Statistik för Biologer
F5: Korrelation och χ^2-test

Shaobo Jin

Matematiska institutionen

Samband

Många frågeställningar inom biologin rör samband av olika slag.

- Samband mellan vattentemperatur och tillväxt hos cyanobakterier
- Temperaturen påverkar tillväxthastighet
- Samband mellan uttrycksnivåerna av två proteiner

Korrelation

Inom statistiken mäter **korrelationen** styrkan på sambandet mellan två numeriska variabler x och y.

Flera olika korrelationsmått förekommer. För alla dessa gäller att:

- Korrelationen ligger mellan -1 och 1.
  - Ju närmare 1 som absolutbeloppet av korrelationen är, desto starkare är sambandet.
- Om det inte finns något samband så är korrelationen 0.
- Om korrelationen är negativ så minskar y då x ökar.
- Om korrelationen är positiv så ökar y då x ökar.

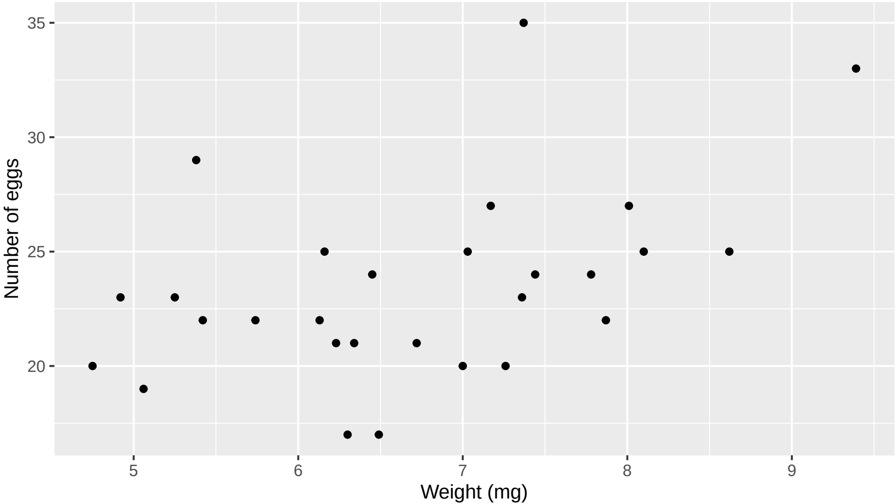
Tångloppevägning

28 tångloppehonor (Platorchestia platensis) samlades in vid en strand. Man räknade antalet ägg som varje hona bar, samt vägde honorna efter frystorkning

Weight (mg)	Eggs
5.38	29
7.36	23
6.13	22
4.75	20
8.10	25
8.62	25
6.30	17
7.44	25
⋮	⋮
6.34	21
6.16	25
5.74	22

Visualisering: Spridningsdiagram (Scatter Plot)

Finns det något samband? Hur starkt är det i så fall?



Pearsonkorrelation

Säg att vi har n parvisa observationer (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Vi kan beräkna:

1. Medelvärden (mean): $\bar{x}$, $\bar{y}$
2. Stickprovsstandardavvikelser (standard deviation):

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

3. **Kovariansen** (covariance):

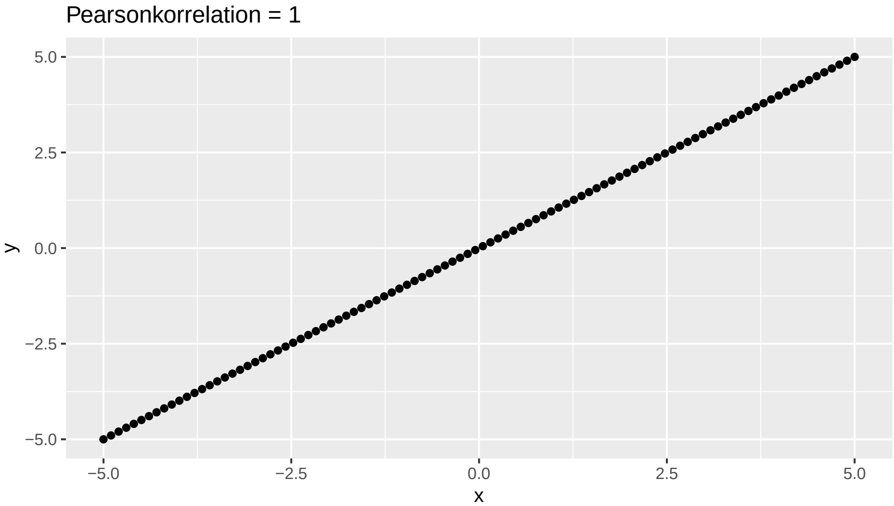
$$c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

4. **Pearsonkorrelationen** (Pearson correlation):

$$\rho = \frac{c}{s_x s_y}.$$

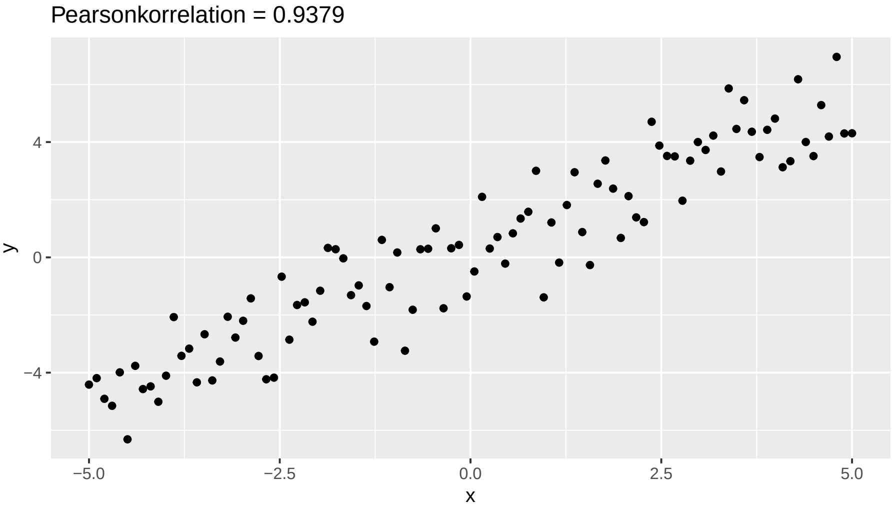
Vad mäter Pearsonkorrelationen?

Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.



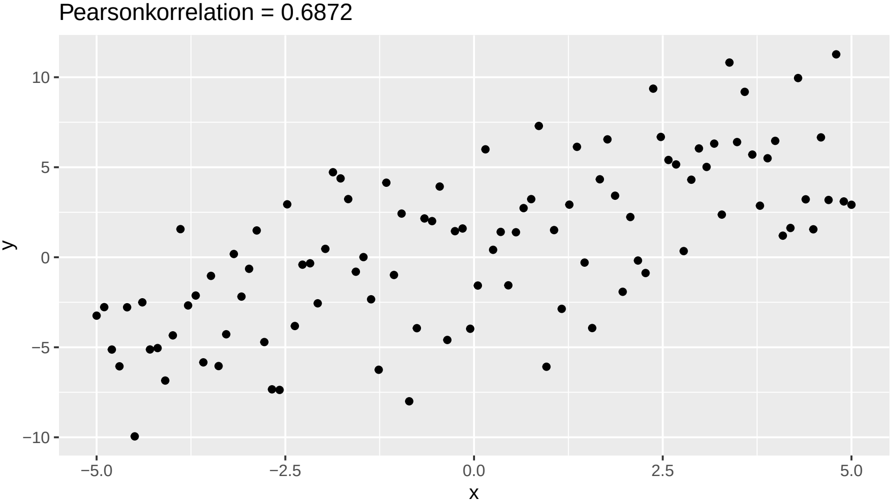
Vad mäter Pearsonkorrelationen?

Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.



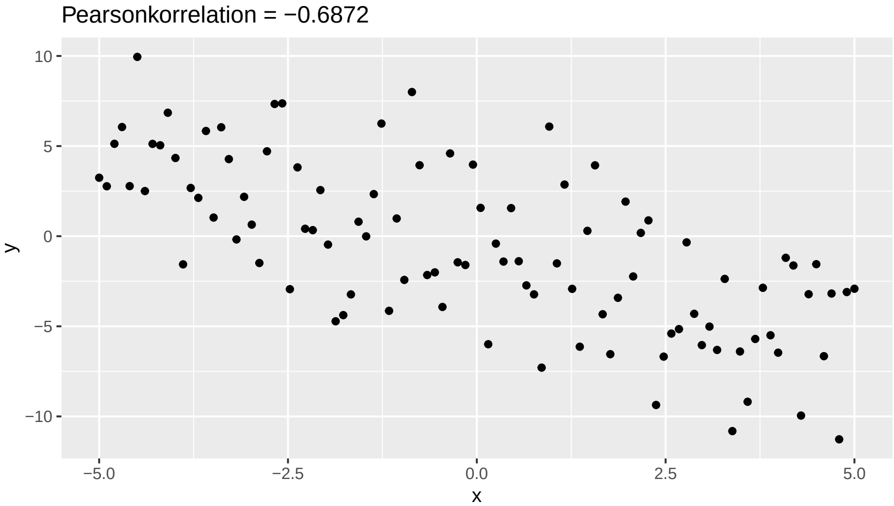
Vad mäter Pearsonkorrelationen?

Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.



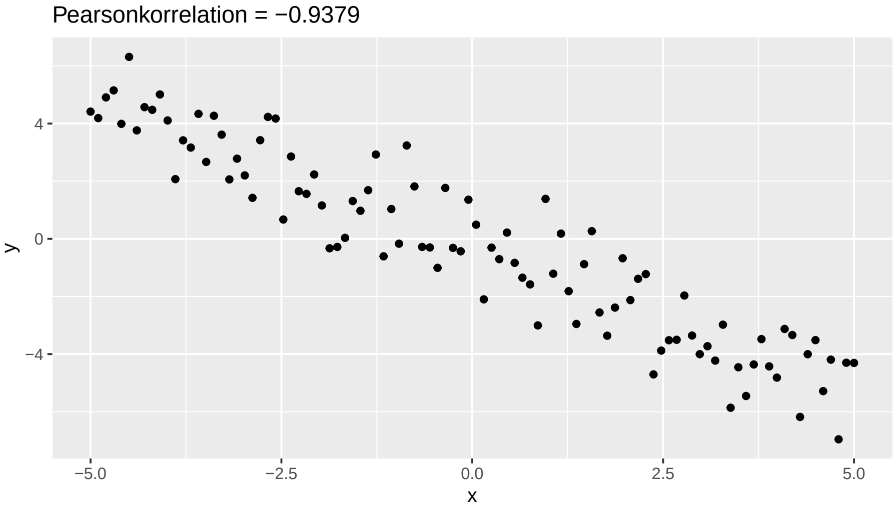
Vad mäter Pearsonkorrelationen?

Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.



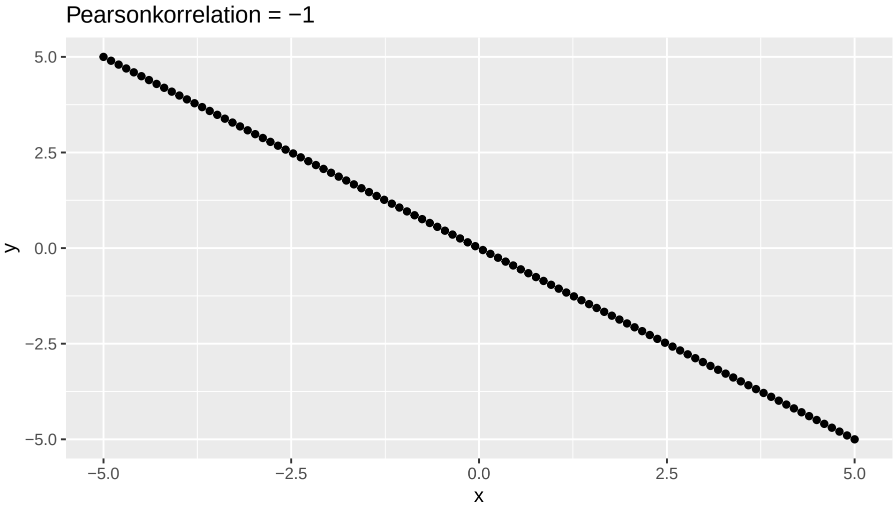
Vad mäter Pearsonkorrelationen?

Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.

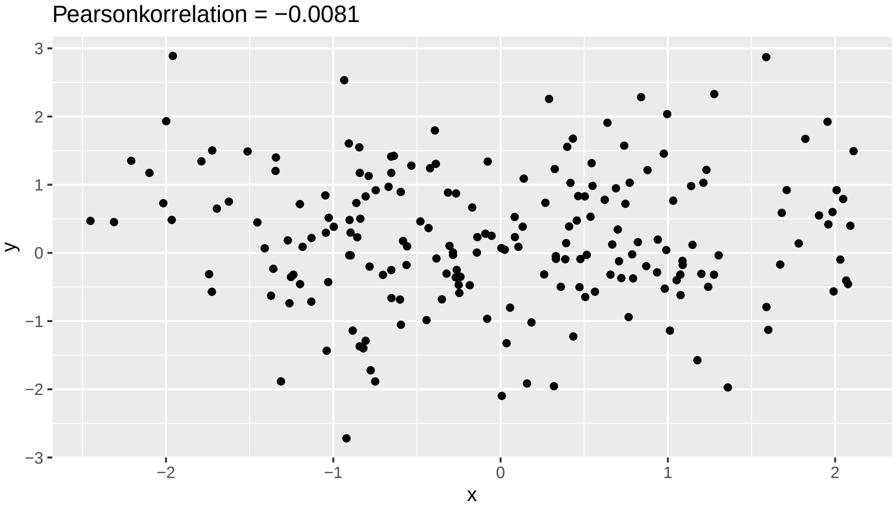


Vad mäter Pearsonkorrelationen?

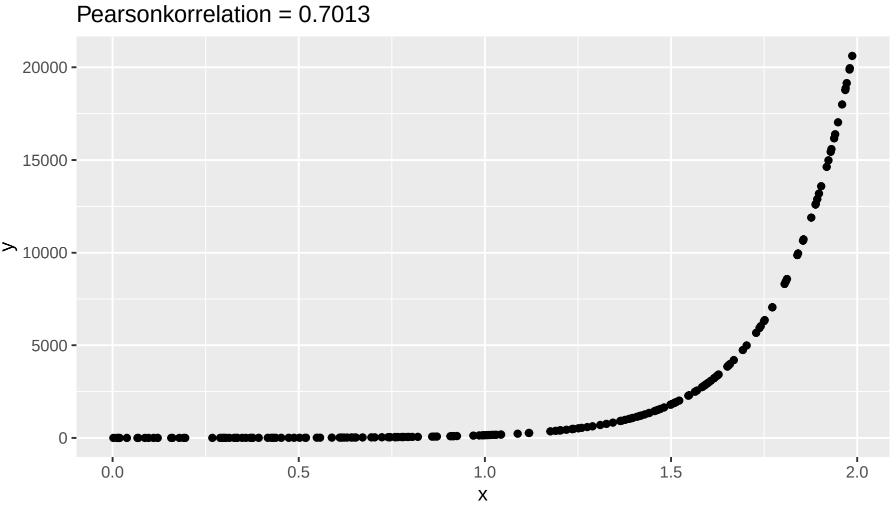
Pearsonkorrelationen mäter graden av **linjärt samband** mellan x och y.



Inget Samband

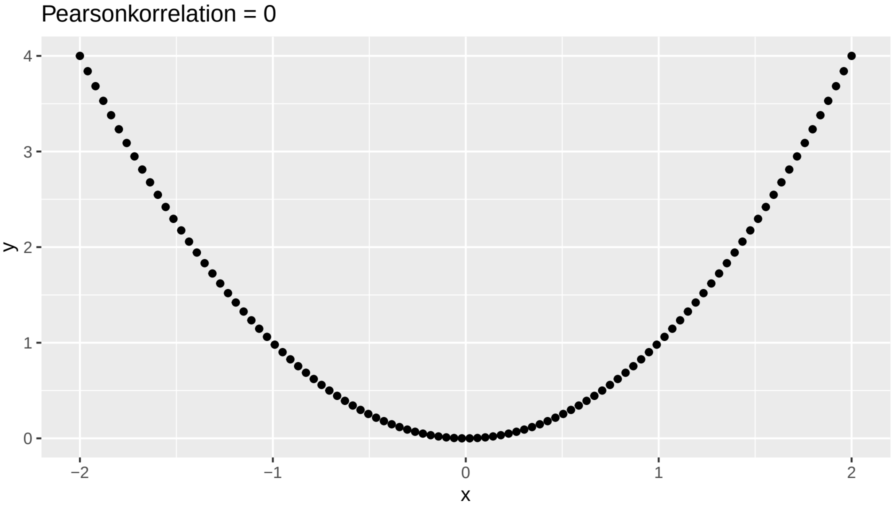


Icke-linjärt samband



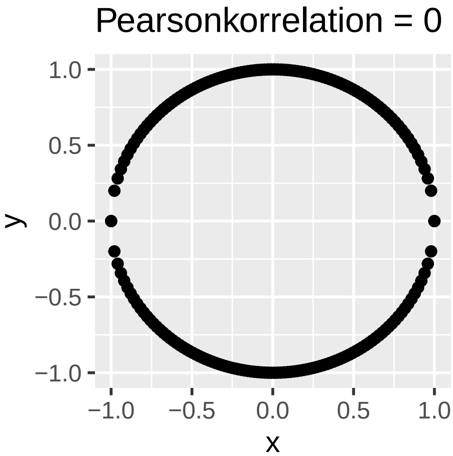
Problemet med Pearsonkorrelation

Pearsonkorrelationen mäter **bara** graden av **linjärt samband** mellan x och y. Om $y = x^2$:

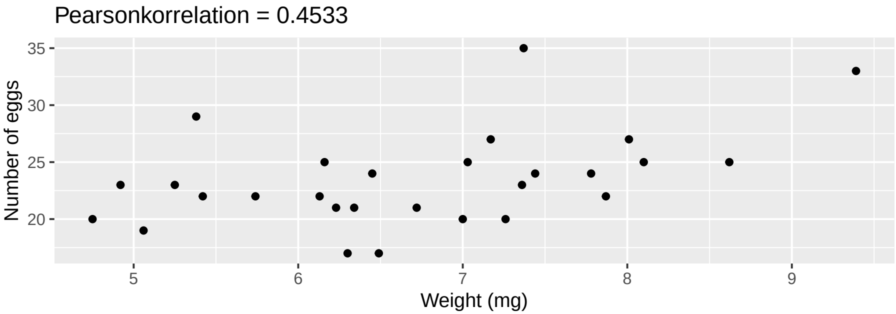


Problemet med Pearsonkorrelation

Pearsonkorrelationen mäter **bara** graden av **linjärt samband** mellan x och y. Om $x^2 + y^2 = 1$:



Tångloppevägning



Beräkning av Pearsonkorrelationen med R:

```
cor(Kraftdjur$Weight, Kraftdjur$Egg)

## [1] 0.4533424
```

Hypotestest för Pearsonkorrelationen

Vi kan utföra ett hypotestest för att testa följande hypoteser:

$$H_0 : \rho = 0 \text{ (det finns inget linjärt samband)}$$
$$H_1 : \rho \neq 0 \text{ (det finns linjärt samband)}$$

Om p-värdet är lågt har vi belägg för att det finns ett linjärt samband.

För att hypotestestet för Pearsonkorrelationen ska fungera ordentligt krävs att:

1. Båda variablerna är normalfördelade, (t.ex. både Kraftdjur$Weight och Kraftdjur$Egg måste vara normaldördelade)
2. Det inte finns några outliers.

Hypotestest för Pearsonkorrelationen

```
cor.test(Kraftdjur$Weight, Kraftdjur$Egg)

##
##  Pearson's product-moment correlation
##
## data:  Kraftdjur$Weight and Kraftdjur$Egg
## t = 2.5934, df = 26, p-value = 0.0154
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09660431 0.70686596
## sample estimates:
##       cor
## 0.4533424
```

Sambandet är signifikant vid 5 % signifikansnivå.

Spearmankorrelation

Ett alternativ är att använda **Spearmankorrelationen** ρ_S.

```
# Pearson
cor(Kraftdjur$Weight, Kraftdjur$Egg)

## [1] 0.4533424

# Pearson
cor(Kraftdjur$Weight, Kraftdjur$Egg, method = "pearson")

## [1] 0.4533424

# Spearman
cor(Kraftdjur$Weight, Kraftdjur$Egg, method = "spearman")

## [1] 0.4473239
```

Konfidensintervall och test

För att beräkna denna rangordnar vi x och y-variablerna var och en för sig, och beräknar sedan Pearsonkorrelationen mellan rangerna.

Weight (mg)	Eggs
5.38	29
7.36	23
6.13	22
4.75	20
8.10	25

$\Rightarrow$

Weight (mg)	Rang	Eggs	Rang
5.38	2	29	5
7.36	4	23	3
6.13	3	22	2
4.75	1	20	1
8.10	5	25	4

Spearmankorrelationen

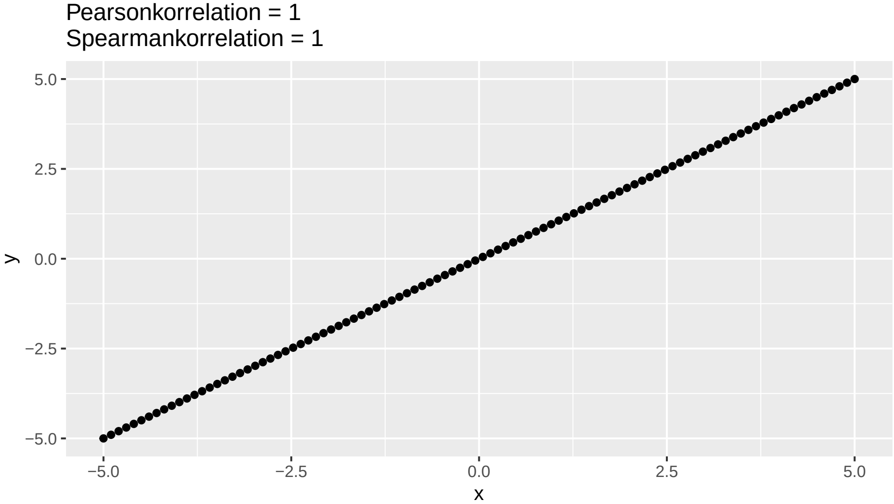
Weight (mg)	Rang	Eggs	Rang
5.38	2	29	5
7.36	4	23	3
6.13	3	22	2
4.75	1	20	1
8.10	5	25	4

```
cor(x = c(5.38, 7.36, 6.13, 4.75, 8.10),
    y = c(29, 23, 22, 20, 25), method = "spearman")

## [1] 0.4
```

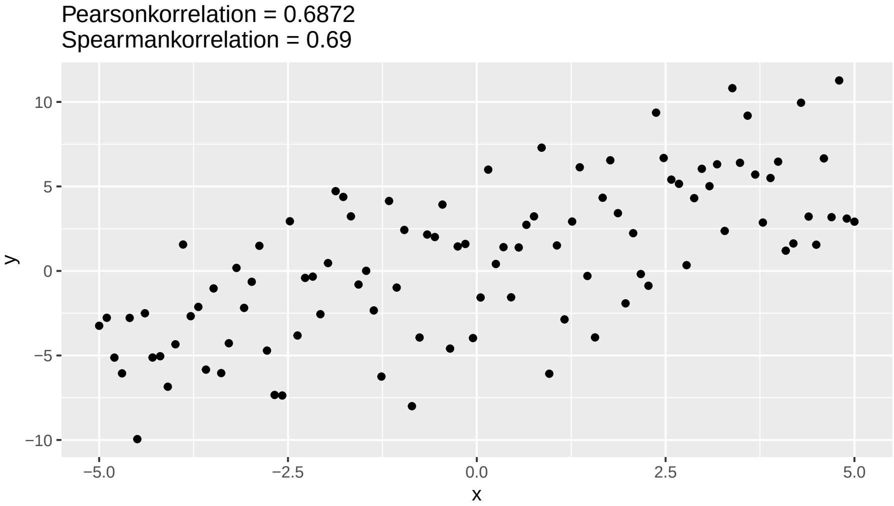
```
cor(x = c(2, 4, 3, 1, 5), y = c(5, 3, 2, 1, 4),
    method = "pearson")

## [1] 0.4
```

Vad mäter korrelationerna?

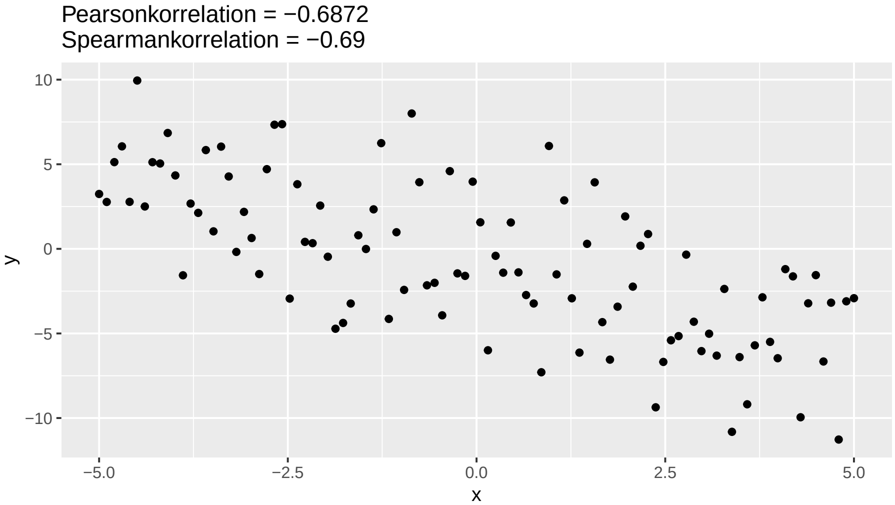
Pearsonkorrelationen mäter graden av **linjärt samband**.
Spearmankorrelationen mäter graden av **monotont samband**.



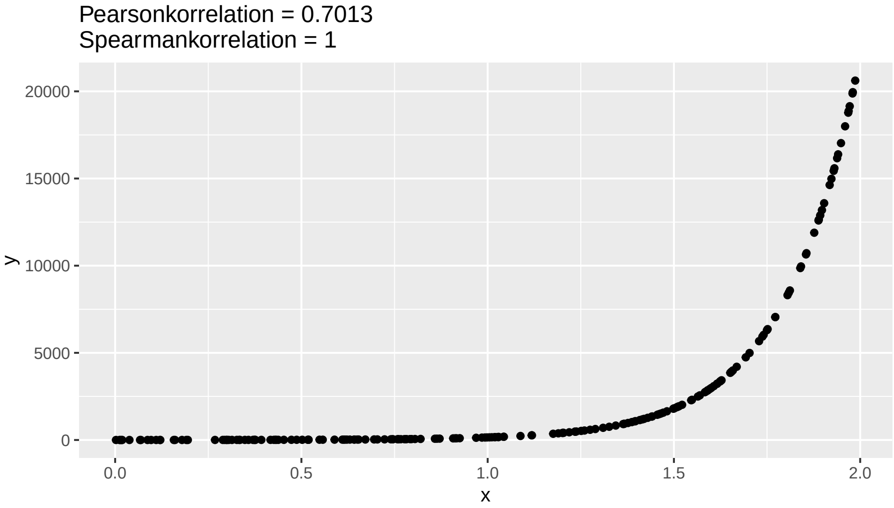
Vad mäter korrelationen?



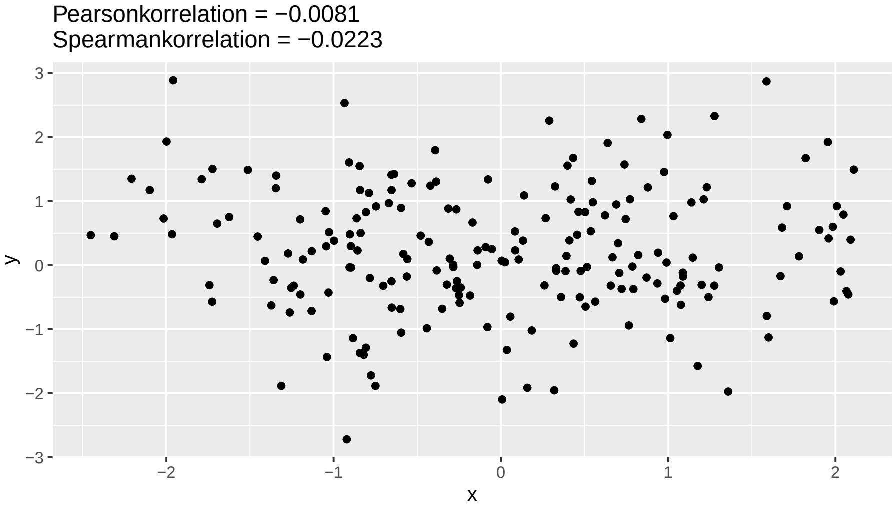
Vad mäter korrelationen?



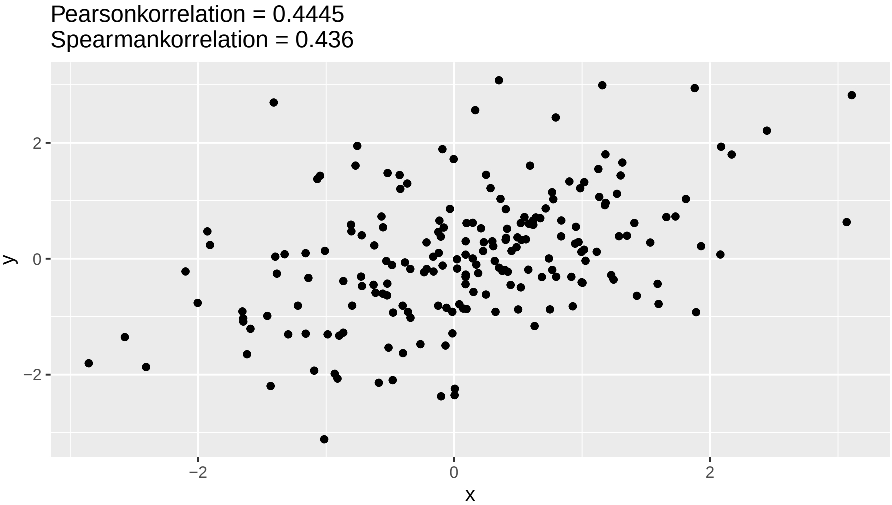
Icke-linjärt samband



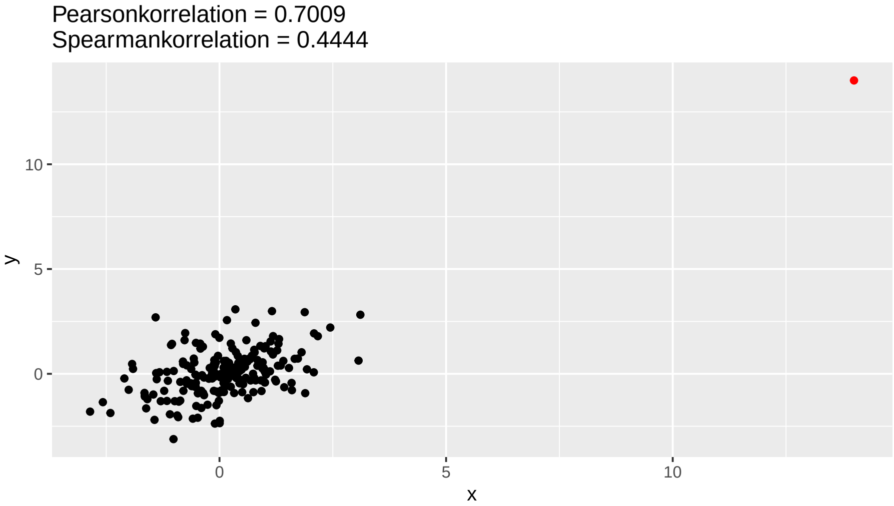
Inget Samband



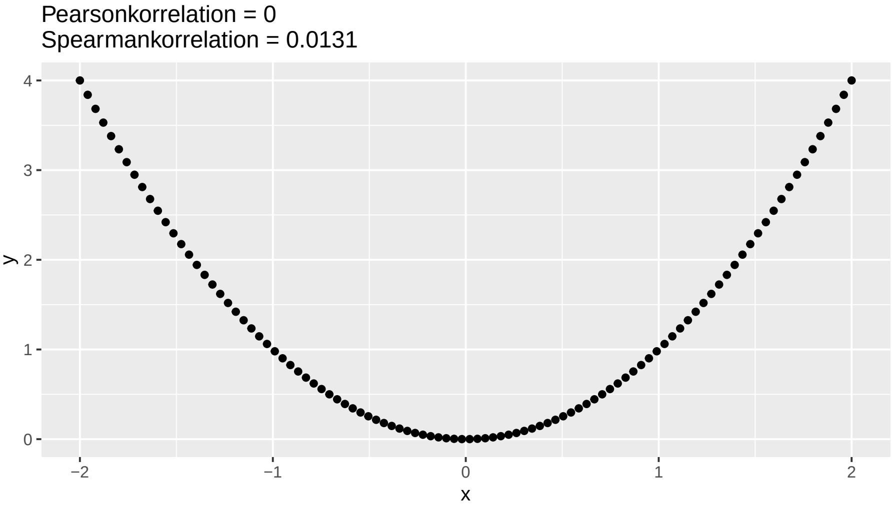
Inte känslig för outliers (Utan outliers)



Inte känslig för outliers (Med outliers)



Problemet med Korrelationerna



Hypotestest för Spearmankorrelationen

$$H_0 : \rho_S = 0 \text{ (det finns inget monotont samband)}$$
$$H_1 : \rho_S \neq 0 \text{ (det finns monotont samband)}$$

```
cor.test(Kraftdjur$Weight, Kraftdjur$Egg, method = "spearman")

##
##  Spearman's rank correlation rho
##
## data:  Kraftdjur$Weight and Kraftdjur$Egg
## S = 2019.5, p-value = 0.017
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4473239
```

Sambandet är signifikant vid 5 % signifikansnivå.

Spearmankorrelationen

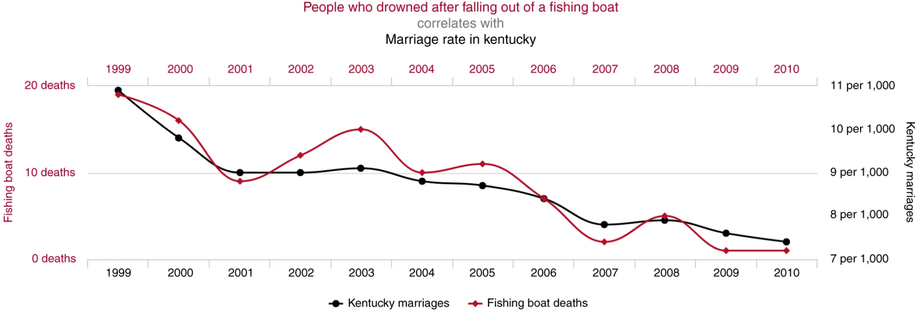
Spearmankorrelationen:

- Kan mäta monotona icke-linjära samband
  - Monoton: går alltid i samma riktning
- Är inte känslig för outliers
- Kräver inte normalfördelning
- Om sambandet är linjärt och data är normalfördelade ger testet som bygger på Pearsonkorrelationen högre styrkan än testet som bygger på Spearmankorrelationen

Annat Problemet med Korrelationerna

Korrelation är inte samma sak som **kausalitet**: alla samband är inte orsakssamband.

- Exempel: Glassförsäljning har en hög korrelation med antalet drunkningsolyckor



Är Svampar Ätliga Eller Inte?

Data: 8124 svampar plockade i Nordamerika på 1980-talet.
Fråga: Säger antalet ringar på svampens fot något om ifall den är ätlig eller oätlig?
Våra data kan sammanfattas med en **frekvenstabell**

	0 ringar	1 ring	2 ringar
Ätlig	0	3680	528
Oätlig	36	3808	72

Finns det ett samband mellan antalet ringar och ätlighet?

$$H_0 : \text{det finns inget samband}$$
$$H_1 : \text{det finns ett samband}$$

Hur får vi fram frekvenstabellen från Data?

Data med 8124 rader ligger i en data.frame

```
##   edible.poisonous ring.number
## 1                p           o
## 2                e           o
## 3                e           o
## 4                p           o
```

Vi kan få fram frekvenstabellen med table():

```
table(Svamp$edible.poisonous, Svamp$ring.number)

##
##         n    o    t
##    e    0 3680  528
##    p   36 3808   72
```

Samband och frekvenstabeller

Idé: om det inte finns något samband (om H_0 stämmer) så är etiketten "ätlig/oätlig" i princip slumpmässig.

	0 ringar	1 ring	2 ringar
Ätlig	0	3680	528
Oätlig	36	3808	72

3916 oätliga svampar (48.2%), 4208 ätliga svampar (51.8%).

1. 36 svampar med 0 ringar:
    1. Förväntat antal oätliga: $36 \cdot 0.482 \approx 17.4$
    2. Förväntat antal ätliga: $36 \cdot 0.518 \approx 18.7$
2. 7488 svampar med 1 ring:
    1. Förväntat antal oätliga: $7488 \cdot 0.482 \approx 3609.4$
    2. Förväntat antal ätliga: $7488 \cdot 0.518 \approx 3878.6$
3. 600 svampar med 2 ringar:
    1. Förväntat antal oätliga:$600 \cdot 0.482 \approx 289.2$
    2. Förväntat antal ätliga: $600 \cdot 0.518 \approx 310.8$

Differens mellan förväntat och observerat

Vi kan beräkna differensen mellan våra observationer och resultatet som förväntas under H_0:

```
##
##               n          o          t
##   e   -18.64697 -198.57016  217.21713
##   p    18.64697  198.57016 -217.21713
```

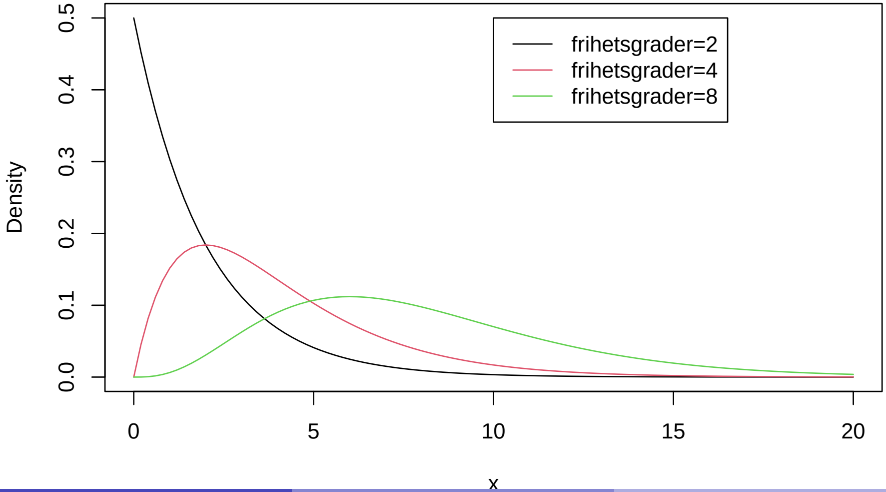
Vi kvadrerar differenserna och skalar om dem genom att dela på antalet förväntade observationer:

$$\frac{(\text{observerat} - \text{förväntat})^2}{\text{förväntat}}$$

Vi summerar alla de här differenserna och kallar summan för Q (ibland kallas den X^2). Stora värden på Q tyder på att H_1 stämmer.

χ^2–Fördelningen

Låt r vara antalet rader i frekvenstabellen, och k vara antalet kolumner. Om H_0 är sann så är Q χ^2-fördelad med $(r-1)(k-1)$ frihetsgrader:



Oberoendetest

Testet kallas för ett (Pearsons) **χ^2-oberoendetest**. Vi testar om det finns ett samband mellan antalet ringar och ätlighet.

```
Table <- table(Svamp$edible.poisonous, Svamp$ring.number)
chisq.test(Table)

##
##  Pearson's Chi-squared test
##
## data:  Table
## X-squared = 374.74, df = 2, p-value < 2.2e-16
```

Fler χ^2-test

Samma resonemang kan användas för flera sorters test:

1. Homogenitetstest: är två (eller fler) populationer lika?
    - Är fördelningen av svamparter densamma i fyra olika områden?
    - Går till på samma sätt som ett **oberoendetest**
    - Vid oberoendetest samlar vi in data och kollar vilka kategorier de tillhör.
    - Vid homogenitetstest samlar vi in data från olika kategorier och kollar en egenskap.
2. Anpassningstest: stämmer en fördelning med på förhand specificerade sannolikheter?
    - Följer andelarna färgblinda/färgseende i Uppsala samma fördelning som i resten av världen (som har kända sannolikheter)?
    - Här behöver vi specificera sannolikheter

Funkar χ^2-test alltid?

χ^2-fördelningen är bara en approximation för fördelningen för Q. Om vi har små stickprov eller sällsynta kombinationer med i vår tabell fungerar approximationen dåligt.

- **Tumregel**: det förväntade antalet observationer i varje cell bör vara minst 5.
- Men det brukar gå bra om någon enstaka cell har färre förväntade observationer än så.

Ett exempel till

Vid ett kliniskt laboratorium samlade man in bakterieprover (dels av Escherichia coli och dels Klebsiella pneumoniae) och undersökte hur många av dessa som var resistenta mot karbapenemer. Resultat:

	Ej resistent	Resistent
E.coli	15	3
K.pneumoniae	17	8

Här har vi ett homogenitetstest!

$$H_0 : \text{andelen resistenta är samma för båda arterna}$$
$$H_1 : \text{andelen resistenta skiljer sig åt mellan arterna}$$

Med R

	Ej resistent	Resistent
E.coli	15	3
K.pneumoniae	17	8

Tabellen sparas med namnet bakterier:

```
bakterier <- matrix(c(15, 3, 17, 8), 2, 2, byrow = TRUE,
                    dimnames = list(c("E.coli", "K. pneumoniae"),
                                    c("Ej resistent","Resistent")))
bakterier

##               Ej resistent Resistent
## E.coli                  15         3
## K. pneumoniae           17         8
```

Vi anger cellvärden, antal rader och kolumner, samt namn på rader och kolumner.

Varning!

Vi kan nu utföra testet:

```
chisq.test(bakterier)

## Warning in chisq.test(bakterier):  Chi-squared
approximation may be incorrect

##
##  Pearson's Chi-squared test with Yates' continuity correctio
##
## data:  bakterier
## X-squared = 0.61249, df = 1, p-value = 0.4339
```

Antal förväntade observationer

Antalet förväntade observationer är mindre än 5 i en cell! Vi kan se detta i R

```
chisq.test(bakterier)$expected

## Warning in chisq.test(bakterier):  Chi-squared
approximation may be incorrect

##                Ej resistent Resistent
## E.coli             13.39535  4.604651
## K. pneumoniae      18.60465  6.395349
```

Fishers exakta test

Vid små stickprov kan **Fishers exakta test** användas istället. Det bygger på exakta beräkningar av hur sannolika olika möjliga resultat i tabellen är.

```
fisher.test(bakterier)

##
##  Fisher's Exact Test for Count Data
##
## data:  bakterier
## p-value = 0.309
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.445298 15.992220
## sample estimates:
## odds ratio
##   2.308043
```

Sammanfattning

1. Korrelation mäter styrkan på samband mellan två numeriska variabler
    1. Ligger mellan -1 och 1
    2. Pearsonkorrelation: linjärt samband
    3. Spearmankorrelation: monotont samband
2. χ^2-test används för att testa hypoteser med frekvenstabeller
    1. Pearsons χ^2-test
    2. Fishers exakta test.