# Statistikk

Per G. Østerlie

20. april 2023

# Innhold

Oppsummering av det vi har gjort i statistikk. Obs! Her kan det være noen feil her og der. Gi meg beskjed hvis du finner noe som er uklart. Dokumentet vil oppdateres etterhvert med korreksjoner, så se på datoen på framsida.

Teksten er skrevet med LaTeX.

# 1 Sannsynlighetsfordelinger

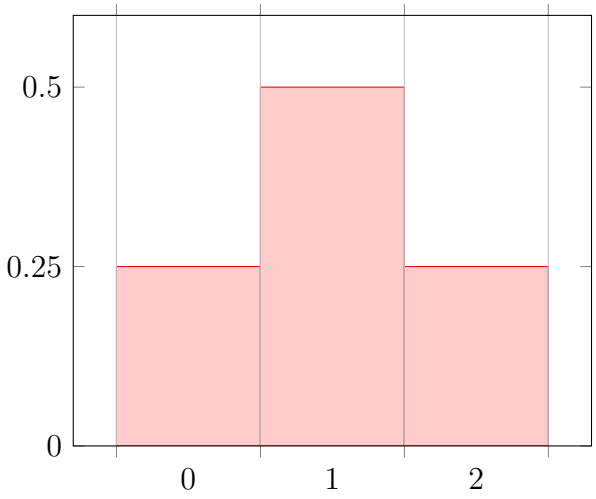
## 1.1 Hva er en sannsynlighetsfordeling?

Vi kan ikke si noe om utfallet av et stokastisk forsøk. Vi vet, som regel, utfallsrommet, men akkurat hvilket utfall er overlatt til tilfeldighetene. Det vi ofte vet noe om er sannsynligheten om hvordan utfallene vil fordele seg. Vi vet *sannsynlighetsfordelingen*.

Et problem vi kjenner fra historien er sannsynligheten for utfallene når vi kaster to mynter[1]. Utfallsrommet vil være $U = \{MM, MK, KK\}$, hvor M står for «mynt» og K for «kron». Vi kan skrive om slik at vi ser på antall «kron». Det gir denne tabellen

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X = x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

Denne tabellen gir sannsynlighetsfordelingen for den stokastiske variabelen $X$. Vi kan også tegne dette som et histogram.




Dette gir oss en grafisk framstilling av sannsynlighetsfordelingen av utfallene. Når vi nå vet sannsynlighetsfordelingen vil det være mulig å slå opp i den for å finne sannsynlighetene. Tabellen gir oss en sammenheng mellom $x$ og sannsynligheten slik at $P(X = x) = p(x)$. Sannsynligheten kan uttrykkes som en funksjon av $x$. I dette tilfellet har vi at $p(0) = \frac{1}{4}$, $p(1) = \frac{1}{2}$ og $p(2) = \frac{1}{4}$.

Sannsynlighetsfordelinger kan være av to typer

**diskrete** når den stokastiske variabelen er diskret. Det betyr at variabelen bare kan være verdier i en mengde av enkeltelementer. Terningkast og myntkast er eksempler

**kontinuerlige** når den stokastiske variabelen er kontinuerlig.

[1] Dette er kjent som diskusjonen mellom Laplace og d'Alembert

## 1.2 Uniform fordeling

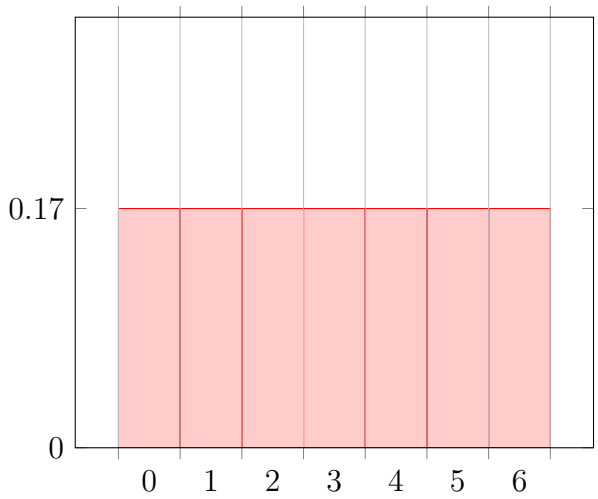
Kaster vi en ideell terning vil det være like sannsynlig hvilken side den lander på. Når alle utfallene er like sannsynlige har vi en uniform sannsynlighetsfordeling. Det samme gjelder myntkast med en ideell mynt. Erstatter vi utfallene «mynt» og «kron» med 0 og 1 får vi

$$P(X = 0) = \frac{1}{2}$$
$$P(X = 1) = \frac{1}{2}$$

Vi kjenner også terningkast med en ideell terning hvor vi har denne sannsynlighetsfordelingen

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $P(X = x)$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

Tegner vi opp denne sannsynlighetsfordelingen får vi dette histogrammet.




Alle sannsynlighetene er like store for alle utfallene. De er *uniforme* og vi har at $P(X = x) = p(x) = \frac{1}{6}$ for alle verdier av $x$. Legg også merke til at summen av alle sannsynlighetene er lik 1 og tilsvarer arealet i histogrammet.

### Forventningsverdi

Hva er forventningsverdien når vi kaster en terning? Hvor mange øyne kan vi forvente å få om vi kaster terningen mange ganger? Svaret kan vi finne ved å gjennomføre forsøket og fordele antall øyne på antall kast. Vi vet at det er like stor sannsynlighet for at terningen lander på alle sidene. Da kan vi finne svaret slik

$$\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

Nå er det ikke mulig å få forventningsverdien i noen av kastene, men den forteller oss hvor mange øyne vi vil få i gjennomsnitt. Med symboler blir utregningen

$$E(X) = \mu = \sum_{x=1}^{6} x \cdot P(X = x) = 3.5$$

$E(X)$ står for *Expected value* eller forventningsverdi på norsk. Summetegnet gir oss en kort måte å skrive alle summene over. Startverdien for $x$ er $x = 1$ og sluttverdien $x = 6$ – akkurat som i summen over.

Dette kan vi også simulere med et programmering hvor vi kaster mange ganger og finner gjennomsnittet av resultatene.

```
1 import random as rd
2 antall_kast = 1000
3  sum = 0
4
5 for i in  range(antall_kast):
6     terning = rd.randint(1,6)
7      sum =  sum + terning
8
9 print( sum/antall_kast)
```

Programkode 1.1: Forventningsverdi ved uniform sannsynlighet

Et resultat ble

```
3.523
```

## 1.3 Binomisk fordeling

Et binomisk forsøk, eller Bernoulli-forsøk[2], kjennetegnes av to utfall, samme sannsynlighet i hvert forsøk og uavhengighet mellom forsøkene. Et myntkast kan være et eksempel på det.

Kaster vi en perfekt mynt seks ganger kan vi finne sannsynligheten antall «kron» på denne måten

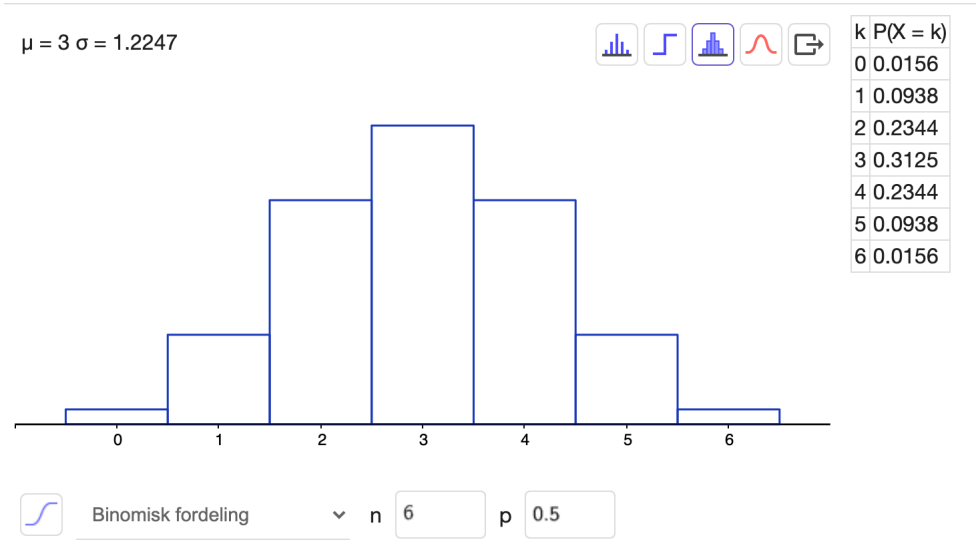
$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}$$

hvor $n$ er antall forsøk, $x$ er antall «kron» og $p$ er sannsynligheten for å oppnå det. Legg merke til at vi her har en funksjon som gir oss sannsynlighetene: $P(X = x) = p(x)$.

La oss si at vi kaster mynten seks ganger $n = 6$ og vi antar at sannsynligheten for å oppnå «kron» er like stor som å oppnå «mynt», $p = \frac{1}{2}$. Da gjør vi et binomisk forsøk og sannsynlighetsfordelingen vil være en binomisk fordeling. Nå kan vi tegne sannsynlighetsfordelingen ved å benytte GeoGebra. Figur 1.1 viser resultatet.

GeoGebra gir oss både et histogram og en tabell for de forskjellige sannsynlighetene. I dette tilfellet ser vi at verdiene blir symmetriske, men det er ikke alltid tilfelle.

[2]Oppkalt etter Jacob Bernoulli (1655 - 1705) som var en kjent matematiker fra en hel familie med matematikere. Han er bl. a. kjent for de store talls lov, som han skrev om i boka Ars Conjectandi.




Figur 1.1: Binomisk sannsynlighetsfordeling med GeoGebra

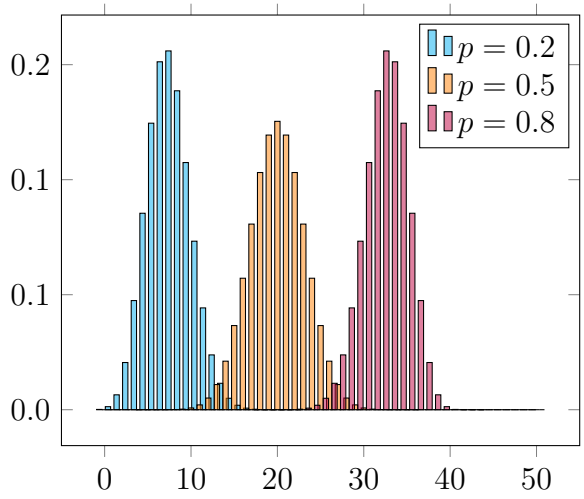
I dette tilfellet sier vi at sannsynlighetene er binomisk fordelt med $n = 6$ og $p = \frac{1}{2}$. Det skrives ofte kortere som at $X$ er $bin\left(6, \frac{1}{2}\right)$ eller $b\left(6, \frac{1}{2}\right)$.

**Binomisk fordeling**

I en binomisk forsøksserie vil den stokastiske variabelen $X$ være binomisk fordelt. Vi finner sannsynligheten for at $X$ inntreffer $k$ ganger ved

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

$X$ er binomisk fordelt med $n$ og $p$ og kan skrives slik: $X \sim bin\,(n, p)$ eller $X \sim b\,(n, p)$




## Forventningsverdi og varians i en binomisk fordeling

Vi kan finne både gjennomsnitt og standardavvik i sannsynlighetsfordelinger på samme måte som vi benytter for relative frekvenser i datamaterialer.

**Gjennomsnittet** finner vi ved

$$\mu = \sum_{x=0}^{n} x \cdot p(x)$$

For å regne det ut er det greit å sette opp denne tabellen (den siste kolonna er med til seinere bruk, så se på de første).

| $x$ | $P(X = x)$ | $x \cdot P(X = x)$ | $(x - \mu)^2 \cdot P(X = x)$ |
|---|---|---|---|
| 0 | 0.01563 | 0 | 0.14063 |
| 1 | 0.09375 | 0.09375 | 0.37500 |
| 2 | 0.23438 | 0.46875 | 0.23438 |
| 3 | 0.31250 | 0.93750 | 0 |
| 4 | 0.23438 | 0.93750 | 0.23438 |
| 5 | 0.09375 | 0.46875 | 0.37500 |
| 6 | 0.01563 | 0.09375 | 0.14063 |
| sum | 1 | 3 | 1.5 |

Da har vi at $\mu = 3$. Egentlig kunne vi spart oss all denne utregningen for GeoGebra har allerede funnet det ut for oss. Se figur 1.1. Nå er dette resultatet også noe vi kan tenke oss til. I hvert delforsøk – hver gang vi kaster mynten - er sannsynligheten for å få «kron», $p = \frac{1}{2}$. Det er dette delforsøket vi gjentar seks ganger. Da kan vi forvente at vi får $6 \cdot \frac{1}{2} = 3$ «kron» når vi kaster mynten seks ganger. Vi ser at utregningene stemmer med det. Å bevise at disse utregningene gir akkurat det resultatet kan vi også gjøre, men vi tar ikke det med her.

Vi kan altså forvente oss tre «kron» i dette tilfellet. Vi sier at forventningsverdien er tre. Denne verdien kan vi regne ut før vi kaster mynter eller utfører andre stokastiske forsøk. «Gjennomsnitt» er et ord som viser til noe som har skjedd. Derfor bruker vi heller «forventningsverdi». Ofte skrives det da som *Expected value*, $E(X)$, i stedet for $\mu$. I en binomisk fordeling har vi at forventningsverdien er

$$E(X) = \mu = n \cdot p$$

**Variansen** finner vi ved

$$Var(X) = \sigma^2 = \sum_{x} (x - \mu)^2 \cdot P(X = x)$$

Ser vi tilbake på tabellen ser vi at det i vårt tilfelle blir $\sigma^2 = 1.5$. Da har vi at $\sigma = \sqrt{1.5} = 1.224744871391589$. Det er det samme som GeoGebra fant for oss i øverst til venstre i figur 1.1.

I en binomisk fordeling kan vi vise at variansen blir

$$Var(X) = \sigma^2 = n \cdot p \cdot (1 - p)$$

Her kan vi se at det vi fant stemmer ved å regne ut

$$Var(X) = \sigma^2 = 6 \cdot \frac{1}{2} \cdot (1 - \frac{1}{2}) = \frac{3}{2}$$

**Forventningsverdi og varians**

Hvis den stokastiske variabelen ($X$ er $bin(n, p)$ har vi at

$$E(X) = \mu = np$$
$$Var(X) = \sigma^2 = np(1 - p)$$

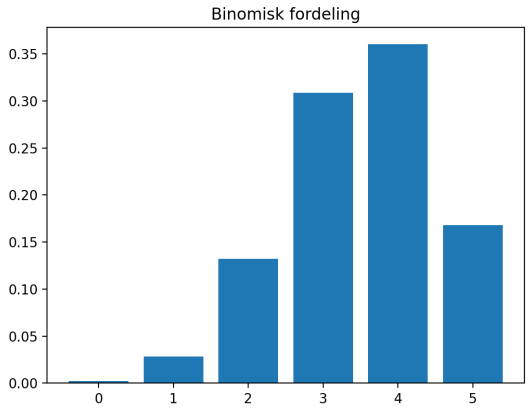
**Oppgave 1**

Eksperimenter med binomisk fordeling i sannsynlighetskalkulatoren i GeoGebra og se hvordan sannsynlighetsfordelingene endrer seg.

**Python**

```python
import matplotlib.pyplot as plt
import math as m

antall = 5
p = 0.7

xverdier = []
yverdier = []

def binomial(a,b):
     bin = m.factorial(a)//(m.factorial(b)*m.factorial(a-b))
    return  bin

def binomisk(n, p, x):
     bin = binomial(n, x)*p**x*(1 - p)**(n-x)
    return  bin

for i in  range(antall+1):
    xverdier.append(i)
    yverdier.append(binomisk(antall,p,i))
    print(i, binomisk(antall,p,i))

plt.title("Binomisk fordeling")
plt.bar(xverdier, yverdier)
plt.show()
```

Programkode 1.2: Binomisk fordeling




```
0 0.0024300000000000016
1 0.028350000000000018
2 0.13230000000000006
3 0.30870000000000003
4 0.36014999999999997
5 0.16806999999999994
```

Eksemplet over kunne vært gjort enklere ved at vi benytter modulen `scipy`. Her finner vi funksjonen `binom.pmf(k,N,p)`, den binomiske tetthetsfunksjonen (*probability mass function*) som regner ut $P(X = k)$ når vi utfører $N$ forsøk med sannsynligheten $p$.

Programkode 1.3 viser hvordan vi kan gjøre det. Resultatet blir samme tabell som den vi fikk i programkode 1.2.

```
1 import scipy.stats as st
2 antall = 5
3 p = 0.7
4
5 for k in  range (6):
6     s = st.binom.pmf(k,antall,p)
7     print(s)
```

Programkode 1.3: Binomisk fordeling

## 1.4 Hypergeometrisk fordeling

Hypergeometrisk forsøk kjennetegnes ved trekking uten tilbakelegging.

Et eksempel er at vi trekker fra ei urne med ti kuler hvor fire er grønne. La oss si at vi trekker tre kuler uten å legge tilbake. Da har vi muligheten for å få ingen, ei, to eller tre grønne kuler. Vi definerer den stokastiske variabelen $X$ som «antall grønne kuler». Sannsynlighetene bestemmes av en hypergeomtrisk fordeling av sannsynlighetene. Vi har vært innom hypergeometriske forsøk og kan regne de forskjellige sannsynlighetene ved

$$P(X = x) = \frac{\binom{s}{x} \cdot \binom{n-s}{r-x}}{\binom{n}{r}}$$

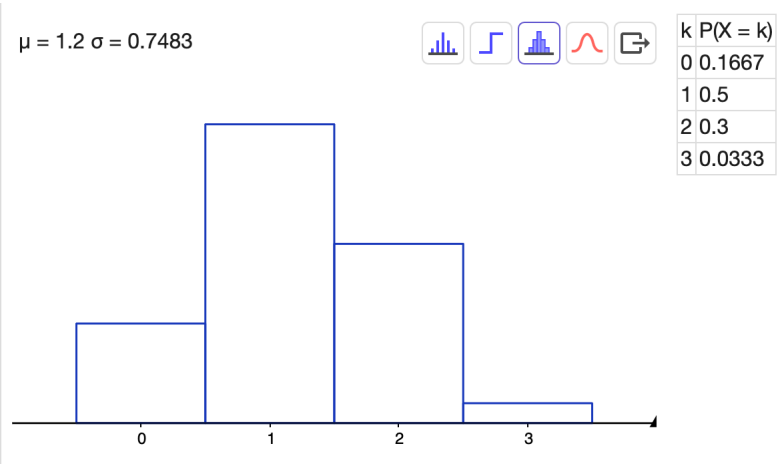
her er $n$ det totale antall kuler, $s$ er antall spesielle og $r$ er antallet i utvalget vi gjør. Sannsynligheten for å trekke to grønne kuler blir da:

$$P(X = 2) = \frac{\binom{s}{x} \cdot \binom{n-s}{r-x}}{\binom{n}{r}} = \frac{\binom{4}{2} \cdot \binom{6}{1}}{\binom{10}{3}} = \frac{3}{10} = 0.3$$

Denne sannsynligheten, og fordelingen for alle de andre sannsynlighetene, kan vi få ved hjelp av GeoGebra og sannsynlighetskalkulatoren. Figur 1.2 viser fordelingen både som tabell og histogram.

Eksemplet gir denne sannsynlighetsfordelingen

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | $\frac{1}{6}$ | $\frac{1}{2}$ | $\frac{3}{10}$ | $\frac{1}{30}$ |




| k | P(X = k) |
|---|---|
| 0 | 0.1667 |
| 1 | 0.5 |
| 2 | 0.3 |
| 3 | 0.0333 |

Figur 1.2: Hypergeometrisk GeoGebra

**Python**

Det samme er også mulig i Python. Da må vi importere modulen `scipy` og benytte funksjonen `hypergeom.pmf(k, n , s, r)`. Her er `k` det samme som $x$ i utregninga over. Programkode 1.4 viser hvordan vi kan få skrevet ut hele tabellen.

```
1 import scipy.stats as st
2
3 s = 4
4 n = 10
5 r = 3
6
7 for k in  range (r+1):
8     p = st.hypergeom.pmf(k,n,s,r)
9     print(p)
```

Programkode 1.4: Hypergeometrisk fordeling

Vi ender opp med dette resultatet

```
0.16666666666666646
0.4999999999999995
0.30000000000000016
0.03333333333333332
```

## Forventningsverdi og varians

**Forventningsverdien** til en stokastisk variabel er gitt ved

$$\mu = E(x) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$

I eksemplet vårt får vi

$$\mu = E(x) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{3}{10} + 3 \cdot \frac{1}{30} = \frac{12}{10} = 1.2$$

**Variansen** kan vi finne ved

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = xi)$$

Regner vi ut får vi

$$\sigma^2 = Var(X) = (0 - \frac{12}{10})^2 \cdot \frac{1}{6} + (1 - \frac{12}{10})^2 \cdot \frac{1}{2} + (2 - \frac{12}{10})^2 \cdot \frac{3}{10} + (3 - \frac{12}{10})^2 \cdot \frac{1}{30} = \frac{14}{25} = 0.56$$

**Standardavviket** finner vi ved

$$\sigma = \sqrt{Var(X)} = \sqrt{0.56} = 0.74833147735479$$

Som vi kan se av figur 1.2 har GeoGebra allerede funnet disse verdiene for oss.

I en hypergeometrisk fordeling fins det en annen måte å finne forventningsverdi og varians på. Vi har da at

$$\mu = E(X) = r \cdot \frac{s}{n}$$
$$\sigma^2 = Var(X) = \frac{n-r}{n-1} \cdot r \cdot \frac{s}{n} \cdot \left(1 - \frac{s}{n}\right)$$

Det gir

$$\mu = E(X) = 3 \cdot \frac{4}{10} = \frac{12}{10} = 1.2$$
$$\sigma^2 = Var(X) = \frac{10-3}{10-1} \cdot 3 \cdot \frac{4}{10} \cdot \left(1 - \frac{4}{10}\right) = \frac{14}{25}$$

**Hypergeometrisk fordeling**

Vi har $n$ elementer hvor $s$ er spesielle. Det trekkes et utvalg av $r$ elementer uten tilbakelegging. La $X$ være *antall spesielle i utvalget*. Da er sannsynlighetsfordelingen gitt ved

$$P(X = x) = \frac{\binom{s}{x} \cdot \binom{n-s}{r-x}}{\binom{n}{r}}$$

Forventningsverdi

$$\mu = E(X) = r \cdot \frac{s}{n}$$

Varians

$$\sigma^2 = Var(X) = \frac{n-r}{n-1} \cdot r \cdot \frac{s}{n} \cdot \left(1 - \frac{s}{n}\right)$$

## 1.5 Normalfordeling

Gjennom historien har vitenskapsmenn oppdaget at målinger i naturen følger en lovmessighet. De fordeler seg på samme måte. Noen verdier er det mange av og de andre fordeler seg på hver side av disse. Settes de opp i hyppighetsdiagram vil de følge en kurve som er ganske symmetrisk og fin. Vi kaller kurven for en Gauss-kurve, oppkalt etter den store matematikeren, fysikeren og astronomen Carl Friedrich Gauss (1777 - 1855). Gauss-kurven viser en representasjon av den den matematiske modellen som kalles Gauss-modellen. Når sannsynligheter er fordelt på samme måte sier vi at vi har en *normalfordeling* eller en *Gaussfordeling*.

Vi har sett på den binomiske fordelinga. Det å regne ut sannsynligheter for store verdier av $n$ var krevende før i tida. Abraham de Moivre[3] tok i bruk et nytt våpen: funksjonsanalyse eller kalkulus. Han fant at den binomiske fordelinga når $p = 0.5$ var omtrent lik en kontinuerlig tetthetsfunksjon som ser slik ut:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}} = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Nå ser den kanskje ikke så enkel ut, her er det mange symboler, men både $e$ og $\pi$ er konstanter og vet vi $\mu$ og $\sigma$, så har vi en funksjon med variabelen $x$. Vi ser at det er en kontinuerlig sannsynlighetsmodell. Funksjonen gjorde det enklere å gjøre beregninger som omtrent tilsvarte en binomisk fordeling med $p = 0.5$. I dag har vi andre verktøy som gjør det enklere å regne med alle sannsynlighetsfordelinger. Denne funksjonen beskriver normalfordelinga.

**Definisjon 1 Normalfordeling**

En stokastisk variabel $X$ er normalfordelt med forventningsverdi $\mu$ og standardavvik $\sigma$ hvis sannsynlighetstettheten er

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}$$

For å skrive at en stokastisk variabel $X$ er normalfordelt med en forventningsverdi $\mu$ og standardavvik $\sigma$ skriver vi: $X \sim Normal(\mu, \sigma)$ eller bare $X \sim N(\mu, \sigma)$.

Noen velger også å skrive $X \sim Normal(\mu, \sigma^2)$ eller $X \sim N(\mu, \sigma^2)$. Bare vi passer på spiller ikke det noen rolle om det er standardavvik eller variansen som benyttes i skrivemåten.

Legg merke til at vi at $X$ er en kontinuerlig variabel og ikke en diskret variabel.

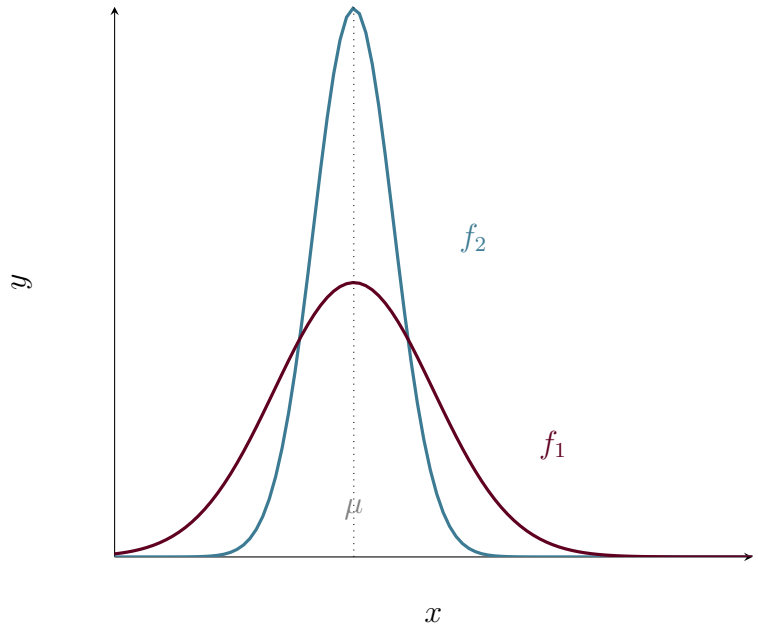
Figur 1.3 viser to grafer med samme $\mu$, men forskjellig $\sigma$. Her har $f_1$ en stor $\sigma$ og $f_2$ en liten.

**Oppgave 2**

Bruk et digitalt verktøy for å tegne grafer som viser normalfordeling. Både GeoGebra og Desmos egner seg godt. Eksperimenter med forskjellige verdier av $\mu$ og $\sigma$.

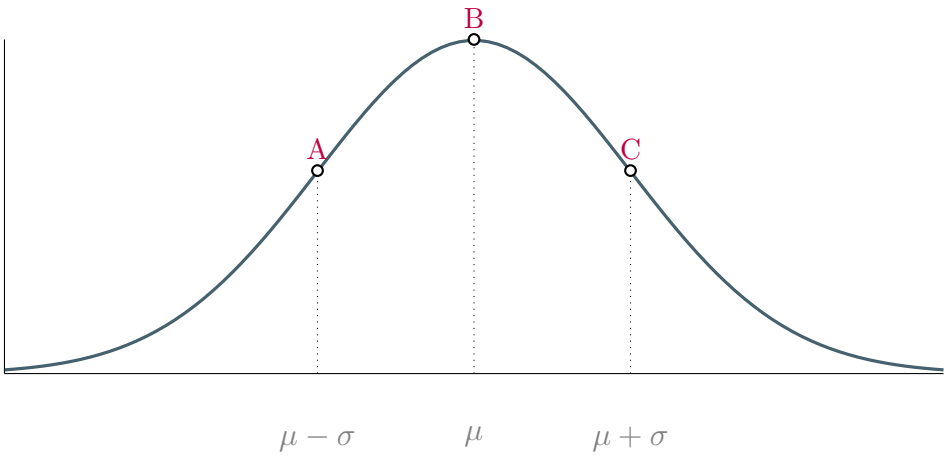
Prøver vi med forskjellige forventningsverdier og standardavvik vil vi se at grafene endrer seg, men bevarer formen som ei klokke. Forventningsverdien flytter grafen langs x-aksen og standardavviket avgjør hvor høy og brei den er. Egenskapene kan beskrives som i figur 1.4 hvor vi ser at vi får et toppunkt for $\mu$. Grafen er symmetrisk om linja for $x = \mu$. Går vi

[3] Abraham de Moivre (1667 - 1754) var en fransk matematiker. Wikipedia: Abraham de Moivre




Figur 1.3: Typisk normalfordeling med samme $\mu$ og forskjellig $\sigma$

ett standardavvik til hver side for linja finner vi vendepunktene for grafen. På figur 1.4 er de markert med $A$ og $C$.




Figur 1.4: Egenskaper ved grafen til normalfordeling

## 1.6 Standardnormalfordeling

Hvis en normalfordeling har forventningsverdi $\mu = 0$ og standardavvik $\sigma = 1$ sier vi at den er standardnormalfordelt. Tidligere var en standardnormalfordeling god å ha siden det gjorde det mulig å slå opp i tabeller for å få utregningene. Har vi allerede en normalfordeling er det ikke noe problem å gjøre den om til en standardnormalfordeling. Det er bare å regne litt om. Har vi en stokastisk variabel $X$ som er normalfordelt: $X \sim Normal(\mu, \sigma)$ kan vi finne en stokastisk variabel $Z$ som er standardnormalfordelt ved

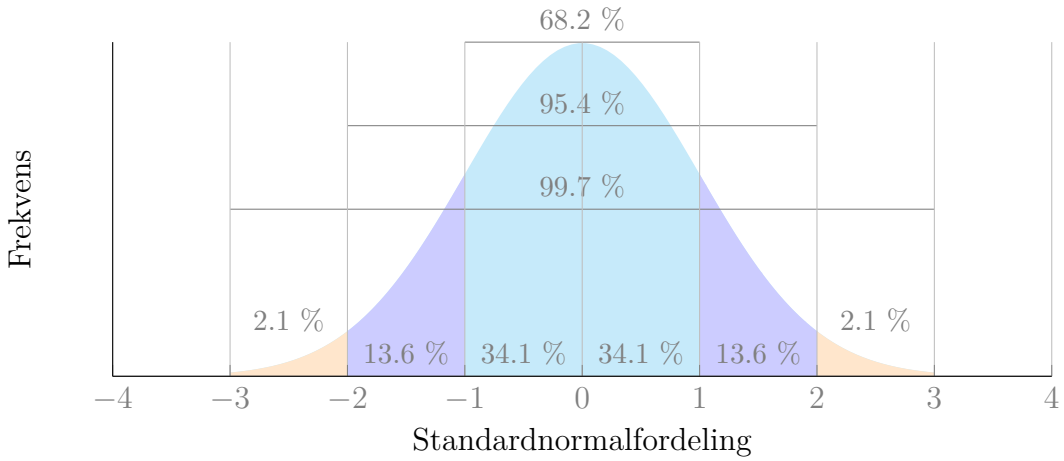
$$Z = \frac{X - \mu}{\sigma}$$

$Z$ vil nå være standardnormalfordelt med forventningsverdi $\mu = 0$ og standardavvik $\sigma = 1$. Vi skriver da

$$Z \sim Normal(0, 1)$$

## 1.7 Sannsynligheter i normalfordelinga

Ser vi på arealet under kurven vil det gi oss sannsynlighetene. Normalfordelinga er symmetrisk og inneholder noen sammenhenger som er interessante. Figur 1.5 viser hvordan sannsynlighetene fordeler seg ut fra standardavvikene i en standardnormalfordeling.




Figur 1.5: Sannsynlighet i normalfordelinga

Figur 1.5 viser en standardnormalfordeling, men gjelder for alle normalfordelinger. Her kan vi se at 95.4 % av arealet under kurven er mellom ett standardavvik under og over forventningsverdien. Går vi to standardavvik i hver retning fra forventningsverdien, vil 68.2 % av hele arealet være med. Hva betyr det? Vet vi forventningsverdi og standardavvik kan vi si noe om grensene for at vi at det er 95.4 % eller 68.2 % sannsynlig for at verdiene er med. Et eksempel kan sikkert klargjøre det bedre.

**Eksempel 1**

Kroppshøydene for norske menn er normalfordelt med $\mu = 179$ cm og $\sigma = 6$ cm. Vi trekker ut en tilfeldig norsk mann. Hvilken kroppshøyde kan vi si med 95.4 % sannsynlighet at denne personen har? Hva med 68.2 % sannsynlighet?

Vi benytter $X$ for kroppshøyden og vet at $X \sim N(179, 6)$. Ut fra egenskapene til normalfordelinga har vi da at grensene blir

| sannsynlighet | nedre grense | øvre grense |
|---|---|---|
| 0.954 | $179 - 2 \cdot 6 = 167$ | $179 + 2 \cdot 6 = 191$ |
| 0.682 | $179 - 1 \cdot 6 = 173$ | $179 + 1 \cdot 6 = 185$ |

Sannsynligheten er 0.954 for at kroppshøyden er i intervallet $[167, 191]$.

Sannsynligheten er 0.682 for at kroppshøyden er i intervallet $[173, 185]$.

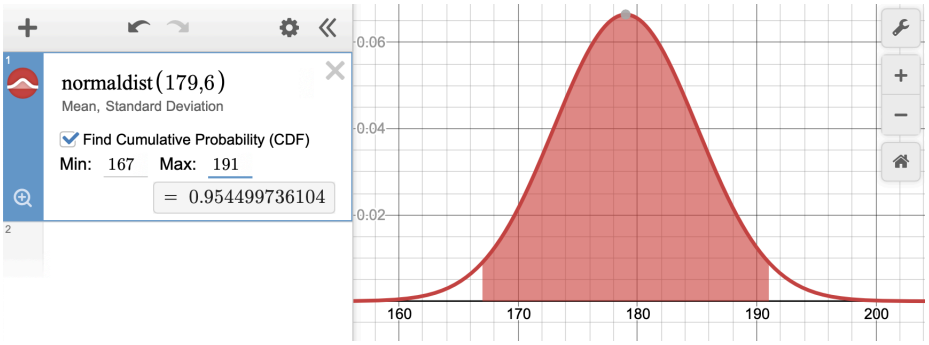
Det samme kan vi skrive sånn

$$P(167 \leq X \leq 191) = 0.954$$
$$P(173 \leq X \leq 185) = 0.682$$

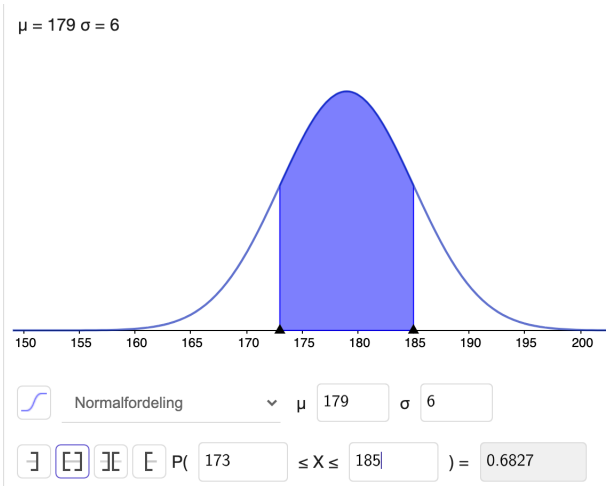
Vi kan regne ut sannsynlighetene for en stokastisk variabel $X$ som er $X \sim N(\mu, \sigma)$ ved å benytte integrasjon

$$P(a \leq X \leq b) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \, dt$$

Dette integralet må vi ha hjelp for å få regna ut siden det ikke kan beregnes eksakt. Heldigvis har vi digitale verktøy til slikt. Figur 1.6 viser bruk av Desmos og figur 1.7 viser GeoGebra.



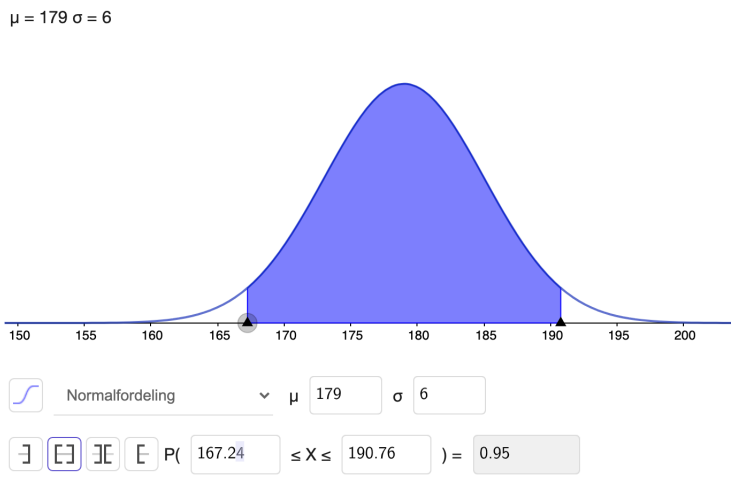

Figur 1.6: Desmos: Normalfordeling




Figur 1.7: GeoGebra

**Spredningsintervall** Vi har sett på hvor sikkert det er at en stokastisk variabel er innafor at gitt intervall. Da sier vi at vi angir et spredningsintervall. Vi fant at det er 95.4 % sikkert at $X$ har en verdi i intervallet $\mu \pm 2 \cdot \sigma$. Arealet avgrensa av disse verdiene utgjør 95.4 % av hele arealet, slik figur 1.9 viser. Ønsker vi å være 99 % sikre må vi utvide arealet til $\mu \pm 2.58 \cdot \sigma$. Vil vi være 99.7 % sikre blir intervallet $\mu \pm 3 \cdot \sigma$. Vi kan observere at verdien vi multipliserer standardavviket med avgjør sannsynligheten. Det er en sammenheng mellom disse to verdiene.

Hva blir intervallet om vi ønsker en sannsynligheten på akkurat 95 % for at kroppshøyden er i intervallet? Vi vet at intervallet må være omtrent $[167, 191]$, så vi kan eksperimentere ved å

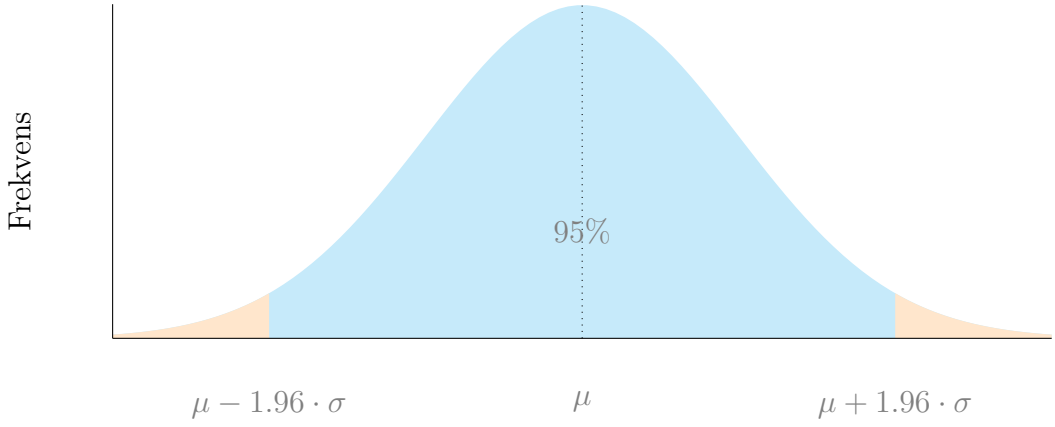



Figur 1.8: GeoGebra

kutte litt på det intervallet. Litt eksperimentering gjør at vi kan komme fram til intervallet $[167.24, 190.76]$. Figur 1.8 viser det.

Da kan vi finne ut hvor langt vi flytter oss til hver side av linja $x = \mu$. Vi får $179 - 167.24 = 11.76$ og $190.76 - 179 = 11.76$. Hvor mange standardavvik blir det? Jo, $11.76/6 = 1.96$. Denne verdien vil bli viktig for oss. Det viser seg at dette gjelder generelt.

$$P(\mu - 1.96 \cdot \sigma \leq X \leq \mu + 1.96 \cdot \sigma) = 0.95$$




Figur 1.9: spredningsintervall for 95 % sannsynlighet

Figur 1.9 viser arealet som gir 95 % sannsynlighet. Den delen av totalarealet som ikke er med utgjør da 5 % og ligger likt fordelt med 2.5 % på hver side.

Ønsker vi et intervall med 99 % sannsynlighet får vi dette intervallet

$$P(\mu - 2.576 \cdot \sigma \leq X \leq \mu + 2.576 \cdot \sigma) = 0.99$$

Vanligvis benytter vi disse verdiene

| $z$ | $P(\mu - z \cdot \sigma \leq X \leq \mu + z \cdot \sigma)$ |
|---|---|
| 1.645 | 0.9 |
| 1.960 | 0.95 |
| 2.326 | 0.98 |
| 2.576 | 0.99 |

Disse verdiene kan vi finne ved å se på standardnormalfordelinga. Kaller vi verdiene for $z$ har vi dette intervallet

$$\mu - z \cdot \sigma \leq X \leq \mu + z \cdot \sigma \tag{1.1}$$

Vi kan få alle normalfordelinger standardnormalfordelt ved å sette $Z = \frac{X-\mu}{\sigma}$. Da har vi

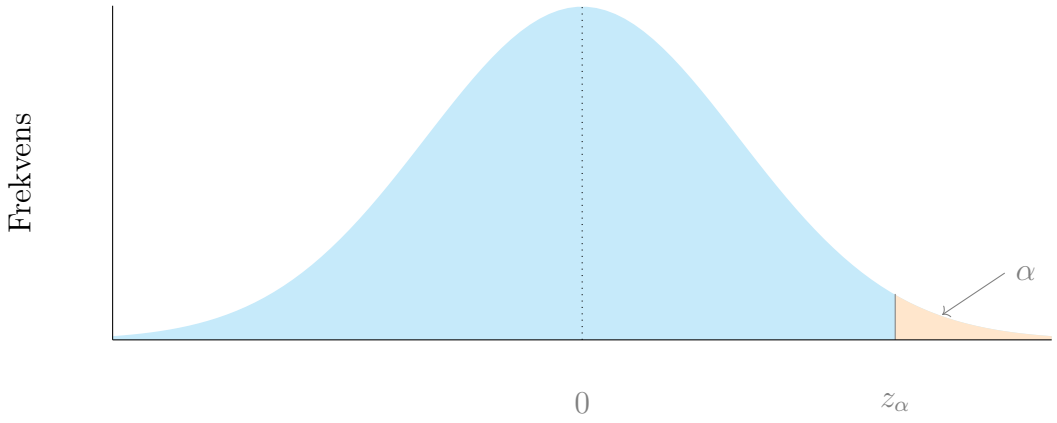
$$Z = \frac{X - \mu}{\sigma} \implies X = \sigma \cdot Z + \mu$$

Setter vi det inn i (1.1) får vi

$$\begin{aligned} \mu - z \cdot \sigma \leq & X \leq \mu + z \cdot \sigma \\ \mu - z \cdot \sigma \leq & \sigma \cdot Z + \mu \leq \mu + z \cdot \sigma \\ -z \leq & Z \leq z \end{aligned}$$

Verdiene vi fant for $z$ ved å eksperimentere kan vi enkelt finne ved regning eller i tabeller. Siden vi ikke har behov for så mange kan vi gå ut fra denne tabellen

| $z$ | $P(-z \leq Z \leq z)$ |
|---|---|
| 1.645 | 0.9 |
| 1.960 | 0.95 |
| 2.326 | 0.98 |
| 2.576 | 0.99 |




Arealet $\alpha$ tabell

| $\alpha$ | $z_\alpha$ |
|---|---|
| 0.100 | 1.282 |
| 0.050 | 1.645 |
| 0.025 | 1.960 |

### Hvorfor er denne så viktig?

Normafordelinga har fått en spesiell plass siden den forekommer så ofte og i så mange sammenhenger. Tidligere var det også sånn at andre fordelinger var vanskelige å regne på, men for normalfordelinga fantes det tabeller som gjorde beregningene mye enklere. Nå har vi andre verktøy som gjør det enklere å regne med all slags sannsynlighetsfordelinger, men det kan være greit å ta med seg både binomisk- og hypergeometriske fordelinger vil kunne være tilnærma normalfordelt under disse betingelsene:

- binomisk fordeling når $np > 5$ og $n(1-p) > 5$
- hypergeometrisk når $rN >> n$ og $np > 5$ og $n(1-p) > 5$. Da er $P = \frac{A}{N}$

**Normalfordeling**

Normalfordelingsfunksjonen er gitt ved:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}} = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Her er $\mu$ forventningsverdien og $\sigma$ standardavviket til fordelinga.
Hvis $X$ er normalfordelt med $\mu$ og $\sigma^2$ og kan skrives slik: $X \sim N(\mu, \sigma^2)$ eller $X \sim N(\mu, \sigma)$

## 1.8 Andre fordelinger

Det fins mange sannsynlighetsfordelinger. De som er presentert så langt er vanligst å ha med i skolematematikken, men det kan være greit å være kjent med noen av de andre også.

### 1.8.1 Geometrisk fordeling

Den geometriske fordelingen knytter seg til binomiske forsøksserier. La oss si at vi kaster terning igjen og registrerer om vi får en sekser eller ikke. Det kjenner vi nå som et binomisk forsøk, men nå stiller vi et annet spørsmål. Hva er sannsynligheten for at vi får en sekser i kast nummer $x$?

Vi kaster terningen og er bare interessert i å finne sannsynligheten for å få sekser i kast $x$

| kast | 1 | 2 | 3 | 4 | 5 | $\cdots$ | x |
|---|---|---|---|---|---|---|---|
| sekser | Nei | Nei | Nei | Nei | Nei | $\cdots$ | Ja |

Dette kan vi regne ut. Vi får

$$(1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdots p = (1-p)^{x-1} \cdot p = p \cdot (1-p)^{x-1}$$

Dette kan vi skrive som

$$P(X = x) = p \cdot (1-p)^{x-1}$$

Hvor mange ganger må vi kaste terningen før en sekser dukker opp?

**Geometrisk fordeling**

$X$ er geomterisk fordelt med sannsynligheten $p$

$$P(X = x) = p \cdot (1 - p)^{x-1}$$

Forventningsverdi og varians

$$E(Y) = \frac{1}{p}$$
$$Var(X) = \frac{1 - p}{p^2}$$

# 2 Estimering og konfidensintervall

## 2.1 Estimering – Hva er det?

Estimering er å prøve å finne en sannsynlig mengde eller størrelse for et eller annet. I statistikken må vi gjøre noe mer enn å bare tippe. Vi må forsøke å finne estimatet på en så god måte som mulig. Hvordan vi gjør det skal vi nå se litt på.

**Oppgave 3**

Finn eksempler på estimat og estimeringer som både du gjør deg og fra nyhetene. Hvordan har en kommet fram til estimatene?

### Punktestimat og intervallestimat

Spørreundersøkelser er populært. Et utvalg av personer, respondenter, gir svarene sine og så vil de som har laget undersøkelsen kunne telle opp svarene. Det kan være at 37 % av respondentene forteller at de skal stemme på en bestemt kandidat i et valg. Det er da gjort et *punktestimat*. Hva det virkelige resultatet blir vet vi ikke før valget er over. Her er noen definisjoner det kan være greit å ha med seg.

**Definisjon 2 Estimator**

En estimator er en stokastisk variabel som kan benyttes for å gi et estimat for den verdien vi prøver å anslå.

**Definisjon 3 Punktestimat**

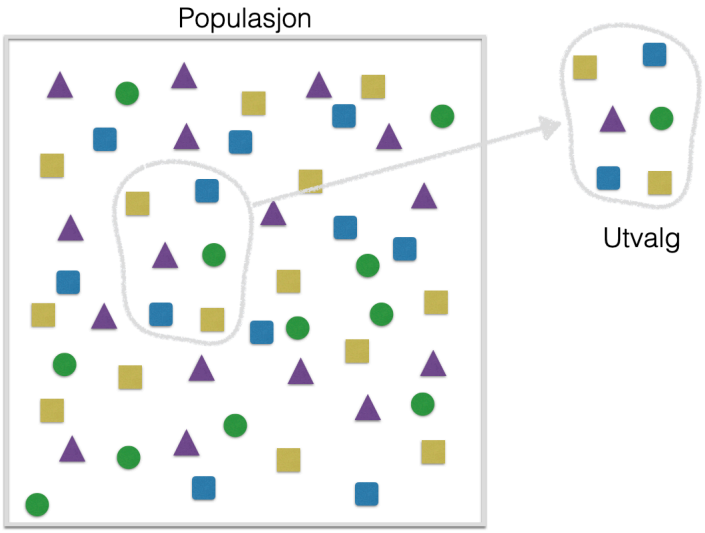
Verdien til en estimator kalles for et punktestimat.

For å komme fram til et punktestimat er det av flere grunner ikke mulig å undersøke hva alle mener. Det kan bli for dyrt, praktisk umulig, eller det tar for lang tid. Forskerne plukker derfor ut et utvalg av hele populasjonen og undersøker utvalget.

Basert på utvalget blir det gjort et forsøk på å beskrive egenskaper ved hele populasjonen: Et estimat. Her er det viktig å legge merke til at en klar feilkilde kan være at utvalget ikke er representativt for hele populasjonen.

Kanskje ønsker den som gjør en undersøkelse en grundigere vurdering av punktestimatet? Med statistikken som verktøy er det mulig å si noe om hvor sikkert dette resultatet er. Statistikerne vil for eksempel kunne si at de er 95 % sikre på at kandidaten får mellom 35% og 39 % av stemmene. Da har de gitt et *intervallestimat* med 95 % sannsynlighet.

Vi har sett på sansynlighetsmodeller hvor vi har en kjent forventningsverdi og et kjent standardavvik. Ofte er det sånn at vi ikke kjenner disse verdiene og ønsker å finne dem. Da må vi estimere dem. I punktestimering anslår vi en verdi. Et konfidensintervall gir et intervall hvor vi med en viss sikkerhet kan si at verdien ligger.




Figur 2.1: Utvalg

## En hatt gjør forskjellen

Det kan benyttes forskjellige symboler for estimatorer. En vanlig symbolbruk er å merke med en hatt.

| | populasjonen | estimator |
|---|---|---|
| standardavvik | $\sigma$ | $\hat{\sigma}$ |
| varians | $\sigma^2$ | $\hat{\sigma}^2$ |
| forventingsverdi | $\mu$ | $\hat{\mu}$ |
| sannsynlighet | $p$ | $\hat{p}$ |

## Hvor sikre kan vi være?

Ved å utføre en estimering finner vi et estimat. Det kan være en valgprognose hvor vi finner en sannsynlighet for at en person skal stemme på et bestemt parti, eller det kan være et estimat for en gjennomsnittslengde i en dyrepopulasjon. Uansett hva det er kan vi ikke være helt sikre på estimatet vi kommer fram til. Det vi kan si noe om er hvor usikre vi er. Har vi en estimator, $\hat{\mu}$, for forventningsverdien, $\mu$, i en populasjon vil den ha en sannsynlighetsfordeling. Som regel kan vi argumentere for at $\hat{\mu}$ er normalfordelt eller tilnærma normalfordelt. Med den antakelsen, og det vi vet om normalfordeling, kan vi uttale oss om hvor sikre vi kan være på estimatet vårt.

Standardavviket til estimatoren kaller vi *standardfeilen* og vi skriver det som $SE(\hat{\mu})$. Figur 2.2 viser intervallet hvor vi kan si at $\hat{\mu}$ ligger i med 95 % sikkerhet.

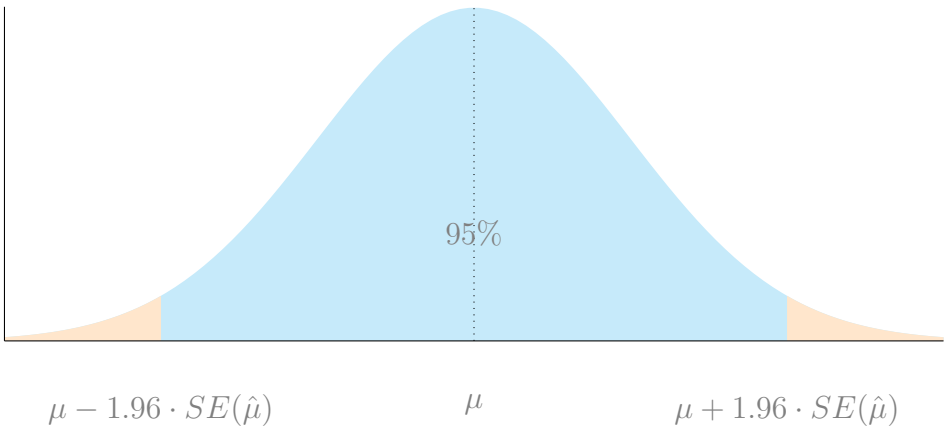
Vi kaller dette for et *konfidensintervall* og kan skrive det som intervallet:

$$[\hat{\mu} - 1.96 \cdot SE(\hat{\mu}) \quad , \quad \hat{\mu} + 1.96 \cdot SE(\hat{\mu})]$$

Mer generelt

$$[\hat{\mu} - z \cdot SE(\hat{\mu}) \quad , \quad \hat{\mu} + z \cdot SE(\hat{\mu})]$$

hvor $z$ er gitt i denne tabellen

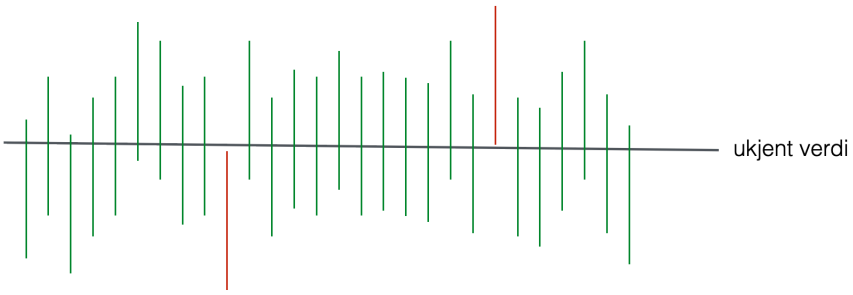



Figur 2.2: Konfidensintervall med 95 % sannsynlighet

| $z$ | Sannsynlighet |
|---|---|
| 1.645 | 0.9 |
| 1.960 | 0.95 |
| 2.326 | 0.98 |
| 2.576 | 0.99 |

Verdiene er gitt ut fra normalfordelingen.

Konfidensintervallet gir oss et intervall hvor vi med en viss sikkerhet kan si at estimatet befinner seg. Det vil bety at også konfidensintervallet vil være stokastisk. Gjør vi flere undersøkelser vil vi kunne finne andre estimat og andre standardfeil slik at konfidensintervallet endrer seg. Figur 2.9 viser flere beregna konfidensintervall og hvordan den ukjente verdien, som skal estimeres, vil ligge innafor konfidensintervallene i de aller fleste tilfellene.




Figur 2.3: Flere konfidensintervaller

**Konfidensintervall**

Hvis vi har en estimator $\theta$ som er normalfordelt har den konfidensintervallet

$$[\theta - z \cdot SE(\theta) \quad , \quad \theta + z \cdot SE(\theta)]$$

Verdien av $z$ bestemmer bredden av konfidensintervallet ved

| $z$ | Sannsynlighet |
|---|---|
| 1.645 | 0.9 |
| 1.960 | 0.95 |
| 2.326 | 0.98 |
| 2.576 | 0.99 |

## 2.2 Estimering av standardavvik i en populasjon

Som regel kjenner vi ikke standardavvik i en populasjon, men ofte krever videre utforsking at vi vet noe om spredningen i populasjonen. Da må vi prøve å finne et estimat for standardavviket $\sigma$ eller variansen $\sigma^2$. Spørsmålet blir hvordan vi kan gjøre det. Kan vi ikke undersøke hele populasjonen blir svaret at vi må ta noen stikkprøver.

### Et eksempel: Kroppshøyder

Vi ser på et eksempel hvor vi har ei liste med høydene til 200 gutter på en videregående skole.

171 174 167 187 174 169 182 180 186 177 178 190 180 176 183 180 169 187 170 181
178 175 181 183 184 179 195 176 176 179 190 165 178 182 179 186 188 170 191 183
179 188 178 177 179 175 186 184 187 174 171 181 171 189 183 174 193 192 181 184
177 194 185 180 173 185 189 189 172 185 180 185 182 173 172 186 177 175 176 184
191 182 172 183 175 184 166 178 176 180 181 188 177 182 168 173 180 181 188 177
182 168 173 182 180 186 177 178 190 180 176 174 167 187 174 169 182 180 186 177
178 171 174 167 187 174 169 182 180 195 176 176 179 190 165 178 182 179 186 188
170 191 183 179 175 186 184 187 174 171 181 171 189 183 174 174 169 182 180 186
177 178 171 174 167 187 174 175 176 184 191 182 172 183 175 184 166 178 176 180
181 169 178 194 175 184 166 178 176 180 172 185 164 191 178 176 180 164 179 181

Verdiene er delt i et regneark på denne adressen : https://bit.ly/3D8UFif og som ei fil her https://bit.ly/3NhEK5x.

Nå kjenner vi hele populasjonen og vi kan finne de statistiske verdiene vi vil. Enten vi gjør det for hånd, eller benytter et verktøy, skal vi få disse verdiene

$$\mu = \bar{X} = 179.2$$
$$\sigma = 6.843$$

**Python**

**Python** kan gjøre dette for oss. Jeg har alle høydene, separert med et semikolon, i ei fil og leser den inn. Måten jeg leser inn på gjør at høydene legges inn i en matrise, eller array, så jeg må konvertere til den til ei liste.

Etter at vi har fått høydene inn i ei liste kan vi benytte kommandoen `mean` til å finne gjennomsnittet og kommandoen `pstdev` til å finne $\sigma$. Vi får de samme verdiene som med andre verktøy.

Programkode 2.1 benytter modulen `statistics` for å finne de statistiske verdiene. Alternative moduler må benyttes hvis en benytter versjoner før 3.8.

```
1  import numpy as np
2  import statistics as st
3
4  data = np.loadtxt("hoyder.txt", delimiter = ";")
5  hoyder = data.tolist()
6
7  gjennomsnitt = st.mean(hoyder)
8  standardavvik = st.pstdev(hoyder)
9
10 print("Gjennomsnitt er: ", gjennomsnitt)
11 print("Standardavvik er: ", standardavvik)
```

Programkode 2.1: Statistiske verdier

```
Gjennomsnitt er:  179.2
Standardavvik er:  6.84324484437025
```

Da har vi fasiten! Vi kjenner både $\mu$ og $\sigma$ i populasjonen. Det er vanligvis verdiene vi er ute etter å estimere. Nå later vi som om vi ikke kjenner noen av dem og forsøker å finne et estimat for $\sigma$.

Nå kan vi ta noen stikkprøver og for at det ikke skal bli for mye å holde styr på tar vi ti stikkprøver med ti verdier i hver. Så få i hver stikkprøve, og så få stikkprøver, er bare valgt for at det skal være mulig å vise utregningene. I praksis bør vi ta med så mange som mulig i hver stikkprøve og gjerne ta så mange stikkprøver som vi kan.

Et mulig resultat kan du se i tabellen hvor gjennomsnittet i hver stikkprøve er funnet.

| utvalg | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 174 | 182 | 168 | 188 | 182 | 186 | 180 | 179 | 177 | 184 |
| 2 | 177 | 172 | 184 | 183 | 172 | 191 | 171 | 183 | 177 | 179 |
| 3 | 180 | 183 | 174 | 194 | 177 | 175 | 184 | 174 | 173 | 195 |
| 4 | 181 | 164 | 176 | 166 | 182 | 178 | 186 | 180 | 174 | 165 |
| 5 | 185 | 180 | 184 | 165 | 176 | 180 | 187 | 181 | 169 | 174 |
| 6 | 166 | 180 | 182 | 175 | 191 | 177 | 181 | 178 | 191 | 179 |
| 7 | 187 | 181 | 183 | 175 | 173 | 176 | 183 | 180 | 184 | 184 |
| 8 | 195 | 184 | 178 | 190 | 178 | 175 | 184 | 174 | 175 | 193 |
| 9 | 191 | 189 | 171 | 181 | 182 | 182 | 180 | 173 | 174 | 182 |
| 10 | 184 | 181 | 176 | 165 | 179 | 167 | 177 | 171 | 178 | 186 |

Vi kan se nærmere på utvalg 1 og finne gjennomsnittshøyde og standardavvik i det.

Gjennomsnittshøyden

$$\bar{X} = \frac{174 + 182 + 168 + 188 + 182 + 186 + 180 + 179 + 177 + 184}{10} = 180.0$$

Standardavviket i utvalg 1 finner vi ved

$$\begin{aligned}
\sigma^2 &= \frac{\sum(x_i - \bar{X})^2}{n} \\
&= \frac{(174 - 180.0)^2 + (182 - 180.0)^2 + \ldots (177 - 180.0)^2 + (184 - 180.0)^2}{10} \\
&= \frac{314}{10} \\
\sigma &= \sqrt{\frac{314}{10}} = 5.60357029044876 \approx 5.60
\end{aligned}$$

Da har vi har et estimat for $\sigma$ i populasjonen. Som vi ser skiller det seg en del fra det vi nå vet er fasiten.

Nå kan statistikerne fortelle oss at vi bør gjøre en liten justering når vi skal estimere ut fra et slikt utvalg. Vi bør ikke dele på $n$, men justere litt ved å heller dele på $n - 1$. Vi prøver det og bruker $s$ for denne måten å regne på.

$$\begin{aligned}
s^2 &= \frac{\sum(x_i - \overline{X})^2}{n - 1} \\
&= \frac{(174 - 180.0)^2 + (182 - 180.0)^2 + \ldots (177 - 180.0)^2 + (184 - 180.0)^2}{9} \\
&= \frac{314}{9} \\
s &= \sqrt{\frac{314}{9}} = 5.90668171555645 \approx 5.91
\end{aligned}$$

Ofte vil vi finne $\sigma_{n-1}$ brukt for denne verdien.

Med denne justeringen ser vi at svaret kom noe nærmere det vi vet er fasiten.

La oss gjøre det for alle stikkprøvene og se hva som skjer

| utvalg | | | | | | | | | | | $s = \sigma_{n-1}$ | $\sigma_n$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 174 | 182 | 168 | 188 | 182 | 186 | 180 | 179 | 177 | 184 | 5.91 | 5.60 |
| 2 | 177 | 172 | 184 | 183 | 172 | 191 | 171 | 183 | 177 | 179 | 6.42 | 6.09 |
| 3 | 180 | 183 | 174 | 194 | 177 | 175 | 184 | 174 | 173 | 195 | 8.11 | 7.70 |
| 4 | 181 | 164 | 176 | 166 | 182 | 178 | 186 | 180 | 174 | 165 | 7.77 | 7.37 |
| 5 | 185 | 180 | 184 | 165 | 176 | 180 | 187 | 181 | 169 | 174 | 7.09 | 6.73 |
| 6 | 166 | 180 | 182 | 175 | 191 | 177 | 181 | 178 | 191 | 179 | 7.32 | 6.94 |
| 7 | 187 | 181 | 183 | 175 | 173 | 176 | 183 | 180 | 184 | 184 | 4.55 | 4.32 |
| 8 | 195 | 184 | 178 | 190 | 178 | 175 | 184 | 174 | 175 | 193 | 7.83 | 7.43 |
| 9 | 191 | 189 | 171 | 181 | 182 | 182 | 180 | 173 | 174 | 182 | 6.49 | 6.15 |
| 10 | 184 | 181 | 176 | 165 | 179 | 167 | 177 | 171 | 178 | 186 | 6.90 | 6.55 |
| snitt | | | | | | | | | | | 6.84 | 6.49 |

I tabellen er det regnet ut standardavvik ved å dele på henholdsvis $n-1$ og $n$. Gjennomsnittet av alle disse standardavvikene viser at vi får et meget godt estimat. Gjennomsnittet av alle standardavvikene beregnet med $n-1$ gir $\bar{s} = 6.84$ og det er jo akkurat det standardavviket som er i populasjonen. Perfekt estimat! Samtidig ser vi at gjennomsnittet av standardavvikene beregna på vanlig måte ikke gir et like godt estimat. Rådet fra statistikken ser ut til å være godt. Vi bør benytte $s$ for å finne et *forventningsrett estimat* for standardavviket i poplulasjonen.

**Python**

Vi kan la Python gjøre det samme for oss.

```
import numpy as np
import statistics as st
import random as rd
# Leser inn høyder og legger i liste
data = np.loadtxt("hoyder.txt", delimiter = ";", dtype = "int")
hoyder = data.tolist()
# Utvalg
ant_utvalg = 10
ant_hoyder = 10
# Lister det blir bruk for
gjsn_liste = []
s_n_liste = []
s_n_1_liste = []
# Tar utvalg og skriver ut
print(f"Utvalg  s          sigma")
for i in  range (ant_utvalg):
    utvalg = rd.sample(hoyder, ant_hoyder)
    s_n = st.pstdev(utvalg)
    s_n_liste.append(s_n)
    s_n_1 = st.stdev(utvalg)
    s_n_1_liste.append(s_n_1)
    print(f"{i + 1:2d} {s_n:10.3f} {s_n_1:10.3f}")

print(f"snitt {st.mean(s_n_1_liste):7.3f} {st.mean(s_n_liste):10.3f}")
```

Programkode 2.2: Estimering med Python

```
Utvalg  s          sigma
 1      5.779      6.092
 2      5.621      5.925
 3      3.494      3.683
 4      8.987      9.473
 5      6.655      7.015
 6      9.421      9.931
 7      6.135      6.467
 8      6.715      7.078
 9      6.818      7.187
10      5.192      5.473
snitt   6.832      6.482
```

I denne koden benyttes formatteringer i utskrifta som bare fungerer med Python 3. Det er bare gjort for å pynte litt og kan godt unngås. Programkoden og fila med høydene kan hentes her: Gist

## Forventningsrett estimat

**Punktestimatorer**

Vi har en populasjon med ukjent gjennomsnitt $\mu$ og ukjent varians $\sigma^2$. Vi gjør $n$ uavhengige observasjoner $X_1, X_2, \cdots, X_n$. Forventningsrette estimatorer for gjennomsnittet og variansen i populasjonen vil da være

$$\hat{\mu} = \bar{X}$$
$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

Hvordan kan vi finne standardavvikene uten å regne ut? Alle hjelpemidler har muligheten for det. Vi kan se på noen.

**Python** krever at vi benytter modulen `statistics`. Da har vi disse kommandoene:

| | | |
|---|---|---|
| `pstdev` | *population standard deviation of data* | $s$ |
| `stdev` | *standard deviation of data* | $\sigma$ |

**Geogebra** har disse to variantene for standardavvik `Standardavvik[Liste]` og `UtvalgStandardavvik[Liste]`. Den første gir oss vanlig standardavvik og den siste det forventningsrette standardavviket.

**Regneark** kan også benyttes. Her kan det variere, men to vanlige kommandoer er `STDAVP` for vanlig standardavvik og `STDAVVIKA` for det forventningsrette.

**R** er et programmeringsspråk for statistikk. Kommandoen for det forventningsrette standardavviket er `sd()`. Da benyttes $n-1$ og vi kan regne om for å få det vanlige standardavviket på

denne måten:

$$\sigma = sd() \cdot \sqrt{\frac{n-1}{n}}$$

## 2.3 Sentralgrenseteoremet

Sentralgrenseteoremet, eller sentralgrensesetningen, er svært viktig for statistikken. Det forteller at tar vi flere stikkprøver i en populasjon og observerer en gjennomsnittsverdi, så vil de observerte gjennomsnittsverdiene være normalfordelt. Kjenner vi standardavviket i populasjon kan vi også vite standardavviket i den normalfordelingen. Egentlig er dette ganske utrolig, men det kan bevises. For oss rekker det å se på et eksempel.

Sentralgrenseteoremet defineres som varianter av de to definisjonene som følger.

**Teorem 1 Sentralgrenseteoremet**

La $X$ være en stokastisk variabel med forventningsverdi $\mu$ og standardavviket $\sigma$. La $\bar{X}$ være gjennomsnittet av $X$ i et utvalg med $n$ elementer. Da er $\bar{X}$ tilnærma normalfordelt. Desto større $n$ er, jo bedre tilnærming. $\bar{X}$ har forventningsverdien $\mu_{\bar{X}} = \mu$ og standardavviket $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

**Teorem 2 Sentralgrenseteoremet**

En populasjon har gjennomsnittsverdi $\mu$ og varians $\sigma^2$. Vi trekker uendelig mange stikkprøver fra populasjonen. Hver stikkprøve er på $n$ uavhengige observasjoner. La $\overline{X}$ være gjennomsnittet i hver av stikkprøvene. Dersom $n$ er tilstrekkelig stor vil $\overline{X}$ være tilnærma normalfordelt $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ og $\bar{X}$ har forventningsverdien $\mu_{\bar{X}} = \mu$ og standardavviket $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

### Hva forteller sentralgrenseteoremet?

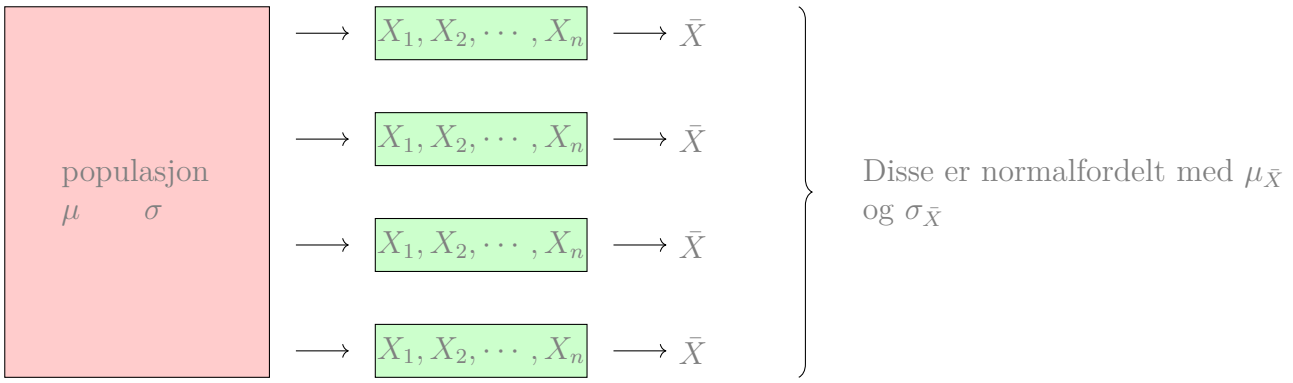
Vi tenker oss en stokastisk variabel, $X$, med en eller annen sannsynlighetsfordeling. Forventningsverdien kaller vi $\mu$ og standardavvik $\sigma$.

$X$ kan være lengden av sild, massen til vågehval eller hva du vil. Hvordan fordelingen av $X$ er spiller mindre rolle.

Vi tar et tilfeldig utvalg av populasjonen og finner gjennomsnittet $\bar{X}$. Dette gjentar vi flere ganger. Alle de $\bar{X}$ vi finner vil da være normalfordelt med forventningsverdi $\mu_{\bar{X}}$ og standardavvik $\sigma_{\bar{X}}$. Her blir det en del symboler, men husk at indeksene som er brukt viser at dette er forventningsverdien og standardavviket til $\bar{X}$. Det er gjort for å ikke blande disse med de andre. I hver stikkprøve er det $n$ elementer.

Vi kan illustrere sentralgrenseteoremet som i figur 2.4.

Vi tar noen stikkprøver fra populasjonen og henter ut verdiene $X_1, X_2, \cdots, X_n$. $\bar{X}$ er gjennomsnittet av verdiene. Gjentar vi det samme flere ganger vil vi få flere $\bar{X}$. Sentralgrenseteoremet forteller oss da at $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Det gir oss disse sammenhengene




Figur 2.4: Stikkprøver fra en populasjon

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Dette gjelder nesten uansett hvilken fordeling det er i populasjonen. Kaller vi fordelinga i populasjonen for $p$ og fordelinga av $\bar{X}$ for $q$ har vi at

$q$ er normalfordelt hvis $p$ er en normalfordelt

$q$ er tilnærmet en normalfordelt hvis $p$ er

- symmetrisk
- ikke har for lange haler
- $n \geq 30$

## Simulering av sentralgrenseteoremet

At sentralgrenseteoremet stemmer kan bevises, men det ligger utafor det vi skal ta for oss. Det vi kan gjøre er å se på et tilfelle der vi simulerer og ser at vi får de resultatene som forventes.

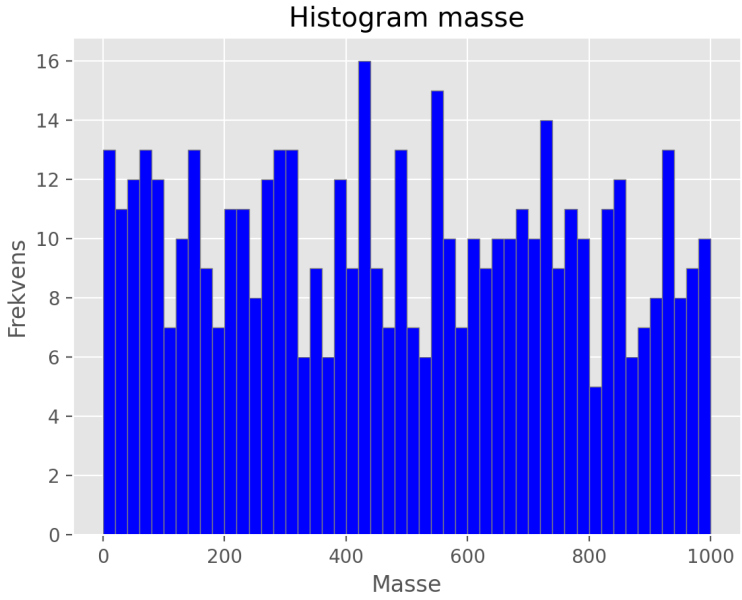
**Fisk i et vann** Vi tenker oss et vann med et antall fisk og at massen [1] til fiskene varierer mellom 1 g og 1 kg. En slik populasjon kan vi simulere med Python. Vi kan også tegne et histogram som viser fordelingen. Programkode 2.3 viser hvordan vi lager 500 fiskemasser og legger de i ei liste. Så tegner vi et histogram.

```
1  import random as rd
2  import matplotlib.pyplot as plt
3  antall = 500
4  # lager liste med fiskemasser
5  masse = [rd.randint(1,1001) for i in  range(antall)]
6  # tegner histogram
7  plt.title("Histogram masse")
8  plt.xlabel("Masse")
9  plt.ylabel("Frekvens")
10 plt.hist(masse, bins = 50,  range = (0,1000))
11 plt.show()
```

[1] masse er det som populært kalles vekta

Programkode 2.3: Fiskemasser i et vann

Histogrammet vil endre seg hver gang vi kjører programmet, men figur 2.5 viser et resultatet.




Figur 2.5: Histogram som viser fiskemassene

Vi kan se at massene er ganske jevnt, og tilfeldig, fordelt mellom 1 g og 1000 g. For å finne gjennomsnittsmasse og standardavvik i populasjonen kan vi legge til to kommandoer i programmet vårt, slik som i programkode 2.4. Det krever at vi importerer modulen `statistics` med kommandoen `import statistics as st`.

```
1 # Statistiske data hele populasjonen
2 print("Statistiske data hele populasjonen")
3 print("Gjennomsnitt er  :", st.mean(masse))
4 print("Standardavvik er :", st.pstdev(masse))
```

Programkode 2.4: Statistiske data i populasjonen

Det kan gi dette resultatet

```
Statistiske data hele populasjonen
Gjennomsnitt er  : 496.574
Standardavvik er : 280.0366128276801
```

Kommandoen `pstdev` står for *population standard deviation of data* og gir den vanlige måten å regne ut standardavvik på.

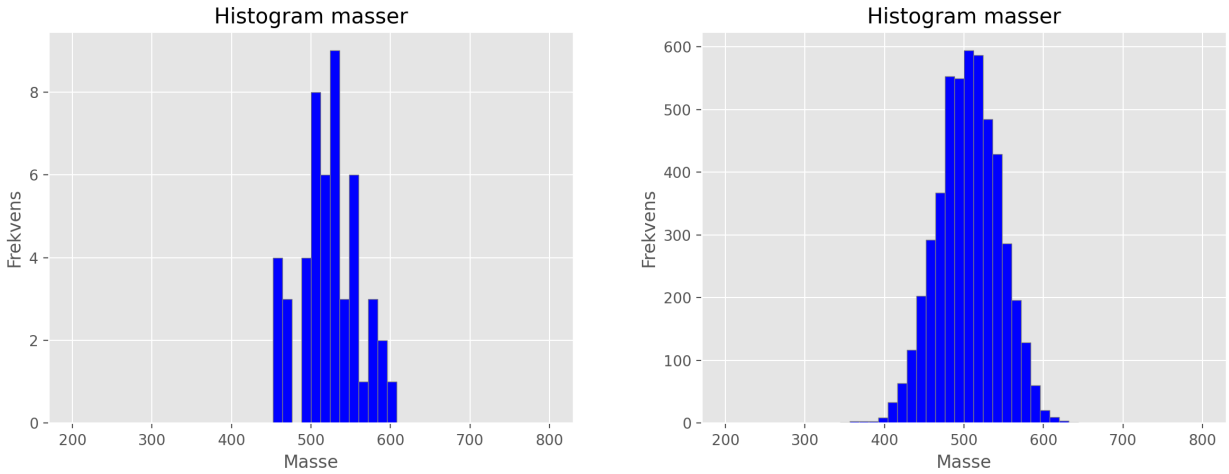
Nå skal vi ta noen stikkprøver. Vi kan ta stikkprøver fra populasjonen vi har fått laget ved å legge til koden i programkode 2.5.

```
1 for i in  range(ant_stikkprover):
2     utvalg = rd.sample(masse, ant_i_utvalg)
3     gjennomsnitt = st.mean(utvalg)
```

```
4     stikkprover.append(gjennomsnitt)
```

Programkode 2.5: Vi tar stikkprøver

Da får vi gjennomsnittet av hver stikkprøve lagt inn i lista `stikkprover`. Nå kan vi tegne et histogram av verdiene i den lista og få noe som likner på figur 2.6.




(a) 50 stikkprøver med 25 i hver

(b) 5000 stikkprøver med 50 i hver

Figur 2.6: Fordelingen av gjennomsnittsverdiene i stikkprøvene

I figure 2.6a ble det kanskje ikke helt normalfordelt, men det er ikke langt unna. Her hadde jeg tatt 50 stikkprøver med 25 i hver. Vi kan i alle fall se at fordelingen likner på en normalfordeling og er en helt annen fordeling enn den i populasjonen. Øker vi antall stikkprøver til 5000 og antallet i hver stikkprøve til 50, blir resultatet slik som i figur 2.6b. Nå begynner det å likne en normalfordeling.

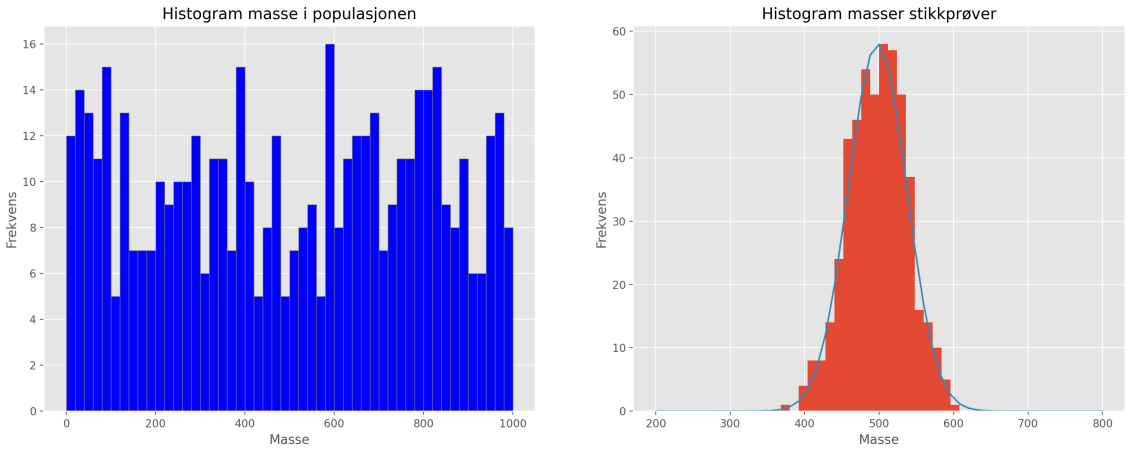
Dette viser sentralgrenseteoremet. I utgangspunktet har vi en nesten uniform fordeling av massene med $\mu = 497$ og $\sigma = 280$. Ved å ta stikkprøver fra denne populasjonen viser simuleringen at vi får en $\bar{X}$ som ser normalfordelt ut Vi har sett sentralgrenseteoremet simulert!

Programkode 2.6 viser et helt program for simulering av sentralgrenseteoremet. Det er noe utvidet og tegner også opp en gauss-kurve over gjennomsnittsmassene i stikkprøvene. Setter vi verdiene slik som i programkoden, vil et typisk resultat av programmet være histogrammene i figur 2.7. Her kan vi se fordelingen av fiskemassene i populasjonen og i stikkprøvene. Samtidig er det vist en gauss-kurve. Som vi kan se stemmer den godt overens med fordelingen i stikkprøvene.

Programmet skrev så ut dette

```
Statistiske data hele populasjonen
Gjennomsnitt er  : 538.81
Standardavvik er : 290.21298713186496
Statistiske data for stikkprøvene
Gjennomsnitt er  : 538.71728
Standardavvik er : 38.29254436834408
Beregna standardavvik:  41.04231423786919
```

Denne simuleringen viser at verdiene i populasjonen ble




Figur 2.7: Histogram

$$\mu = 538.8$$
$$\sigma = 290.2$$

Sentralgrenseteoremet for teller at $\bar{X}$ skal være normalfordelt med

$$\mu_{\bar{X}} = \mu = 538.8$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{290.2}{\sqrt{50}} = 41.04$$

Når vi beregne $\mu_s$ og standardavvik $\sigma_s$ i de stikkprøvene som er tatt (indeksen s er der for å vise at dette er i stikkprøvene) får vi

$$\mu_s = 538.7$$
$$\sigma_s = 38.2$$

Det er ikke så langt unna de verdiene som er beregna.

I statistikken benytter vi sentralgrenseteoremet i flere sammenhenger. At vi vet at stikkprøvegjennomsnittene er normaltfordelt gjør at vi kan fortelle noe om hvor sikker en stikkprøve er. For å gjøre det må vi vite noe om populasjonen. Det er to problem: Vi må kjenne standardavviket i populasjonen for å vite hvordan stikkprøvene er normalfordelt og vi må ha ganske mange i hver stikkprøve. Hvis standardavviket i populasjonen er ukjent, og det er det vanligvis, må vi estimere det.

## Hele programmet litt utvidet

Her er ei lenke til programkoden: programmet på gist.

```
import random as rd
import statistics as st
import matplotlib.pyplot as plt
import numpy as np
antall = 500
ant_i_utvalg = 50
ant_stikkprover = 500
stikkprover = []

# definerer funksjon for gauss-fordelingen av massene
def gauss(x,m,s,a):
    g = a*np.exp(-(x-m)**2/(2*s**2))
    return g

# lager liste med fiskemasser
masse = [rd.randint(1,1001) for i in  range(antall)]

# tar stikkprøver og finner statistiske data
for i in  range(ant_stikkprover):
    utvalg = rd.sample(masse, ant_i_utvalg)
    gjennomsnitt = st.mean(utvalg)
    stikkprover.append(gjennomsnitt)

# tegner histogram
plt.style.use('ggplot')
x = np.linspace(200,800,500)
# hele populasjonen
plt.subplot(1,2,1)
plt.title("Histogram høyder i populasjonen")
plt.xlabel("Høyder")
plt.ylabel("Frekvens")
plt.hist(masse, bins = 50,  range = (0,1000))
# stikkprøvene
plt.subplot(1,2,2)
plt.title("Histogram høyder stikkprøver")
plt.xlabel("Høyder")
plt.ylabel("Frekvens")
y, x,_= plt.hist(stikkprover, bins = 50,  range = (200,800))
# tegner gauss-fordelingen
plt.plot(x,gauss(x,st.mean(stikkprover),st.pstdev(stikkprover),y. max()))
plt.show()

# Statistiske data hele populasjonen
print("Statistiske data hele populasjonen")
print("Gjennomsnitt er  :", st.mean(masse))
print("Standardavvik er :", st.pstdev(masse))

#Statistiske data for stikkprøvene
print("Statistiske data for stikkprøvene")
print("Gjennomsnitt er  :", st.mean(stikkprover))
print("Standardavvik er :", st.pstdev(stikkprover))
print("Beregna standardavvik: ", st.pstdev(masse)/ant_i_utvalg**0.5)
```

Programkode 2.6: Simulering av sentralgrenseteoret

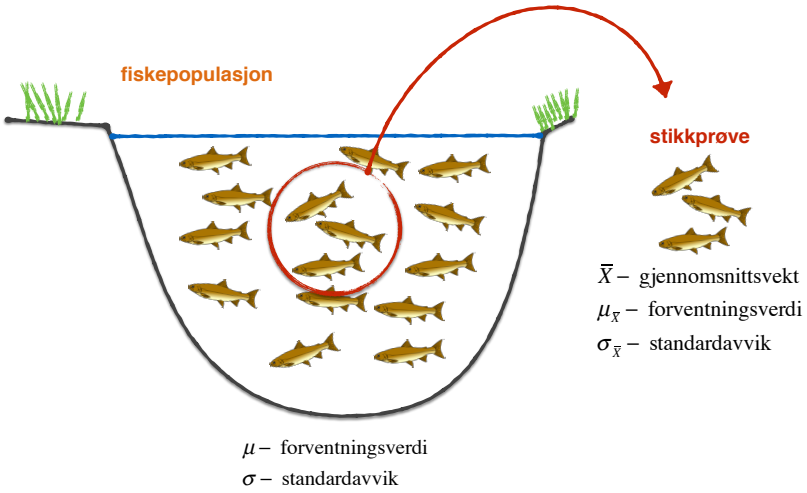
## 2.4 Punktestimering av gjennomsnitt i en populasjon

### En undersøkelse av en fiskepopulasjon

Vi kan ta utgangspunkt i fiskene vi allerede har sett på og ser på hvordan vi ville brukt noe slikt i en praktisk situasjon hvor vi ønsker å undersøke hvordan vekta fordeler seg hos fjellørret i fjellvannet Bratjern. De fiskene som er i vannet vet vi ikke så mye om. Den sikreste metoden vil være å plukke ut hver eneste fisk og legge den på vekta. Det vil ikke være mulig for oss! Da har vi behov for en lur metode. Vi tar en stikkprøve ved å sette ut et garn. Stikkprøven ønsker vi skal være så representativ for alle fiskene som er i vannet som mulig. Det må vi prøve å tenke på når vi setter ut garnet. Det vil også være lurt å gjenta garnfangsten noen ganger.

Alle fiskene som er i vannet kaller vi *populasjonen* og stikkprøven vil være et *utvalg*.

I populasjonen sier vi at massene i populasjonen vil ha forventningsverdien $\mu$ og standardavviket $\sigma$




Hvordan kan vi gå fram for å kunne si noe om gjennomsnittsmassen i hele populasjonen? Vi kan ikke veie alle fiskene, så da blir løsningen å ta en stikkprøve.

**En stikkprøve** Vi fanger $n$ fisker fra vannet, legger alle på vekta og finner massen til hver av fiskene i stikkprøven. Vi benytter variabelen $X$ for massen til en tilfeldig valgt fisk, og vi finner

$$X_1, X_2, X_3, \cdots, X_n$$

Nå kan vi finne gjennomsnittsmassen i stikkprøven

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

Vi må altså anta at stikkprøven gir et godt bilde på hele populasjonen. I så fall vil gjennomsnittsmassen være en estimator for $\mu$. For å skille de to kaller vi estimatoren for $\hat{\mu}$ og har

$$\hat{\mu} = \mu_{\bar{X}} = \bar{X}$$

Etter å ha regnet ut den har vi et estimat for forventningsverdien i hele populasjonen.

### Simulering med Python

Vi lar igjen teknologien hjelpe oss for å lage en populasjon med ørret i et fjellvann. I vannet vårt er det 500 ørreter som veier mellom 1 g og 1 kg. Så tar vi stikkprøver i «vannet vårt» og finner gjennomsnittet. Programkode 2.7 er den samme som vi har sett på tidligere og den viser en måte å gjøre det på.

```
import random as rd
import statistics as st
antall = 500
ant_i_utvalg = 25
ant_stikkprover = 50
stikkprover = []
# lager liste med fiskemasser
masse = [rd.randint(1,1001) for i in  range(antall)]
# tar stikkprøver og finner statistiske data
for i in  range(ant_stikkprover):
    utvalg = rd.sample(masse, ant_i_utvalg)
    gjennomsnitt = st.mean(utvalg)
    stikkprover.append(gjennomsnitt)

# Statistiske data hele populasjonen
print("Statistiske data hele populasjonen")
print("Gjennomsnitt er  :", st.mean(masse))
print("Standardavvik er :", st.pstdev(masse))

#Statistiske data for stikkprøvene
print("Statistiske data for stikkprøvene")
print("Gjennomsnitt er  :", st.mean(stikkprover))
```

Programkode 2.7: Stikkprøver

Denne gjennomføringen ga resultatene

```
Statistiske data hele populasjonen
Gjennomsnitt er  : 501.472
Standardavvik er : 283.14324504744945
Statistiske data for stikkprøvene
Gjennomsnitt er  : 501.1896
```

Her har vi tatt stikkprøver ved hjelp av kommandoen `sample` fra modulen `random`. Simuleringen viser at vi får estimatet

$$\hat{\mu} = \mu_{\bar{x}} = 501.19$$

Vi sitter med fasiten som er $\mu = 501.47$ og kan si at det var ikke så aller verst – særlig når vi tenker på at dette er verdier som varierer mellom 1 g og 1 kg.

## 2.5 Estimatorer for en populasjon

Fra før av har sitt sett hvordan vi kan estimere standardavviket i populasjonen. Vi tar med det vi fant der og har disse punktestimatorene for populasjonen

**Punktestimatorer for $\mu$ og $\sigma^2$**

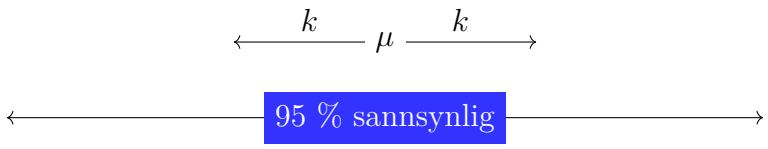
Vi har en populasjon med gjennomsnitt $\mu$ og varians $\sigma^2$ hvor disse verdiene er ukjente. Vi gjør $n$ uavhengige observasjoner fra populasjonen $X_1, X_2, \cdots, X_n$.
Forventningsrette estimatorer vil da være

$$\hat{\mu} = \bar{X}$$
$$\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

Spørsmålet nå er om hvor sikre vi kan være på at vi har kommet fram til riktig verdi. For å avgjøre det må vi innom noe som heter *konfidensintervall*

## Konfidensintervall for $\mu$

Vi kan ikke være helt sikre på estimatene vi kommer fram til, men vi kan si noe om hvor sikre vi er. Vi finner et intervall hvor vi kan uttale oss om sannsynligheten. La oss si at vi ønsker å være 95 % sikre




Vi vil finne et intervall slik at

$$P(\bar{X} - k \leq \mu \leq \bar{X} + k) = 0.95$$

Sentralgrenseteoremet forteller at gjennomsnittsverdiene i stikkprøvene vil være normalfordelt med $\mu$ og $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Fra normalfordelingen vet vi at 95 % av alle verdiene vil ligge mellom 1.96 standardavvik større og mindre enn forventningsverdien. Det gir at

$$P(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$$

Det gir dette intervallet

$$[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}]$$

La oss ta fiskene som et eksempel igjen. Vi tar vi bare en stikkprøver med 25 fisker og får $\bar{X} = 560.58$. Jukser vi litt og går ut fra at vi vet standardavviket i populasjonen har vi at det er $\sigma = 283.14$ g. Da kan vi finne et intervall hvor vi er 95 % sikre på at estimatet vårt ligger. Intervallet blir

$$\begin{aligned}
[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad &, \quad \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}] \\
[560.58 - 1.96 \cdot \frac{283.14}{\sqrt{25}} \quad &, \quad 560.58 + 1.96 \cdot \frac{283.14}{\sqrt{25}}] \\
[449.60 \quad &, \quad 671.57]
\end{aligned}$$

Ved hjelp av sentralgrenseteoremet og det vi vet om normalfordeling har vi nå kommet fram til at vi kan være 95 % sikre på at gjennomsnittsmassen til fiskene i populasjonen vil ligge i det intervallet.

**Konfidensintervall**

Et konfidensintervall for $\mu$ basert på $n$ uavhengige observasjonene er gitt ved

Hvis vi kjenner $\sigma$:

$$[\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{X} + z \cdot \frac{\sigma}{\sqrt{n}}]$$

Hvis vi ikke kjenner

$$[\bar{X} - z \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad , \quad \bar{X} + z \cdot \frac{\hat{\sigma}}{\sqrt{n}}]$$

Verdien for $z$ kan vi hente fra det vi vet om normalfordelingen.

| $z$ | $P(-z \leq Z \leq z)$ |
|---|---|
| 1.645 | 0.9 |
| 1.960 | 0.95 |
| 2.326 | 0.98 |
| 2.576 | 0.99 |

Ønsker vi et 95 % konfidensintervall er det bare å bytte ut $z$ med 1.96

## Eksempel: Høyder

Vi ser på høydene til guttene igjen

171 174 167 187 174 169 182 180 186 177 178 190 180 176 183 180 169 187 170 181
178 175 181 183 184 179 195 176 176 179 190 165 178 182 179 186 188 170 191 183
179 188 178 177 179 175 186 184 187 174 171 181 171 189 183 174 193 192 181 184
177 194 185 180 173 185 189 189 172 185 180 185 182 173 172 186 177 175 176 184
191 182 172 183 175 184 166 178 176 180 181 188 177 182 168 173 180 181 188 177
182 168 173 182 180 186 177 178 190 180 176 174 167 187 174 169 182 180 186 177
178 171 174 167 187 174 169 182 180 195 176 176 179 190 165 178 182 179 186 188
170 191 183 179 175 186 184 187 174 171 181 171 189 183 174 174 169 182 180 186
177 178 171 174 167 187 174 175 176 184 191 182 172 183 175 184 166 178 176 180
181 169 178 194 175 184 166 178 176 180 172 185 164 191 178 176 180 164 179 181

I denne populasjonen hadde vi

$$\mu = 179.2$$
$$\sigma^2 = 47.07$$
$$\sigma = 6.86$$

Nå kan vi prøve å estimere disse verdiene. Vi glemmer dem et øyeblikk og ser på dette eksemplet på en oppgave.

**Eksempel 2**

Finn estimat for $\mu$ og $\sigma$ ved å ta en stikkprøve fra alle kroppshøydene.

Vi starter med å ta en stikkprøve på 50 tilfeldige høyder.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 194 | 185 | 180 | 173 | 185 | 188 | 179 | 180 | 172 | 187 |
| 167 | 187 | 174 | 169 | 182 | 194 | 176 | 181 | 185 | 174 |
| 174 | 167 | 187 | 174 | 175 | 170 | 177 | 178 | 177 | 195 |
| 182 | 179 | 186 | 188 | 170 | 181 | 185 | 190 | 180 | 182 |
| 176 | 176 | 179 | 190 | 165 | 169 | 183 | 176 | 182 | 165 |

Ut fra stikkprøven får vi

$$\hat{\mu} = \bar{X} = 179.4$$
$$\hat{\sigma} = s = 7.61$$

Her er det brukt den forventningsrette standardavviket

$$s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

**Eksempel 3**

Sett opp et 95 % konfidensintervall for forventningsverdien.

Da har vi estimert både $\mu$ og $\sigma$ og kan ut fra det finne et 95 % konfidensintervall.

Vi setter $z = 1.96$, bruker estimatene og regner ut

$$\begin{aligned} [\bar{X} - z \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad &, \quad \bar{X} + z \cdot \frac{\hat{\sigma}}{\sqrt{n}}] \\ [179.4 - 1.96 \cdot \frac{7.61}{\sqrt{50}} \quad &, \quad 179.4 + 1.96 \cdot \frac{7.61}{\sqrt{50}}] \\ [177.3 \quad &, \quad 181.5] \end{aligned}$$

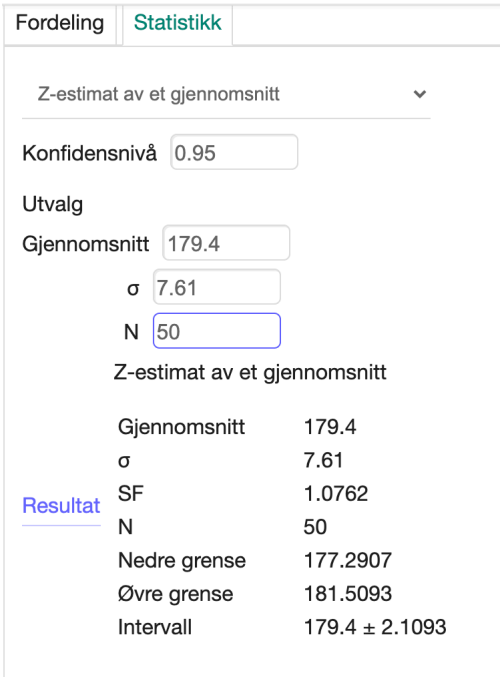
**GeoGebra** kan gjøre den jobben for oss ved at vi benytter sannsynlighetskalkulatoren på den måten som figur 2.8 viser.

## Standardfeil

I forbindelse med konfidensintervall introduseres ofte uttrykket *standardfeil*. Det har vi allerede benyttet i utregningene våre. Standardfeil, eller på engelsk *Standard Error*, $SE$, defineres slik

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}$$

hvor $s$ er det forventningsrette standardavviket i utvalget (der vi delte på $(n-1)$). Vi har brukt det uten å vite hva det heter. Siden ordet dukker opp innimellom kan det være greit å kjenne til det. Legger merke til at GeoGebra regner det ut for oss. I figur 2.8 kan vi se at det står `SF = 1.0762`. I den norske versjonen av GeoGebra benyttes SF for SE. Legg merke til at det er det samme som $\frac{7.61}{\sqrt{50}} = 1.076216520965925$.




Figur 2.8: Konfidensintervall med GeoGebra

## Oppgaver med løsning

**Oppgave 4**

Vi ønsker å finne gjennomsnittshøyden for alle guttene som går siste året på videregående skole i en by. Standardavviket for høyden på gutter ved sesjon er 7 cm. Vi antar at det gjelder for disse guttene også. 100 gutter plukkes ut i en stikkprøve og vi finner at gjennomsnittshøyden er 180.7 cm

Finn et konfidensintervall for gjennomsnittshøyden med konfidensnivå 0.95

**Løsningsforslag**

Vi setter $X$ for høyden til en tilfeldig valgt gutt og antar at $X$ er normalfordelt $N(\mu, 7^2)$

$\bar{X}$ er gjennomsnittshøyden til guttene i stikkprøven på $n = 100$ tilfeldig valgte gutter. Sentralgrenseteoremet sier at $\bar{X}$ er normalfordelt $N(\mu, \frac{7}{\sqrt{100}}) = N(\mu, 0.7)$. Da har vi

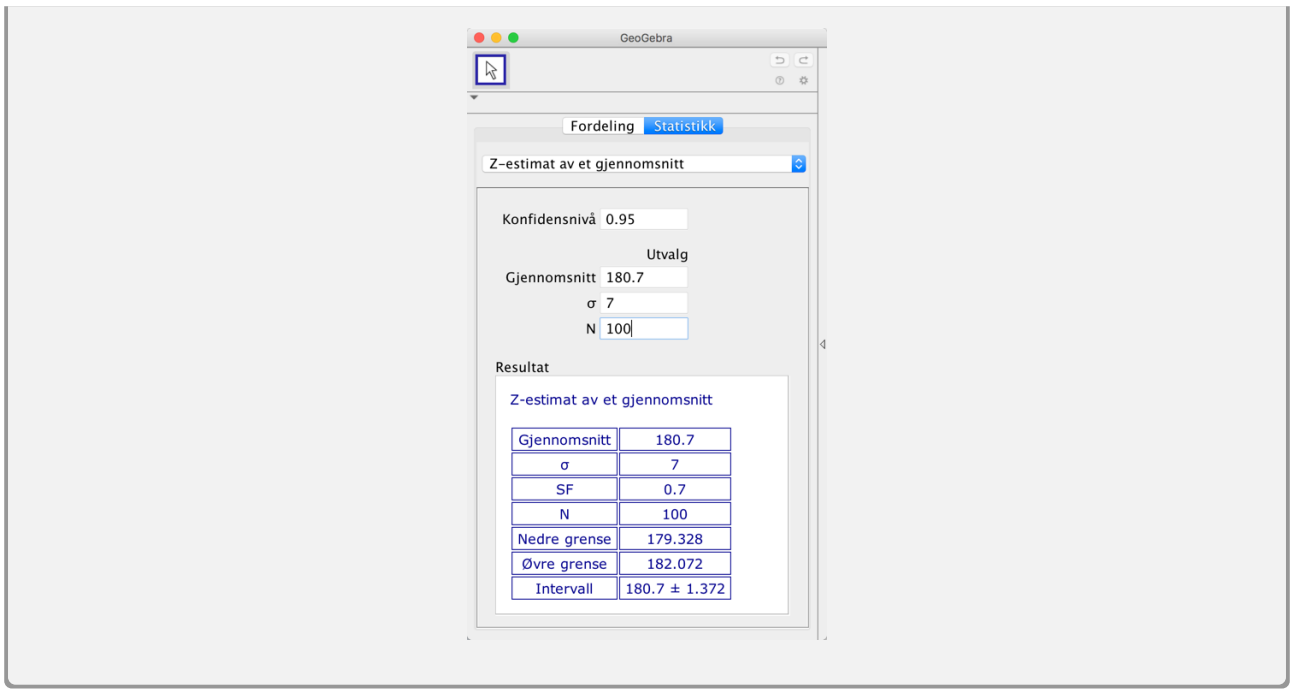
$$P(\bar{X} - 1.96 \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96 \cdot \sigma_{\bar{X}}) = 0.95$$
$$P(180.7 - 1.96 \cdot 0.7 \leq \mu \leq 180.7 + 1.96 \cdot 0.7) = 0.95$$
$$P(179.3 \leq \mu \leq 182.1) = 0.95$$

Konfidensintervallet

$$[179.3, 182.1]$$

Husk at dette er et 95 % konfidensintervall for forventningsverdien, $\mu$, for alle høydene.

Vi kan få det samme svaret med GeoGebra




## 2.6 Punktestimering for sannsynligheter

Vi skal nå se på hvordan vi kan estimere sannsynligheter i binomiske sannsynlighetsfordelinger.

En binomisk sannsynlighetsfordeling får vi når det utføres binomiske forsøksserier. Da har vi sett at sannsynligheten kan beregnes ved

$$P(S = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Ofte kjenner vi ikke $p$, men vi ønsker å finne et estimat, $\hat{p}$. Nå skal vi se mer på hvordan vi kan gjøre det og hvordan vi kan uttale oss om hvor godt dette estimatet er. Før det må se litt mer på forventningsverdi og varians.

### Forventningsverdi og varians til $\hat{p}$

Når vi skal estimere sannsynligheten $p$ vil vi observere det binomiske forsøket og finne hvor mange ganger suksess-utfallet inntreffer. Utfører vi forsøket $n$ ganger og finner at det skjer $X$ ganger har vi

$$\hat{p} = \frac{X}{n}$$

Vi vet også at $X$ er binomisk fordelt og at forventningsverdien og variansen i en binomisk sannsynlighetsfordeling er

$$\begin{aligned} \mathrm{E}(X) &= \mu = np \\ \mathrm{Var}(X) &= \sigma^2 = np(1-p) \end{aligned}$$

Det gjør at vi kan regne ut det samme for estimatoren

$$\mathrm{E}(\hat{p}) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{1}{n} \cdot \mathrm{E}(X) = \frac{1}{n} \cdot np = p$$

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \mathrm{Var}(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$

Her er det satt inn det vi har for $\hat{p}$ og utledet. Uten at vi går inn på det er det greit å legge merke til at $\mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \mathrm{Var}(X)$.

Legg også merke til at dette er en forventningsrett varians som avtar desto flere vi velger. Det kan vi se av uttrykket hvor vi dividerer med $n$. Store verdier av $n$ gir mindre varians. Vi kan få den så nøyaktig vi ønsker.

Vi ender opp dette resultatet

**Forventningsverdi og varians for $\hat{p}$**

Hvis $\hat{p}$ er en estimator for en sannsynlighet i en binomisk sannsynlighetsfordeling har vi at

$$\mathrm{E}(\hat{p}) = p$$
$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

## Et eksempel: Valgprognose

Et eksempel kan vise dette og da er en spørreundersøkelse nærliggende. Slikt utføres på et utvalg for å si noe om en hel populasjon.

**Eksempel 4**

I en gallupundersøkelse undersøkes det hvor mange som vil stemme Arbeiderpartiet ved neste valg. 1500 personer er plukket ut og av dem svarer 447 personer at de vil stemme på Arbeiderpartiet.
Estimer oppslutningen Arbeiderpartiet vil få ved neste valg og drøft usikkerheten.

Noen har gjort en undersøkelse for oss og vi kan finne et estimat for sannsynligheten for at en tilfeldig valgt person skal stemme på Arbeiderpartiet ved neste valg. Ut fra eksemplet er vårt beste estimat for sannsynligheten

$$\hat{p} = \frac{447}{1500} = 0.298$$

Undersøkelsen gir oss et estimat for at 29.8 % av stemmene vil gå til Arbeiderpartiet.

Nå skjønner vi at det er vanskelig å spå oppslutningen i det endelige valget. De som ble spurt kan endre oppfatning og alle de som ikke ble spurt kan ha helt andre syn på saken. Det viktige er at det er gjort et representativt utvalg når de 1500 respondentene ble plukket ut.

Vi gjør også en forutsetning om at det å spørre noen om de vil stemme på Arbeiderpartiet er et binomisk forsøk. Det kan diskuteres og det er viktig å være klar over alle forutsetninger vi gjør.

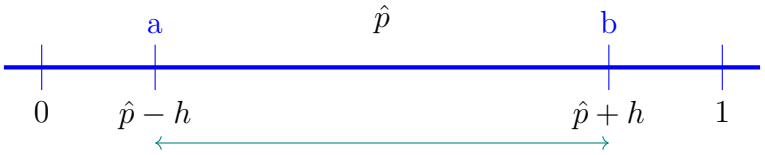
Vi får holde oss til dette eksemplet når vi skal uttale oss om usikkerheten ved estimatet. Vi vil finne et konfidensintervall for estimatet.

## Konfidensintervall for $\hat{p}$

Vi fant et estimat

$$\hat{p} = \frac{X}{n}$$

Vi ønsker å finne et konfidensintervall som kan si noe om estimatet vi kom fram til. Vi vil fram til noen verdier hvor vi kan si at vi er f. eks. 95 % sikre på at sannsynligheten vil ligge. Figur 2.9 illustrerer det.




Figur 2.9: Et konfidensintervall for $\hat{p}$

For å finne et konfidensintervall må vi gjøre en antakelse om at $\hat{p}$ er normalfordelt. En binomisk fordeling kan jo nesten være det!

Vi har sett at variansen til $\hat{p}$ er

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Nå kjenner vi ikke $p$ i høyre side av dette uttrykket, men vi har funnet en estimator for $p$ som er $\hat{p}$. Bruker vi den får vi

$$\text{Var}(\hat{p}) \approx \frac{\hat{p}(1-\hat{p})}{n}$$

Det er variansen. Standardavviket, som er rota av variansen, har fått et eget navn: *standardfeilen*

**Standardfeilen til $\hat{p}$**

Standardfeilen er standardavviket til en estimator for en binomisk sannsynlighet er gitt ved

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Konfidensintervallet for $p$ er

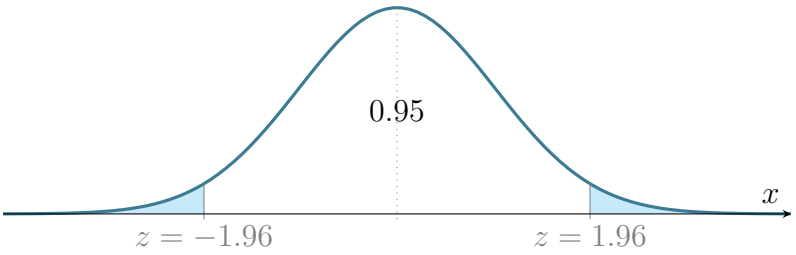
$$\langle \hat{p} - z \cdot \text{SE}(\hat{p}) \quad , \quad \hat{p} + z \cdot \text{SE}(\hat{p}) \rangle$$

hvor

$$\hat{p} = \frac{X}{n} \qquad \text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Det er verdien av $z$ som avgjør bredden på intervallet. Figur 2.10 minner oss på hvordan sannsynlighetene fordelte seg i normalfordelingen.

De mest brukte verdiene for $z$ fant vi i denne tabellen




Figur 2.10: Normalfordeling

| $z$ | konfidensintervall |
|---|---|
| 1.65 | 90 % |
| 1.96 | 95 % |
| 2.58 | 99 % |

**Konfidensintervallet** i eksemplet vårt forteller noe om hvor sikre vi kan være på estimatet vi fant. Vi kom fram til at

$$\hat{p} = 0.298$$

Nå ønsker vi å finne et 95% konfidensintervall for den sanne stemmeandelen Arbeiderpartiet vil få ved neste valg.

Standardfeilen

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.298(1-0.298)}{1500}} = 0.011809$$

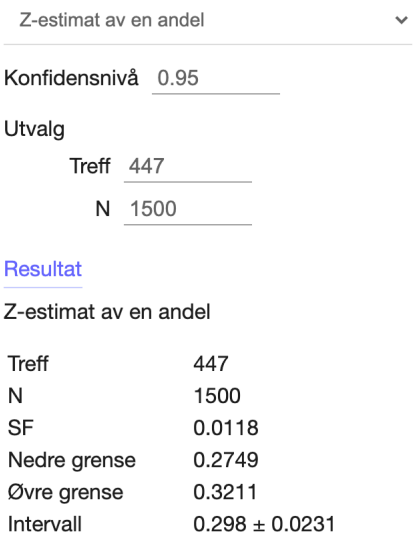
I et 95% konfidensintervall er $z = 1.96$ og intervallet blir

$$\begin{aligned} KI_{95} &= \langle \hat{p} - 1.96 \cdot \text{SE}(\hat{p}) \quad , \quad \hat{p} + 1.96 \cdot \text{SE}(\hat{p}) \rangle \\ &= \langle 0.298 - 1.96 \cdot 0.012, 0.298 + 1.96 \cdot 0.012 \rangle = \langle 0.275, 0.321 \rangle \end{aligned}$$

Ut fra det vi har funnet kan vi si at vi kan være 95 % sikre på at stemmeandelen vil være mellom 27.5 % og 32.1 %.

Her har vi funnet det ved regning. GeoGebra kan gjøre den jobben for oss. I figure 2.11 ser vi at vi får de samme svarene når vi velger at vi vil ha et *Z-estimat av en andel*. De 447 personene som svarte at de ville stemme på Arbeiderpartiet må vi skrive inn som antall *treff* i det totale antallet som er *N*. Vi må også si fra om at vi ønsker et *konfidensnivå* som er 0.95.

GeoGebra gir oss den nedre- og den øvre grensa i intervallet. I tillegg ser vi også at vi har fått regna ut standardfeilen, *SF*, som er den samme som vi fant tidligere.




Figur 2.11: Konfidensintervallet med GeoGebra

Kan vi anta et dette er et binomisk forsøk? Den antakelsen krever at det er samme sannsynlighet hver gang for å plukke ut en person som stemmer på Arbeiderpartiet. Det betyr i så fall at vi velger ut personer med tilbakelegging! I en praktisk situasjon ville vi vel ikke risikert å spørre samme person to ganger? I så fall vil spørreundersøkelsen vår være et hypergeometrisk forsøk.
For en hypergeoemtrisk situasjon kan det vises at:

$$\begin{aligned}\text{Var}(\hat{p}) &\approx \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \frac{X}{n} \cdot \left(1 - \frac{X}{n}\right) \\ &= \frac{1}{n} \cdot \frac{N-n}{N-1} \cdot \hat{p} \cdot (1-\hat{p})\end{aligned}$$

I en binomisk situasjon har vi sett at

$$\text{Var}(\hat{p}) \approx \frac{\hat{p}(1-\hat{p})}{n}$$

Det som skiller disse to beregningene er verdien av $\frac{N-n}{N-1}$
La oss gå ut fra at det er 5 millioner stemmeberettige i Norge. Da har vi dette regnestykket

$$\frac{N-n}{N-1} = \frac{5000000 - 1500}{5000000 - 1} = 0.99970019994004$$

Med alle andre feilkilder ser vi at denne verdien ikke vil gjøre noe stort utslag. Uansett om vi velger å se det som en binomisk eller en hypergeometrisk situasjon vil konfidensintervallet bli omtrent det samme hvis $n$ er liten i forhold til $N$.

## Flere eksempler

Vi kan se på en typisk oppgave i neste eksempel.

**Eksempel 5**

Vi skal finne sannsynligheten $p$ for at ei fyrstikkeske lander på ei av de «største» sidene.
La $X$ være at eska lander på «stor side»
Så utfører vi et forsøk hvor vi kaster eska 500 ganger og finner at den lander med den «store» sida opp 263 ganger

Finn et estimat for $p$ og et intervall som med 99 prosent sannsynlighet inneholder $p$

Her får vi vite at

$$X = 263 \qquad n = 500$$

Det gir

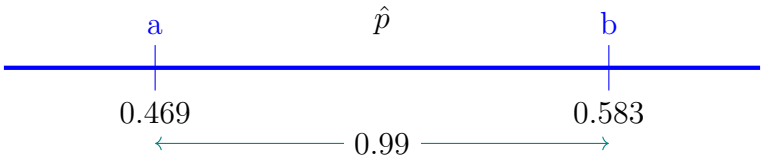
$$\hat{p} = \frac{X}{n} = \frac{263}{500} = 0.526$$

Standardfeilen

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.526(1-0.526)}{500}} = 0.02233$$

I et 99% konfidensintervall er $z = 2.58$ og intervallet blir

$$\begin{aligned} KI_{99} &= \langle \hat{p} - 2.58 \cdot \mathrm{SE}(\hat{p}) \quad , \quad \hat{p} + 2.58 \cdot \mathrm{SE}(\hat{p}) \rangle \\ &= \langle 0.526 - 2.58 \cdot 0.022 \quad , \quad 0.526 + 2.58 \cdot 0.022 \rangle \\ &= \langle 0.469 \quad , \quad 0.583 \rangle \end{aligned}$$




$$KI_{99} = \langle 0.469 \quad , \quad 0.583 \rangle$$

Bruker vi GeoGebra får vi resultatet i figur 2.12.

Konklusjon: Ut fra eksperimentet kan vi med 99 % sannsynlighet si at sannsynligheten for at fyrstikkeska lander med «stor side» ned er mellom 0.469 og 0.583.

**Eksempel 6**

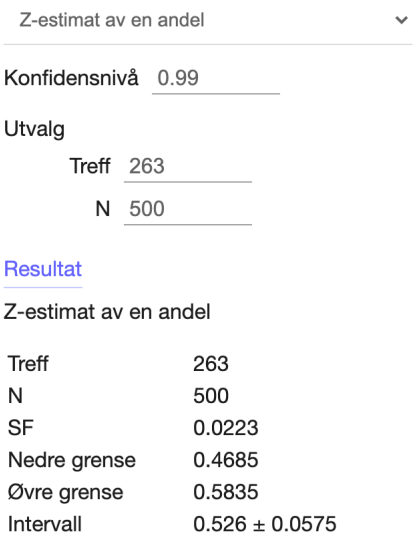
Vi kaster en mynt hundre ganger og får 64 kron. Er det noe rart med denne mynten?

Ut fra eksperimentet finner vi et estimat for sannsynligheten for å få «kron» med denne mynten.

$$\hat{p} = \frac{X}{n} = \frac{64}{100} = 0.64$$

For å finne et konfidensintervall for estimatet må vi finne standardfeilen

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.64(1-0.64)}{100}} = 0.048$$




Figur 2.12: Konfidensintervallet med GeoGebra

Da kan vi sette opp et 95 % konfidensintervall

$$\begin{aligned} &\langle \hat{p} - z \cdot \mathrm{SE}(\hat{p}) \quad , \quad \hat{p} + z \cdot \mathrm{SE}(\hat{p}) \rangle \\ &\langle 0.64 - z \cdot 0.048 \quad , \quad 0.64 + z \cdot 0.048 \rangle \end{aligned}$$

Velger vi et 95 % konfidensintervall blir $z = 1.96$ og vi får

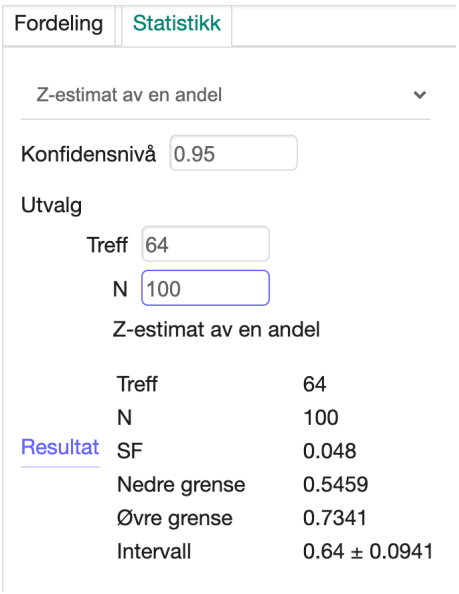
$$\langle 0.54592 \quad , \quad 0.73408 \rangle$$

Vi kan finne dette konfidensintervallet med GeoGebra også. Velg Z-estimat av andel og skriv inn konfidensnivå. Resultatet vises i figur 2.13.

I et 99% konfidensintervall er $z = 2.58$ og intervallet blir

$$\langle 0.51616 \quad , \quad 0.76384 \rangle$$

Ingen av konfidensintervallene omfatter sannsynligheten for å få «kron» med en ideell mynt, som er $p = 0.5$. Her må det være noe rart!




Figur 2.13: Konfidensintervallet med GeoGebra

# 3 Hypotesetesting

## 3.1 Hypotesetesting

En hypotese er en ubekrefta antakelse – noe vi tror er sant. I vitenskaplig arbeid ønsker vi å finne ut mer om denne antakelsen. Hvor sikker kan vi være på at den stemmer? For å avgjøre det må vi teste hypotesen.

**Definisjon 4 Hypotese**

En statistisk hypotese er en antakelse eller påstand om en eller flere populasjoner som kan enten være sanne eller usanne.

Når vi kommer med en hypotese vil det alltid kunne formuleres en alternativ hypotese, som er den motsatte. Påstår jeg at «jorda er rund», så vil den vil den alternative hypotesen være: «jorda er ikke rund». Vi vil kalle de to hypotesene for

$$H_0 : \text{Nullhypotese}$$
$$H_1 : \text{Alternativ hypotese}$$

I hypotesetesting vil det være $H_1$, den alternative hypotesen vi ønsker å teste. Det gjør vi med utgangspunkt i at $H_0$, nullhypotesen, stemmer. I likhet med grunnleggende prinsipp i rettssaler antar vi at den tiltalte er uskyldig inntil det motsatte er bevist. I hypotesetestingen blir det at vi må ha tilstrekkelig med bevis for lande på den alternative hypotesen. En utfordring er hva som er tilstrekkelig. I hypotesetestingen må vi gjøre noen valg for det også. Vi vil derfor aldri kunne fullstendig bekrefte eller avkrefte hypotesen vår, men vi kan si noe om hvor sannsynlig den er.

Vanligvis kan vi ikke undersøke en hel populasjon. Det kan være flere grunner til det. Vi må gjøre et utvalg og basere oss på det vi kan finne ut om utvalget. Det betyr at det vil være et estimat vi undersøker.

### Eksempel: Klatretau

Vi tenker oss en produsent av klatretau som sier at tauene vil tåle en belastning på 2500 kg. Vi plukker ut 49 tau og undersøker hvor stor belastning de tåler før de ryker. Vi finner ut at de i gjennomsnitt tåler 2400 kg og at standardavviket er 350 kg. Kan vi med det si at påstanden til produsenten er feil?

Vi vet ikke noe om alle tauene. Bare om utvalget vi har undersøkt. Hvordan skal vi gå fram for å komme til en konklusjon basert på en statistisk undersøkelse?

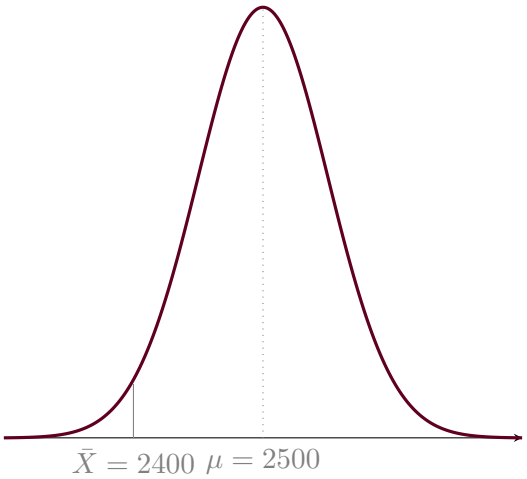
Vi har altså 49 tau i utvalget. Populasjonen, alle klatretauene, har en gjennomsnittlig bruddgrense, $\mu$. Den er oppgitt til å være 2500 kg og vi kan anta at den er normalfordelt. Den gjennomsnittlige bruddgrensen i utvalget er $\bar{X} = 2400$ kg og standardavviket i utvalget er $s = 350$ kg.

Nå skal vi sette opp hypotesene. Vi kaller den ene for *nullhypotesen* og den andre for *alternativ hypotese*. En vanlig måte å skrive det på er slik

$H_0 : \mu = 2500$ Produsenten oppgir korrekt belastningsevne

$H_1 : \mu < 2500$ Belastningsevnen er mindre enn det som er oppgitt

Da har vi satt opp de to hypoteser som vi vil undersøke. Figur 3.8 viser situasjonen. Produsenten har oppgitt $\mu = 2500$. Vi har gjort et utvalg og funnet $\bar{X} = 2400$. Det er mindre enn det oppgitte, men er det innafor en normal variasjon?




Figur 3.1: Typisk normalfordeling

Sentralgrenseteoremet gir at $\bar{X}$ er normalfordelt med $\mu_{\bar{X}} = \mu = 2500$. Standardavviket må vi estimere siden det ikke er oppgitt. Vi kom fram til at $s = 350$. Da har vi at standardfeilen er

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{350}{\sqrt{49}} = 50$$

Det gir at $\bar{X} \sim N(2500, 50)$. Hva er da sannsynligheten for å få resultatet i utvalget? Vi må finne

$$P(\bar{X} \leq 2400)$$

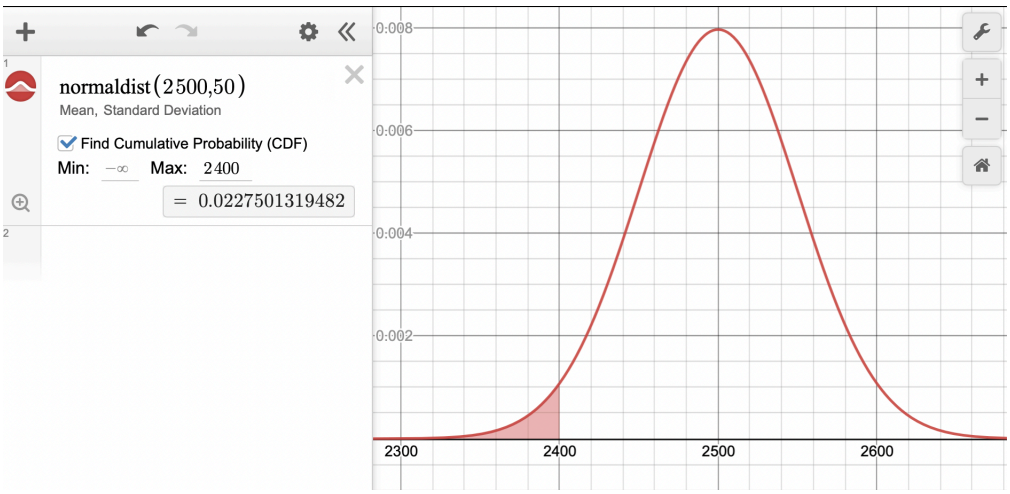
Desmos eller GeoGebra kan gi oss svaret. Figur 3.9 viser hvordan vi kan bruke Desmos

Sannsynligheten blir

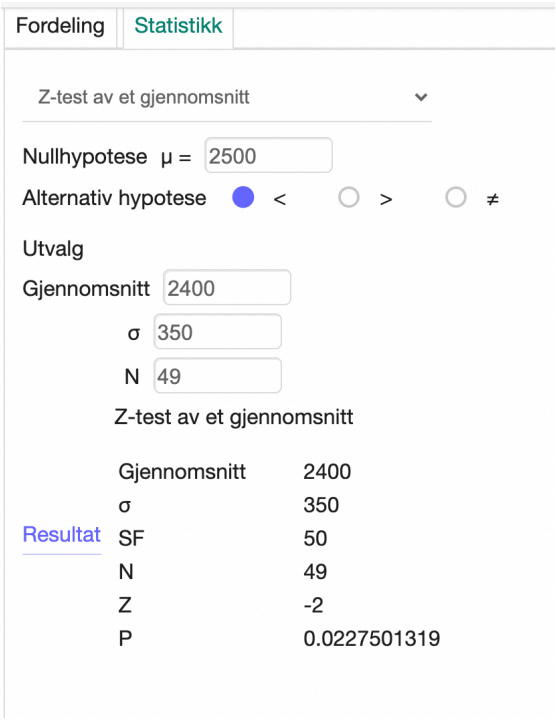
$$P(\bar{X} \leq 2400) = 0.02275$$

Ut fra beregningene våre er verdien vi har funnet ved å undersøke utvalget ikke så veldig sannsynlig gitt at $\mu = 2500$. Som regel setter vi ei grense på 5 %, men den kan vi velge som vi vil. Her ser vi at det bare er 2.23 % sannsynlig at nullhypotesen stemmer. Det gjør at vi forkaster nullhypotesen. Vi tror ikke at det stemmer det produsenten oppgir. Vi tror at bruddevnen er mindre enn det. Legg merke til at vi ikke kan bevise det, bare sannsynliggjøre at det er tilfelle.

**GeoGebra** har en egen kalkulator for hypotesetesting. I dette tilfellet vi vi velge Z-test av et gjennomsnitt. Skriver vi inn verdiene vi har får vi resultatet i figur 3.10. Resultatet oppgir en P-verdi vi kjenner igjen. Den er den samme som sannsynligheten vi fant tidligere. Her finner vi også standardfeilen vi regna ut. GeoGebra bruker SF for standardfeilen. I kalkulatoren har vi skrevet inn nullhypotesen og at den alternative hypotesen er verdier mindre enn det.




Figur 3.2: Sannsynligheten med Desmos




Figur 3.3: Hypotesetesting med GeoGebra

## 3.2 Feiltyper

Når vi forkaster en nullhypotesen kan vi naturligvis gjøre en feil. I så fall gjør vi en *forkastningsfeil*, ofte kalt en feil av type 1. Vi kan også godta nullhypotesen om den ikke stemmer. Da gjør vi en *godtakingsfeil*, som også kalles feil av type 2. Det hele kan summeres opp i tabellen under.

| | $H_0$ er sann | $H_1$ er sann |
|---|---|---|
| beholder $H_0$ | korrekt | godtakingsfeil |
| forkaster $H_0$ | forkastningsfeil | korrekt |

I eksemplet med klatretauet fant vi at tauene i utvalget tålte 2400 kg i gjennomsnitt. Vi måtte vurdere denne verdien mot produsentens oppgitte bruddstyrke på 2500 kg. Vi fant at sannsynligheten for at et utvalg med det antall vi valgte skulle tåle bare 2400 kg var godt under

5 %. Vi valgte å forkaste nullhypotesen, men vi kan ikke være helt sikre. Det kan hende vi gjør en forkastningsfeil. Grensa vi satte på 5 % gjør at sannsynligheten for å gjøre en forkastningsfeil er nettopp 5 %. Hvis denne feiltypen var den eneste bekymringen vår kunne vi satt grensa lavere. Hadde vi valgt den til 2 %, ville vi, så vidt, ikke kunne forkast nullhypotesen. Da oppstår et annet problem: Muligheten for en godtakingsfeil ville lettere oppstå.

En analogi til denne problematikken er røykvarslere. Røykvarslere kan være irriterende. Den begynner å ule om en åpner steikovnen eller svir maten. Den tror det er brann uten at det er det. I vår terminologi forkaster den nullhypotesen om at det ikke brenner. Den gjør en forkastningsfeil. En enkel løsning på problemet er å fjerne batteriene. Ulinga i tide og utide stopper, men nå oppstår problemet om at det kan skje en godtakingsfeil: Alarmen går ikke om det brenner. Her er det et dilemma. En hypersensitiv røykvarsler gjør aldri godtakingsfeil, men ofte forkastningsfeil. I statistikken må vi finne en balanse ved å finne et passende signifikansnivå. En god justering av røykvarsleren som unngår begge typer feil i størst mulig grad.

## 3.3 Signifikansnivå

I eksemplet vårt bestemte vi oss for å forkaste nullhypotesen siden sannsynligheten var så lav. Vi satte et *signifikansnivå* på 5 %. Det skriver vi som

$$\alpha = 0.05$$

**Definisjon 5 Signifikansnivå**

Signifikansnivå $\alpha$ er hvor stor sannsynligheten for forkastningsfeil vi er villig til å akseptere

Vi kan skrive det som

$$\alpha = P(\text{forkaste } H_0 \text{ når } H_0 \text{ er riktig}) = P(\text{forkaste } H_0 \mid H_0) = P(\text{forkastningsfeil} \mid H_0)$$

Måten å skrive det på er kanskje ikke helt korrekt siden dette egentlig ikke er betinga sannsynlighet og $H_0$ er ikke en hending. Meningen kommer fram.

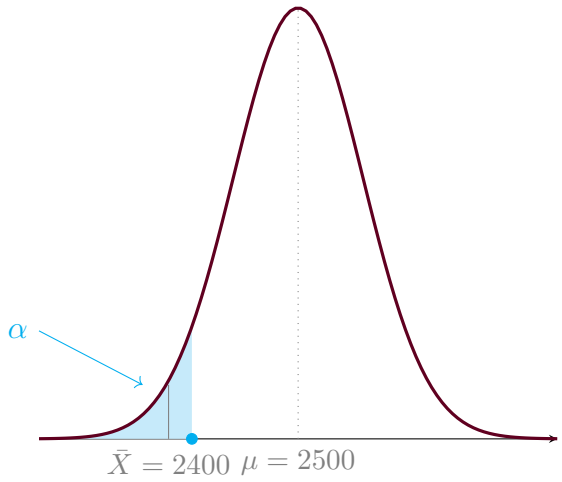
I eksemplet vårt kan vi markere $\alpha$ som et areal under normalfordelingskurven. Figur 3.4 viser det. Dette området kalles en kritisk region. I eksemplet satte vi den alternative hypotesen til å være $H_1 : \mu < 2500$. Det gjør at den kritiske regionen havner til venstre. I andre tilfeller kunne det vært området lengst til høyre. Det skjer når vi utfører ensidige hypotesetester.

Noen ganger ønsker vi å gjøre tester som er dobbelsidige. Figur 3.5 viser det.

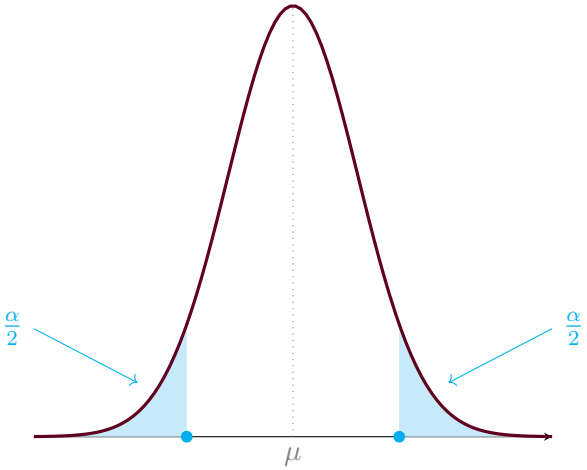
Som en introduksjon til hypotesetesting vil vi holde oss til ensidige tester, men tankegangen vil være den samme for andre typer hypotesetesting.

## 3.4 Hypotesestest av gjennomsnitt

Eksemplet med klatretau viser en hypotesetest av et gjennomsnitt vi fant i et utvalg. I dette eksemplet visste vi heller ikke standardavviket i populasjonen. Egentlig vil det være problematisk siden vi først må finne et estimat for standardavviket. I statistikken vil vi da gjøre noen




Figur 3.4: Signifikansnivå




Figur 3.5: Dobbelsidig

justeringer i testen. Det kan vi se på seinere og heller fortsette med et eksempel hvor vi tester en gjennomsnittsverdi i et utvalg med oppgitt standardavvik.

Les først eksemplet og tenk på hvordan vi kan gå fram.

**Eksempel 7**

En bedrift produserer esker med sjokolade. Sjokoladeeskene skal ha ei normalfordelt vekt på 35 hg. Standardavviket er oppgitt til 3 hg. Bedriften har mistanke om at vekta er lavere enn 35 hg og tar en stikkprøve på ti esker. Resultatet er

35 30 32 37 29 31 37 32 31 36

Hva kan vi si om bedriftens mistanke? Stemmer den?

Vi starter med å setter opp disse hypotesene

$H_0 : \mu = 35$ Sjokoladeeskene er 35 hg med gitte avvik
$H_1 : \mu < 35$ Sjokoladeeskene veier mindre enn 35 hg

Gjennomsnittet i stikkprøven gir

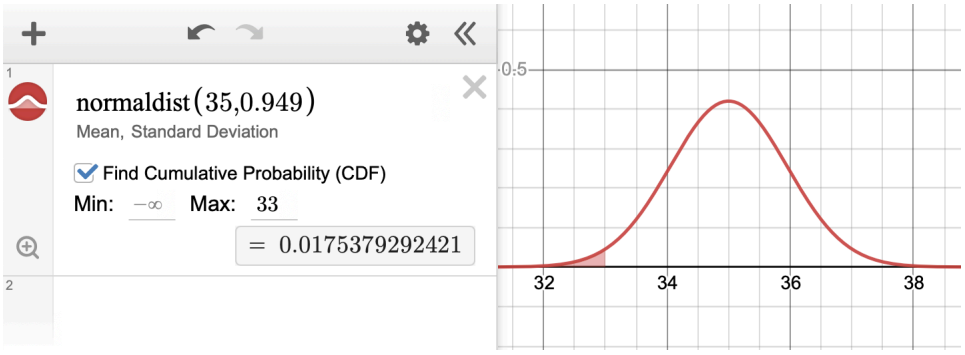
$$\bar{X} = \frac{35 + 30 + 32 + 37 + 29 + 31 + 37 + 32 + 31 + 36}{10} = 33$$

Vi ser at $\bar{X}$ er mindre enn det som er oppgitt, men er denne forskjellen signifikant? Det kan vi si noe om ved hypotesetesting. Før vi går videre bestemmer vi oss for å sette signifikansnivå til 5 %. Vi får også oppgitt at vekta er normalfordelt.

Sentralgrenseteoremet forteller at utvalget er normalfordelt med $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. Da har vi at standardfeilen er

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{10}} = 0.949$$

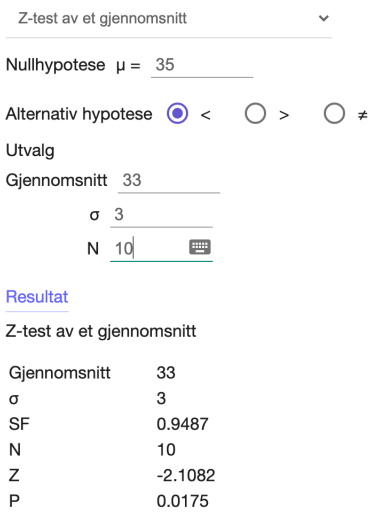
Vi kan finne sannsynligheten for at vi kan få $\bar{X} = 33$ med et verktøy. Figur 3.6 viser hvordan vi kan benytte Desmos.




Figur 3.6: Sannsynligheten med Desmos

Sannsynligheten for at vi får en gjennomsnittsverdi på $\bar{X} = 33$ i en stikkprøve er 1.75 %. Vi har valgt et siginifikansnivå på 5 %, så konklusjonen blir at vi forkaster nullhypotesen. Sannsynligheten for en forkastningsfeil blir 1.75 %.

Figur 3.7 viser hvordan vi kan benytte GeoGebra ved å utføre en `Z-test av et gjennomsnitt` i sannsynlighetskalkulatoren. Vi fyllet inn opplysningene og GeoGebra regner ut.




Figur 3.7: Hypotesetest med GeoGebra

Legg merke til at vi må skrive inn standardavviket i populasjonen og at GeoGebra regner ut standardfeilen $SF = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{10}} = 0.949$

P-verdien forteller oss sannsynligheten for at gjennomsnittet i utvalget, $\bar{X}$, vil være 33, eller mindre, gitt at nullhypotesen stemmer. At vi også inkluderer sannsynligheten for at gjennomsnittet kan være mindre vil øke sannsynligheten for at det inntreffer. Det vil være til støtte for nullhypotesen og minke sannsynligheten for forkastningsfeil.

## 3.5 Hypotesetest av sannsynlighet

Vi skal også se på hypotesetesting av sannsynlighet, eller andeler.

**Eksempel 8**

Etter å ha vært på kasino får James mistanke om at terningene som ble brukt var laget slik at de ga seksere for ofte. En slik manipulering kan bli gjort ved å flytte tyngdepunktet. James vil ha hjelp til å undersøke om det er tilfelle.

Hvordan kan vi konkludere med at terningene er fikset eller ikke? Hvis den ikke er fikset kan vi anta at sannsynligheten for å få en sekser er $\frac{1}{6}$.

Vi setter opp to hypoteser

| | | | |
|---|---|---|---|
| $H_0$ | Nullhypotesen | Terningen er ikke manipulert | $p = \frac{1}{6}$ |
| $H_1$ | Alternativ hypotese | Terningen er manipulert | $p > \frac{1}{6}$ |

Vi vil undersøke om vi kan forkaste nullhypotesen. Det kan vi gjøre om vi greier å skaffe oss terningen. Da kan vi kaste den og sjekke resultatet. Vi får tak i terningen og kaster den 100 ganger. Resultatet blir 26 seksere. Er det for mange for en ideell terning?

Å kaste terning er et binomisk forsøk. Forventningsverdien i en binomisk fordeling kjenner vi. Den forteller at vi kan forvente oss å få

$$E(X) = n \cdot p = 100 \cdot \frac{1}{6} = 16.67$$

Resultatet er langt over det. Det neste vi kan gjøre er å finne ut mer om sannsynligheten. For å være på den sikre sida finner vi sannsynligheten for å få 26 eller flere seksere

$$P(X \geq 26)$$

Heldigvis vet vi at dette er et binomisk forsøk og regne ut sannsynligheten. Vi gjør det med GeoGebra og sannsynlighetskalkulatoren. Se figur 3.8.

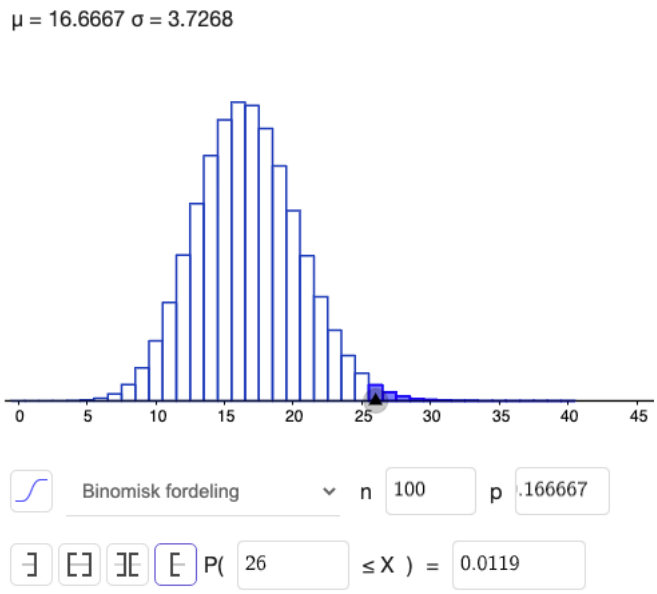
Vi får at

$$P(X \geq 26) = 0.0119$$

Sannsynligheten for å få 26, eller flere, seksere på 100 kast er altså 0.0119 eller 1.19%. Det er ikke veldig sannsynlig, men er det såpass liten sannsynlighet at vi kan forkaste nullhypotesen vår? Vi finner ikke et klart svar på det, men vi kan si noe om hvor sikre vi ønsker å være. Vi må bestemme et signifikansnivå.

La oss si at vi setter det til 5%. Det betyr at P(påstå juks | han jukser ikke) $\leq 0.05$. Vi må altså finne ut om

$$P\left(X \geq 26 \mid p = \frac{1}{6}\right) \leq 0.05$$




Figur 3.8: Sannsynlighetene i GeoGebra

Vi har funnet at

$$P\left(X \geq 26 \mid p = \frac{1}{6}\right) = 0.0119$$

Da har vi at sannsynligheten er mindre enn signifikansnivået og at vi forkaster nullhypotesen på det grunnlaget.

Vi ser på et eksempel til [1].

**Eksempel 9**

Et meningsmålingsinstitutt gjennomfører en spørreundersøkelse for et bestemt politisk parti.
I undersøkelsen blir 1500 tilfeldig valgte personer spurt om de ville ha stemt på partiet dersom det var valg.
I undersøkelsen svarer 321 personer at de ville ha stemt på partiet. Ved forrige valg stemte 19,8 % av velgerne på partiet.
Bruk det du har lært i statistikk, til å vurdere om partiet har hatt framgang siden forrige valg. Begrunn resonnementet ditt med beregninger.

Vi lar $X$ være antall personer av de 1500 personene som ville stemt på partiet. Dette er et binomisk forsøk og $X$ er binomisk fordelt med $p = 0.198$ og $n = 1500$.

Hypotesene blir

$$\begin{aligned} H_0: &\quad p = 0.198 \quad \text{partiet har ikke hatt framgang siden forrige valg} \\ H_1: &\quad p > 0.198 \quad \text{partiet har hatt framgang siden forrige valg} \end{aligned}$$

[1] Eksemplet er henta fra eksamen i S2 høsten 2009

Andel som sier de vil stemme på partiet:

$$\frac{321}{1500} = 0.214$$

I undersøkelsen er det altså 21.4 % som sier at de vil stemme på partiet. Vi må finne ut om det skiller seg signifikant fra resultatet ved forrige valg. Det gjør vi ved å teste hypotesene våre.

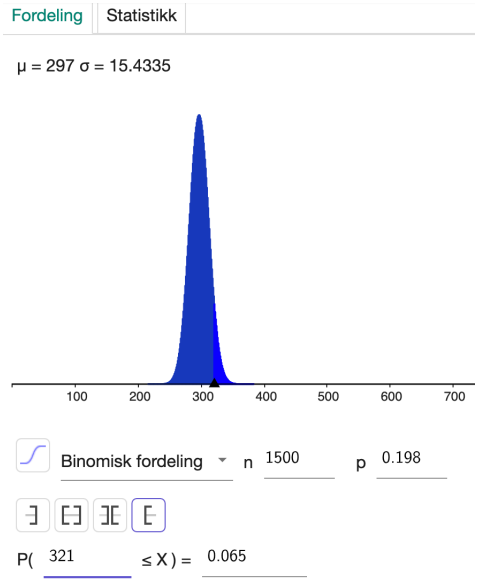
Vi velger et signifikansnivå på 5 % og setter $\alpha = 0.05$. Da vil sannsynligheten for å forkaste $H_0$ på feil grunnlag være 5 %.

Nå blir spørsmålet: Hva er sannsynligheten for at minst 321 personer i et utvalg på 1500 sier at de vil stemme på partiet gitt at 19.8 % prosent av befolkningen virkelig stemmer på partiet?

Vi velger å betrakte dette som et binomisk forsøk hvor vi spør om en person stemmer på partiet eller ikke. Det gjentar vi 1500 ganger.

Sannsynligheten kan vi regne ut

$$P(X \geq 321) = \sum_{x=1}^{1500} = \binom{1500}{x} \cdot 0.198^x \cdot (1 - 0.198)^{1500-x} \approx 0.065$$




Figur 3.9: Sannsynligheten i GeoGebra

For å forkaste nullhypotesen har vi stilt dette kravet

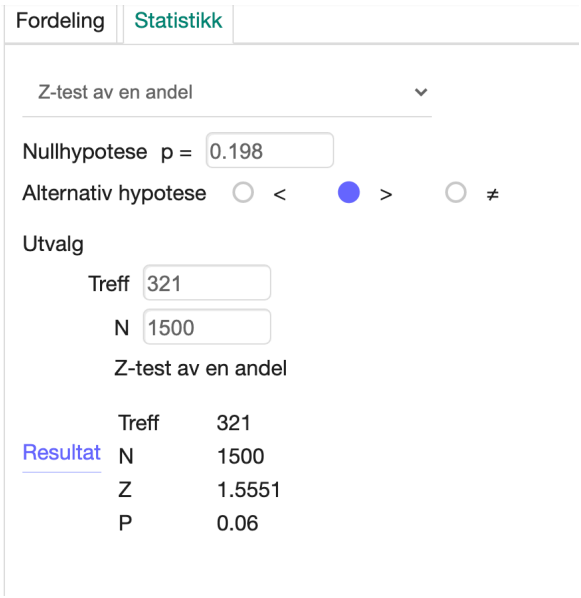
$$P(X \geq 321) \mid p = 0.198) \leq 0.05$$

Resultat ble

$$P(X \geq 321) \approx 0.065$$

Kravet var at denne sannsynligheten skulle være mindre enn 0.05 for å forkaste nullhypotesen. Vi har ikke grunnlag for å si at partiet har gått fram siden forrige valg. Vi beholder nullhypotesen med det valgte signifikansnivået.

Det fins en kalkulator i GeoGebra som kan benyttes for hypotesetesting av andel. Figur 3.10 viser hvordan vi kan skrive inn nullhypotesen og alternativ hypotese for å få ut en P-verdi. Vi må velge *Z-test av andel* og GeoGebra benytter en tilnærma normalfordeling for å regne ut. Det gjør at vi ikke får nøyaktig samme sannsynlighet som ved utregning. Fordelen med å regne ut i en binomisk sannsynlighetsmodell er at vi både får et mer nøyaktig svar og at det egentlig er enklere.




Figur 3.10: Hypotesetesting i GeoGebra

# Pythonkode