Statistik 1 för biologer, logopeder och psykologer

Föreläsningar, del 2

Innehåll

Innehåll

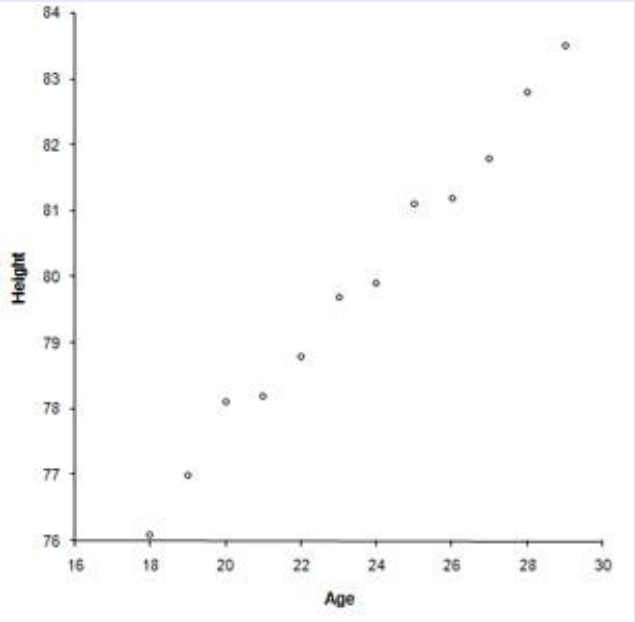
Spridningsdiagram

- Då ett datamaterial består av två (eller flera) variabler är man ofta intresserad av att veta om det finns ett samband mellan variablerna.
- Om variablerna är kvantitativa kan man bilda sig en uppfattning om ett eventuellt samband genom att grafiskt märka ut observationerna i ett koordinatsystem.
- En sådant diagram kallas kallas spridningsdiagram eller *scatter plot*.

Spridningsdiagram

Exempel.

Ett spridningsdiagram som visar sambandet mellan ålder och längd hos ett antal individer. Vi ser ett klart *linjärt* samband mellan variablerna.



Korrelation

- Korrelation anger styrkan och riktningen av sambandet mellan två variabler.
- Om vi anpassar en rät linje genom punkterna i ett spridningsdiagram anger korrelationen hur stor eller liten spridningen kring linjen är.
- Fast inte alla samband är linjära, ger korrelationen ändå *ofta* (men inte alltid!) en god uppfattning om huruvida ett samband finns eller inte.

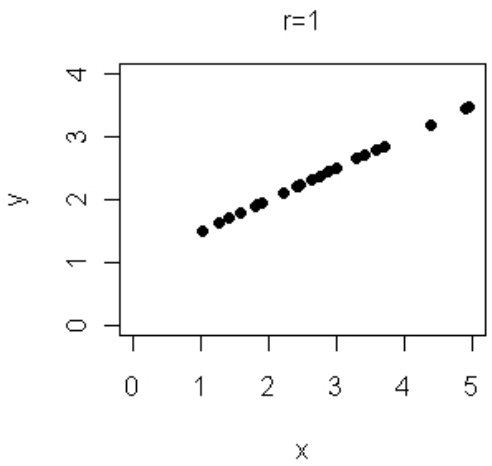
Pearson's korrelationskoefficient

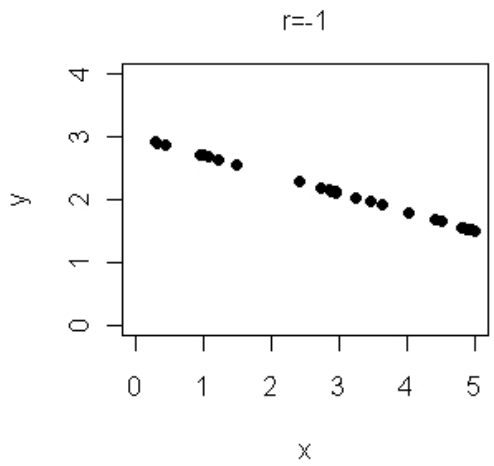
- Pearson's korrelationskoefficient definieras som

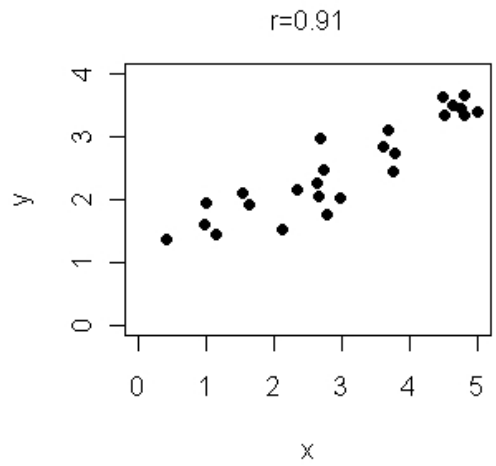
$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1) \cdot s_x \cdot s_y},$$

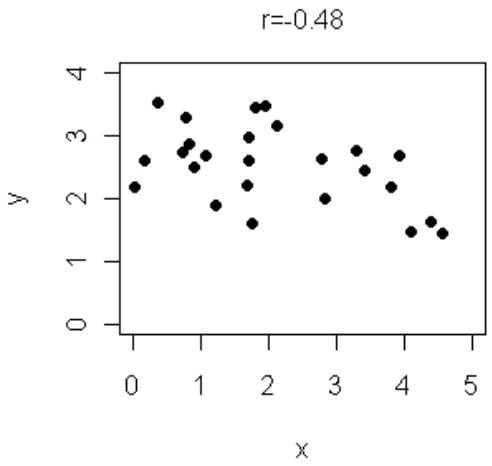
  där s_x är standardavvikelsen av talen x_i, s_y är standardavvikelsen av talen y_i och n är antalet observationspar.
- Korrelationskoefficienten r får ett värde mellan -1 och 1:
  - om $r = -1$, befinner sig alla punkter på en linje med negativ lutning
  - om $r = 0$, finns inget linjärt samband mellan variablerna
  - om $r = 1$, befinner sig alla punkter på en linje med positiv lutning.

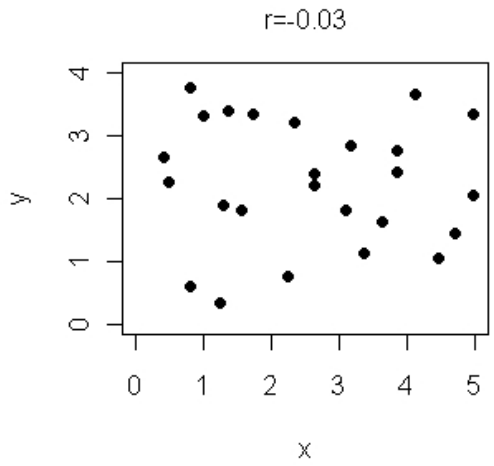
Pearson's korrelationskoefficient

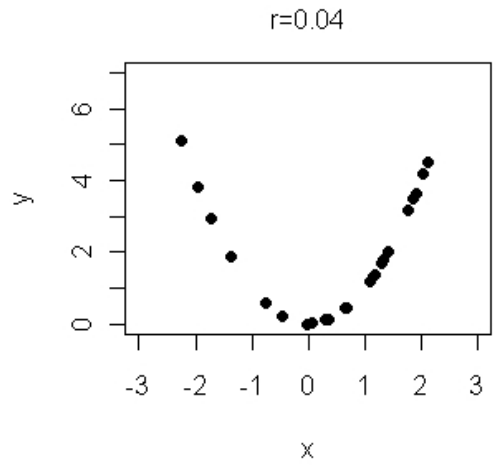












Regression

- Ett samband mellan två variabler kan vidare granskas genom regressionsanalys.
- Regression är en allmän benämning på modeller där värdet på en sk. *beroende* variabel förklaras med hjälp av en eller flera sk. *förklarande* variabler.
- Den vanligaste formen av regression är *linjär regression*, där sambandet mellan variablerna antas vara linjärt.

Enkel linjär regression

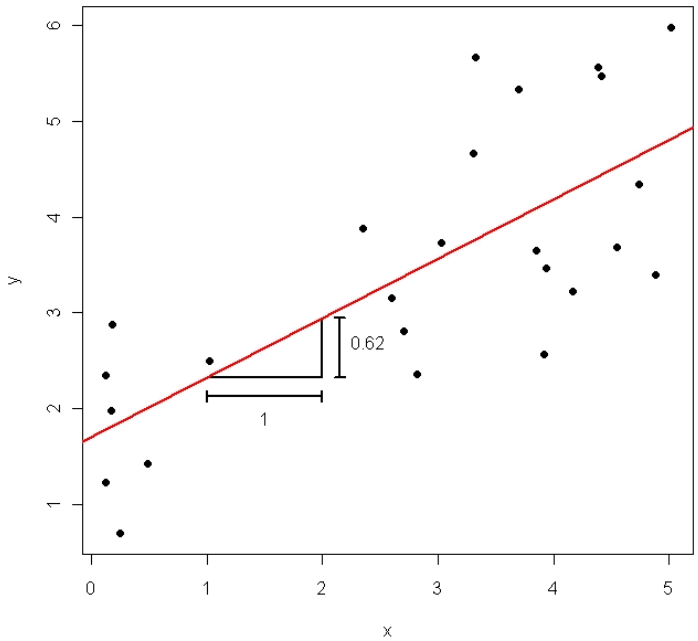
- Aningen förenklat kan ekvationen för enkel linjär regression (linjär regression med *en* förklarande variabel) skrivas som

$$y = a + b \cdot x \ ,$$

  där y är den beroende variabeln och x är den förklarande variabeln.
- Konstanterna a (skärningspunkten med y-akseln) och b (riktningskoefficienten) skattas från datamaterialet men hjälp av *minsta kvadrat-metoden*.

Regressionskoefficienten

- Riktings- eller regressionscoefficienten b talar om för oss hur mycket den beroende variabeln y ändras då den förklarande variabeln x ökar med 1.
- T.ex. i den linjära modellen $y = a + b \cdot x = 1.70 + 0.62x$ leder en ökning av enhetslängd i variabeln x till en ökning av 0.62 i variabeln y.



Förklaringsgrad

- Förklaringsgraden anger vilken andel (ofta i procent) av den beroende variablens värde i en regressionsmodell som kan förklaras med den förklarande variabeln.
- Förklaringsgraden betecknas R^2 och definieras som

$$R^2 = r^2 \cdot 100\% \, ,$$

  där r är korrelationskoefficienten.
- Om korrelationen är -1 eller 1, blir $R^2 = 100\%$. Då kan datamaterialets y-värden fullständigt förklaras med x.
- Om korrelationen är 0 blir $R^2 = 0\%$. Då kan datamaterialets y-värden över huvudtaget inte förklaras med x.

En del vanliga misstolkningar

- Även om y delvis kan förklaras med x kan vi inte säga att y *förorsakas* av x.
- Ett synbarligen starkt samband mellan två orelaterade variabler kan uppstå om båda påverkas av en tredje variabel:
  - t.ex. kan man påvisa ett positivt samband mellan glassförsäljning och antalet drunkningsolyckor. Fast variablerna är totalt orelaterade påverkas de båda av årstiden.

Partiell korrelation och justering

Ibland kan ett samband mellan två variabler x och y på något sätt förvrängas av en tredje variabel z. Såväl korrelations- som regressionskoefficienten kan då få felaktiga värden.

- En korrelationskoefficient mellan x och y där inverkan av z har beaktats, kallas partiell korrelation och betecknas $r_{xy \cdot z}$ (se definition samt exempel `http://faculty.vassar.edu/lowry/ch3a.html`).
- Regressionsmodellen $y = a + bx$ justeras för inverkan av z genom att man lägger till z som förklarande variabel i modellen. I den nya modellen $y = a + bx + cz$ får koefficenten b ett värde där inverkan av z har beaktats.

Rangkorrelation

- Pearson's korrelationskoefficient (dvs. "vanlig" korrelation) lämpar sig endast för intervall- och kvotskalevariabler.
- För ordinalskalevariabler kan man i stället andvända sig av Spearman's rangkorrelationskoefficient

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \,,$$

där d_i är differensen mellan rangtalen av observationerna för individ i och n är antalet observationspar (individer).

- Ett motsvarande rangkorrelationsmått är Kendall's tau, se `http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient`.

Spearman's rangkorrelation

Exempel.

Två vinsmakare A och B ordnar sex vinprover i rangordning från den bästa (1) till den sämsta (6). Hur väl stämmer åsikterna överens?

vinprov	rang		d_i	d_i^2
	A	B		
a	5	3	2	4
b	2	1	1	1
c	1	2	-1	1
d	4	4	0	0
e	3	5	-2	4
f	6	6	0	0

Rangkorrelationen blir då

$$r_s = 1 - \frac{6 \cdot 10}{6(36-1)} = 1 - \frac{2}{7}$$
$$\approx 0.71 \, .$$