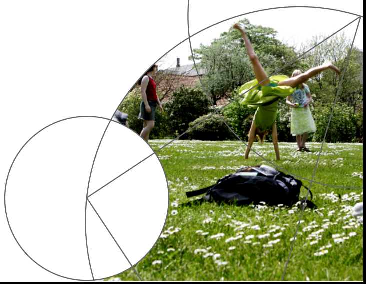
Faculty of Life Sciences

Egenskaber ved normalfordelingen

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk



Program

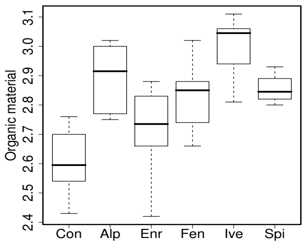
- Repetition fra i mandags
- Er data normalfordelt?
- Egenskaber ved normalfordelingen (jf. opgave 4.3 fra mandag)
- Standard normalfordelingen
- Diverse normalfordelingssandsynligheder

Afsnit 4.3.2 om transformation: læs selv ifm. case.

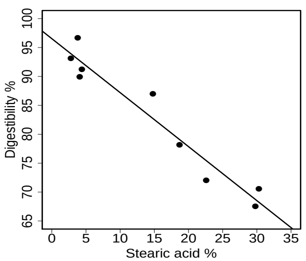
Hvorfor skal vi lære om normalfordelingen (nu)?

Har set tre typer af data/eksperimenter med kontinuerte data:

Ensidet ANOVA



Lineær regression



En stikprøve:

Blood pressure								
96	119	119	108	126	128	110	105	94

Vi skal bruge normalfordelingen for alle tre forsøgstyper/datatyper!

Hvorfor netop normalfordelingen?

Normalfordelingen er vigtig af flere grunde

- Passer ofte godt til (biologiske) data
- Pæne matematiske egenskaber — får pæne resultater for estimater, statistiske test mm.
- Centrale grænseværdisætning (CLT): gennemsnit af næsten hvad som helst er approksimativt normalfordelt?

Kaldes også den Gaussiske fordeling.

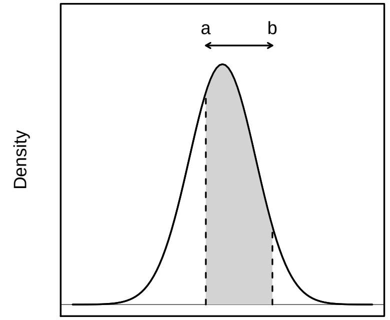
Opkaldt efter Carl Friedrich Gauss — tysk matematiker og fysiker, 1777–1855.

Gauss

Tæthed og sandsynligheder

Tæthed for normalfordelingen med middelværdi μ og spredning σ:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$



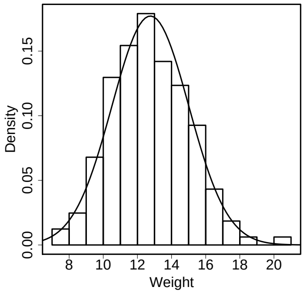
$Y \sim N(\mu, \sigma^2)$ hvis sandsynligheden for at Y havner mellem a og b er lig areal fra a til b under tætheden

$$P(a < Y < b) = \int_a^b f(y)\,dy$$

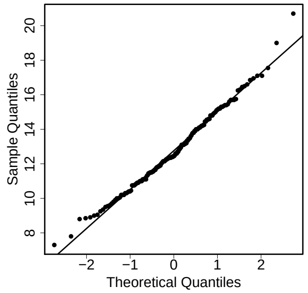
Er data normalfordelt?

Vægt af 162 krabber: $y_1, \ldots, y_{162}$. Tidligere: $\bar{y} = 12.76$ og $s = 2.25$.

Histogram og tæthed



QQ-plot



- Histogram sammen med tæthed for $N(\bar{y}, s^2)$
- QQ-plot: quantile-quantile plot. Sammenlign med ret linie med skæring $\bar{y}$ og hældning σ. R: `qqnorm(wgt);` `qqline(wgt)/abline`

Egenskaber ved normalfordelingen — simulation

- Histogrammer og QQ-plots
- $Y \sim N(\mu, \sigma^2)$: middelværdi $\mathrm{E}Y = \mu$, spredning $\mathrm{sd}(Y) = \sigma$
- $Y_1 \sim N(\mu_1, \sigma_1^2)$ og $Y_2 \sim N(\mu_2, \sigma_2^2) \Rightarrow Y_1 + Y_2 \sim$
  Specielt $\mathrm{sd}(Y_1 + Y_2) =$
- $Y \sim N(\mu, \sigma^2) \Rightarrow a + b \cdot Y \sim$
  Specielt $\mathrm{sd}(a + b \cdot Y) =$
- $Z = \frac{Y-\mu}{\sigma} \sim$
- $P(\mu - \sigma < Y < \mu + \sigma) =$
  $P(\mu - 2\sigma < Y < \mu + 2\sigma) =$

Egenskaber ved normalfordelingen — simulation

	Normal?	$\bar{y}$	s	s^2
`y1`				
`y2`				
`z=y1+y2`				
`v=4+2*y1`				
`u=(y2-8)/2`				

Standard normalfordelingen

Hvis $Y \sim N(\mu, \sigma^2)$, så er $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$ — standard N-fordelingen.

Sandsynligheder om $N(\mu, \sigma^2)$ kan omskrives til sandsynligheder om $N(0,1)$. For eksempel

$$P(Y \leq a) = P\left(\frac{Y-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

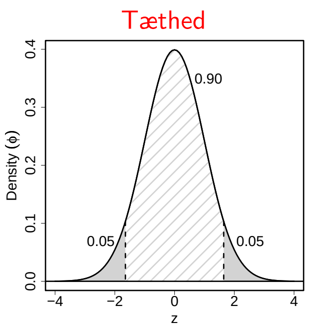
Tæthed for $N(0,1)$:

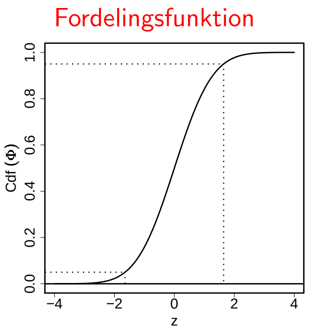
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Fordelingsfunktion — "areal til venstre for z":

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x)\,dx$$

Standard normalfordelingen





- 95%-fraktilen er 1.6449: $\Phi(1.6449) = P(Z \leq 1.6449) = 0.95$
  $P(-1.6449 \leq Z \leq 1.6449) =$
- 97.5%-fraktilen er 1.960: $\Phi(1.960) = P(Z \leq 1.960) = 0.975$
  $P(-1.96 \leq Z \leq 1.96) =$

Standard normalfordelingen

Intet eksplicit udtryk for Φ, men Φ er tabelleret i statistiske tabeller.

I R:

- fordelingsfunktion: `pnorm`, fx.
  ```
  > pnorm(1.6449)
  [1] 0.9500048
  ```
- Fraktiler: `qnorm`, fx.
  ```
  > qnorm(0.975)
  [1] 1.959964
  ```

Tabel: Appendix C2, p. 400 i bogen.

N-sandsynligheder

Krabbedata: antag at krabbevægt $Y \sim N(12.76, 2.25^2)$

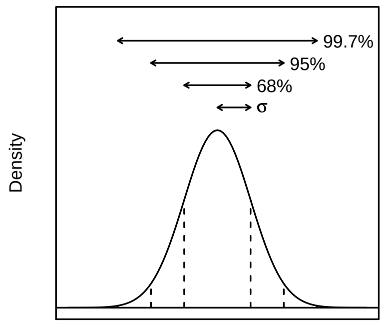
Beregn sandsynligheden for at

- en tilfældigt udtaget krabbe vejer højst 14 gram?
- en tilfældigt udtaget krabbe vejer mindst 10 gram?
- en tilfældigt udtaget krabbe vejer mellem 10 og 14 gram?
- summen af to tilfældigt udvalgte krabber vejer mindst 26 gram?

```
> pnorm(1.227)
[1] 0.8900887
> pnorm(-1.227)
[1] 0.1099113
> pnorm(0.151)
[1] 0.5600121
```

```
> pnorm(0.5511)
[1] 0.7092174
> pnorm(-0.5511)
[1] 0.2907826
```


Sandsynligheder for $\mu \pm k \cdot \sigma$



$$P(\mu - \sigma < Y < \mu + \sigma)$$

$$= P\left(\frac{\mu - \sigma - \mu}{\sigma} < Z < \frac{\mu + \sigma - \mu}{\sigma}\right)$$

$$= P(-1 < Z < 1)$$

$$= 0.68$$

- 68% mest centrale obs. ligger i intervallet $\mu \pm \sigma$
- 95% mest centrale obs. ligger i intervallet $\mu \pm 2 \cdot \sigma$

Dette gælder for alle normalfordelinger!

Dagens hovedpunkter

- Er data normalfordelt?
  - Sammenlign histogram med tæthed for $N(\bar{y}, s^2)$.
  - Sammenlign QQ-plot med ret line
- Summer af normalfordelte variable er normalfordelt
- Lineær transformation af normalfordelt variabel er normalfordelt
- Standard normalfordelingen, $N(0,1)$
  - $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$
  - Fraktiler og sandsynligheder for $N(0,1)$
  - Alle *N*-sandsynligheder kan beregnes vha. $N(0,1)$
- Sandsynligheder for $\mu \pm \sigma$, $\mu \pm 2 \cdot \sigma$, osv.

Ordliste

Engelsk	Dansk
Central limit theorem (CLT)	centrale grænseværdisætning
cumulative distribution function (cdf)	fordelingsfunktion
density	tæthed
drawn at random	tilvældigt udvalgt
Gaussian distribution	normalfordelingen
mean	gennemsnit el. middelværdi
quantile	fraktil
sample	stikprøve
standard deviation	spredning