

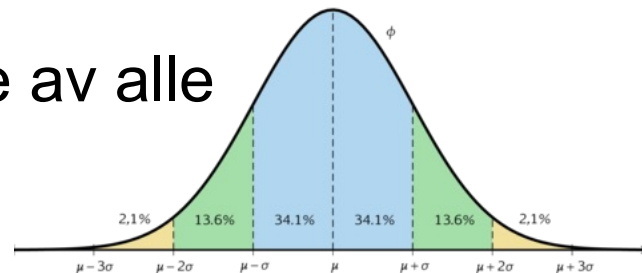
STK1100 våren 2024

Normalfordelingen

Svarer til avsnitt 4.3 og deler av
avsnitt 4.6 i læreboka

Matematisk institutt
Universitetet i Oslo

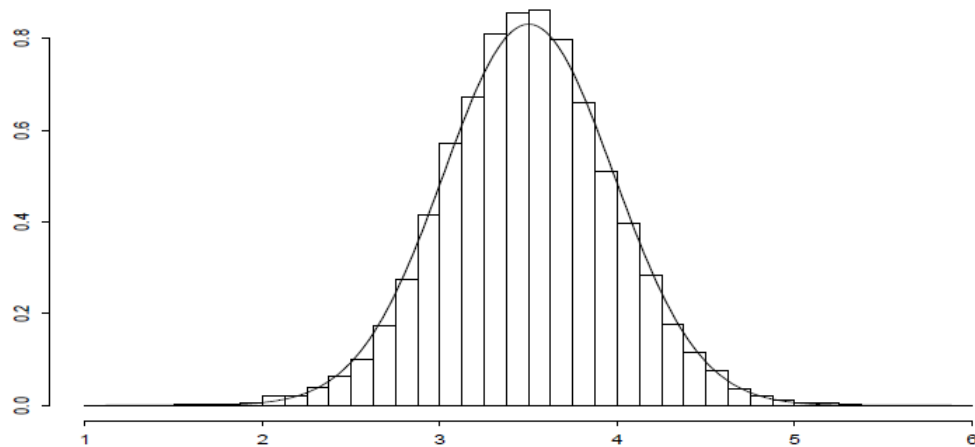
Normalfordelingen er den viktigste av alle sannsynlighetsfordelinger.



Normalfordelingen kan brukes til å beskrive variasjonen i numeriske observasjoner (f.eks. vektene til nyfødte jenter).

Normalfordelingen brukes også for ulike skalaer som menneskene selv har lagd (f.eks. IQ).

Normalfordelingen kan også brukes som en god tilnærming til andre fordelinger (jf. Sentralgrensesetningen i avsnitt 6.2).



Vi har sett at histogrammet til vekten for 20000 nyfødte jenter kan tilnærmes med funksjonen

$$f(x) = \frac{1}{0.48\sqrt{2\pi}} e^{-\frac{1}{2 \cdot 0.48^2}(x-3.50)^2}$$

$f(x)$ kalles **sannsynlighetstettheten** til X og svarer til histogrammet for «uendelig mange» fødselsvekter

Vi har at $P(a \leq X \leq b) = \int_a^b f(x) dx$

Vi sier at en kontinuerlig stokastisk variabel X er **normalfordelt** med forventning μ og standardavvik σ (varians σ^2) hvis den har sannsynlighetstetthet

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{for} \quad -\infty < x < \infty$$

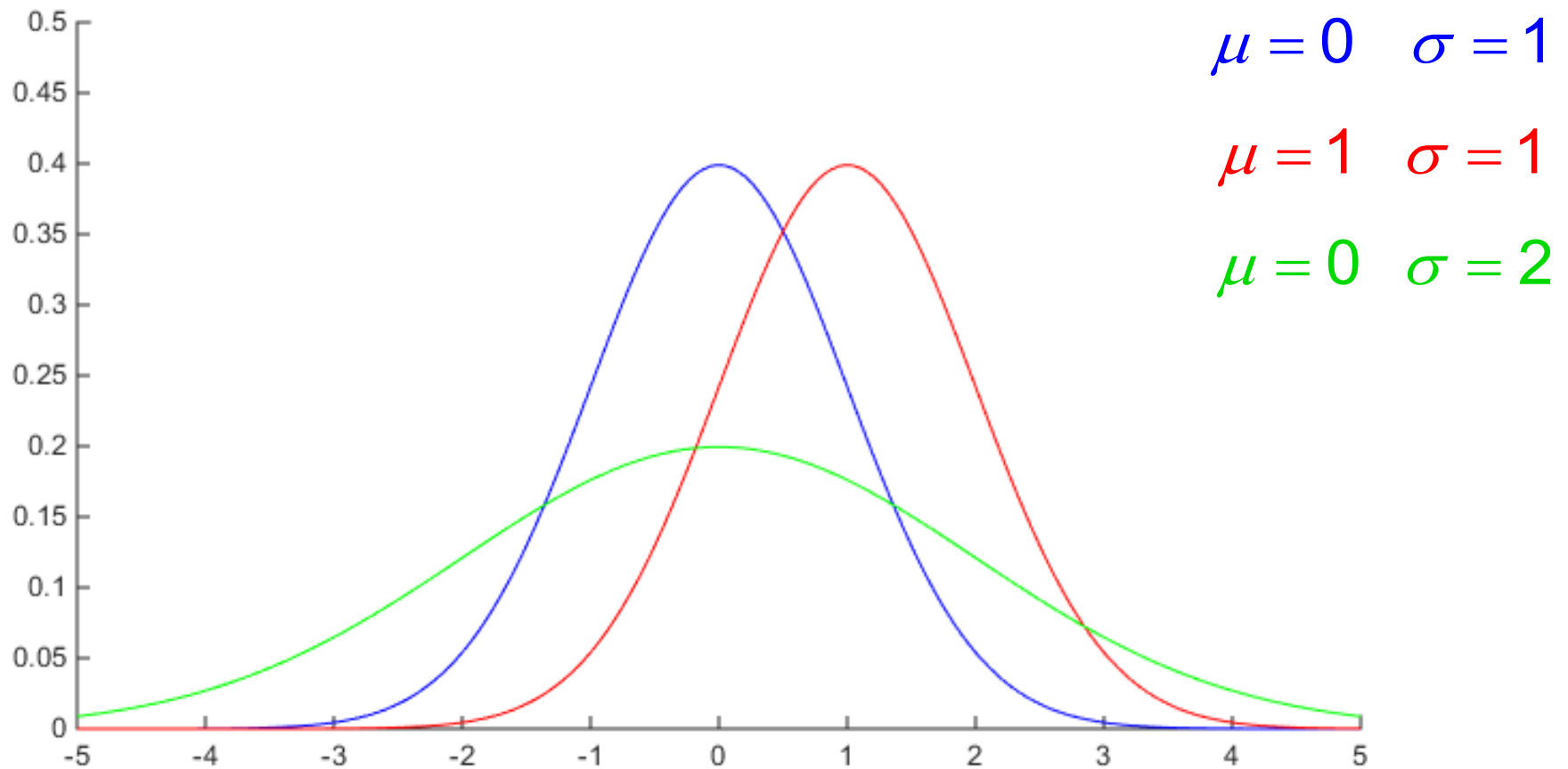
Kort skriver vi: $X \sim N(\mu, \sigma^2)$

Det kan vises at

$$\int_{-\infty}^{\infty} f(x; \mu, \sigma) dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

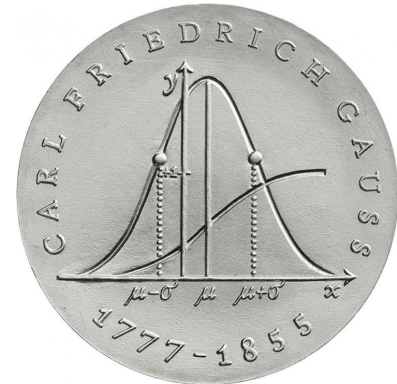
slik det skal være for en sannsynlighetstetthet

Normaltettheter for ulike verdier av μ og σ





Normalfordelingen kalles også
Gaussisk fordeling etter
Carl Friedrich Gauss
(1777-1855)

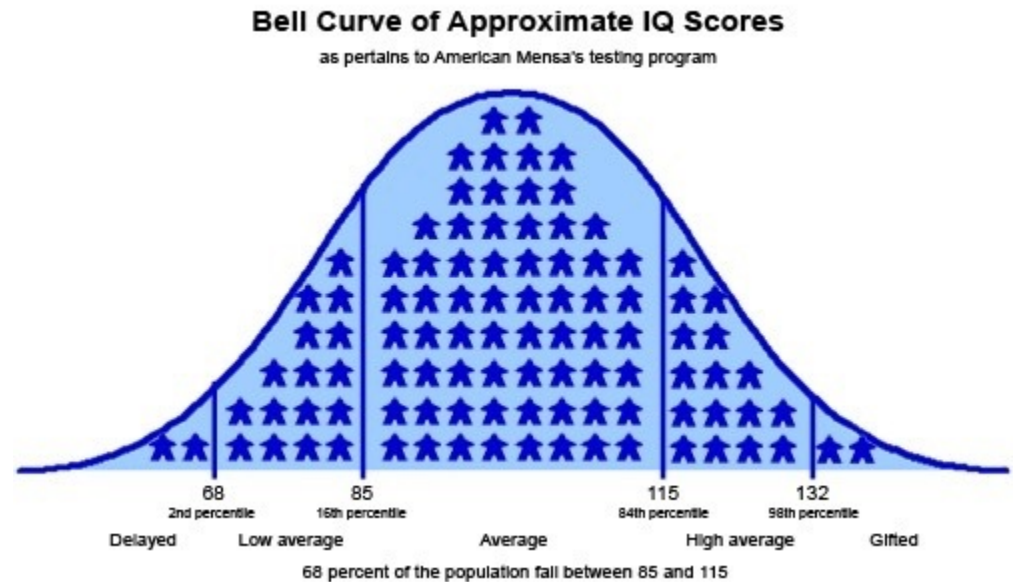


Eksempler på Gauss' vedvarende aktualitet, 2024 (ikke ment som produktplassering):



GAUSS LABS
We normalize AI.

Eksempel: IQ-score



IQ-skalaen er lagd slik at hvis X er IQ-score til en tilfeldig valg person, så er $X \sim N(100, 15^2)$

Forventning og varians

Momentgenererende funksjon (vises senere):

$$M_X(t) = E(e^{tX}) = e^{\mu t + \sigma^2 t^2 / 2}$$

Det gir

$$R_X(t) = \ln M_X(t) = \mu t + \sigma^2 t^2 / 2$$

Herav følger det at

$$E(X) = R'_X(0) = \mu$$

$$V(X) = R''_X(0) = \sigma^2$$

Standardnormalfordelingen

Hvis en kontinuerlig stokastisk variabel Z er normalfordelt med forventning $\mu = 0$ og standardavvik $\sigma = 1$ sier vi at Z er **standardnormalfordelt**

Sannsynlighetstetthet:

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{for} \quad -\infty < z < \infty$$

Kort skriver vi: $Z \sim N(0, 1)$

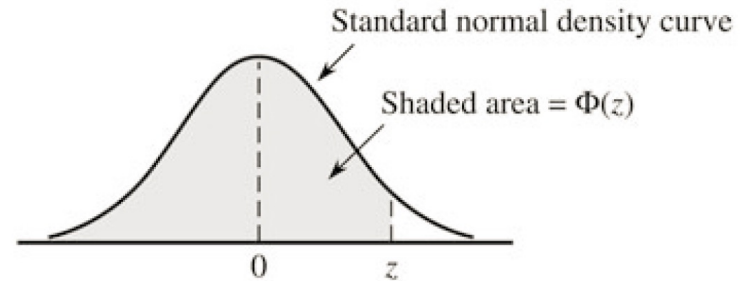
Kumulativ fordeling:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

Tabell A.3 bak i boka gir $\Phi(z)$ for verdier av z fra -3.49 til 3.49 (i trinn på 0.01)

$$\Phi(z) = P(Z \leq z)$$

Table A.3 Standard Normal Curve Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319

Eksempel:

$$P(Z \leq -3.02) = 0.0013$$

$$P(Z > 1.25) = 1 - P(Z \leq 1.25) = 1 - 0.8944 = 0.1056$$

Persentiler

Vi kan finne persentiler ved å bruke tabellen «baklengs»:

<i>z</i>	.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
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Eksempel:

97.5-persentilen = 1.96

Kritiske verdier z_α

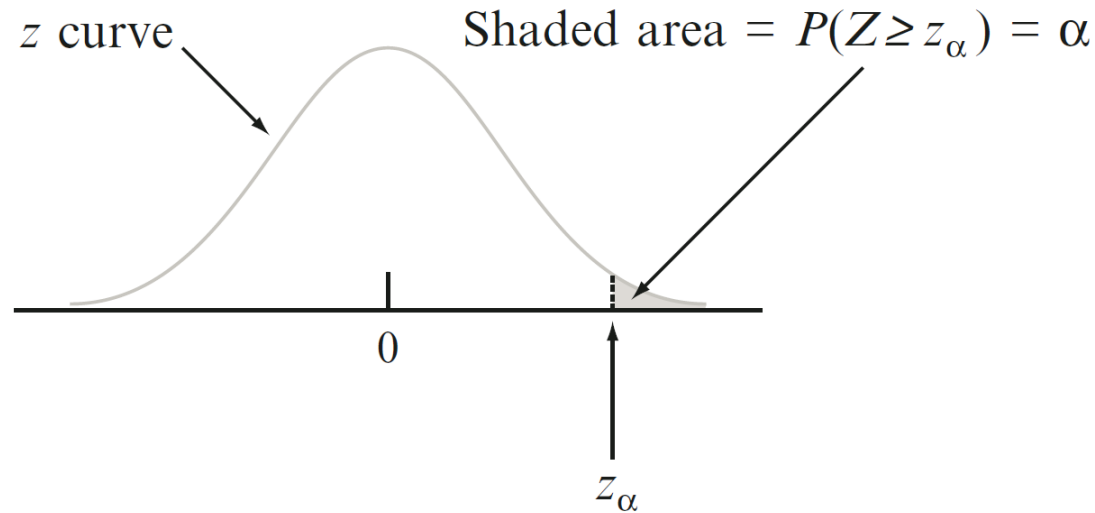


Table 4.1 Standard normal percentiles and critical values

Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1 - \alpha)$ th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

$$\text{Hvis } X \sim N(\mu, \sigma^2) \text{ så er } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Bevis:

Kumulativ fordeling til Z :

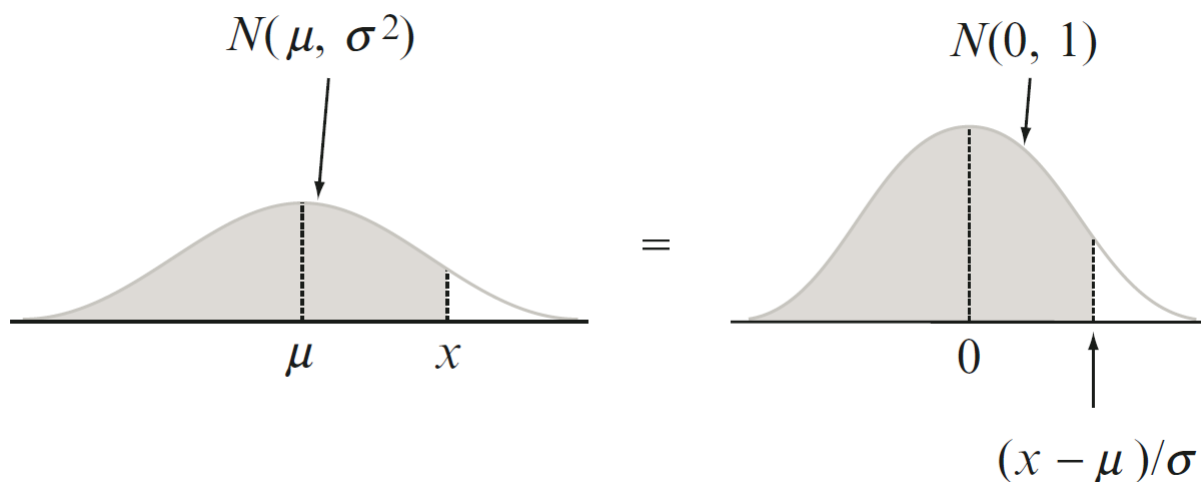
$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P(X \leq \sigma z + \mu) \\ &= \int_{-\infty}^{\sigma z + \mu} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \quad (\text{subst. } y = (x - \mu) / \sigma) \end{aligned}$$

Tetthet til Z :

$$f_Z(z) = F'_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Vi kan bruke standardnormalfordelingen til å finne sannsynligheter for $X \sim N(\mu, \sigma^2)$:

$$\begin{aligned} P(X \leq x) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) \end{aligned}$$



Normalfordeling med Python

Sannsynlighetstetthet: $f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Python: `import scipy.stats as stats`
`stats.norm.pdf(x,m,s)`

Kumulativ fordeling: $F(x; \mu, \sigma) = \int_{-\infty}^x \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$

Python: `stats.norm.cdf(x,m,s)`

Persentiler: $\eta(p)$ er gitt ved at $F(\eta(p); \mu, \sigma) = p$

Python: `stats.norm.ppf(p,m,s)`

Se forøvrig
Jupyter notebook

Anta at $X \sim N(\mu, \sigma^2)$

Da er

$$\begin{aligned} P(\mu - k\sigma \leq X \leq \mu + k\sigma) &= P\left(-k \leq \frac{X - \mu}{\sigma} \leq k\right) \\ &= P(-k \leq Z \leq k) = \Phi(k) - \Phi(-k) \end{aligned}$$

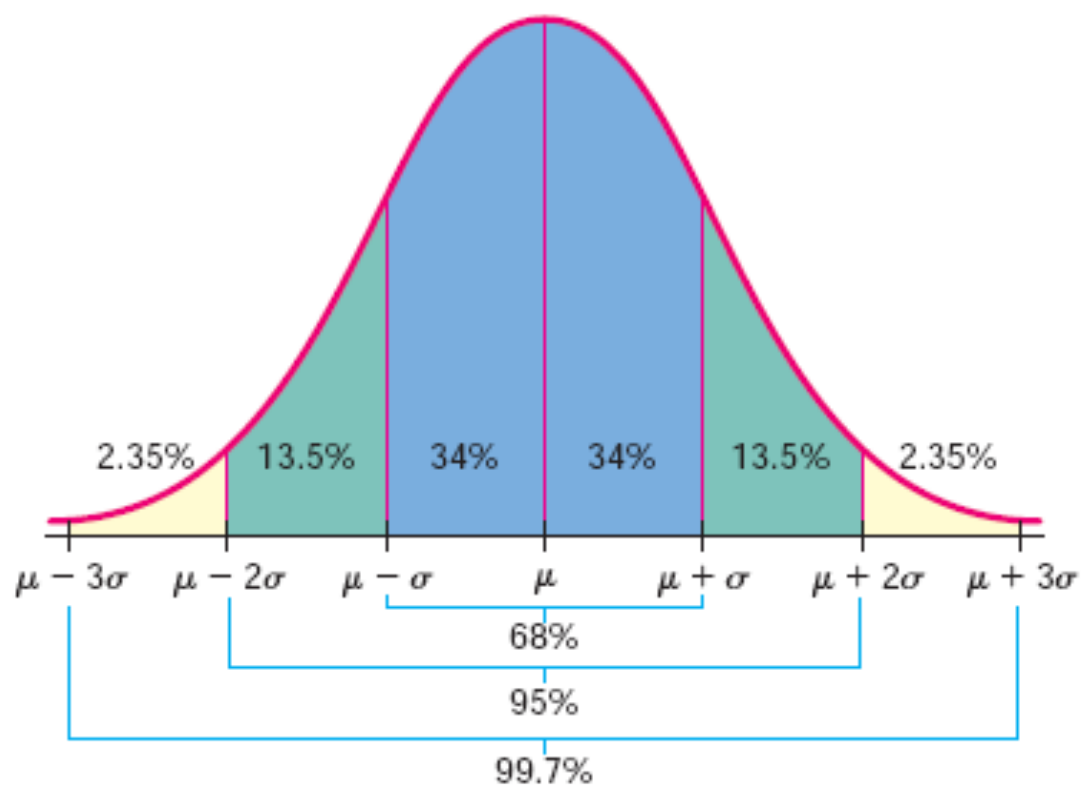
Vi finner sannsynlighetene for $k = 1, 2$ og 3 .

Det gir følgende:

If the population distribution of a variable is (approximately) normal, then

1. Roughly 68% of the values are within 1 SD of the mean.
 2. Roughly 95% of the values are within 2 SDs of the mean.
 3. Roughly 99.7% of the values are within 3 SDs of the mean.
-

Area Under a Normal Curve



Momentgenererende funksjon

Vil bestemme den momentgenererende funksjonen til X .

Bruker at hvis $X \sim N(\mu, \sigma^2)$ så er $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

Vi viser først at (jf. forelesningen):

$$M_Z(t) = e^{t^2/2}$$

Av dette finner vi at:

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E(e^{t(\sigma Z + \mu)}) = E(e^{t\mu} e^{(t\sigma)Z}) \\ &= e^{\mu t} E(e^{(t\sigma)Z}) = e^{\mu t} M_Z(t\sigma) = e^{\mu t + \sigma^2 t^2/2} \end{aligned}$$

Persentiler

La $\eta_X(p)$ og $\eta_Z(p)$ være $100p$ -persentilene til X og Z

Vi vil finne sammenhengen mellom dem.

Vi har at $P(Z \leq \eta_Z(p)) = p = P(X \leq \eta_X(p))$

Nå er $P(X \leq \eta_X(p)) = P\left(Z \leq \frac{\eta_X(p) - \mu}{\sigma}\right)$

Det gir $\frac{\eta_X(p) - \mu}{\sigma} = \eta_Z(p)$

Derfor er $\eta_X(p) = \mu + \eta_Z(p) \cdot \sigma$

Normalfordelingsplott

Sammenhengen

$$\eta_x(p) = \mu + \eta_z(p) \cdot \sigma$$

kan brukes til å motivere en metode for å sjekke om observasjoner er (omtrent) normalfordelte.

Eksempel:

Hvilepuls til 10 kvinnelige studenter:

67 58 105 97 81 62 73 75 87 68



Er det rimelig å anta at dataene er normalfordelte?

Idé: Plott empiriske persentiler mot persentilene i standardnormalfordelingen.

Empiriske persentiler med n observasjoner:

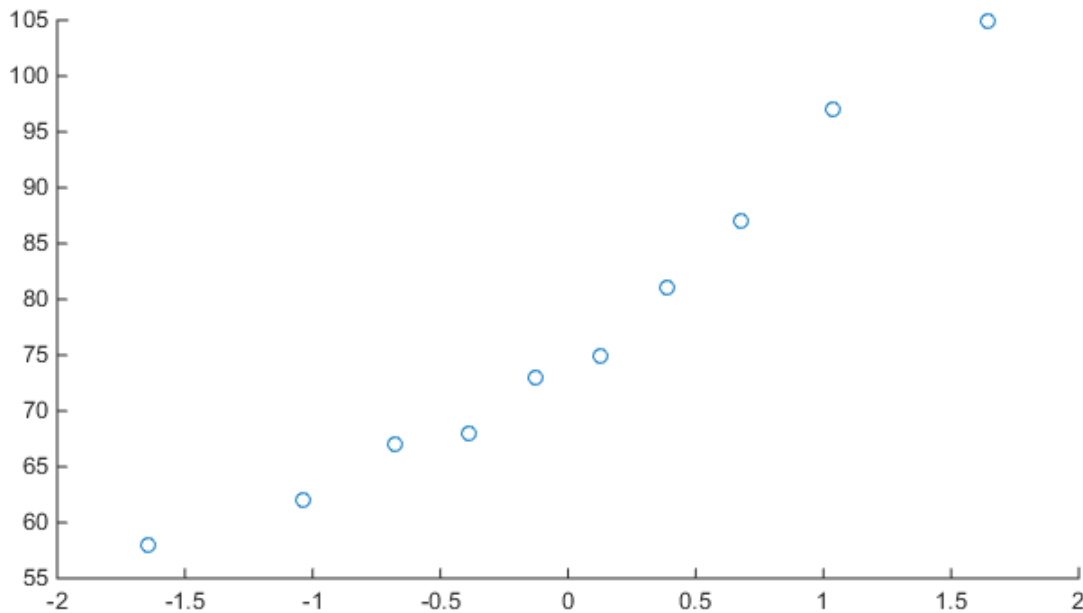
Skriv observasjonene i stigende rekkefølge fra den minste til den største.

Da svarer den i -te minste observasjonen til den

empiriske $100\left(\frac{i-1/2}{n}\right)$ -persentilen.

Når $n = 10$ svarer observasjonene til følgende empiriske persentiler:

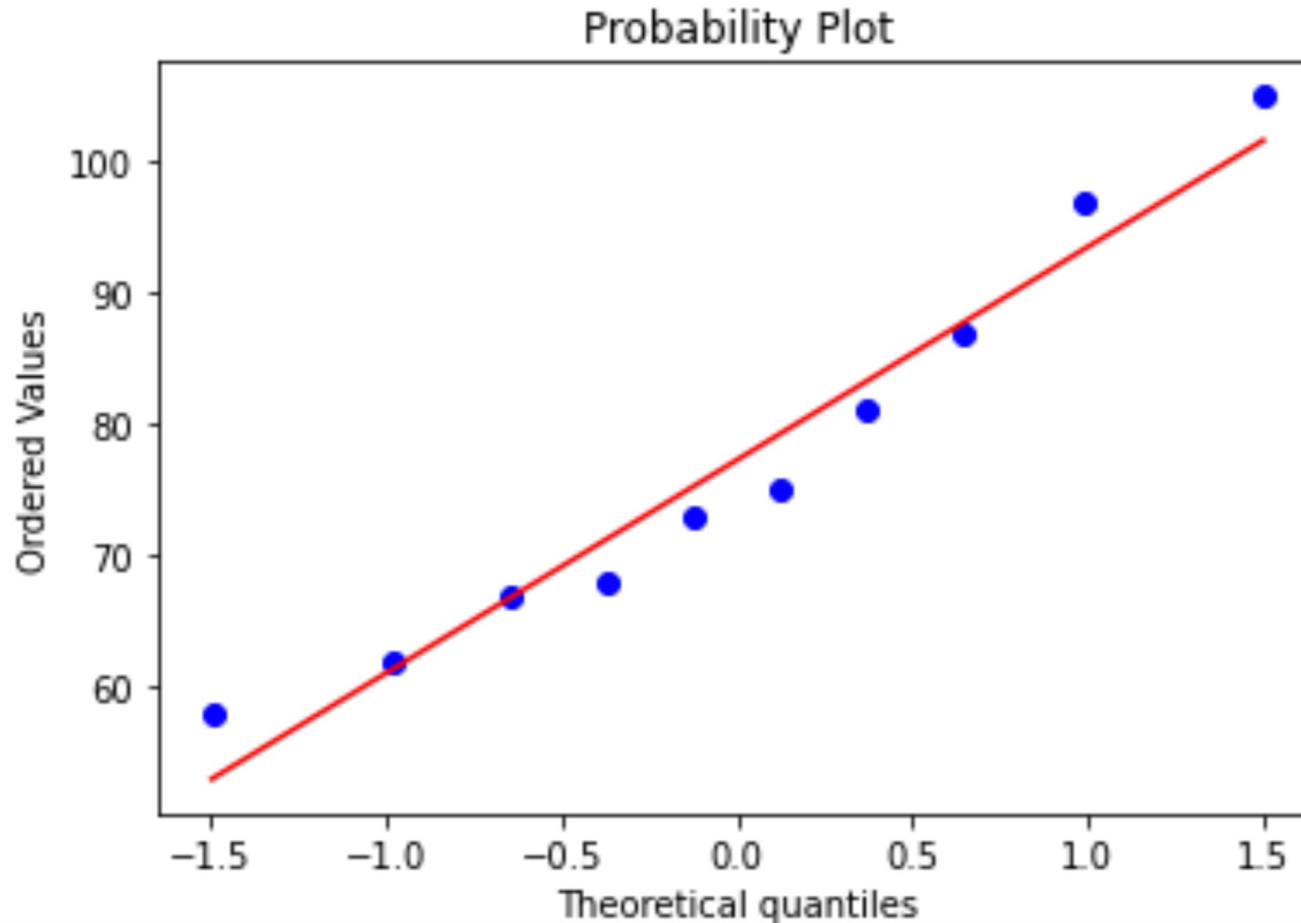
5, 15, 25, 35, 45, 55, 65, 75, 85, 95



Punktene ligger omtrent på en rett linje, så normalfordeling er rimelig.

Python kommandoer:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
x=np.array([67,58,105,97,81,62,73,75 ,87,68])
n=10
ii=np.arange(1,n+1)
pers=(ii-0.50)/n
plt.scatter(stats.norm.ppf(pers,0,1),np.sort(x))
```

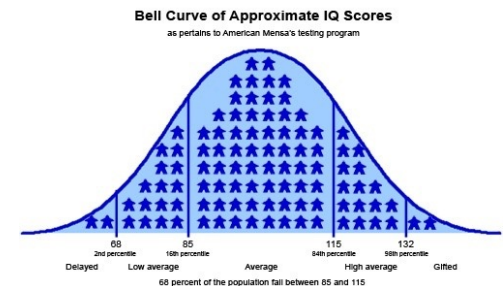


Ferdig-kommandoen `stats.probplot(x,plot=plt)` i Python lager et tilsvarende plott som på forrige slide

Normalfordeling som tilnærming til diskrete fordelinger

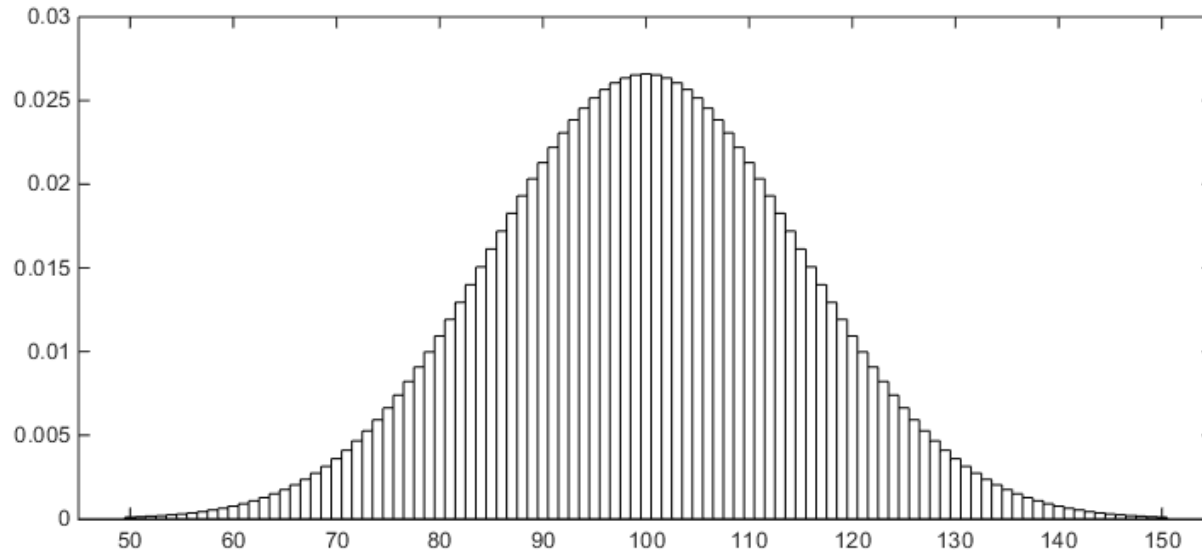
Normalfordelingen brukes mange ganger som en tilnærming for diskrete fordelinger.

Eksempel: IQ:

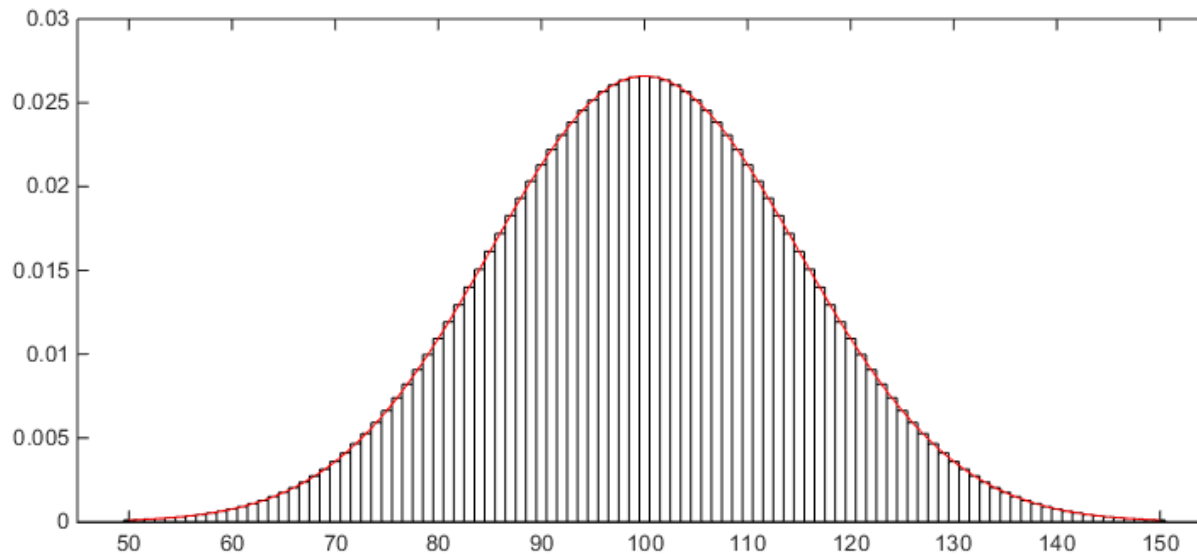


IQ-skalen er lagd slik at IQ i en befolkning er normalfordelt med $\mu = 100$ og $\sigma = 15$ (selv om en ikke regner IQ som desimaltall).

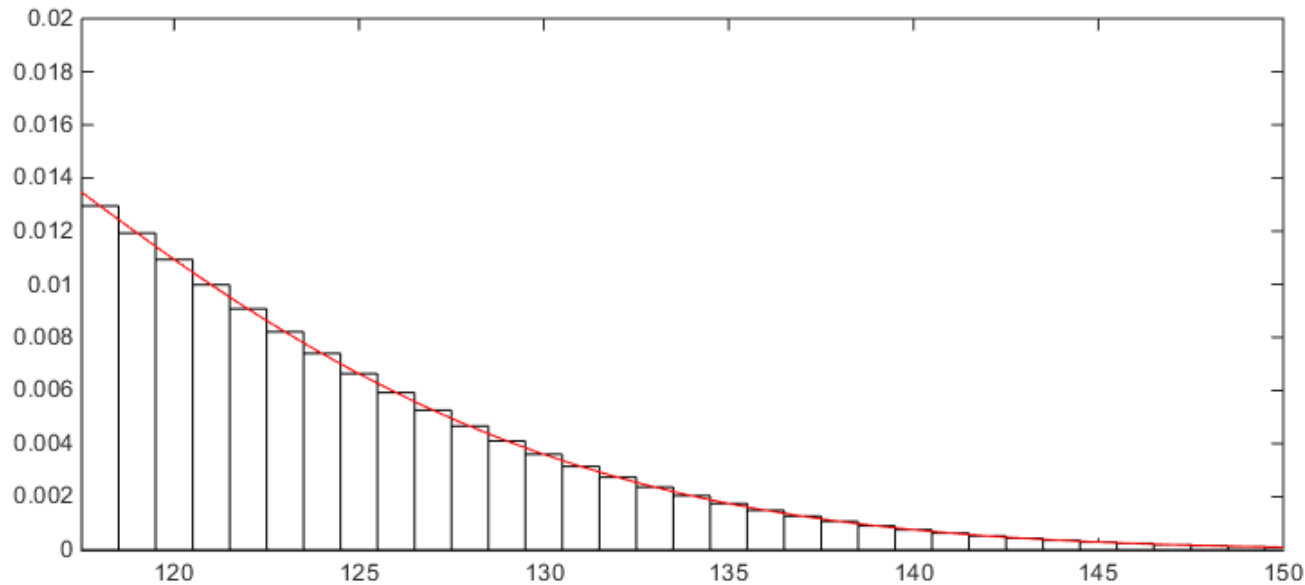
Hvor stor del av befolkningen har IQ minst 120?



Sannsynlighets-
histogram for
diskret IQ-fordeling



Rød kurve gir
tilsvarende
normalfordeling



Utsnitt av
nederste figur
på forrige slide

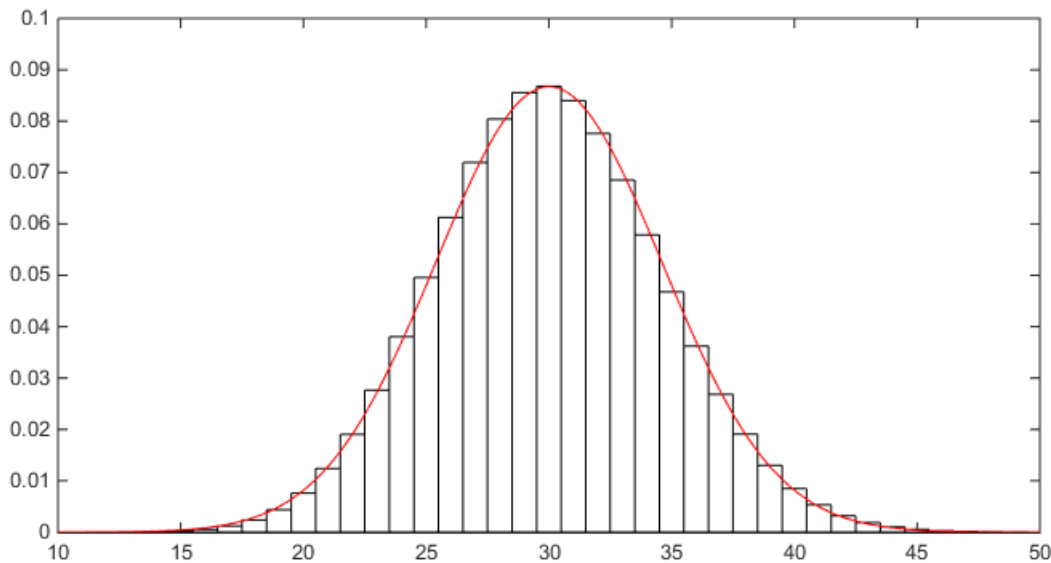
For å få med hele stolpen som svarer til IQ lik 120 må vi bestemme arelet under normalfordelingskurven til høyre for 119.5.

Det svarer til sannsynlighet 9.7%

Normalfordeling som tilnærming til binomisk fordeling

La X være binomisk fordelt med $n=100$ forsøk og sannsynlighet $p=0.30$. Da er $E(X) = np = 30$

og $SD(X) = \sqrt{np(1-p)} = 4.58$



Figuren viser den binomiske punktsannsynligheten og normalfordeligstettheten med $\mu = 30$ og $\sigma = 4.58$

Generelt har vi følgende resultat:

Anta at X er binomisk fordelt med n forsøk og sannsynlighet p .

Den kumulative fordelingen er

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^x \binom{n}{y} p^y (1-p)^{n-y}$$

Hvis $np \geq 10$ og $n(1-p) \geq 10$, så er

$$B(x; n, p) \approx \Phi\left(\frac{x + 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Senere vil vi se at dette resultatet er et spesialtilfelle av sentralgrensesetningen (avsn. 6.2)