

Et lite notat om og rundt normalfordelingen.

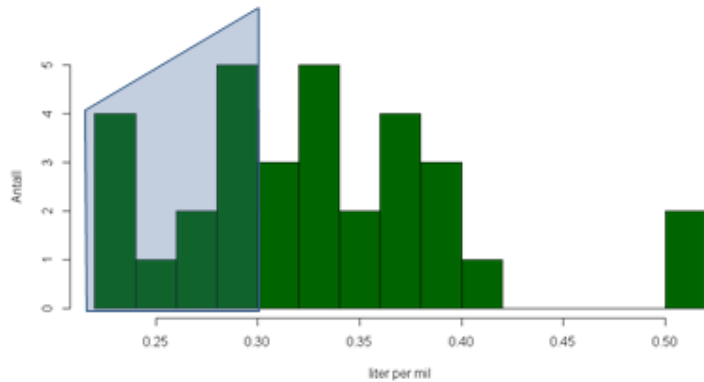
Anta at vi har kontinuerlige data. Hva er likt og ulikt for histogrammer og fordelingskurver?

Boka (Ch 1.4) motiverer dette ved å gå fra histogrammer til tetthetskurver.

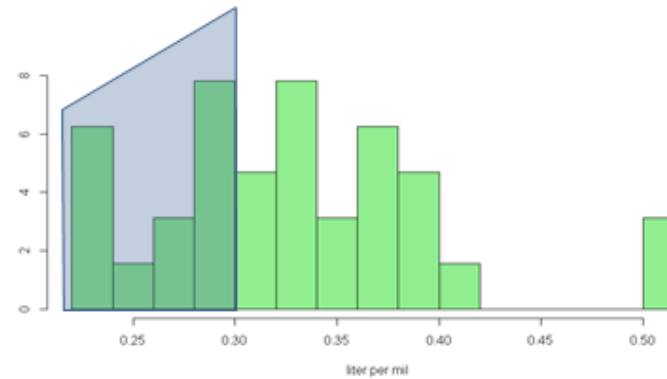
Tetthetskurver

Eksempel: Drivstofforbruk hos 32 biler

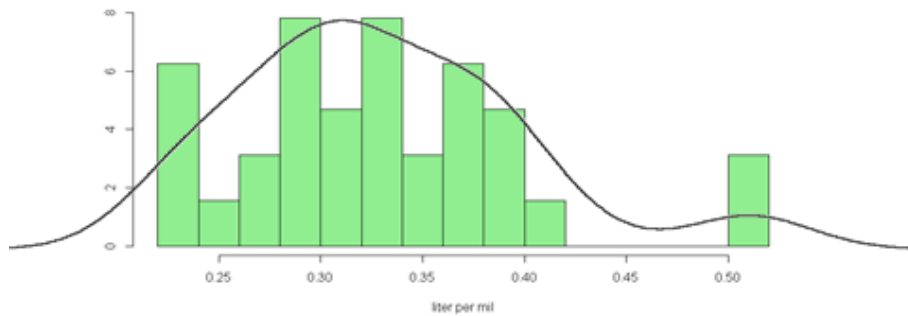
Histogram over drivstofforbruket hos 32 biler. Antall på y-aksen.
Andel biler som bruker mindre enn 0.3 liter per mil:
 $(4+1+2+5)/32 = 0.375 \approx 0.38$



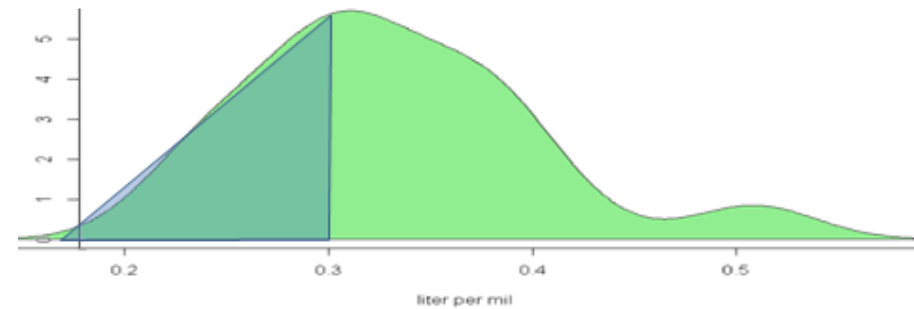
Histogram over drivstofforbruket hos 32 biler. y-aksen er skalert slik at andel biler i hvert intervall tilsvarer arealet av søylen i det intervallet. Andelen biler som bruker mindre enn 0.3 liter per mil blir da: $(6.3+1.6+3.1+7.8) \cdot 0.02 \approx 0.38$



Histogram over drivstofforbruket hos 32 biler.
Tetthetskurve (glattet histogram)



Tetthetskurve for drivstofforbruket.
Andel biler som bruker mindre enn 0.3 liter per mil \approx
areal under kurven i det aktuelle intervallet:
 $(5.5 \cdot (0.3-0.15))/2 = 0.4125 \approx 0.41$

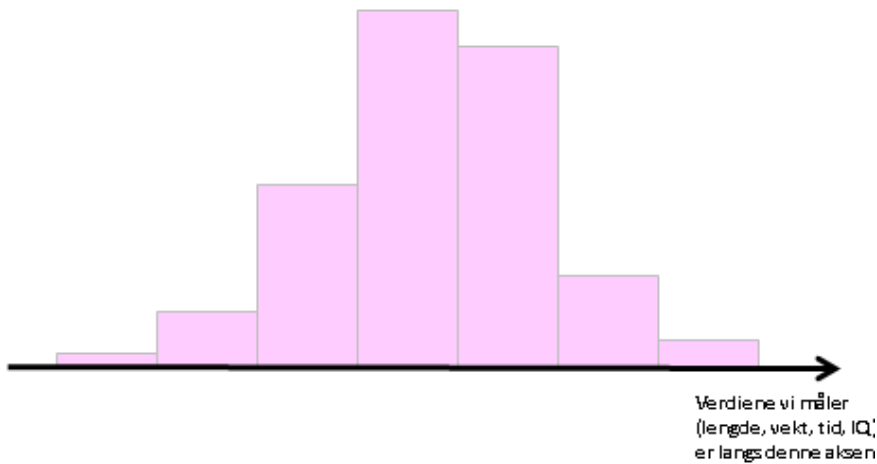
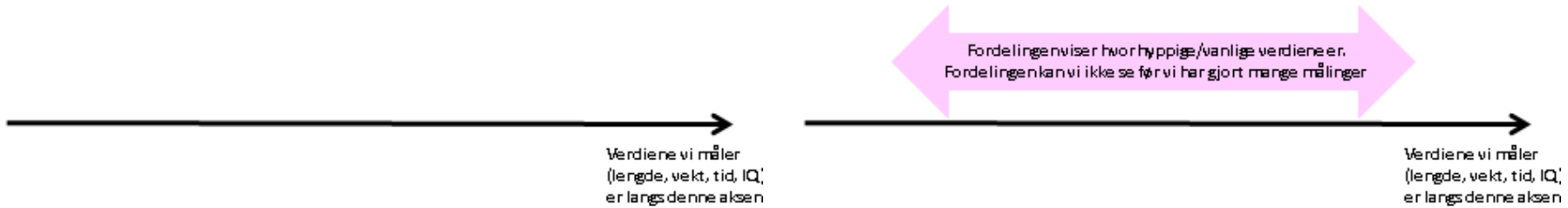


Et lite notat om og rundt normalfordelingen.

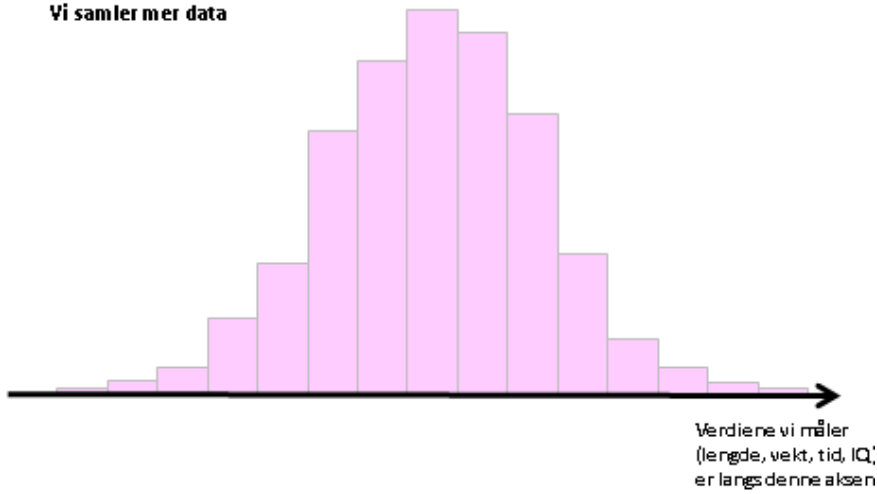
Anta at vi har kontinuerlige data. Hva er likt og ulikt for histogrammer og fordelingskurver?

Observasjoner Histogram	Teoretisk/tenkt Fordelingskurve
Viser fordelingen av faktiske observerte verdier som søyler i intervaller	Viser den idealiserte fordelingen av verdier som en kontinuerlig kurve
Kan være symmetrisk, skjev, ha en eller flere topper, lette eller tunge haler (outliere/ekstremverdier)	Kan være symmetrisk, skjev, ha en eller flere topper, lette eller tunge haler
Gjennomsnitt: \bar{x} Tyngdepunktet for histogrammet	Forventningsverdi: μ Tyngdepunktet for kurven
Median: Verdien som har like mange observasjoner på hver side	Median: Verdien som har like mye areal under kurven på hver side
Kvartiler: Deler inn observasjonene i fire grupper med like mange observasjoner i hver gruppe	Kvartiler: Deler inn x-aksen i fire intervaller, så det er like stort areal under kurven i hvert intervall.
(Empirisk) standardavvik: sd, SD	Standardavvik: σ
Varsians	Varsians: σ^2
	Toppunkt = Mode
Histogrammer kan enten vise antall på y-aksen, eller skaleres slik at arealet av hver søyle tilsvarer andelen av observasjoner som er i det gitte intervallet.	Fordelingskurver har totalt areal lik 1. Velger vi et intervall på x-aksen, vil arealet under kurven i dette intervallet tilsvare andelen verdier som havner i intervallet.

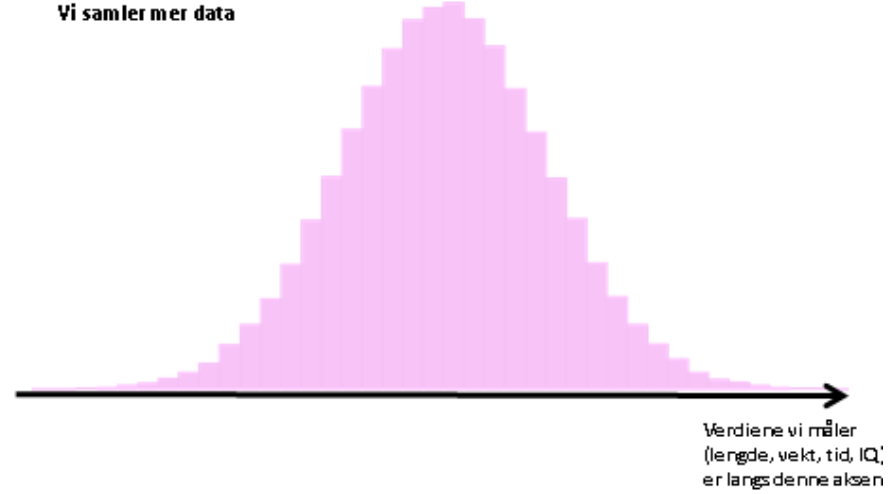
Innsamling av kontinuerlige data



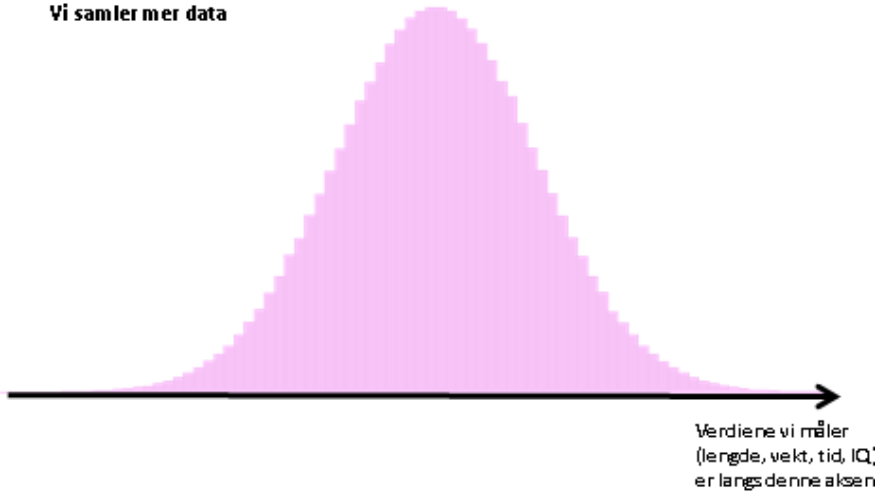
Vi samler mer data



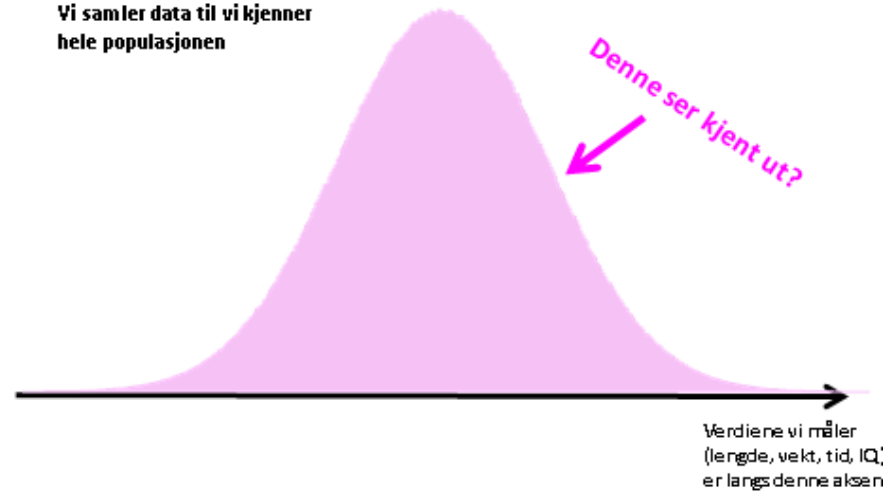
Vi samler mer data



Vi samler mer data



Vi samler data til vi kjenner hele populasjonen



Normalfordeling

Hvordan ser en normalfordeling ut?

OBS: Teoretisk sannsynlighetstetthet

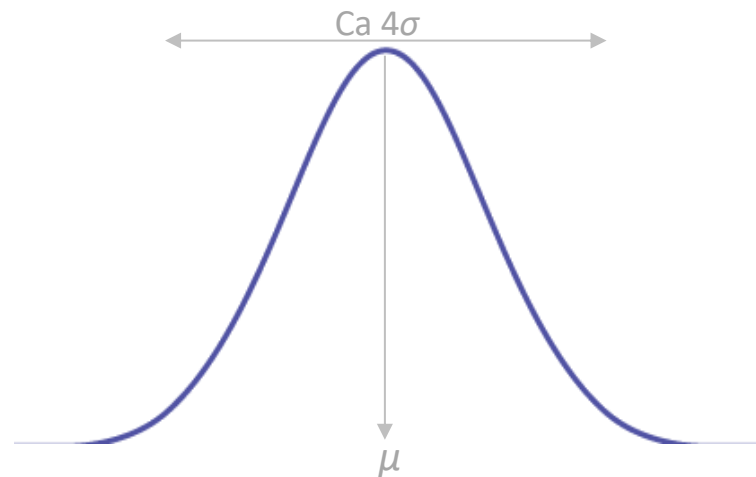
Entoppet, symmetrisk, lette haler. «Bell-shaped»

Formelen for normalfordeling:

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

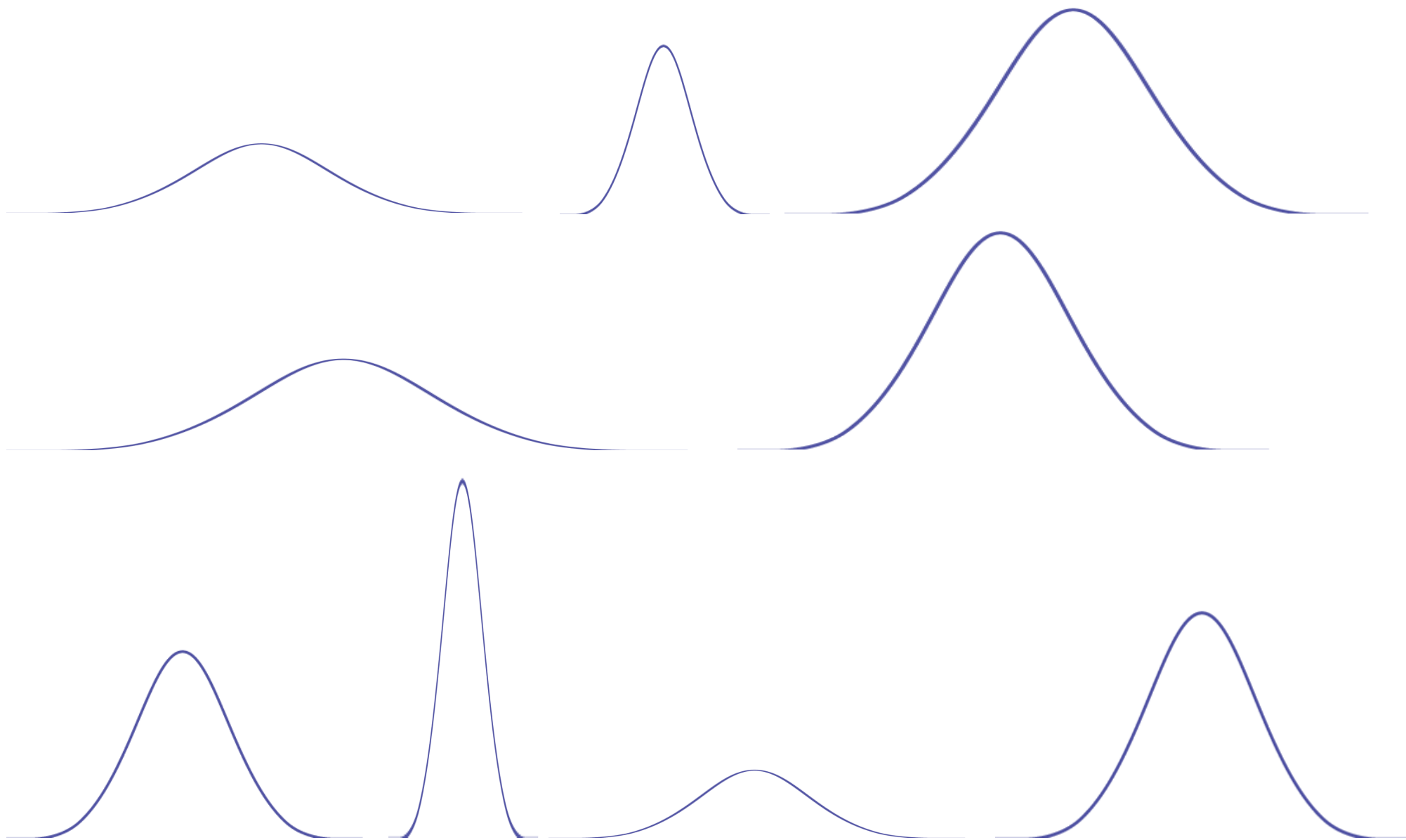
μ = Mean

σ = Standard Deviation



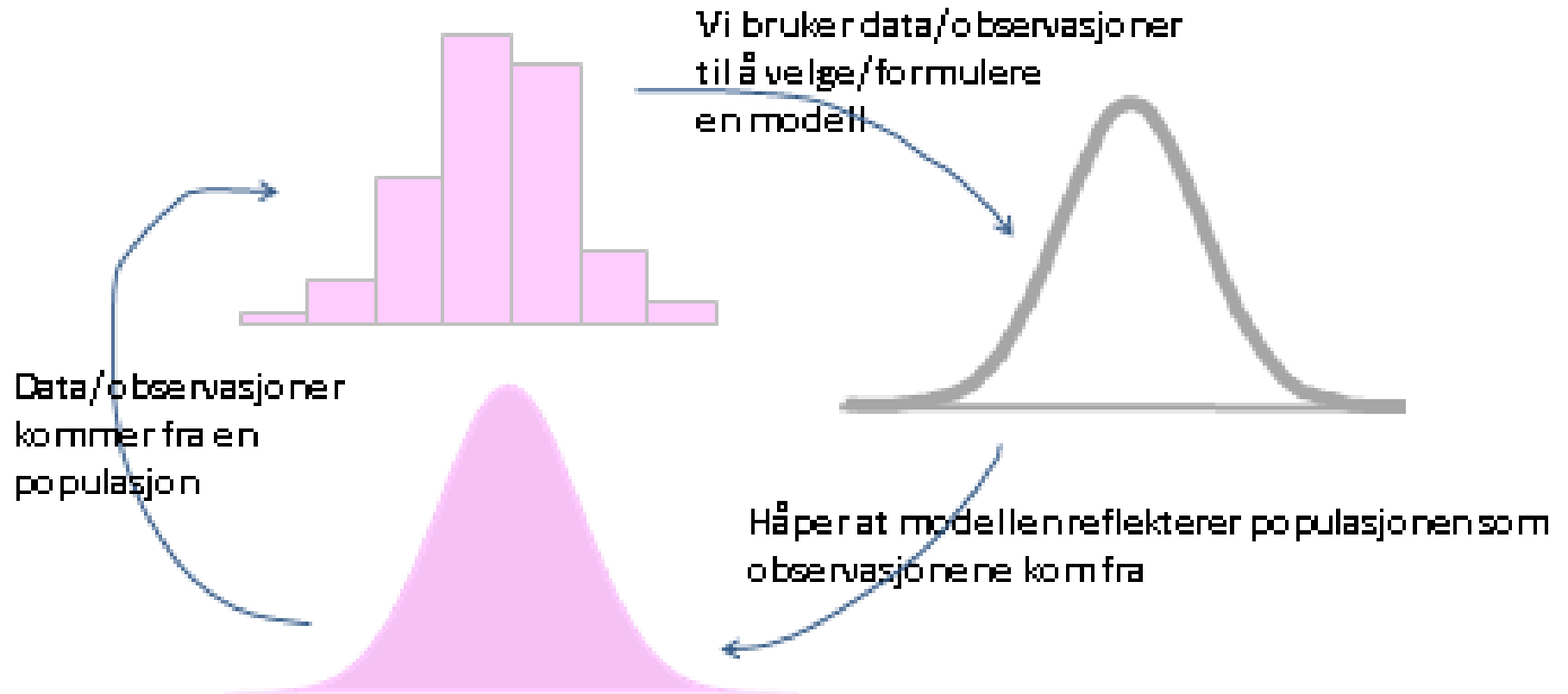
Normalfordelingen og dens far: Johann Carl Friedrich Gauß (1777–1855)

Hei på deg, normalfordeling! Finn μ og σ

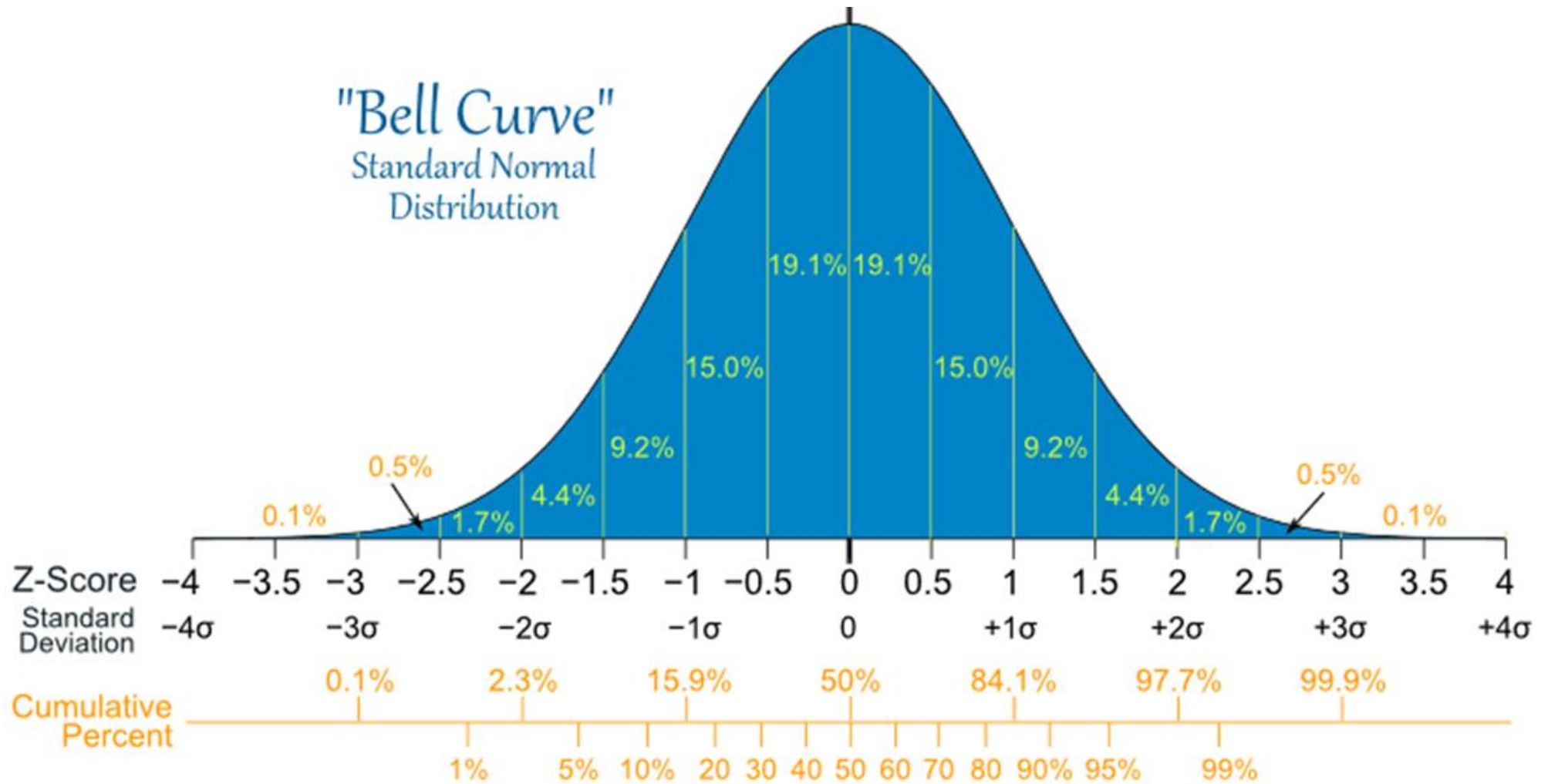


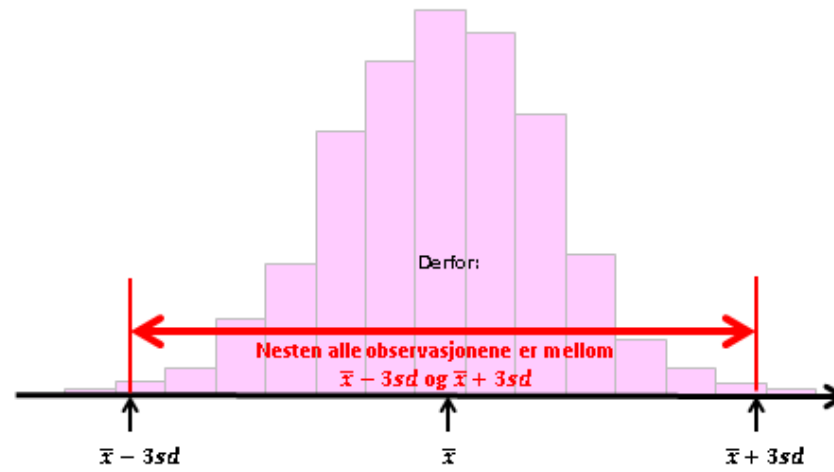
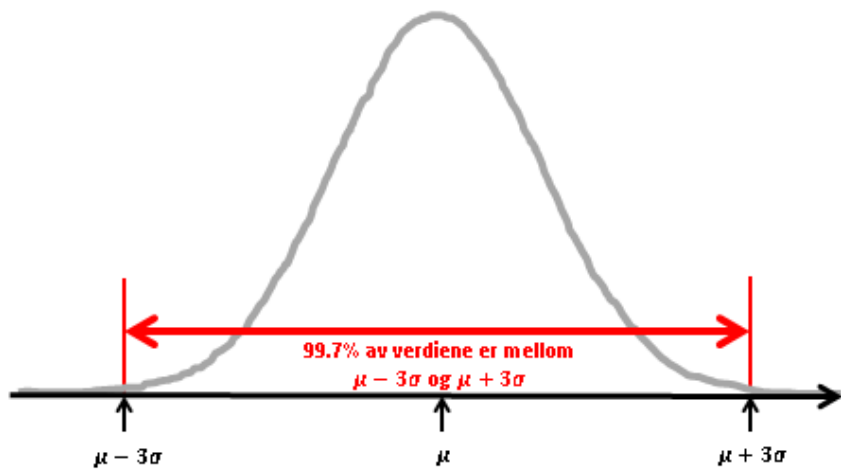
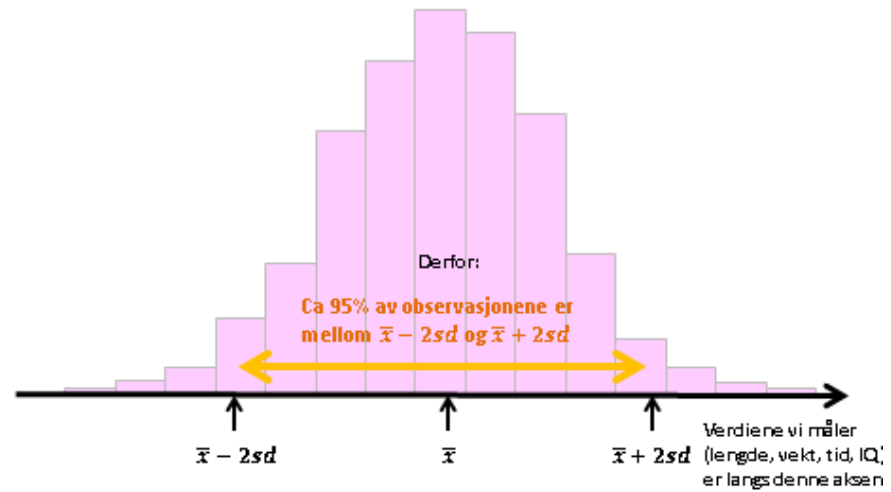
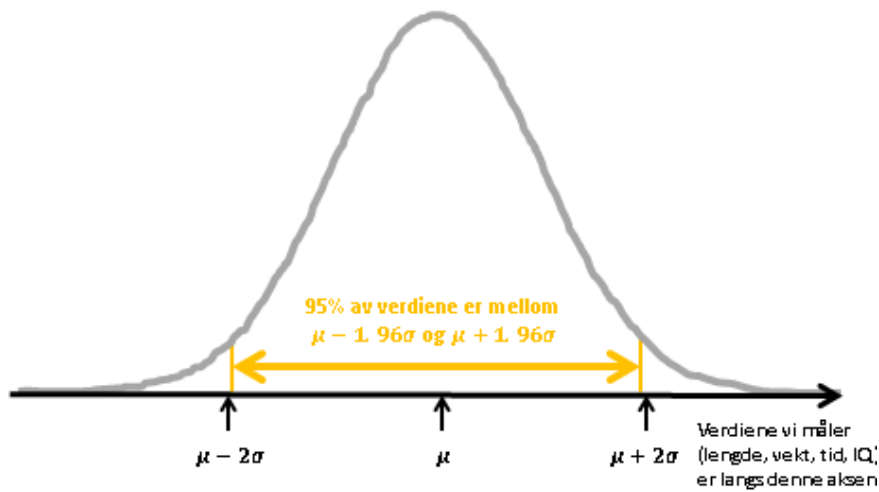
Hva bruker vi den til, og hvorfor er den viktig?

Normalfordelingen brukes (blant annet) som modell for observerte data, når fordelingen til observasjonene er entoppet, symmetrisk og har lette haler:



Normalfordelingen er formulert som en formel (!), og kan regnes på:





Normalfordelingen og deskriptiv statistikk

Praktisk anvendelse av det vi vet om normalfordelingen er nyttig når vi skal oppgi og tolke oppsummeringstall for det vanlige/typiske/senteret i en fordeling av kontinuerlige data (av og til diskrete data):

Standardavviket er et utmerket oppsummeringstall for spredningen når datasettet ligner på en normalfordeling, altså når fordelingen er symmetrisk, entoppet, og har lette haler, fordi:

Da vil intervallet $\bar{x} \pm 2sd$ inneholde omtrent 95% av observasjonene, og så godt som ingen observasjoner befinner seg utenfor $\bar{x} \pm 3sd$.

Hvis man oppgir \bar{x} og sd som oppsummeringstall, har man altså implisitt sagt at fordelingen ligner på en normalfordeling, at \bar{x} er omtrent midt i fordelingen, og at de fleste observasjonene ligger i intervallet som strekker seg 2 (3) standardavvik ut fra midten på begge sider.

Hvis jeg forteller deg at i mitt datasett, så kan

a) deltakernes fastende blodsukkernivå oppsummeres med $\bar{x} = 6.1$ og $sd = 1.6$

b) deltakernes BMI oppsummeres med $\bar{x} = 31$ og $sd = 2.4$

Tegn en skisse av fordelingen i a) og fordelingen i b).

Standardisering: Fra originalobservasjoner til z-scorer.

Normalfordelingen er gitt ved

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

μ = Mean

σ = Standard Deviation

der y viser høyden på y -aksen, og x er verdiene til variabelen som er normalfordelt.

Siden μ kan være et hvilket som helst tall på tallinja, og σ kan være et hvilket som helst tall større eller lik null, finnes det uendelig mange normalfordelinger.

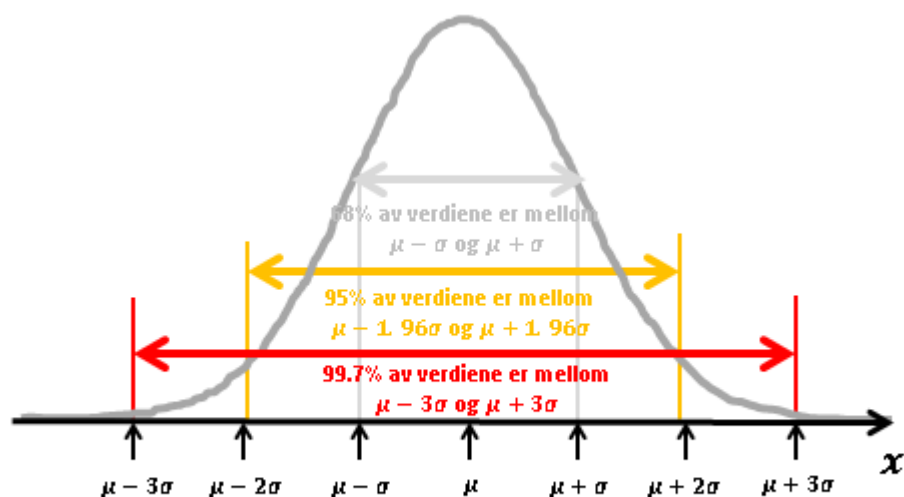
Men alle normalfordelte verdier kan regnes om til en såkalt standard normalfordeling, som har forventning 0 og standardavvik 1.

Hvis vi vet hva μ og σ er, finner vi z -scoren, altså den standardiserte verdien av x , slik:

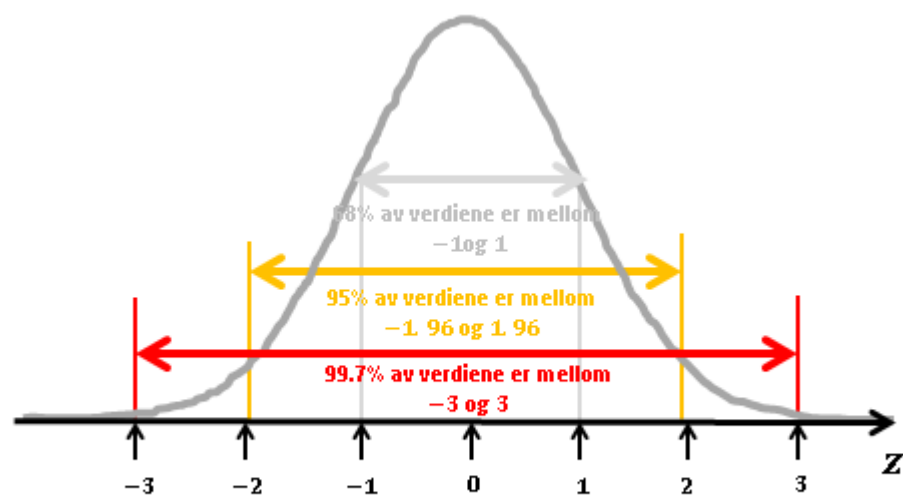
$$z = \frac{x - \mu}{\sigma}$$

Vi bruker notasjonen $N(\mu, \sigma)$ for en generell normalfordeling, og $N(0, 1)$ for en standard normalfordeling. Ofte skriver vi også $X \sim N(\mu, \sigma)$, og $Z \sim N(0, 1)$. Slik ser fordelingene ut:

$$X \sim N(\mu, \sigma)$$

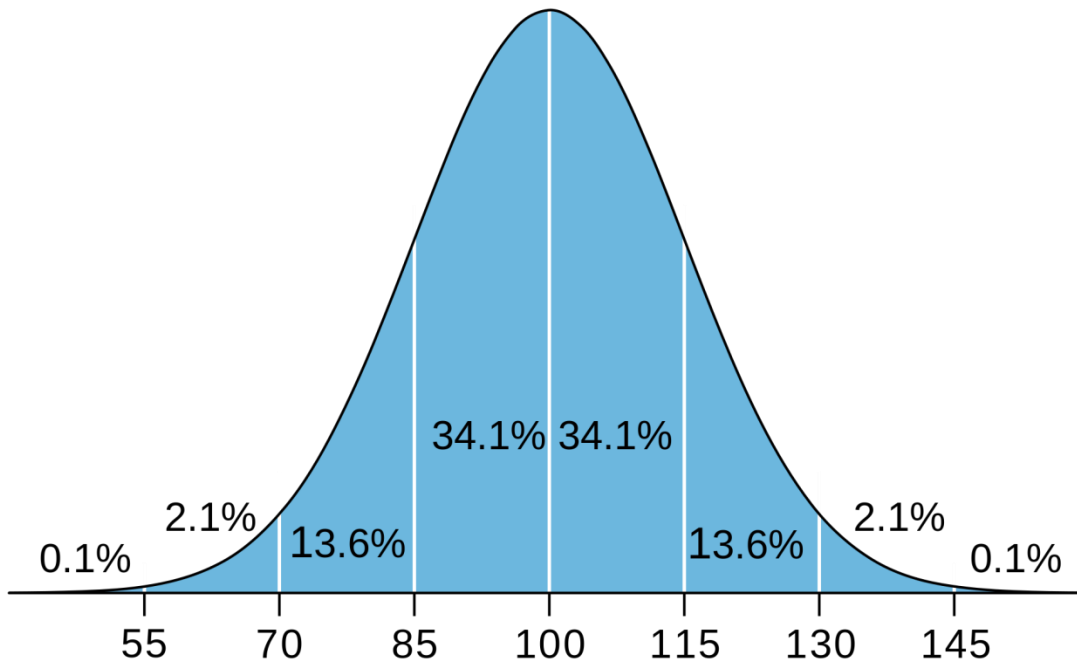


$$Z \sim N(0, 1)$$



Eksempel: IQ (Ukas oppgave)

IQ-tester er konstruert for å ha et populasjonsgjennomsnitt (μ) på 100, og standardavvik (σ) på 15:



Jeg tok en IQ-test på www.funeducation.com, og endte med et resultat på 115.

Den tilsvarende z-scoren blir da

$$z = \frac{x - \mu}{\sigma} = \frac{115 - 100}{15} = 1$$

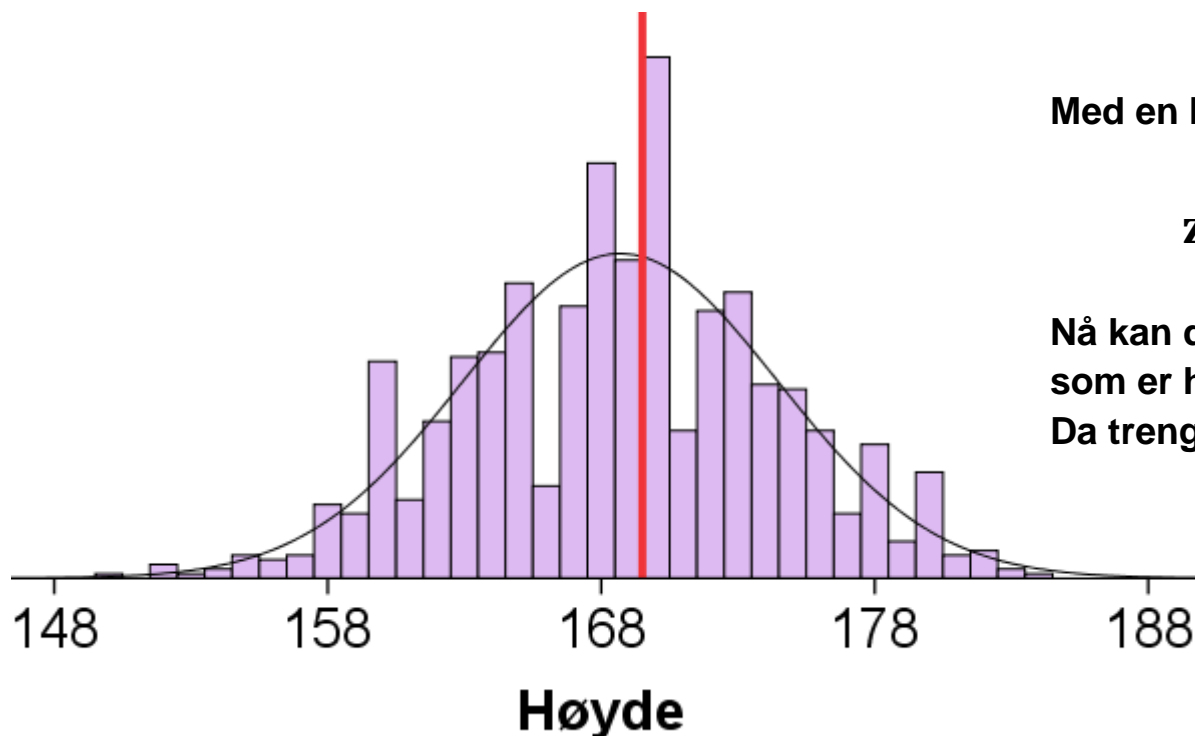
Ser på figuren at siden z-scoren er nøyaktig 1, kan det leses rett ut av figuren at 15.8% av befolkningen (13.6%+2.1%+0.1%) er smartere enn meg.

Ta samme IQ-test som jeg har tatt, og regn om til z-score.

Hvis vi ikke vet hva μ og σ er, kan vi erstatte dem med \bar{x} og sd, og allikevel lage z-scorer.

Eksempel: Høyde:

I en fil med høyde for 1016 kvinner som har født barn i perioden 2000-2007, er gjennomsnittshøyden 168.7 cm, og sd 5.8 cm



Med en høyde på 169.5 cm har KFF en z-score på

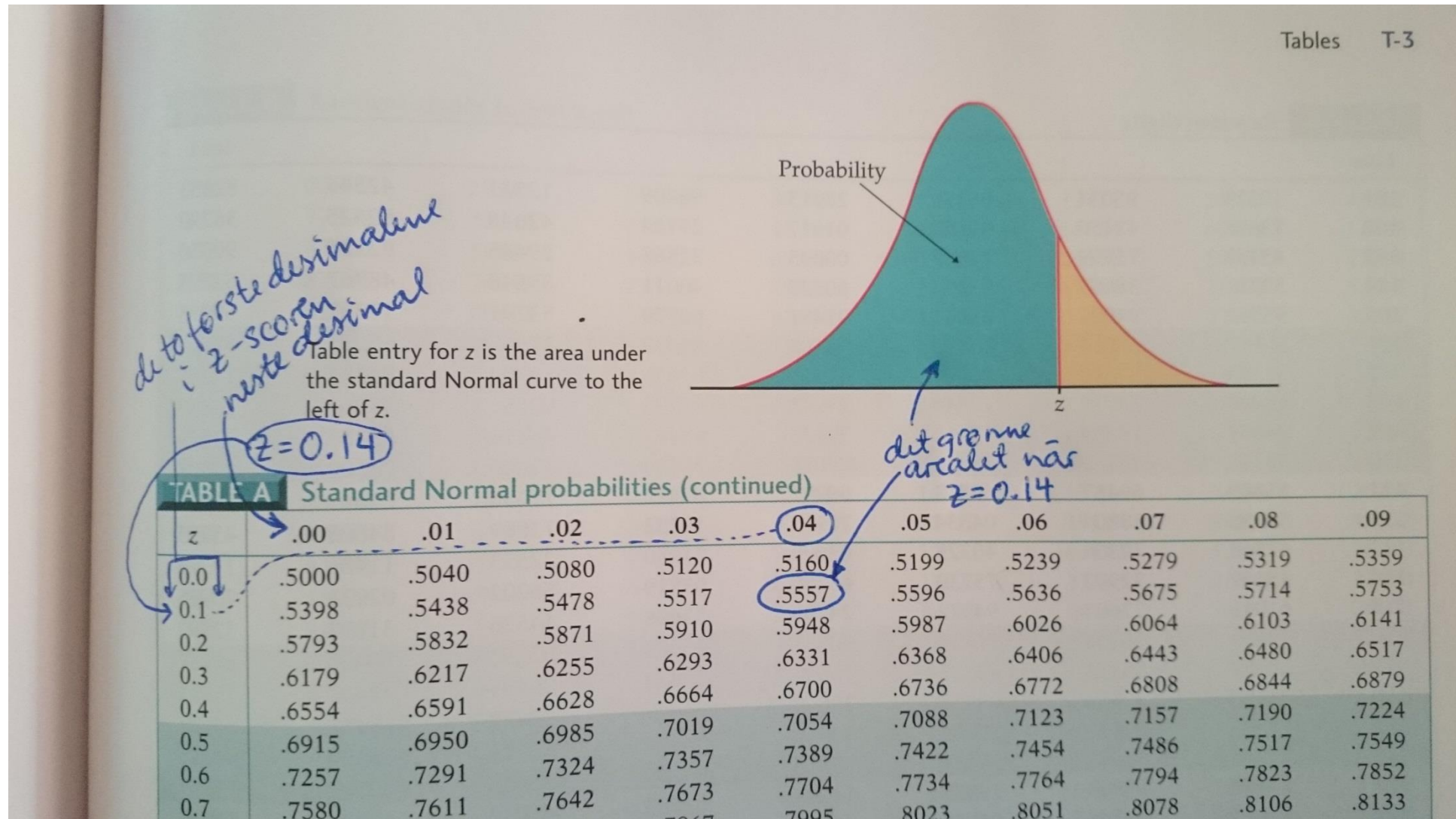
$$z = \frac{x - \mu}{\sigma} = \frac{169.5 - 168.7}{5.8} = 0.14$$

Nå kan det ikke leses rett ut av figuren hvor stor andel som er høyere eller lavere enn meg.

Da trenger vi en normalfordelingstabell!

Etter eksemplet: Bruk opplysningene over samt en normalfordelingstabell til å regne ut z-scoren for din egen høyde (hvis du er kvinne), din drømmekvinne (reell eller ikke), eller hun som sitter nærmest deg akkurat nå.

Table A, side 720-721 i boka til Moore & McCabe:

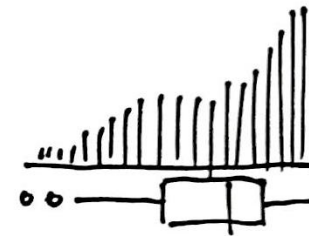
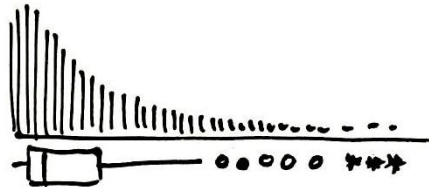
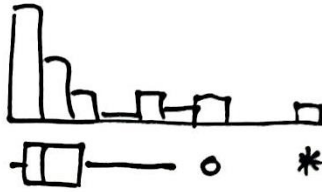
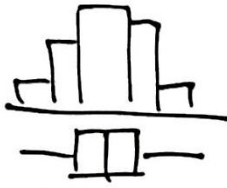


Med andre ord: 55.6% av norske kvinner er lavere enn meg.

```
R: > pnorm(0.14)
[1] 0.55567
```


Hvordan sjekker vi om noe «er normalfordelt»?

Lag histogram, boksplott: **Sjekk symmetri og haler**



Fortsatt usikker? Lag normalfordelingsplott (QQ-plott):

Hvis normalfordelt:

Rett linje

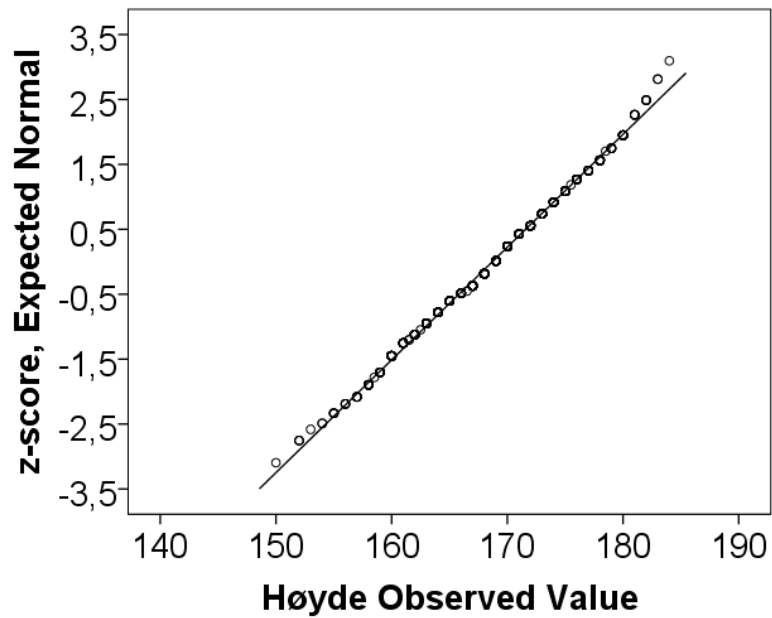
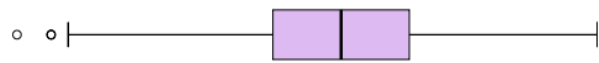
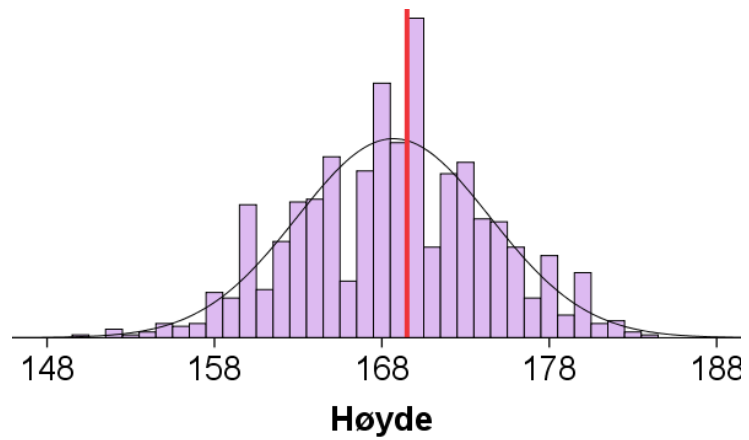
Hvis ikke rett linje:

Ikke normalfordelt

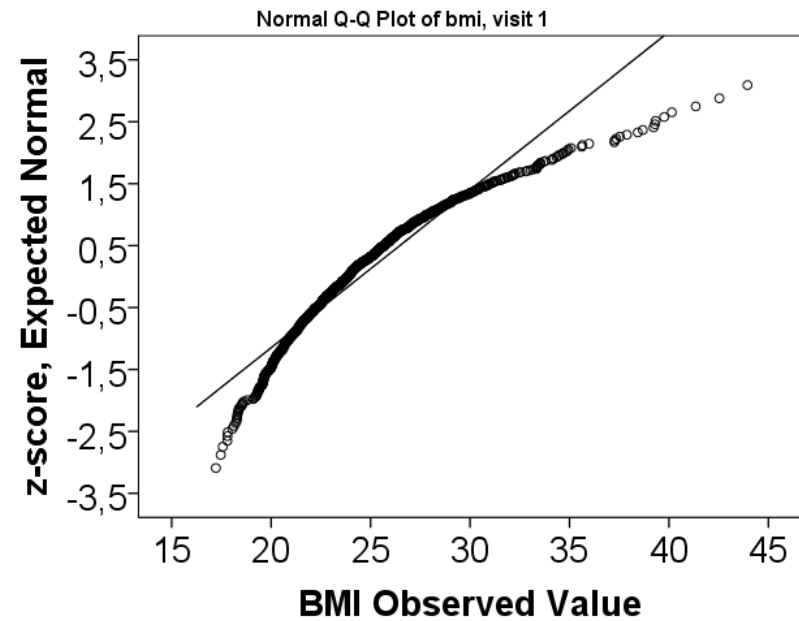
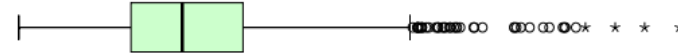
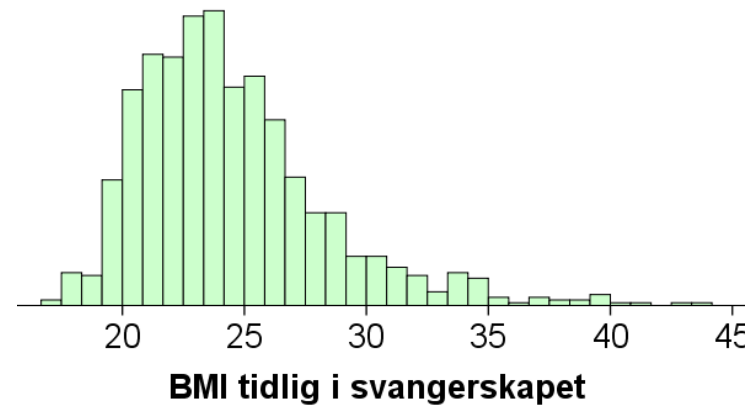
QQ-plott/Normal quantile plot lages ikke for hånd, vi bruker alltid programvare. Da gjøres dette i bakgrunnen:

- ✓ Sortér alle observasjonene dine fra lavest til høyest, og gjør dem om til percentiler, altså verdier som sier hvor stor andel av de observerte verdiene som er mindre eller lik den verdien du kikker på.
- ✓ Finn hvilken z-score de ulike percentilene ville hatt hvis de var i en $N(0,1)$ -fordeling (Bruk tabellen motsatt vei)
- ✓ Plott observasjonene mot de tilsvarende z-scorene

Hvis normalfordelt: QQ-plottet viser en rett linje



QQ-plottet viser ikke en rett linje: Ikke normalfordelt



(Fortsatt:) Hva bruker vi den til, og hvorfor er den viktig?

Mange matematiske resultater (kall det gjerne formler) er basert på at enten dataene selv, eller en avledning av dataene, er normalfordelt.

For eksempel er formelen for 95% konfidensintervall for populasjonsgjennomsnittet (forventningen) μ :

$$\bar{x} \pm 1.96SE(\bar{x}),$$

Denne formelen stammer fra *sentralgrenseteoremet*, der normalfordelingen også dukker opp.

Dessuten: Normalfordelingen er en god approksimasjon til binomisk fordeling med stor n , og til poissonfordeling med høy frekvens, og det gjør utregningene av tester og konfidensintervaller mye enklere.

Vi får da også at 95% konfidensintervall for en andel p i populasjonen:

$$\hat{p} \pm 1.96SE(\hat{p})$$

Så: Mange statistiske analyser forutsetter at normalfordelingen er til stede, enten at observasjonene er normalfordelte, eller at noe avledet av data er normalfordelt.

I de to formelene på denne siden finner vi tallet 1.96. Hvor kommer det fra?

Analyseoversikt

Versjon 1

Hva skal vi gjøre?

Oppsummere/presentere/
beskrive data

Inferens

Beregne
konfidens-intervall

Hypotesetesting:

Sammenligne to eller flere grupper
(Undersøke om det er sammenheng
mellom to variabler, der den ene er en
en kategorisk variabel)

H_0 : Gruppene er like
Konfidensintervall kan også brukes

Undersøke om det er sammenheng
mellom to variabler, der den ene er en
kontinuerlig variabel
 H_0 : Ingen sammenheng

Analysere forskjell på par av data
 H_0 : Ingen forskjell

Kvantifisere samsvar.
Utgangspunkt: Det er samsvar
Reliabilitet

Analysere sammenheng mellom en
responsvariabel og en eller flere
kovariater (forklaringsvariabler eller
prediktorer)
 H_0 : Ingen effekt/ingen prediktiv verdi

H v i l k e n t y p e d a t a h a r v i ?

Kategoriske data

Tabeller, andeler
Søyle(stolpe)/kakediagram

95% KI for p , som er
populasjonens andel

$$\hat{p} \pm 1.96 \cdot SE(\hat{p})$$

«Tabellanalyse»
Krysstabeller
Andeler

Differanse av andeler m/ 95% KI
Relativ risiko (RR) m/ 95% KI
Odds-ratio (OR) m/ 95% KI
Pearsons χ^2 -test (kji-kvadrat-test)

To kategorier: t -test /Wsr-test
Flere kategorier: ANOVA/KW

Krysstabeller
Andelen samsvar & McNemars test

Krysstabeller
Cohens kappa

Binær responsvariabel:
Logistisk regresjon
Effekt mål: OR m/ 95% KI
 H_0 : OR = 1

Kontinuerlige data

Skjeve



Median, kvartiler

95% KI for
populasjonens median

Tabeller? Bootstrapping?

To grupper:
Median, kv i hver gruppe
Wilcoxon rank sum test

Flere grupper:
Median, kv i hver gruppe
Kruskal-Wallis test
Posthoc-tester=Wrs-tester

Spearman's
korrelasjonskoeffisient

Skjevfordelte differanser:
Wilcoxon signed-rank test

Symmetriske



\bar{x} , SD

95% KI for μ , som er
populasjonens gjennomsnitt

$$\bar{x} \pm 1.96 \cdot SE(\bar{x})$$

To grupper:
 \bar{x} , SD i hver gruppe
Diff. av gj.sn. m/ 95% KI
To-utvalgs t -test

Flere grupper:
 \bar{x} , SD i hver gruppe
Enveis ANOVA. Hvis $p < 0.05$
Posthoc-tester = t -tester

Pearson's
korrelasjonskoeffisient r

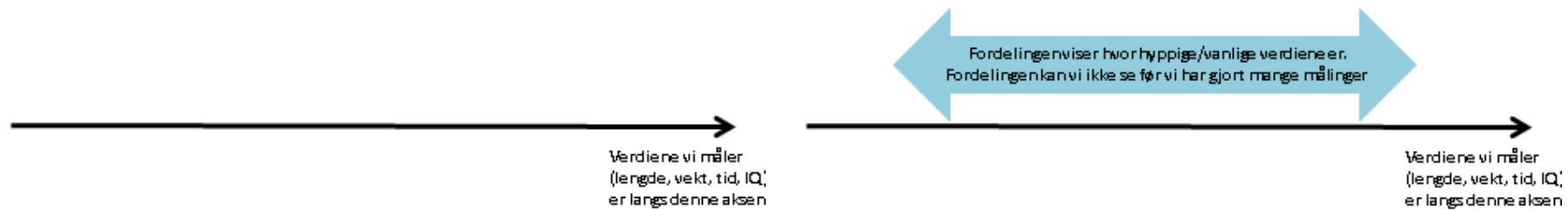
Normalfordelte differanser:
Paret t -test

Scatterplott med $y=x$,
Bland-Altman-plott
ICC

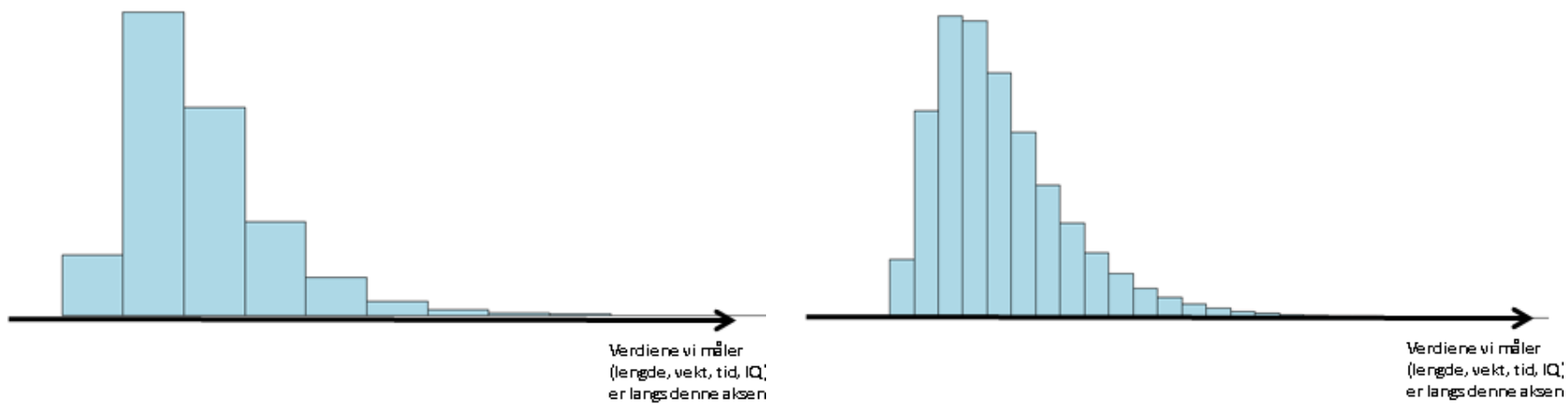
Kontinuerlig responsvariabel:
Lineær regresjon
Effekt mål: B
 H_0 : B = 0
(Normalfordelte residualer)

Mange data er ikke normalfordelte. Hva gjør vi med dem?

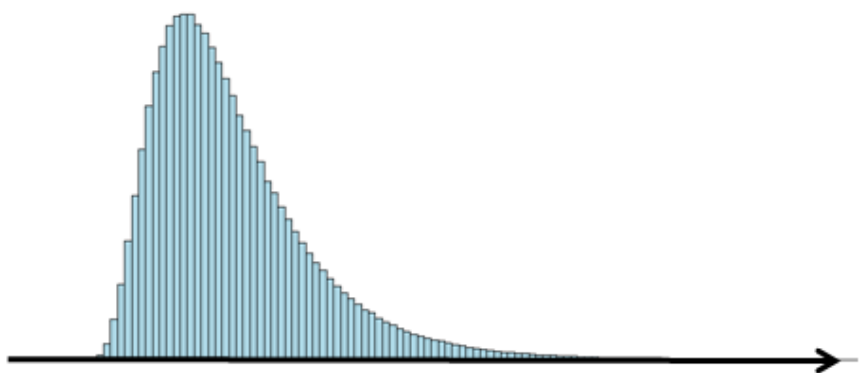
Andre data



Vi samler mer data

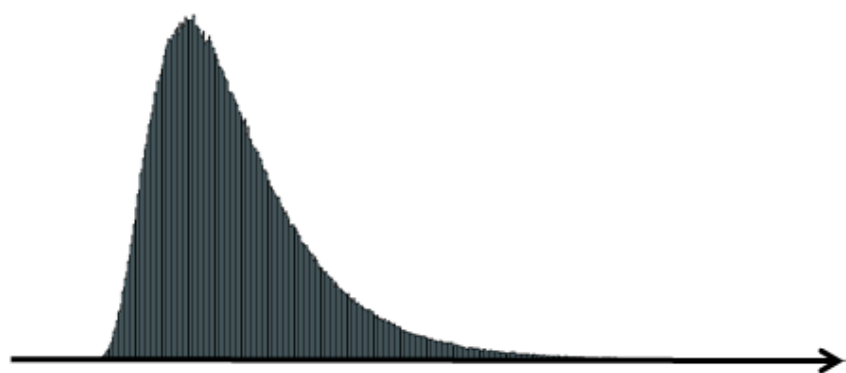


Vi samler mer data



Verdiene vi måler
(lengde, vekt, tid, IQ,
er langs denne akse

Vi samler mer data



Verdiene vi måler
(lengde, vekt, tid, IQ,
er langs denne akse

Tetthetskurve for en skjev fordeling

