Beskrivande statistik

Numeriska värden

Innehåll

1. Skalor
2. Lägesmått
3. Spridningsmått
4. Beroendemått

1. Skalor

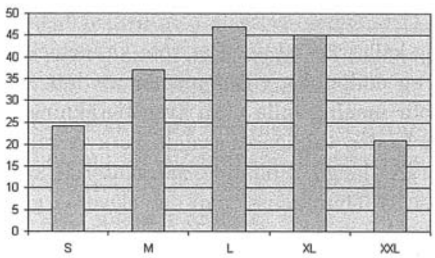
Nominalskala

glassort	antal
Nogger	39
88:an	127
Magnum Classic	212
Magnum vit	198
Skattkista	145
Solero shots	103
Iglo	99
Piggelin	120

Data i 1:a kolumnen!

Det finns ingen storleksordning
Olika kategorier, inget mer

Ordinalskala



Sålda T-tröjor

Storleksordning finns, men differenser saknar betydelse
XXL – XL = ?

Intervallskala

- 0°C, 10°C, 20°C, 30°C
- Differenser har betydelse
- … men inte kvoter
  - 20°C är inte dubbelt så varmt som 10°C
  - Ingen objektiv nollpunkt

Kvotskala

Längder hos elever i skolår 6 på Thunmansskolan

Pojkar	159	155	148	150	175	151	153	146	168
	153	138	161	164	157	146	148	143	
Flickor	155	156	161	158	156	154	160	150	158
	147	160	164	161	157	157	149	143	151
	159	173	161	158	152	152	164	160	157

- En objektiv nollpunkt finns →
- Differenser och kvoter är meningsfulla
- Volym, area, vikt, längd ….

Varför lägesmått, spridningsmått och något sådant ?

Grupp 1:
85, 87, 150, 100, 100, 90, 70, 72, 75, 70, 85, 143, 100, 121, 92, 66, 70, 69, 75, 80, 140, 92, 130, 83, 70, 68, 67, 75, 83, 149, 95, 130, 80, 68, 85, 75, 73, 78, 140, 90, 124, 86, 69, 70, 75, 77, 110, 165, 110, 150, 110, 115, 80, 75, 75, 98, 172, 110, 145, 110, 95, 52, 80, 96, 110, 168, 110, 145, 110, 80, 80, 75, 89, 95, 170, 110, 145, 120, 89, 72, 79, 75, 95

Grupp 2:
220, 100, 149, 100, 110, 80, 85, 80, 90, 165, 103, 135, 95, 77, 76, 85, 80, 88, 155, 103, 120, 85, 79, 78, 82, 75, 85, 150, 103, 135, 90, 75, 85, 78, 75, 88, 150, 95, 130, 90, 70, 76, 89, 82, 95, 145, 100, 133, 90, 77, 89, 79, 80, 90, 165, 103, 135, 95, 77, 86, 80, 85, 100, 160, 120, 140, 100, 90, 79, 92, 70, 100, 165, 120, 140, 100, 120, 86, 71, 95, 100, 155, 120, 139, 100, 89, 86, 78, 78, 110, 158, 122, 145, 108, 95, 95, 78

2. Lägesmått

Det aritmetiskta medelvärdet

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \quad \dots \quad + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Antal syskon: 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 3 4

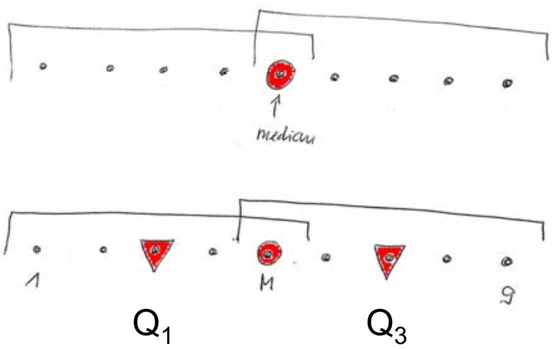
$$\bar{x} = \frac{1}{25}(0 + 0 + 0 + 0 + \quad \dots \quad + 3 + 4) = \frac{27}{25} = 1.08$$

antal syskon (x)	0	1	2	3	4
Antal elever (f)	7	12	4	1	1

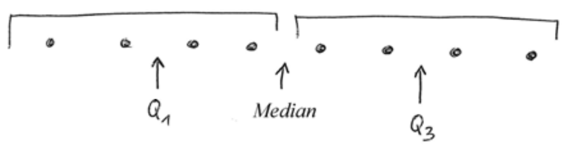
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot x_i = \frac{1}{25}\sum_{i=1}^{5} f_i \cdot x_i$$

$$\bar{x} = \frac{1}{25}(7 \cdot 0 + 12 \cdot 1 + 4 \cdot 2 + 1 \cdot 3 + 1 \cdot 4) = \frac{27}{25} = 1.08$$

Median och kvartiler; n=9 (ojämt)



Median och kvartiler; n=8 (jämt)



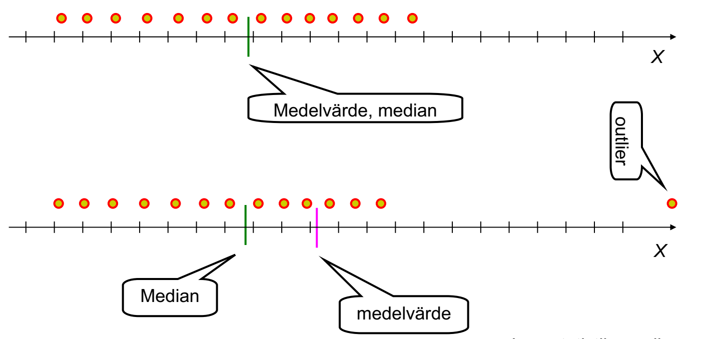
Flera definitioner för kvartiler !

The following table summarizes a number of common methods for computing the position of the first and third quartiles from a sample size n (P. Stikker, pers. comm., Jan. 24, 2005). In the table, $[x]$ denotes the nearest integer function.

method	1st quartile	1st quartile	3rd quartile	3rd quartile
	n odd	n even	n odd	n even
Minitab	$\frac{n+1}{4}$	$\frac{n+1}{4}$	$\frac{3n+3}{4}$	$\frac{3n+3}{4}$
Tukey (Hoaglin et al. 1983)	$\frac{n+3}{4}$	$\frac{n+2}{4}$	$\frac{3n+1}{4}$	$\frac{3n+2}{4}$
Moore and McCabe (2002)	$\frac{n+1}{4}$	$\frac{n+2}{4}$	$\frac{3n+3}{4}$	$\frac{3n+2}{4}$
Mendenhall and Sincich (1995)	$\left[\frac{n+1}{4}\right]$	$\left[\frac{n+1}{4}\right]$	$\left[\frac{3n+3}{4}\right]$	$\left[\frac{3n+3}{4}\right]$
Freund and Perles (1987)	$\frac{n+3}{4}$	$\frac{n+3}{4}$	$\frac{3n+1}{4}$	$\frac{3n+1}{4}$

Källa: *Wolfram Research*

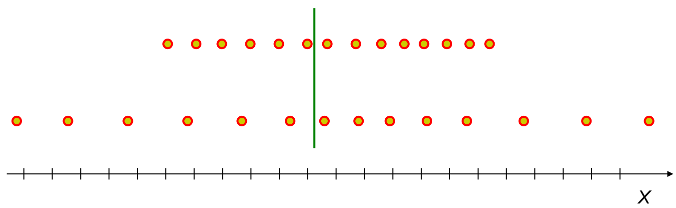
Medelvärde eller median ?



Lönestatistik: median

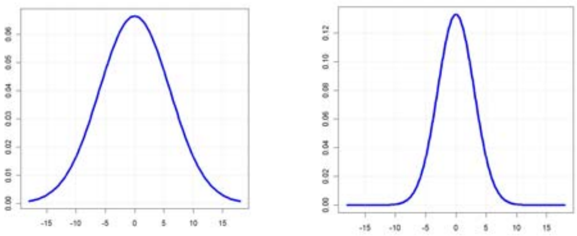
3. Spridningsmått

Medelvärdet räcker inte till



Samma medelvärdet, men ändå helt olika

Medelvärdet räcker inte till



Samma medelvärdet, men ändå helt olika

Variationsbredd

67 69 71 75 75 75 75 75 78 78 78 79 79 80 80 82 85 85 89 92

$$V = v_{\max} - v_{\min} = 92 - 67 = 25$$

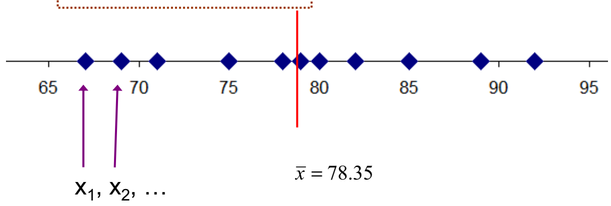
Mellankvartilvariation

67 69 71 75 75 | 75 75 75 78 78 | 78 79 79 80 80 | 82 85 85 89 92

$$\Delta Q_{1,3} = Q_3 - Q_1 = 81 - 75 = 6$$

Stickprovsvarians

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \textit{stickprovsvarians}$$



$$\bar{x} = 78.35$$

$$x_1 = 67 \qquad x_2 = 69 \qquad x_3 = \ldots$$

$$(x_1 - \bar{x})^2 = (67 - 78.35)^2 = 128.82$$

$$(x_2 - \bar{x})^2 = (69 - 78.35)^2 = 87.42$$

$$\ldots$$

Standardavvikelse

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{stickprovs varians}$$

$$s = \sqrt{s^2} \qquad \textit{standardavvikelse}$$

$$s^2 = \frac{1}{n-1}S_{xx} \qquad S_{xx} = \sum x_i^2 - n \cdot \bar{x}^2$$

Miniräknare: σ_{n-1} = (stickprovs-) standardavvikelse

Standardavvikelse för klassindelade (grupperade) data

antal syskon (x)	0	1	2	3	4
Antal elever (f)	7	12	4	1	1

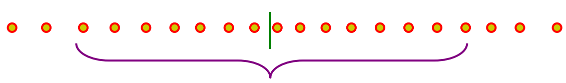
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2 \qquad (k \;\; klasser)$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i x_i^2 - n\bar{x}^2\right]$$

Standardavvikelse – att kolla:

- Måste vara positiv
- Minst 75% av alla observationer i intervallet:

$$\boxed{(\bar{x} - 2 \cdot s \, , \, \bar{x} + 2 \cdot s)}$$



4 standardavvikelser runt medelvärdet innerhåller minst 75% av mätpunkterna

4. Beroendemått

Förkortningar (som är nyttiga)

$$S_{xy} = \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n \cdot \bar{x}^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n \cdot \bar{y}^2$$

Förkortningar

$$c_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n-1} S_{xy} \qquad \textit{kovarians}$$

$$s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} S_{xx} \qquad \textit{stickprovsvarians för } x$$

$$s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{1}{n-1} S_{yy} \qquad \textit{stickprovsvarians för } y$$

Korrelationskoefficient

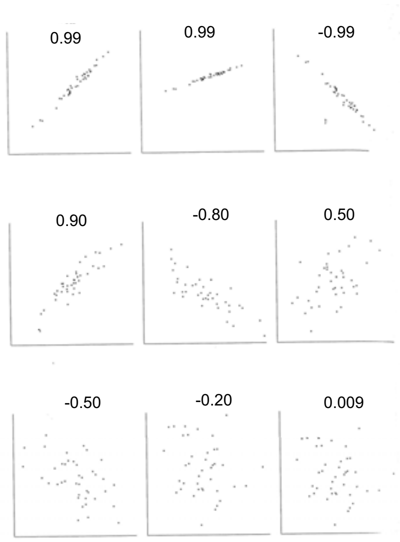
- Finns ett linjärt samband mellan x och y?
- Hur beroende är x och y av varandra?

$$r = \frac{S_{xy}}{\sqrt{S_{yy} \cdot S_{xx}}} \qquad -1 \le r \le 1$$

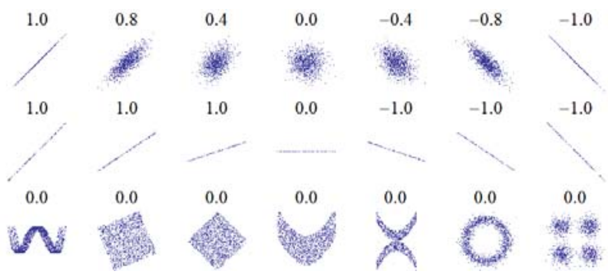
$$r = \frac{c_{xy}}{s_x \cdot s_y} \qquad c_{xy} = \frac{1}{n-1} S_{xy} \qquad (\textit{kovarians})$$

Korrelationskoefficient

- *Korrelation positiv:* höga värden på den ena variabeln hänger ihop med höga värden på den andra variabeln
- *Korrelation negativ:* höga värden på den ena variabeln hänger ihop med låga värden på den andra variabeln

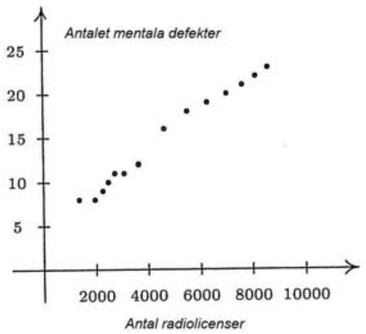


Korrelationskoefficient



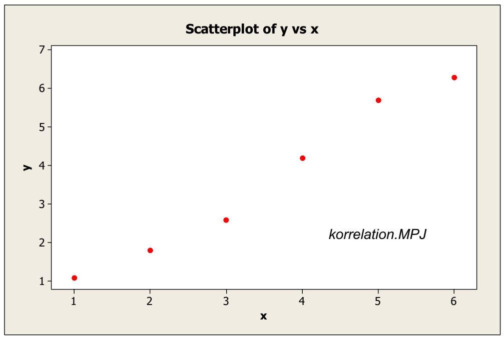
Källa: Wikipedia

Korrelation ≠ Kausalitet !



Korrelation: Räkneexempel

x	y
1	1.1
2	1.8
3	2.6
4	4.2
5	5.7
6	6.3



Räkneexempel

x	y
1	1.1
2	1.8
3	2.6
4	4.2
5	5.7
6	6.3

$$\bar{x} = \tfrac{1}{6}(1+2+3+4+5+6) = \tfrac{21}{6} = 3.5$$

$$\sum x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 = 91$$

$$\bar{y} = \tfrac{1}{6}(1.1+1.8+2.6+4.2+5.7+6.3) = \tfrac{21.7}{6} = 3.6167$$

$$\sum y_i^2 = 1.1^2 + 1.8^2 + 2.6^2 + 4.4^2 + 5.7^2 + 6.3^2 = 101.03$$

$$\sum x_i \cdot y_i = 1 \cdot 1.1 + 2 \cdot 1.8 + 3 \cdot 2.6 + 4 \cdot 4.2 + 5 \cdot 5.7 + 6 \cdot 6.3) = 95.6$$

$$S_{xy} = \sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y} = 95.6 - 6 \cdot 3.5 \cdot 3.6167 = 19.65$$

$$S_{xx} = \sum x_i^2 - n \cdot \bar{x}^2 = 91 - 6 \cdot 3.5^2 = 17.5$$

$$S_{yy} = \sum y_i^2 - n \cdot \bar{y}^2 = 101.03 - 6 \cdot 3.6167^2 = 22.547$$

$$r = \frac{S_{xy}}{\sqrt{S_{yy} \cdot S_{xx}}} = \frac{19.65}{\sqrt{22.547 \cdot 17.5}} = 0.989$$

Räkneexempel - Minitab

x	y
1	1.1
2	1.8
3	2.6
4	4.2
5	5.7
6	6.3

- *Corrdata.MTW*
- Stat / Basic Statistics / Correlation
- Variables: ”x y” (mellanslag)

Correlations: x, y

```
Pearson correlation of x and y = 0.989
P-Value = 0.000
```

Räkneexempel

$$c_{xy} = \frac{1}{n-1} S_{xy} = \frac{1}{5} \cdot 19.65 = 3.93 \qquad \textit{covarians}$$

$$s_x = \frac{1}{n-1} S_{xx} = \frac{1}{5} \cdot 17.5 = 3.5 \qquad \textit{stickprovsvarians för } x$$

$$s_y = \frac{1}{n-1} S_{yy} = \frac{1}{5} \cdot 22.547 = 4.51 \qquad \textit{stickprovsvarians för } y$$

Stat / Basic Statistics / Covariance
Variables: ”x y”

Covariances: x, y

```
          x          y
x   3.50000
y   3.93000    4.50967
```