definisjon (Empirisk middelverdi)

Den empiriske middelverdien (utfallsmiddelverdien) til observasjonene $x_1, x_2, \ldots, x_n$, betegnet med $\bar{x}$, er gitt ved

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Eksempler

- Empirisk middelverdi til 1, 2, 3, 4, 5 er 3
- Empirisk middelverdi til 1, 1, 1, 2, 2, 2 er 1.5
- Empirisk middelverdi til 1, 2, 3, 4, 20 er 6

Sentralmål som unngår ekstreme observasjoner

Hvis det er ekstreme observasjoner som har sterk påvirkning på den empiriske middelverdien er empirisk median et bedre mål på senteret av de observerte verdiene

Definisjon (Empirisk median)

Den empiriske medianen (utfallsmedianen) til observasjonene $x_1, x_2, \ldots, x_n$, betegnet med $\tilde{x}$, er gitt ved

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{for } n \text{ odde}, \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{for } n \text{ like}, \end{cases}$$

Eksempler

- Empirisk median til 1, 2, 3, 4, 5 er 3
- Empirisk median til 1, 1, 1, 2, 2, 2 er 1.5
- Empirisk median til 1, 2, 3, 4, 20 er 3

Sammenligning av sentrene

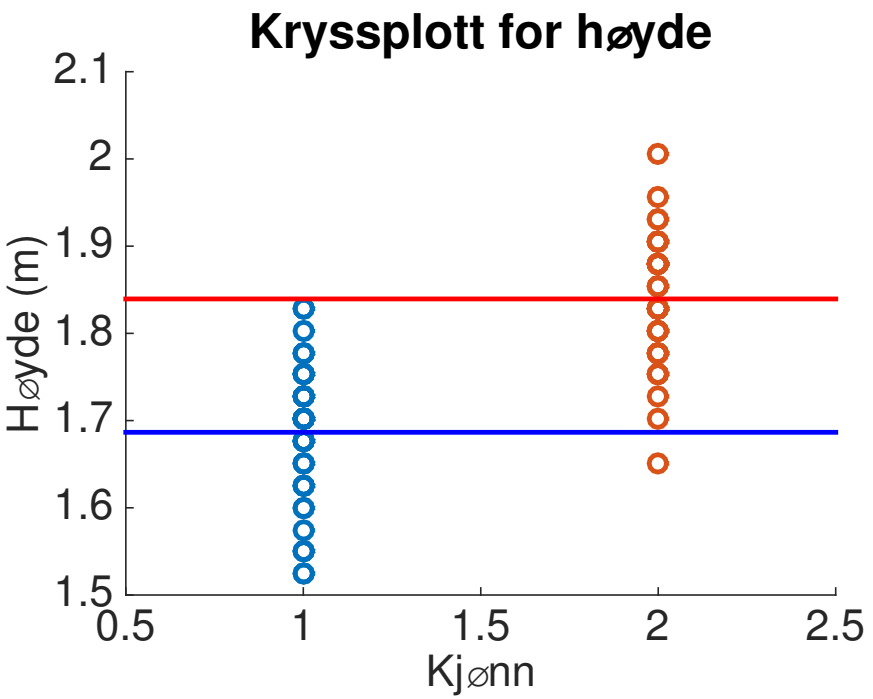
	Høyde		Kroppstemperatur	
Sentralmål	Kvinne	Mann	Kvinne	Mann
Empirisk middelverdi	1.687	1.840	36.72	36.89
Empirisk median	1.702	1.842	36.72	36.89

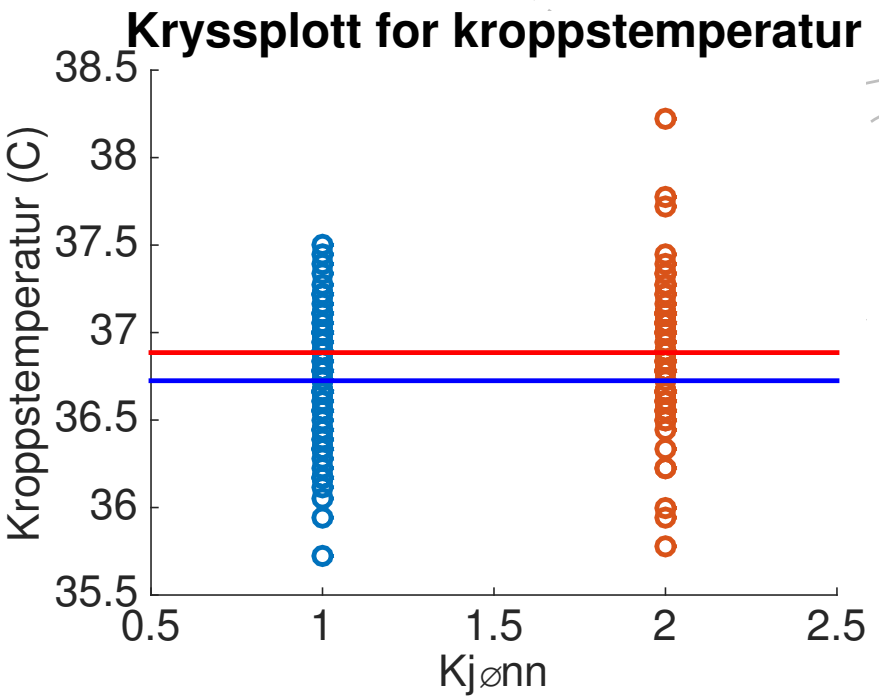
Sammenligning av sentrene

	Høyde		Kroppstemperatur	
Sentralmål	Kvinne	Mann	Kvinne	Mann
Empirisk middelverdi	1.687	1.840	36.72	36.89
Empirisk median	1.702	1.842	36.72	36.89

Empirisk middelverdi er høyre for menn enn for kvinner for både høyde og kroppstemperatur, men er det nok til å hevde at det faktisk er en forskjell?

Middelverdi og median forteller ikke hele historien





Empiriske middelverdier er markerte med horisontale streker

Variabilitet om middelverdien

Vi må også vurdere variabilitetene til de observerte verdiene for å kunne trekke konklusjoner om forskjeller for menn og kvinner

Variabilitet om middelverdien

Vi må også vurdere variabilitetene til de observerte verdiene for å kunne trekke konklusjoner om forskjeller for menn og kvinner

For høyde ser forskjellen ut til å være stor sammenlignet med variabiliten, men for kroppstemperatur ser forskjellen ut til å være liten sammenlignet med variabiliteten

Variabilitet om middelverdien

Vi må også vurdere variabilitetene til de observerte verdiene for å kunne trekke konklusjoner om forskjeller for menn og kvinner

For høyde ser forskjellen ut til å være stor sammenlignet med variabiliten, men for kroppstemperatur ser forskjellen ut til å være liten sammenlignet med variabiliteten

Men vi er nødt til å tallfeste forskjellene!

Mål på variabilitet

Det mest vanlige målet på variabilitet er empirisk varians

Definisjon (Empirisk varians)

Den empiriske variansen (utfallsvariansen) til observasjonene $x_1, x_2, \ldots, x_n$, betegnet med s^2, er gitt ved

$$s^2 = \frac{1}{n-1}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Eksempler

- Empirisk varians til 1, 2, 3, 4, 5 er 2.5
- Empirisk varians til 1, 1, 1, 2, 2, 2 er 0.3
- Empirisk varians til 1, 2, 3, 4, 20 er 62.5

Mål på variasjon

Empirisk varians kan være vanskelig å tolke fordi den er uttrykt i kvadratiske enheter. Empirisk standardavvik er uttrykt i samme enheter som dataene.

Definisjon (Empirisk standardavvik)

Det empiriske standardavviket (utfallsstandardavviket) til observasjonene $x_1, x_2, \ldots, x_n$, betegnet med s, er gitt ved

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

Eksempler

- Empirisk standardavvik til 1, 2, 3, 4, 5 er 1.58
- Empirisk standardavvik til 1, 1, 1, 2, 2, 2 er 0.548
- Empirisk standardavvik til 1, 2, 3, 4, 20 er 7.91

Sammenligning av menn og kvinner

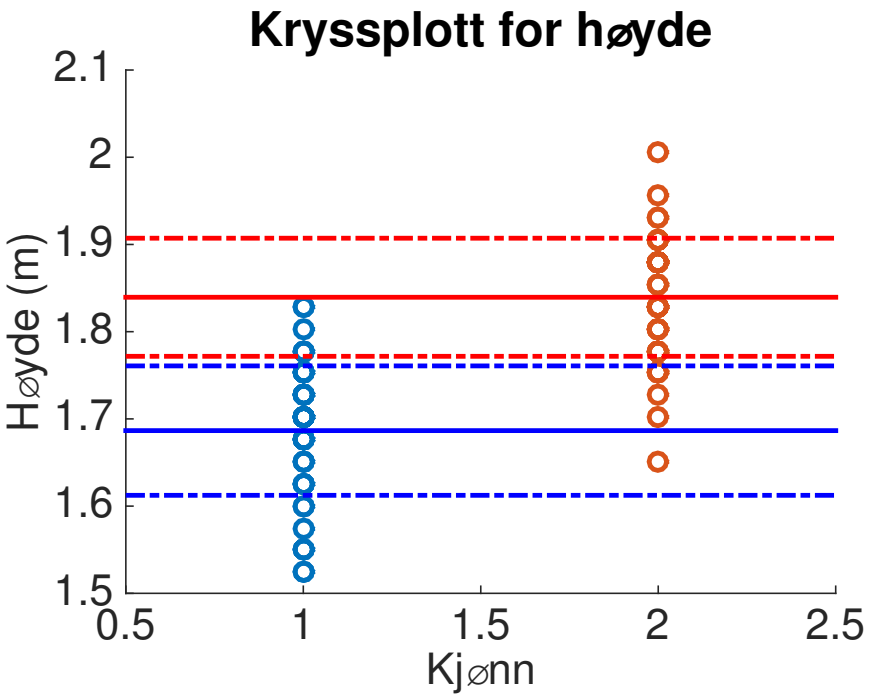
	Høyde		Kroppstemperatur	
	Kvinne	Mann	Kvinne	Mann
Empirisk middelverdi	1.687	1.840	36.72	36.89
Empirisk standardavvik	0.074	0.068	0.39	0.41

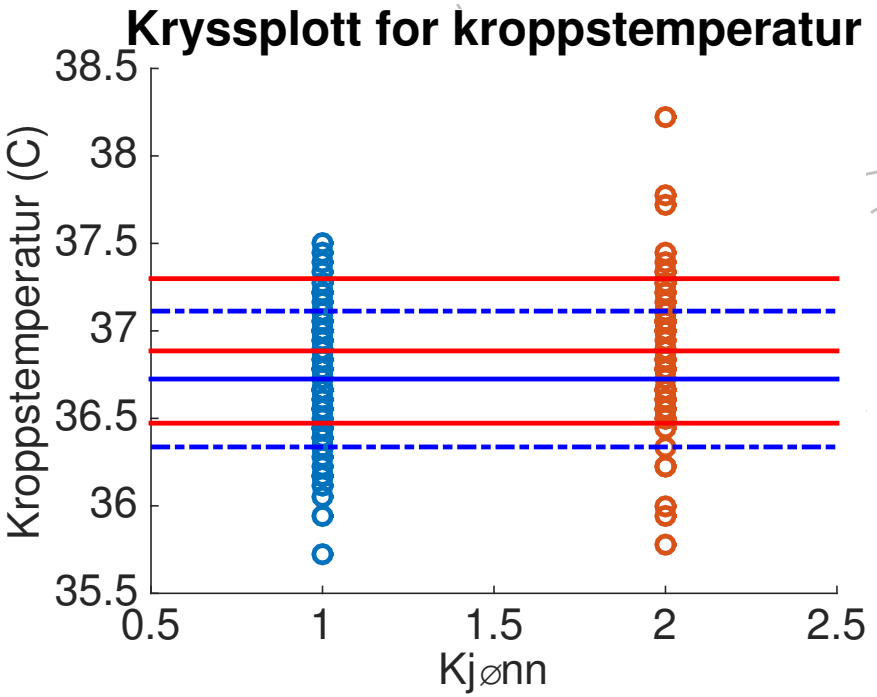
Sammenligning av menn og kvinner

	Høyde		Kroppstemperatur	
	Kvinne	Mann	Kvinne	Mann
Empirisk middelverdi	1.687	1.840	36.72	36.89
Empirisk standardavvik	0.074	0.068	0.39	0.41

Forskjellen i kroppstemperatur er mindre enn ett standardavvik, men forskjellen i høyde er større enn to standardavvik

Sentrum og variabilitet

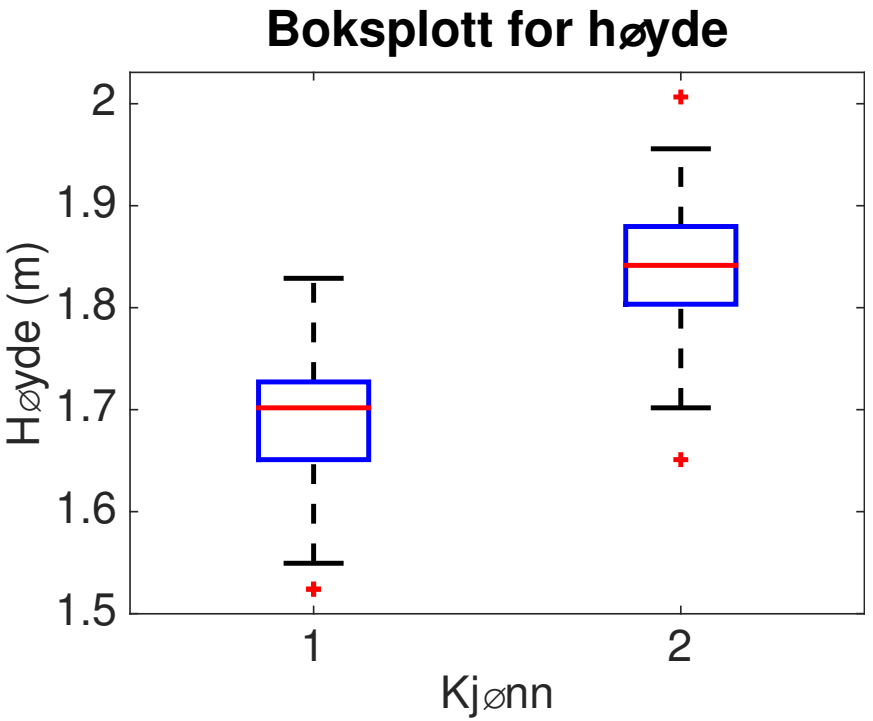


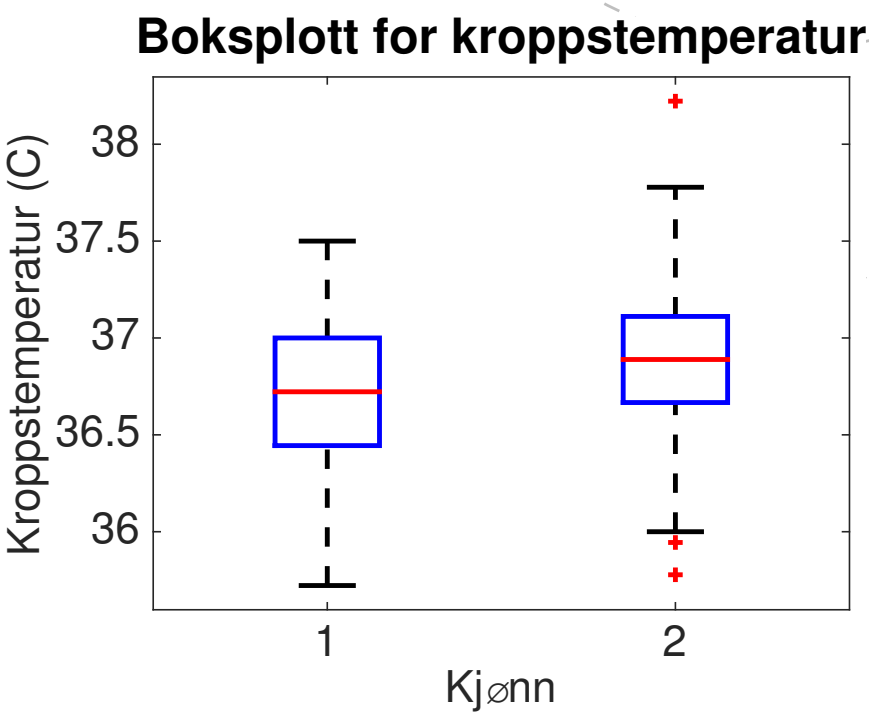


Empiriske middelverdier er markerte med horisontale streker og $\pm$ ett empirisk standardavvik er markerte med stiplede horisontale linjer

Boksplott

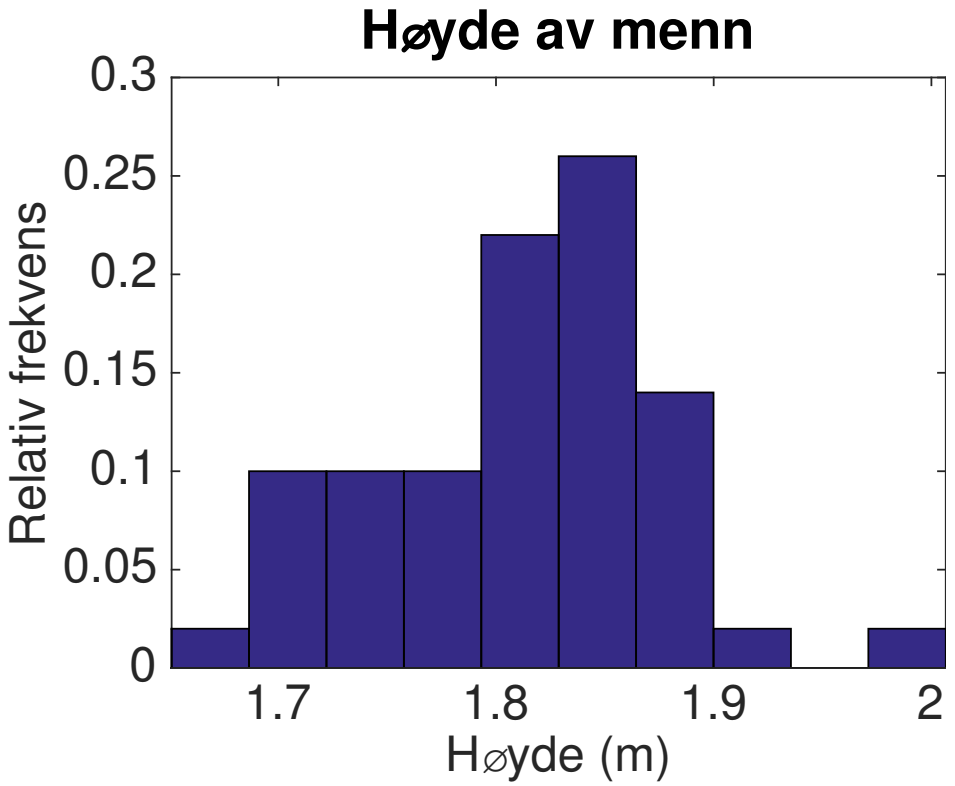
Denne sammenligningen gjøres bedre i et boksplott





Histogram

Hvis man ønsker å se hvordan de observerte verdiene er spredt rundt middelverdien for en av kategoriene er det vanlige å bruke et histogram



Mangler ved deskriptiv statistikk

- Ingen bruk av sannsynlighet. Vi har ingen tall på hvor sikker vi er på at menn er høyere enn kvinner.
- Veldig subjektivt om man anser det for å være en forskjell eller ikke
- Deskriptiv statistikk egner seg best som et startpunkt for å utforske de observerte verdiene

Mangler ved deskriptiv statistikk

- Ingen bruk av sannsynlighet. Vi har ingen tall på hvor sikker vi er på at menn er høyere enn kvinner.
- Veldig subjektivt om man anser det for å være en forskjell eller ikke
- Deskriptiv statistikk egner seg best som et startpunkt for å utforske de observerte verdiene

I morgen begynner vi reisen mot **statistisk inferens** hvor vi gjennom sannsynlighetsteori kan tallfeste vår sikkerhet til konklusjonene vi gjør