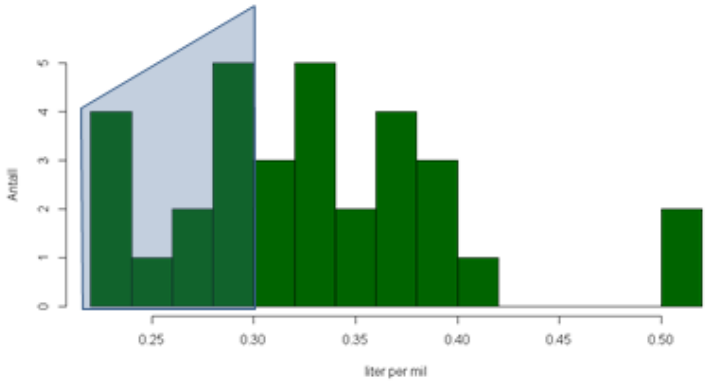
STK1000 Uke 36, 2016. Studentene forventes å lese Ch 1.4 (+ 3.1-3.3 + 3.5) i læreboka (MMC).

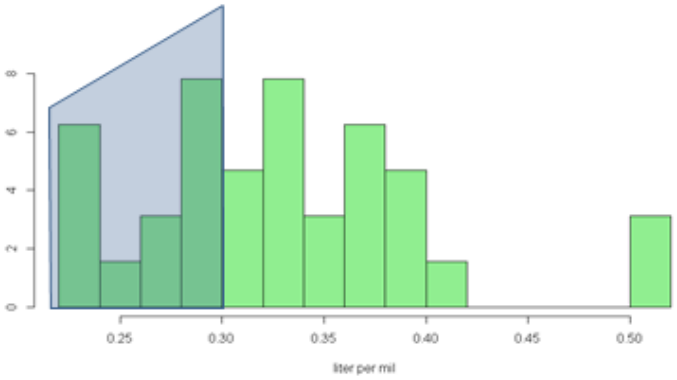
Tetthetskurver

Eksempel: Drivstofforbruk hos 32 biler

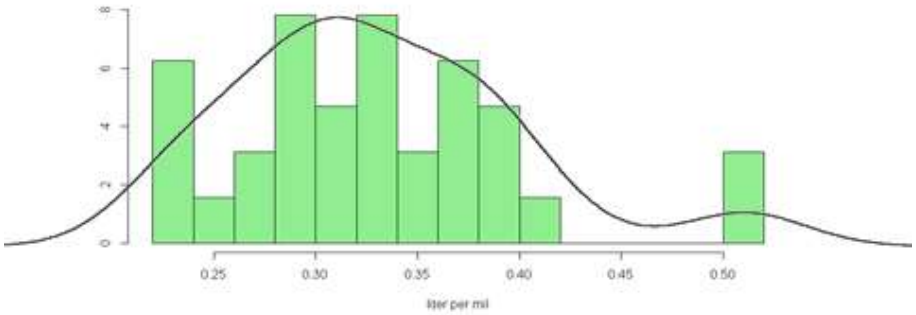
Histogram over drivstofforbruket hos 32 biler. Antall på y-aksen.
Andel biler som bruker mindre enn 0.3 liter per mil:
(4+1+2+5)/32 = 0.375 ≈ 0.38



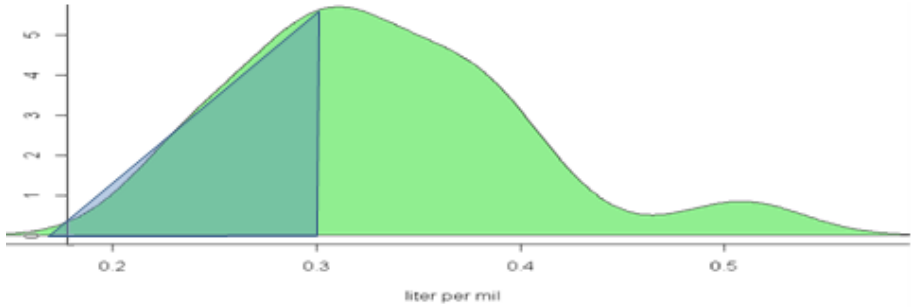
Histogram over drivstofforbruket hos 32 biler. y-aksen er skalert slik at andel biler i hvert intervall tilsvarer arealet av søylen i det intervallet. Andelen biler som bruker mindre enn 0.3 liter per mil blir da: (6.3+1.6+3.1+7.8) · 0.02 ≈ 0.38



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



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

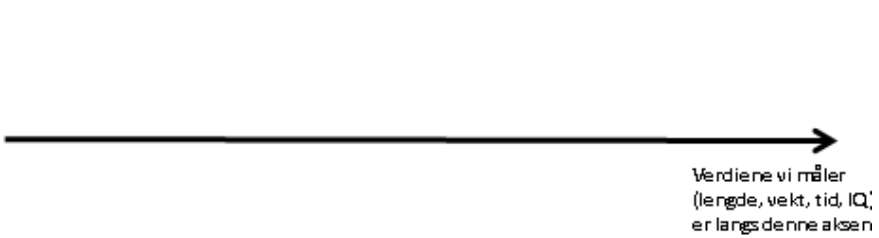


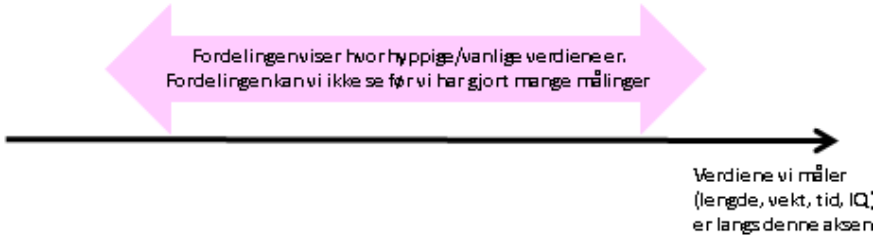
Fra histogram til tetthetskurver

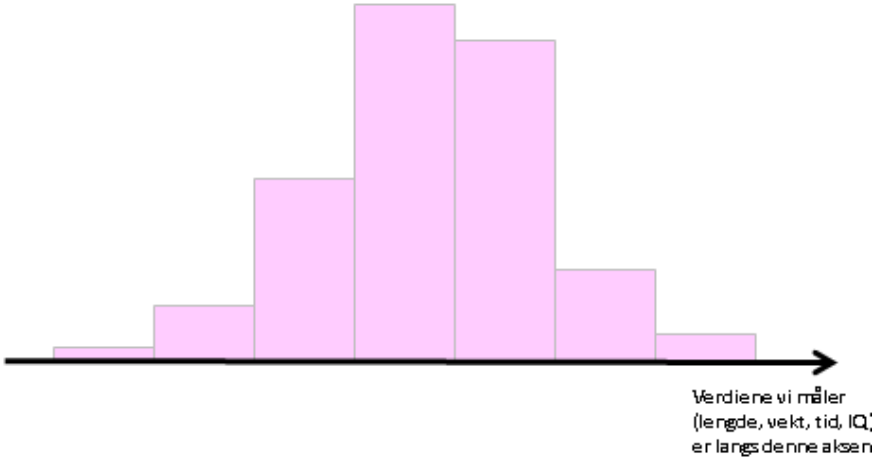
Anta at vi har kontinuerlige observasjoner

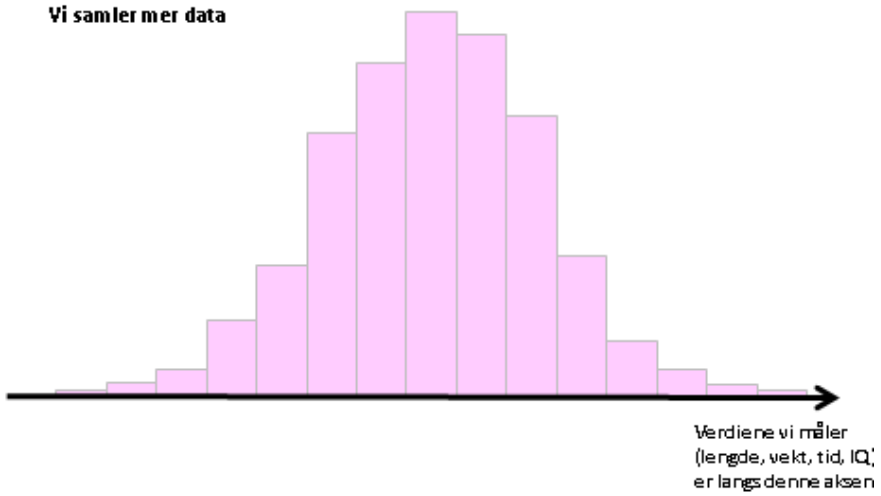
Observasjoner	**Teoretisk/tenkt**
Histogram	**Tetthetskurve**
Viser fordelingen av faktiske observerte verdier som søyler i intervaller	**Viser den idealiserte fordelingen av verdier som en kontinuerlig kurve**
Kan være symmetrisk, skjev, ha en eller flere topper, lette eller tunge haler (outliere/ekstremverdier)	**Kan være symmetrisk, skjev, ha en eller flere topper, lette eller tunge haler**
Gjennomsnitt: $\bar{x}$ Tyngdepunktet for histogrammet	**Forventningsverdi: μ Tyngdepunktet for tetthetskurven**
Median: Verdien som har like mange observasjoner på hver side.	**Median: Verdien som har like mye areal under tetthetskurven 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**
Varians	**Varians: σ^2**
	Toppunkt = Mode
Histogrammer kan enten vise antall på y-aksen, eller skaleres slik at arealet av hver søyle reflekterer andelen av observasjonene som er i det gitte intervallet.	**Tetthetskurver har totalt areal lik 1. Velger vi et intervall på x-aksen, vil arealet under en tetthetskurven i dette intervallet tilsvare andelen observasjoner som havner i intervallet.**

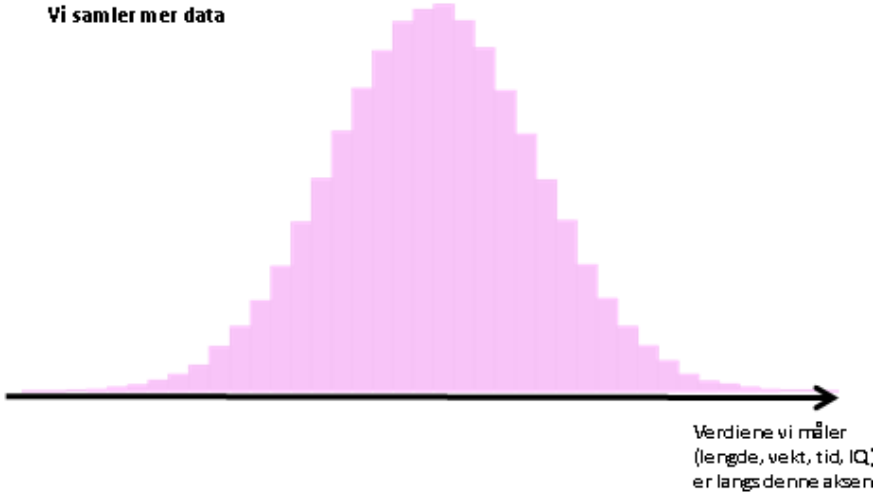
Innsamling av kontinuerlige data

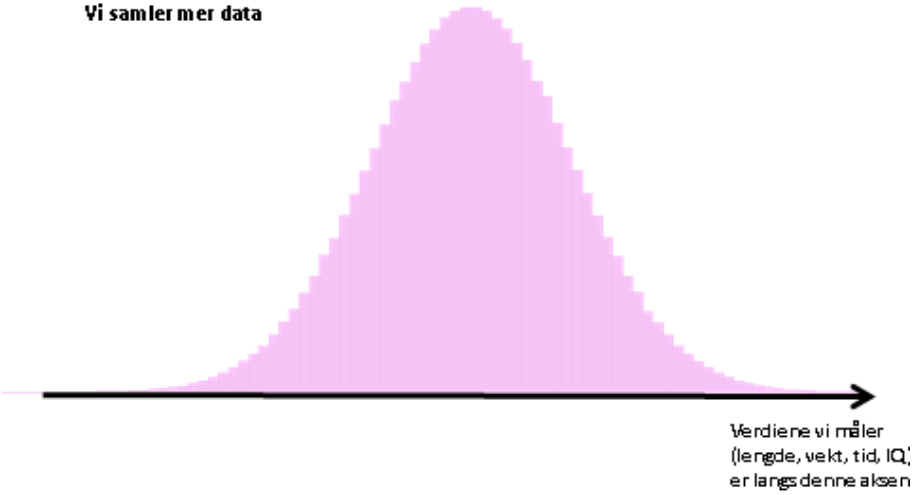






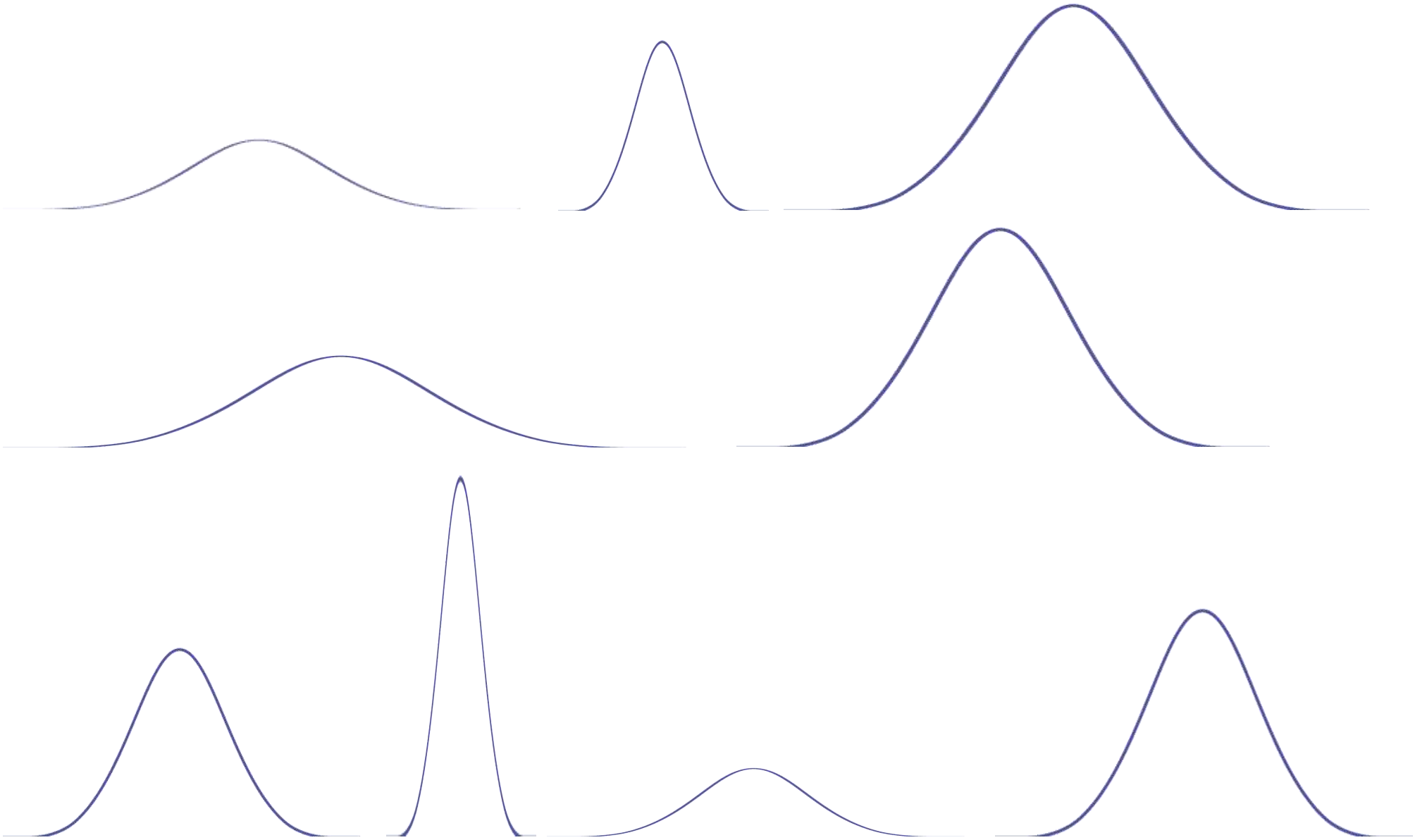

Vi samler mer data
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen


Vi samler mer data
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen


Vi samler mer data
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen


Vi samler data til vi kjenner
hele populasjonen
Denne ser kjent ut?
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen

Hei på deg, normalfordeling! (Finn μ og σ)



Normalfordeling

Hvordan ser en normalfordeling ut?

OBS: Teoretisk sannsynlighetstetthet

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

Formelen for normalfordeling:

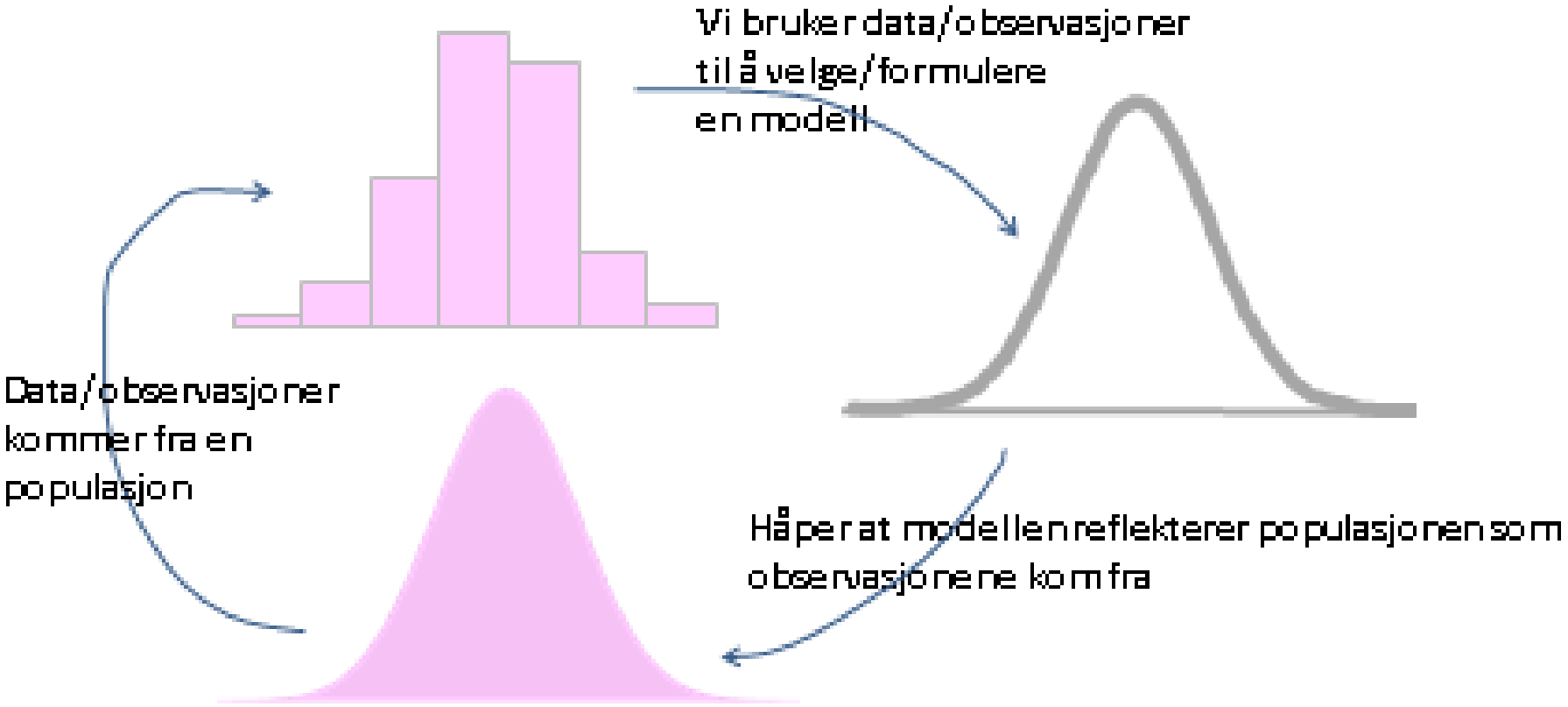
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

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

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:



En praktisk anvendelse, husk fra forrige uke:

Oppsummeringstall for det vanlige/typiske/senteret i en fordeling av kontinuerlige data (av og til diskrete data):

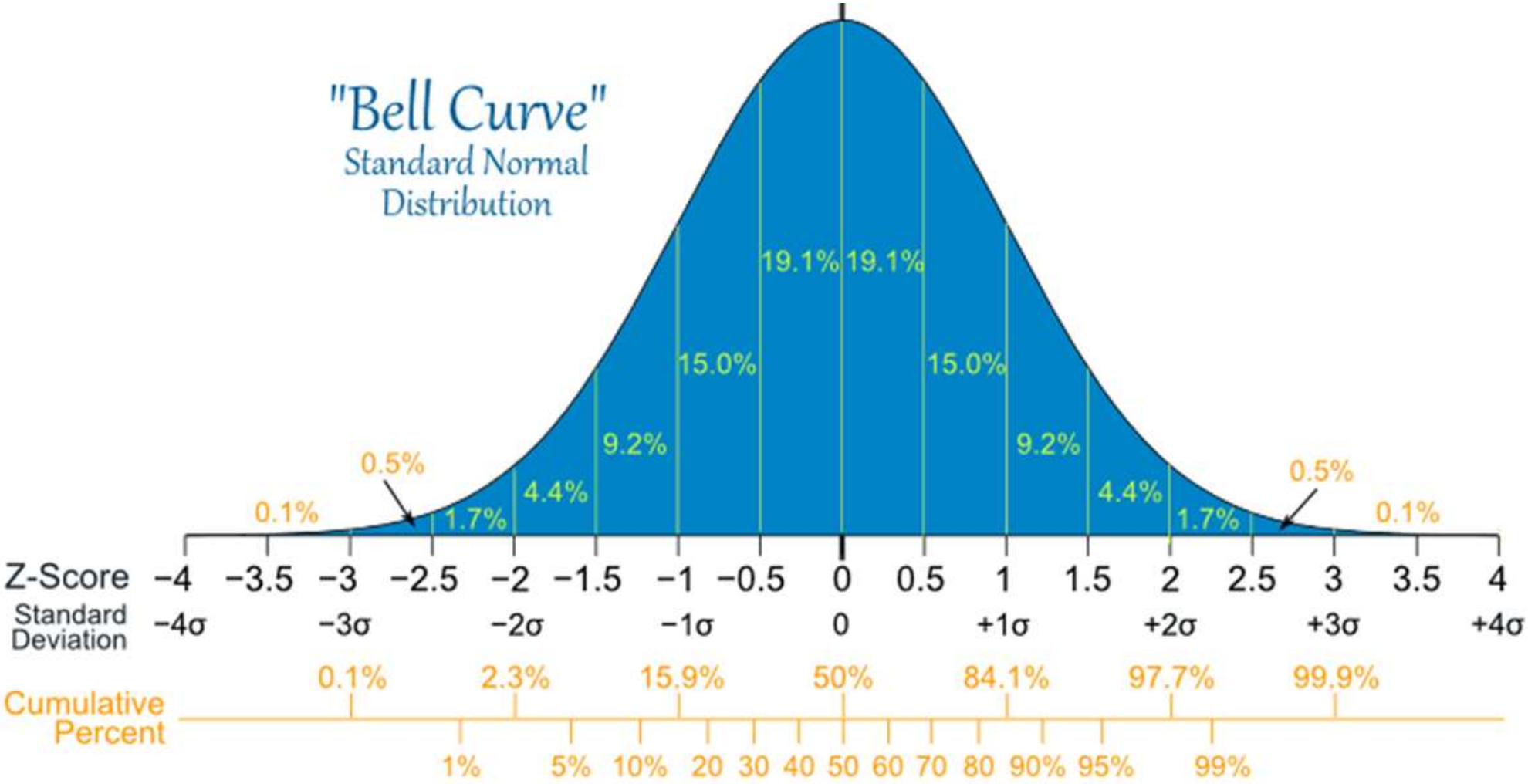
Standardavviket er et utmerket oppsummeringstall for spredningen når datasettet ligner på en normalfordeling, altså når fordelingen er symmetrisk, entoppet, og har lette haler, fordi:

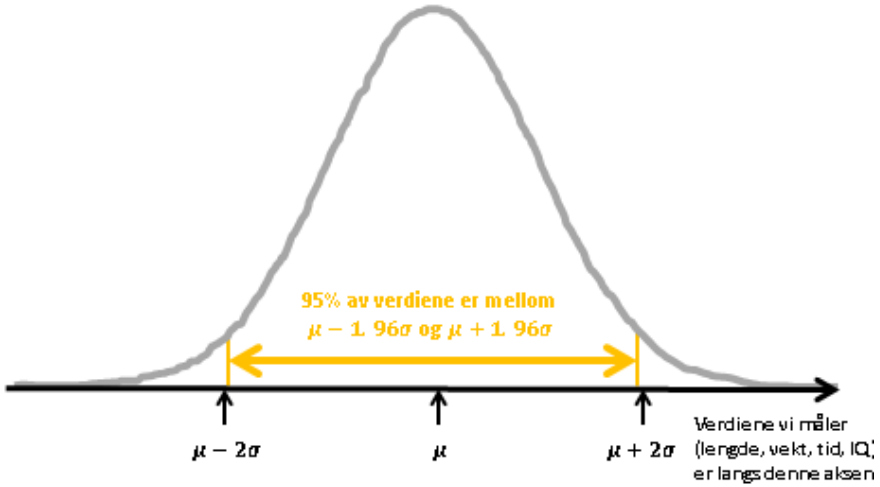
Da vil intervallet $\bar{x} \pm 2sd$ ***inneholde omtrent 95% av observasjonene, og så godt som ingen observasjoner befinner seg utenfor*** $\bar{x} \pm 3sd$***.***

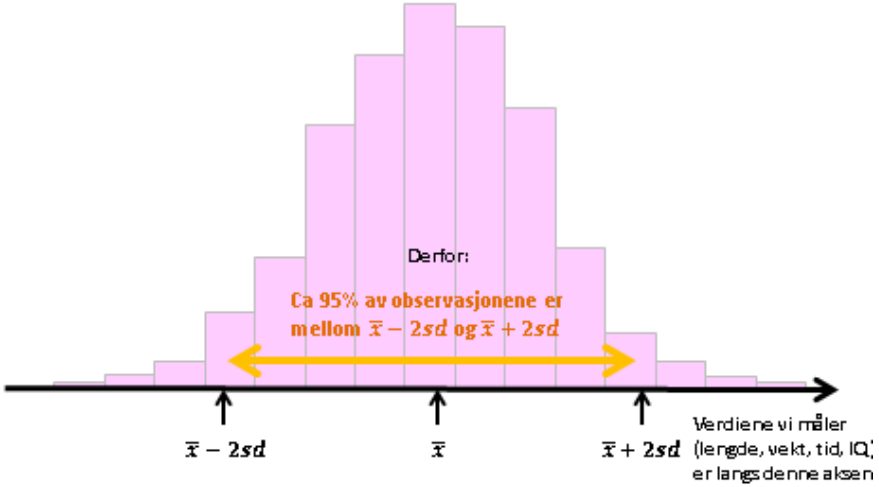
Hvis man oppgir $\bar{x}$ ***og*** sd ***som oppsummeringstall, har man altså implisitt sagt at fordelingen ligner på en normalfordeling, og at de fleste observasjonene ligger mellom 2 til 3 standardavvik fra midten.***

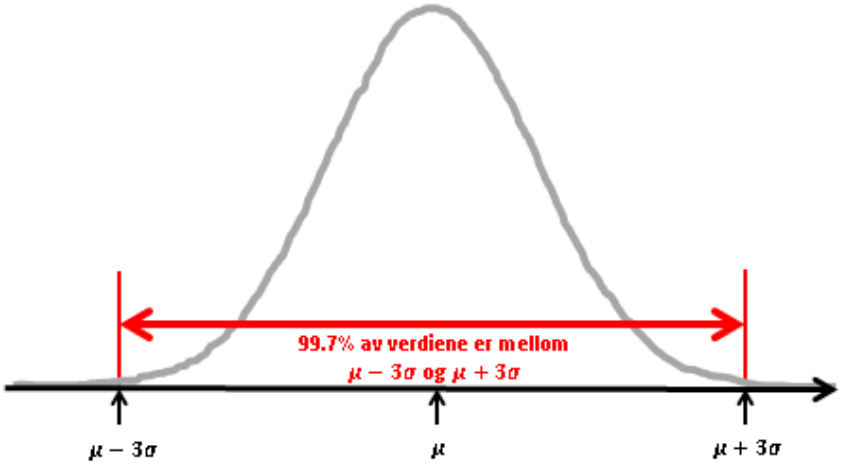
Hvorfor?

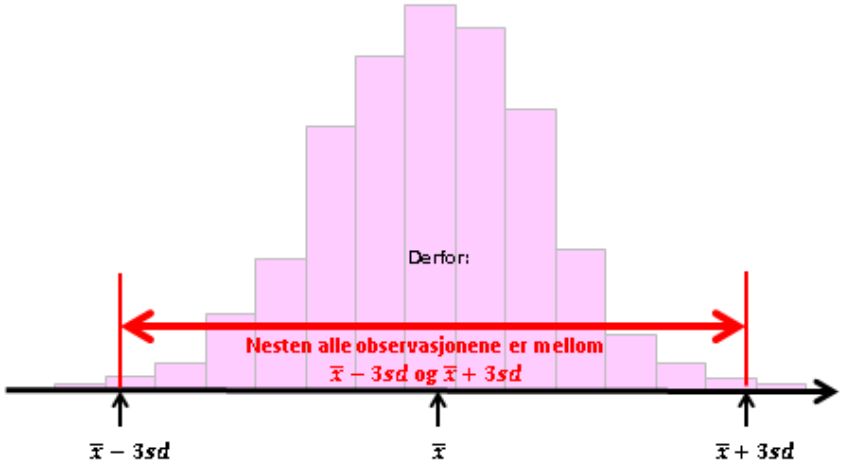
Normalfordelingen er formulert som en formel (!), og kan regnes på:


"Bell Curve"
Standard Normal Distribution
19.1%
19.1%
15.0%
15.0%
9.2%
9.2%
4.4%
4.4%
1.7%
1.7%
0.5%
0.5%
0.1%
0.1%
Z-Score
−4 −3.5 −3 −2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5 3 3.5 4
Standard Deviation
−4σ −3σ −2σ −1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percent
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
1% 5% 10% 20 30 40 50 60 70 80 90% 95% 99%


95% av verdiene er mellom
$\mu - 1.96\sigma$ og $\mu + 1.96\sigma$
$\mu - 2\sigma$
μ
$\mu + 2\sigma$
Verdiene vi måler (lengde, vekt, tid, IQ) er langs denne aksen


Derfor:
Ca 95% av observasjonene er mellom $\bar{x} - 2sd$ og $\bar{x} + 2sd$
$\bar{x} - 2sd$
$\bar{x}$
$\bar{x} + 2sd$
Verdiene vi måler (lengde, vekt, tid, IQ) er langs denne aksen


99.7% av verdiene er mellom
$\mu - 3\sigma$ og $\mu + 3\sigma$
$\mu - 3\sigma$
μ
$\mu + 3\sigma$


Derfor:
Nesten alle observasjonene er mellom
$\bar{x} - 3sd$ og $\bar{x} + 3sd$
$\bar{x} - 3sd$
$\bar{x}$
$\bar{x} + 3sd$

Standardisering: Fra originalobservasjoner til z-scorer.

Normalfordelingen er gitt ved

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

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

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.

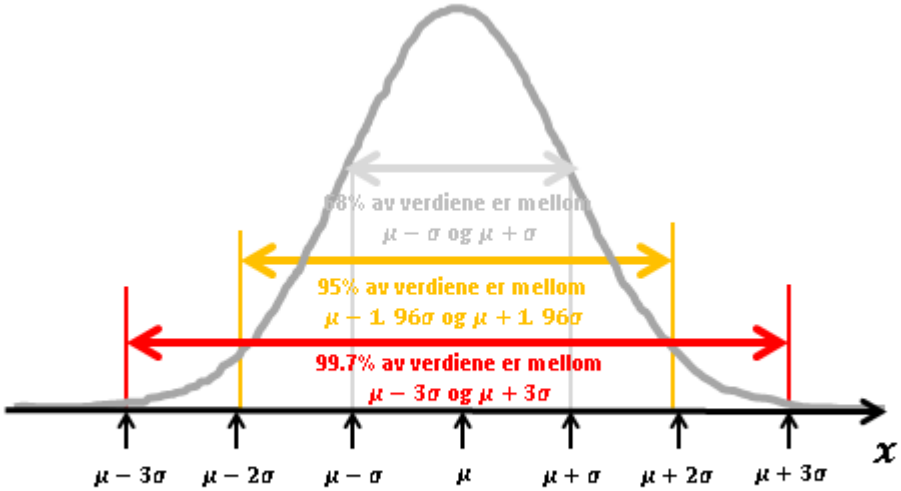
Men alle normalfordelte verdier kan regnes om til en såkalt standard normalfordeling, som har forventning 0 og standardavvik 1.

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

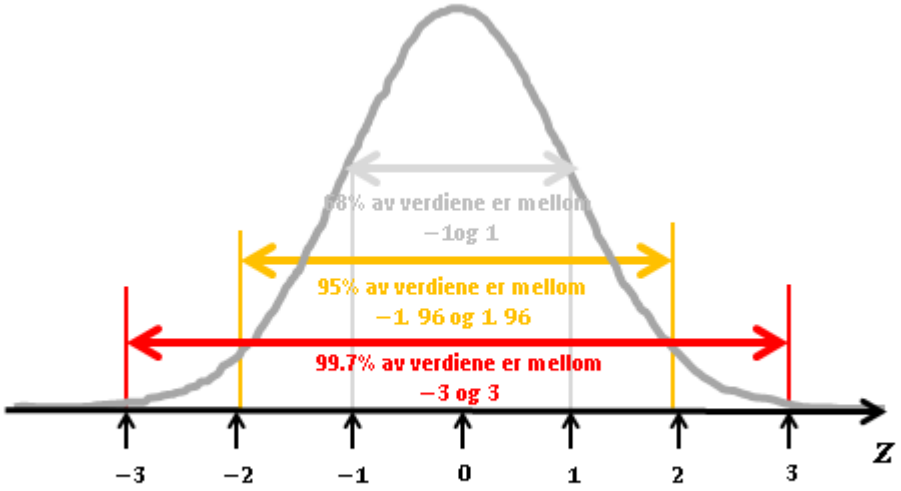
$$\boldsymbol{z = \frac{x - \mu}{\sigma}}$$

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:

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

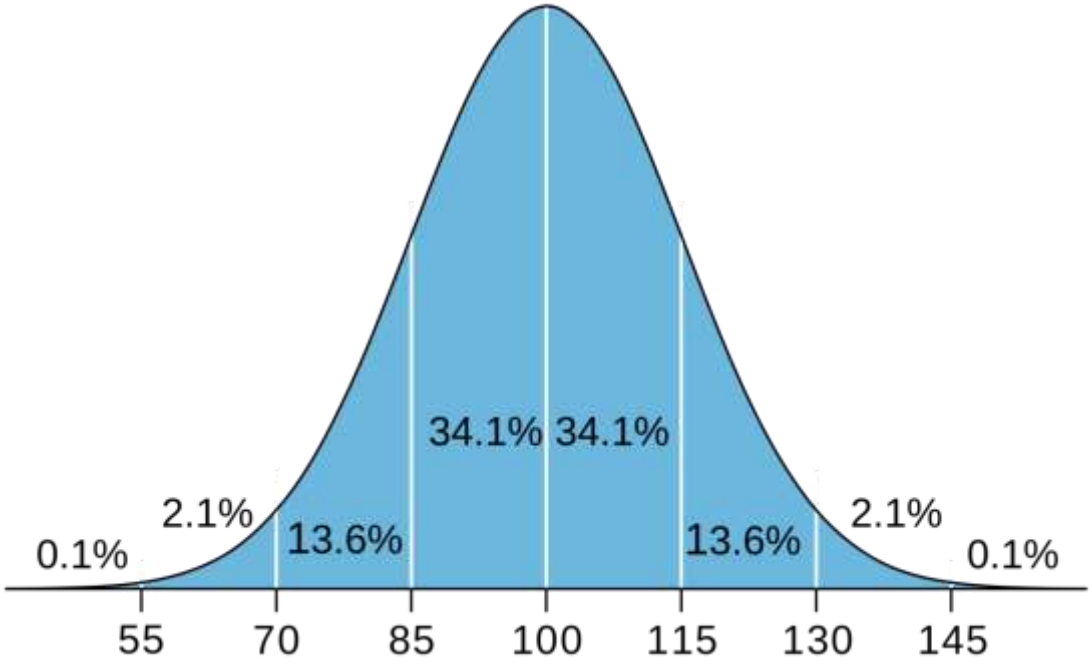


$$Z \sim N(0,1)$$



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

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

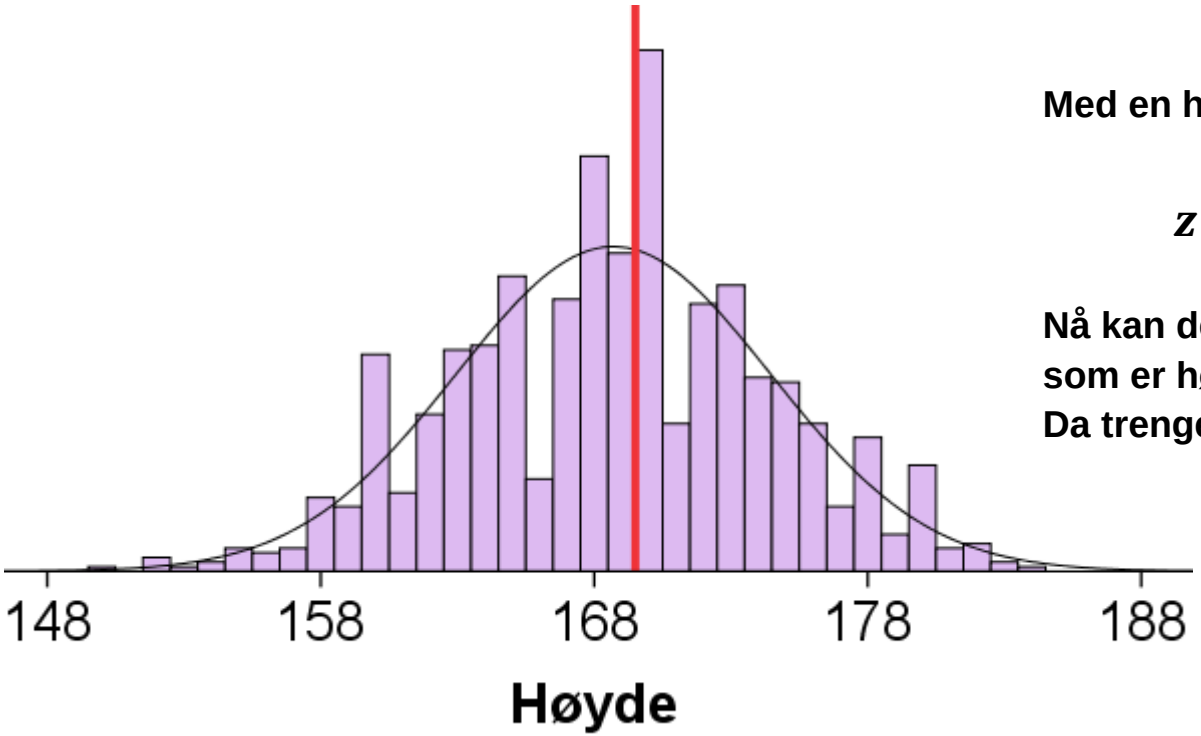
Ser på figuren at siden z-scoren er nøyaktig 1, kan det leses rett ut av figuren at 15.8% av befolkningen (13.6%+2.1%+0.1%) er smartere enn meg.

I ukas oppgave skal dere ta samme IQ-test som jeg har tatt, regne om til z-score, og rapportere begge tall (med litt ekstra informasjon) i nettskjema. Disse resultatene skal vi bruke senere i kurset.

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

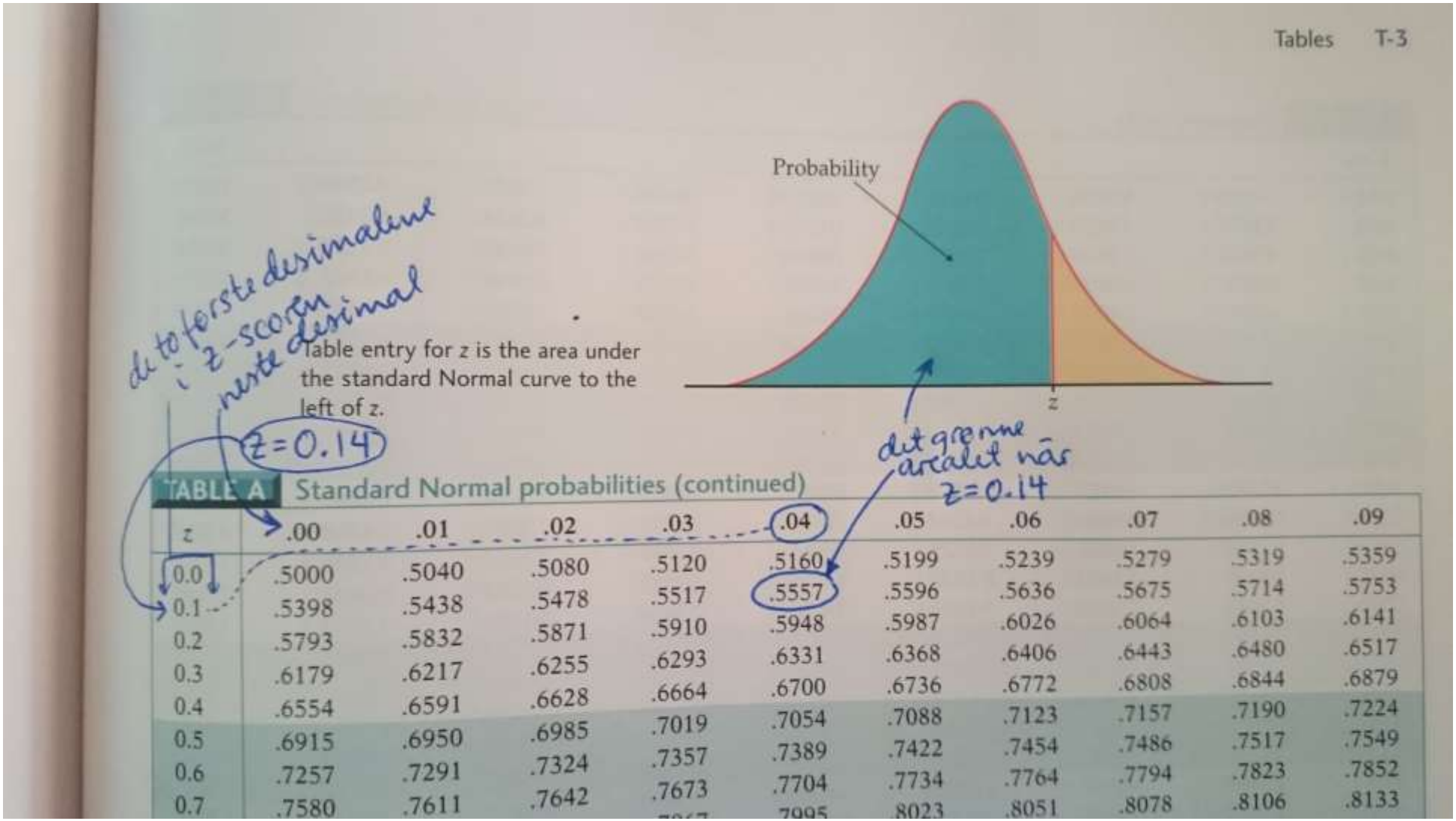


Med en høyde på 169.5 cm har KFF en z-score på

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

Nå kan det ikke leses rett ut av figuren hvor stor andel som er høyere eller lavere enn meg.
Da trenger vi en normalfordelingstabell!

Table A, side 720-721 i boka:



Tables T-3

Probability

z

de to første desimalene i z-scoren

neste desimal

Table entry for z is the area under the standard Normal curve to the left of z.

z = 0.14

det grønne arealet når z = 0.14

TABLE A Standard Normal probabilities (continued)

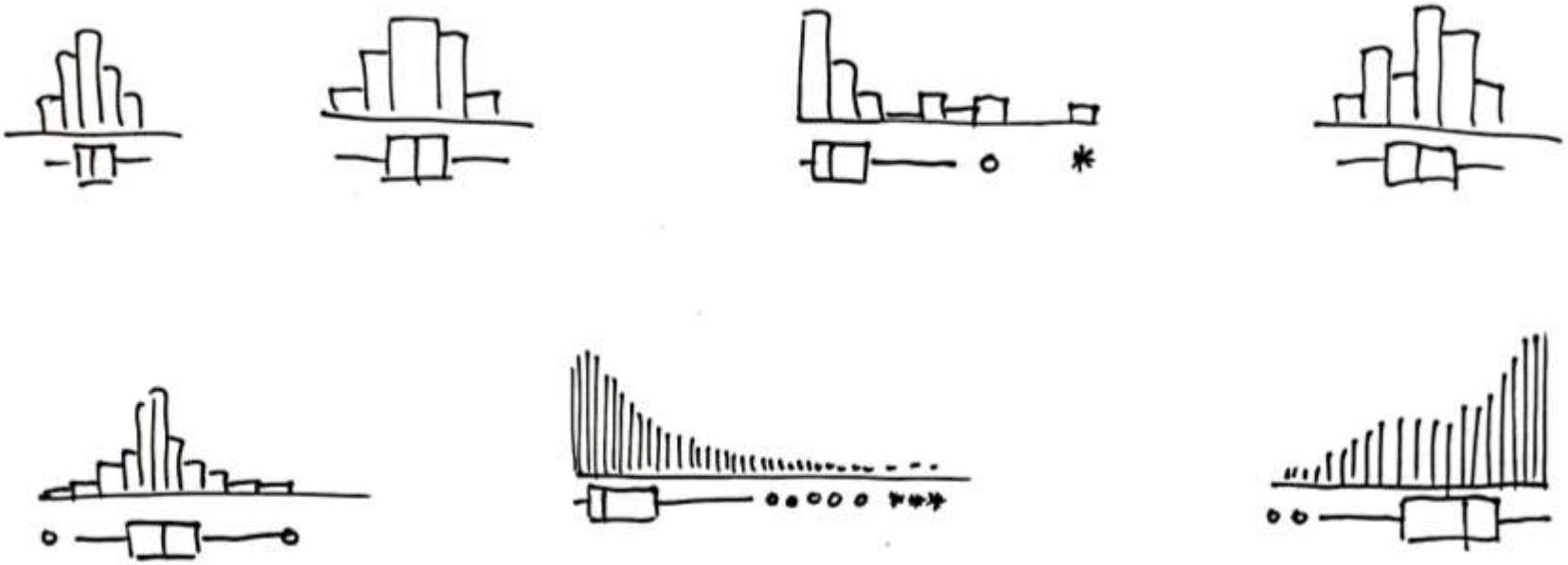
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
					7995	.8023	.8051	.8078	.8106	.8133

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**



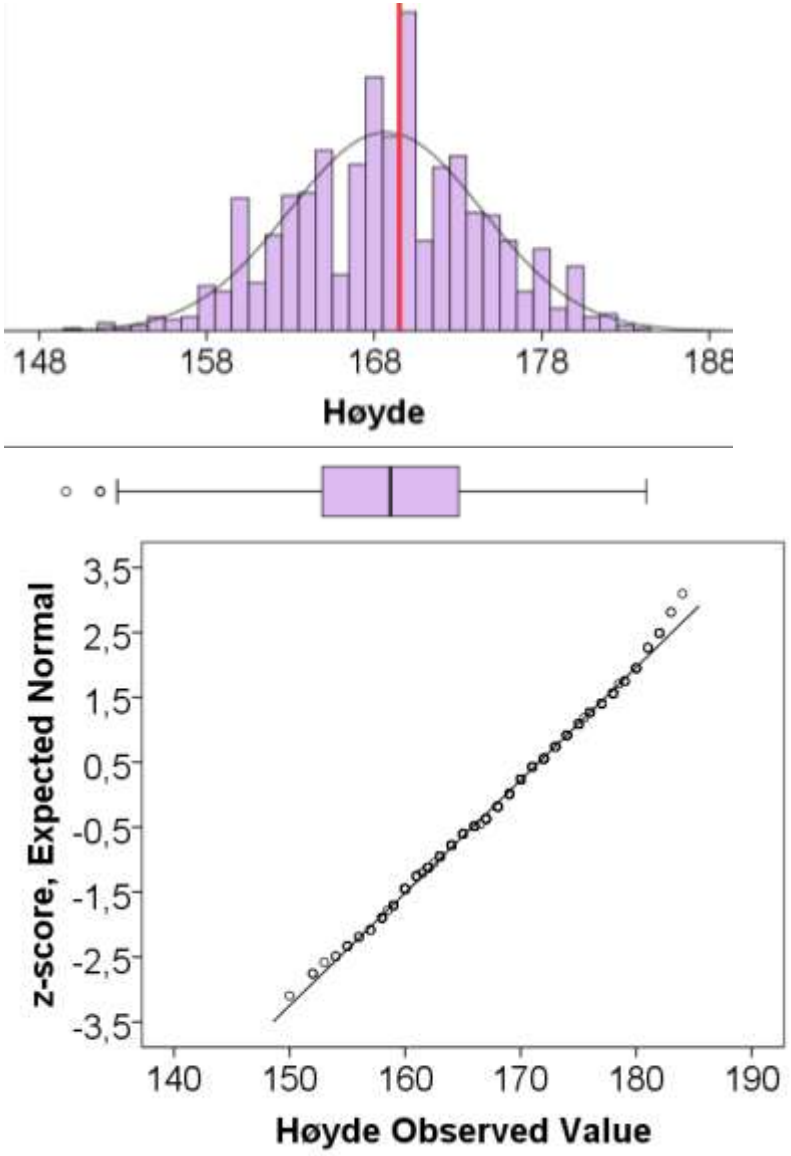
Fortsatt usikker? Lag normalfordelingsplott (QQ-plott):

Hvis normalfordelt: **Rett linje**
Hvis ikke rett linje: **Ikke normalfordelt**

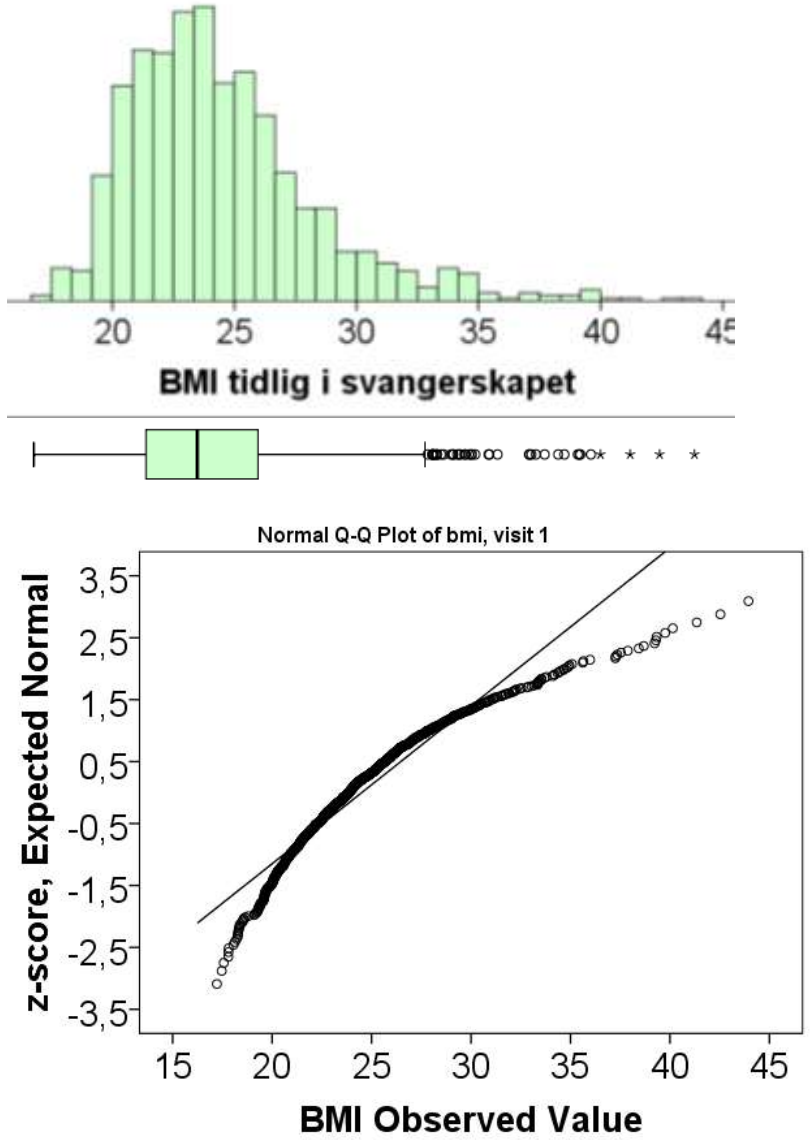
QQ-plott/Normal quantile plot lages ikke for hånd, vi bruker alltid programvare. Da gjøres dette i bakgrunnen:

- ✓ **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:

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

Denne formelen stammer fra *sentralgrenseteoremet*, der normalfordelingen også dukker opp. Dette skal vi se på senere i kurset.

Dessuten: Normalfordelingen er en god approksimasjon til binomisk fordeling med stor n, og 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:

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

Dette kommer vi også til senere.

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

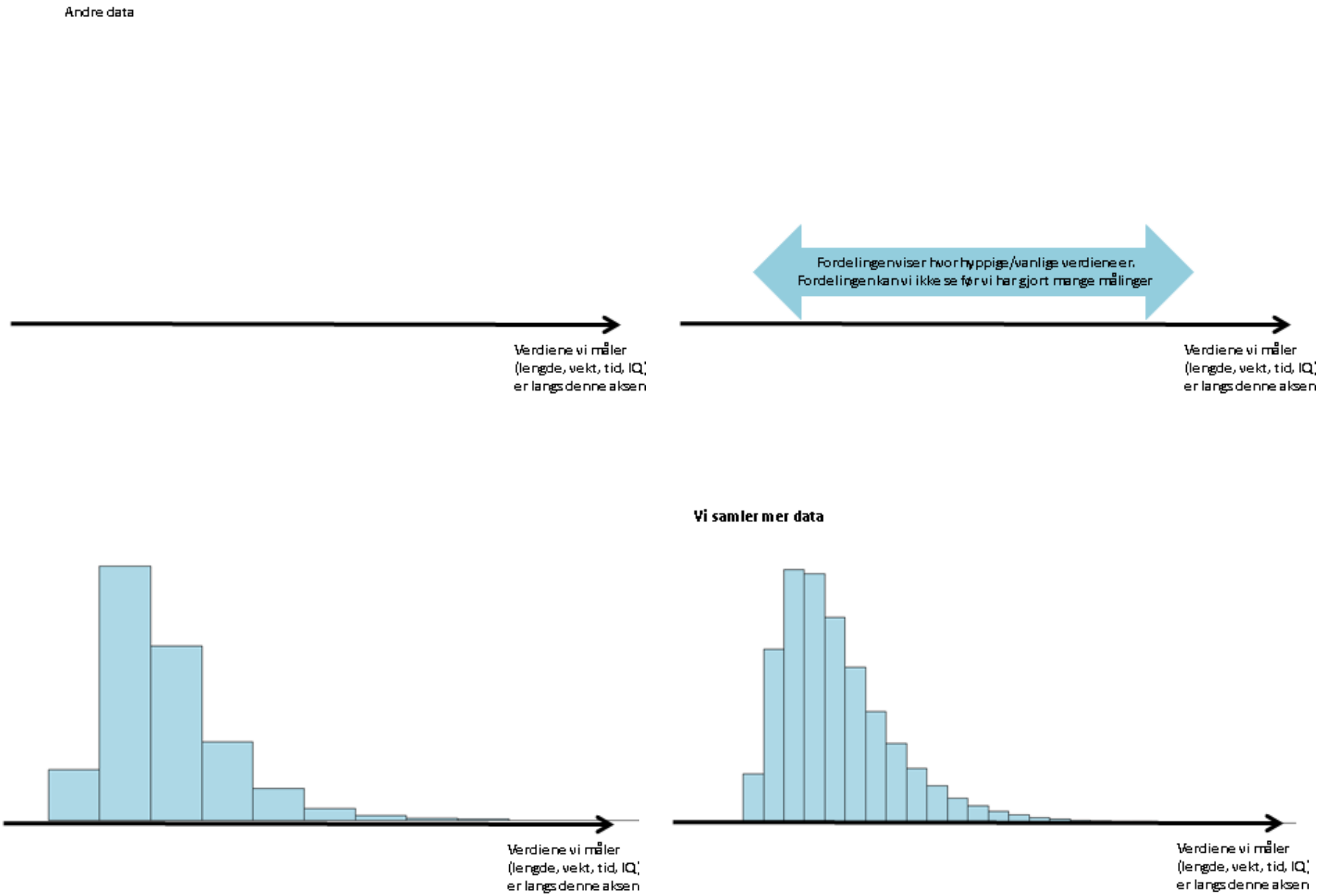
Analyseoversikt, Uke 35

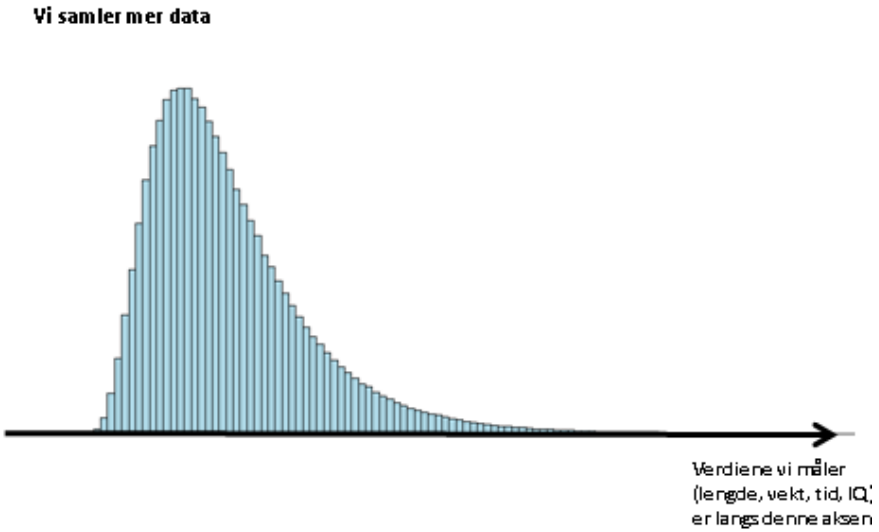
Versjon 1

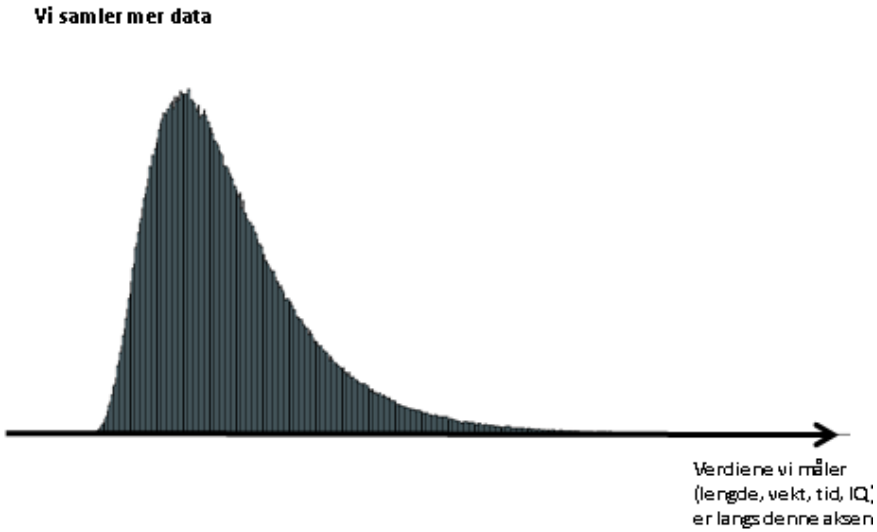
Hvilken type data har vi?

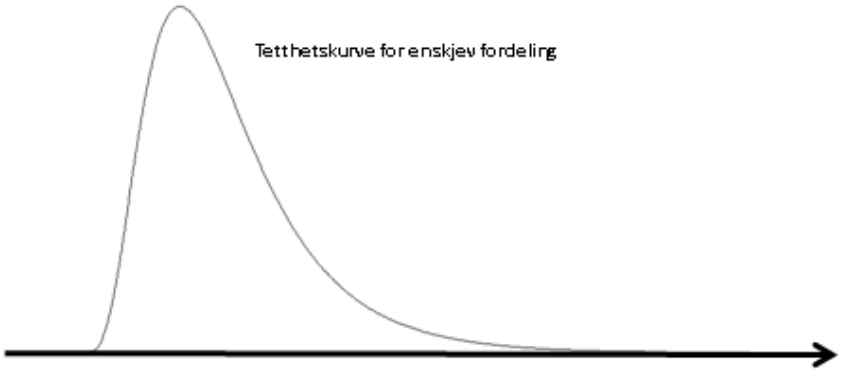
Hva skal vi gjøre?	Kategoriske data	Kontinuerlige data: Skjeve	Kontinuerlige data: Symmetriske
Oppsummere/presentere/ beskrive data	Tabeller, andeler Søyle(stolpe)/kakediagram	Median, kvartiler	$\bar{x}$, *SD*
Inferens Beregne konfidens-intervall	95% KI for *p*, som er populasjonens andel $\hat{p} \pm 1.96 \cdot SE(\hat{p})$	95% KI for populasjonens median Tabeller? Bootstrapping?	95% KI for μ, som er populasjonens gjennomsnitt $\bar{x} \pm 1.96 \cdot SE(\bar{x})$
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	«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 grupper: Median, kvt i hver gruppe Wilcoxon rank sum test Flere grupper: Median, kvt i hver gruppe Kruskal-Wallis test Posthoc-tester=Wrs-tester	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
Undersøke om det er sammenheng mellom to variabler, der den ene er en kontinuerlig variabel H_0: Ingen sammenheng	To kategorier: *t*-test /Wsr-test Flere kategorier: ANOVA/KW	Spearmans korrelasjonskoeffisient	Pearsons korrelasjonskoeffisient *r*
Analysere forskjell på par av data H_0: Ingen forskjell	Krysstabeller Andelen samsvar & McNemars test	Skjevfordelte differanser: Wilcoxon signed-rank test	Normalfordelte differanser: Paret *t*-test
Kvantifisere samsvar. Utgangspunkt: Det *er* samsvar Reliabilitet	Krysstabeller Cohens kappa		Scatterplott med y=x, Bland-Altman-plott ICC
Analysere sammenheng mellom en responsvariabel og en eller flere kovariater (forklaringsvariabler eller prediktorer) H_0: Ingen effekt/ingen prediktiv verdi	Binær responsvariabel: Logistisk regresjon Effektmål: OR m/ 95% KI H_0: OR = 1		Kontinuerlig responsvariabel: Lineær regresjon Effektmål: B H_0: B = 0 (Normalfordelte residualer)

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




Vi samler mer data
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen


Vi samler mer data
Verdiene vi måler
(lengde, vekt, tid, IQ)
er langs denne aksen


Tetthetskurve for en skjev fordeling

Neste uke:

STK1000 Uke 37, 2016. Studentene forventes å lese Ch 3.1-3.3 + 3.5 i læreboka (MMC).

Planlegging av studier, design, og innsamling av data