Statistik för bioteknik sf1911
Föreläsning 2

TK

01.11.2017

KTH Matematik

Föreläsning 2: lärandemål

- mer om korrelationskoefficienten
- mer om linjär regression
- kvantiler, kvartiler, utliggare, lådagram (boxplot)
- histogram
- klockkurva (normalkurva)

KTH Matematik

Basic Concepts of Correlation

We are looking at paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We can look at them as a table, too.

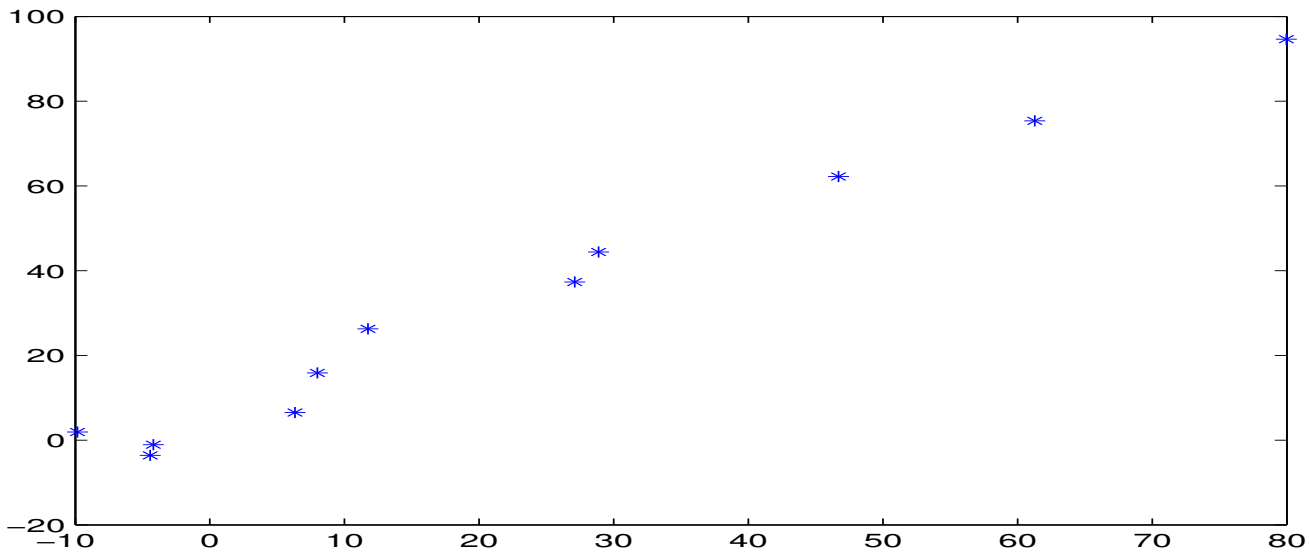
A **correlation** exists between two variables when one of them is related to the other in some way.

	Sample			
	1	2	...	n
Variable x	x_1	x_2	...	x_n
Variable y	y_1	y_2	...	y_n

KTH VETENSKAP OCH KONST
KTH Matematik

Scatterplot

A **scatterplot** *is a graph in which the paired samples* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *are plotted with a horizontal* x *-axis and a vertical* y*-axis. Each individual* (x, y) *pair is plotted as a single point (*$*$ *in the figure).*



KTH Matematik

Covariance

The *covariance* between x- and y-values in $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ is

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

and where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

The intuitive interpretation is: if large sample values of x tend to go together with large sample values of y, then $(x_i - \bar{x})(y_i - \bar{y})$ is most often positive, and c_{xy} will tend to be positive.

The other way, if small sample values of x tend to go together with large sample values of y, then $(x_i - \bar{x})(y_i - \bar{y})$ is most often negative, and c_{xy} will tend to be negative.

KTH Matematik

Covariance and Correlation Coefficient

We standardize (to get back to the original units of measurement)

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

with s_x and s_y,

$$s_x = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}, s_y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

and get the *correlation coefficient* as

$$r \stackrel{\text{def}}{=} \frac{c_{xy}}{s_x s_y},$$

KTH VETENSKAP OCH KONST
KTH Matematik

The Correlation Coefficient

The **correlation coefficient** *r measures the strength of* **linear association** *between paired x and y sample values.*

KTH Matematik

The Correlation Coefficient: requirements for validity

- random sample
- visual examination of the scatterplot must confirm that the points approximate a straight line pattern.
- Any **outliers** ((utliggare)=points lying far away from the other data points) must be removed if they are errors. The effects of any other outliers should be considered

KTH Matematik

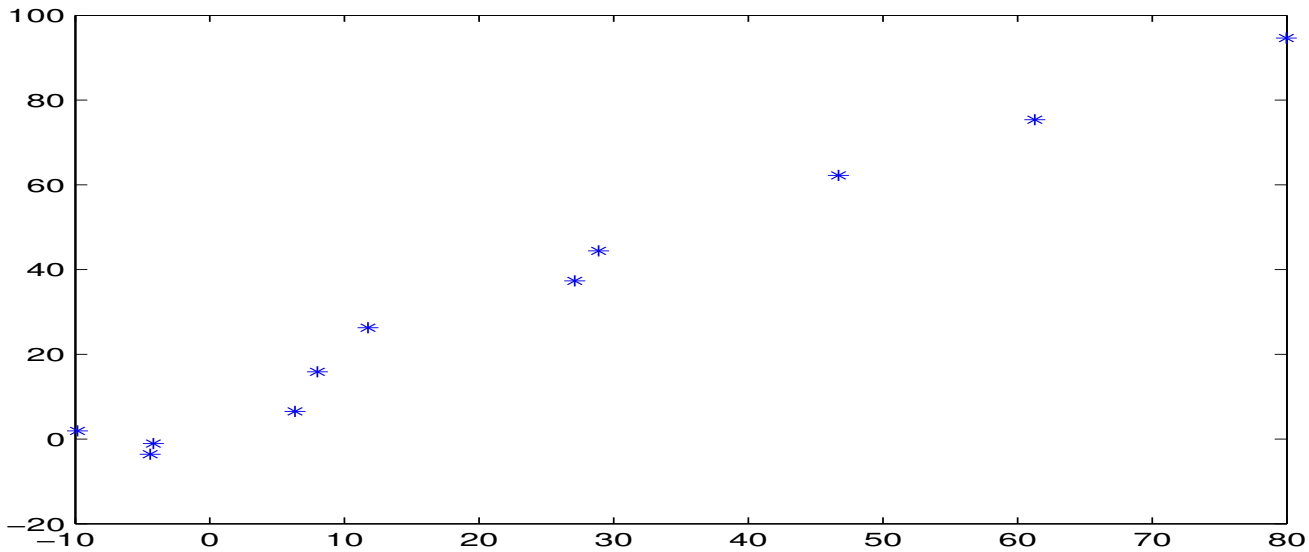
Coefficient of correlation: general properties

- *r is always between -1 and 1, i.e., $-1 \leq r \leq 1$. $r = 1$: perfect positive correlation, $r = -1$: perfect negative correlation, $r = 0$: x and y are non-correlated. (Recall the discussion of the sign of c_{xy})*
- *the value of r does not change if all values of the other variable are converted to a different scale*
- *r is not changed if x values are interchanged with y values.*
- ***r measures the strength of a linear association, it is not designed measure the strength of an association that is not linear.***

KTH Matematik

Coefficient of correlation

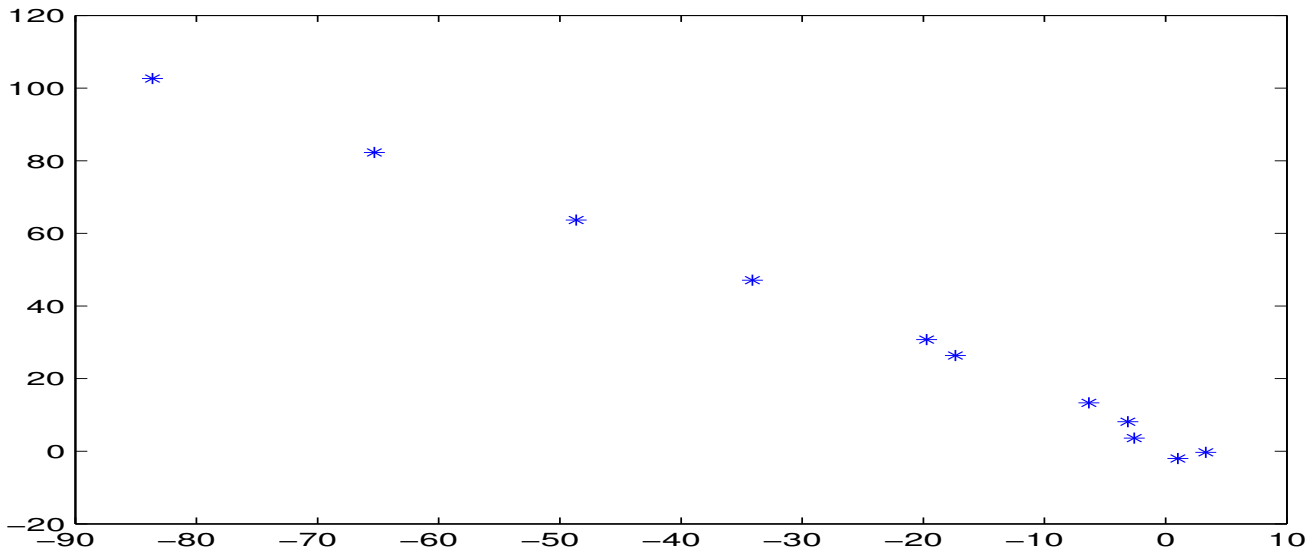
Here $r = 0.9989$



KTH Matematik

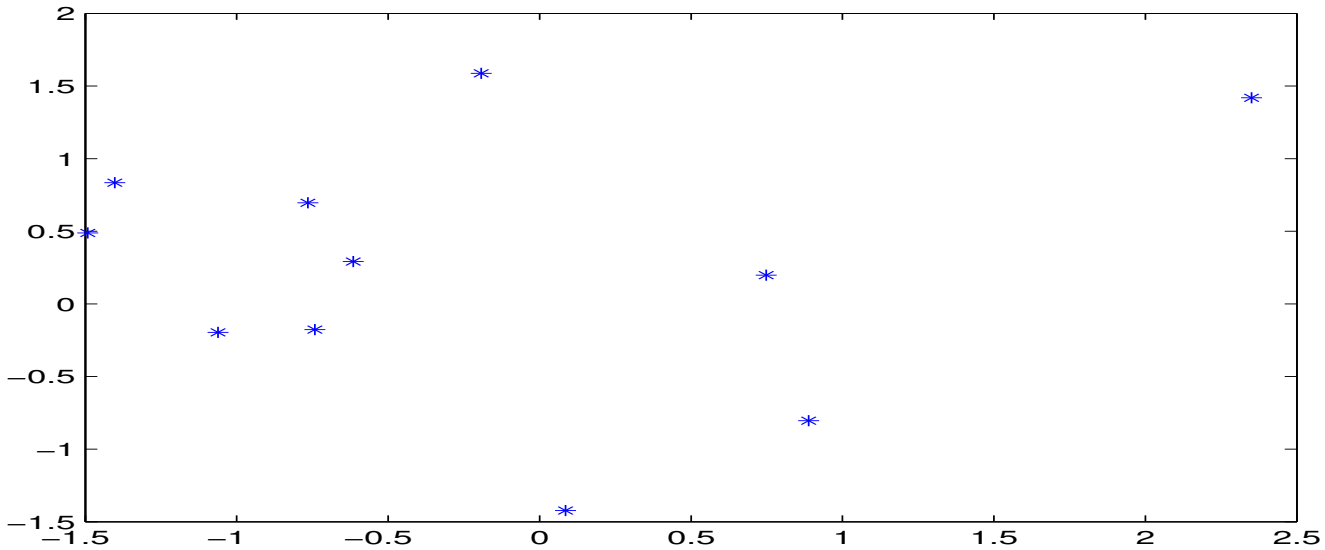
Coefficient of correlation

Here $r = -0.9977$



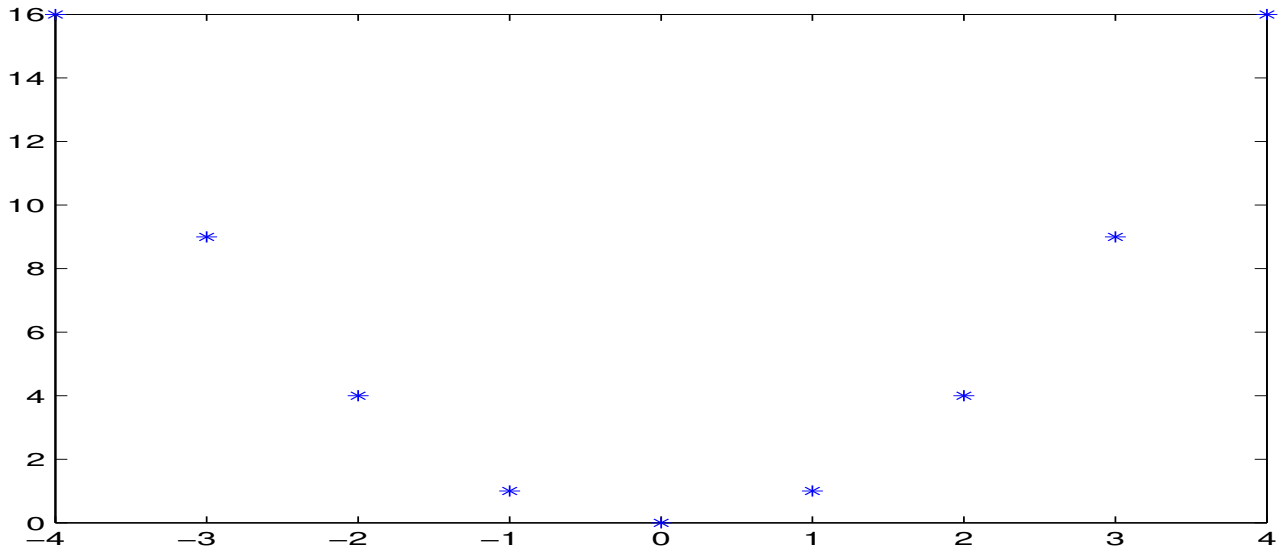
Coefficient of correlation

Here $r = 0.063$



A non-linear association and the coefficient of correlation

Here $y = x^2$ is plotted for $x = -4, -3, \ldots 3, 4$. $r = 0.000$!



KTH VETENSKAP OCH KONST
KTH Matematik

Correlation and Causation

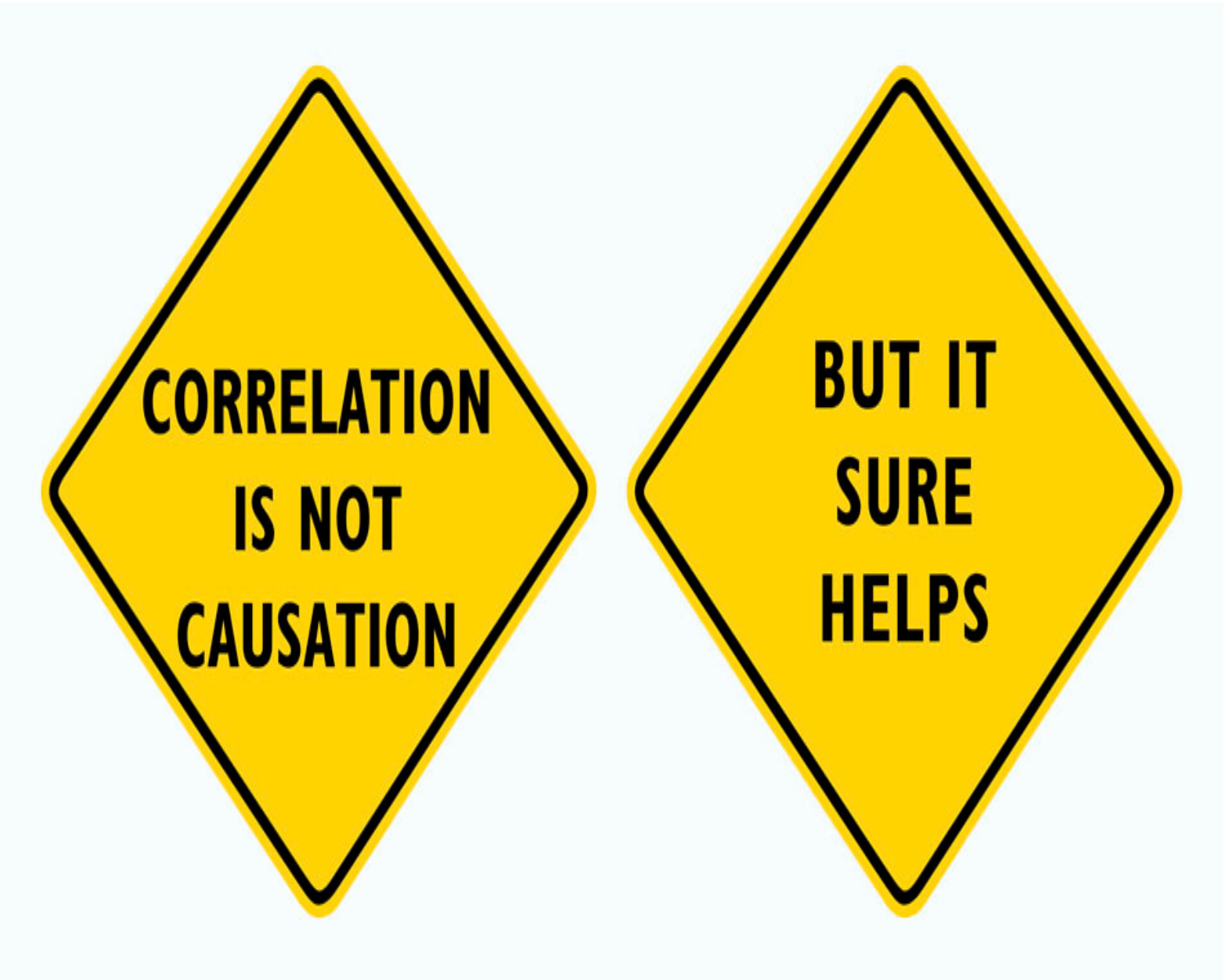
Correlation does not imply causation *is a phrase used to emphasize that a correlation between two variables does not necessarily imply that one causes the other. The counter assumption, that correlation proves causation, is considered a questionable cause logical fallacy in that two events occurring together are taken to have a cause-and-effect relationship.*

For a more thoroughgoing analysis, see

Bill Shipley: Cause and Correlation in Biology, *Cambridge University Press 2000*

KTH VETENSKAP OCH KONST
KTH Matematik

Correlation and Causation



KTH Matematik

Epidemiological studies showed that women taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading to the conclusion that HRT was protective against CHD.

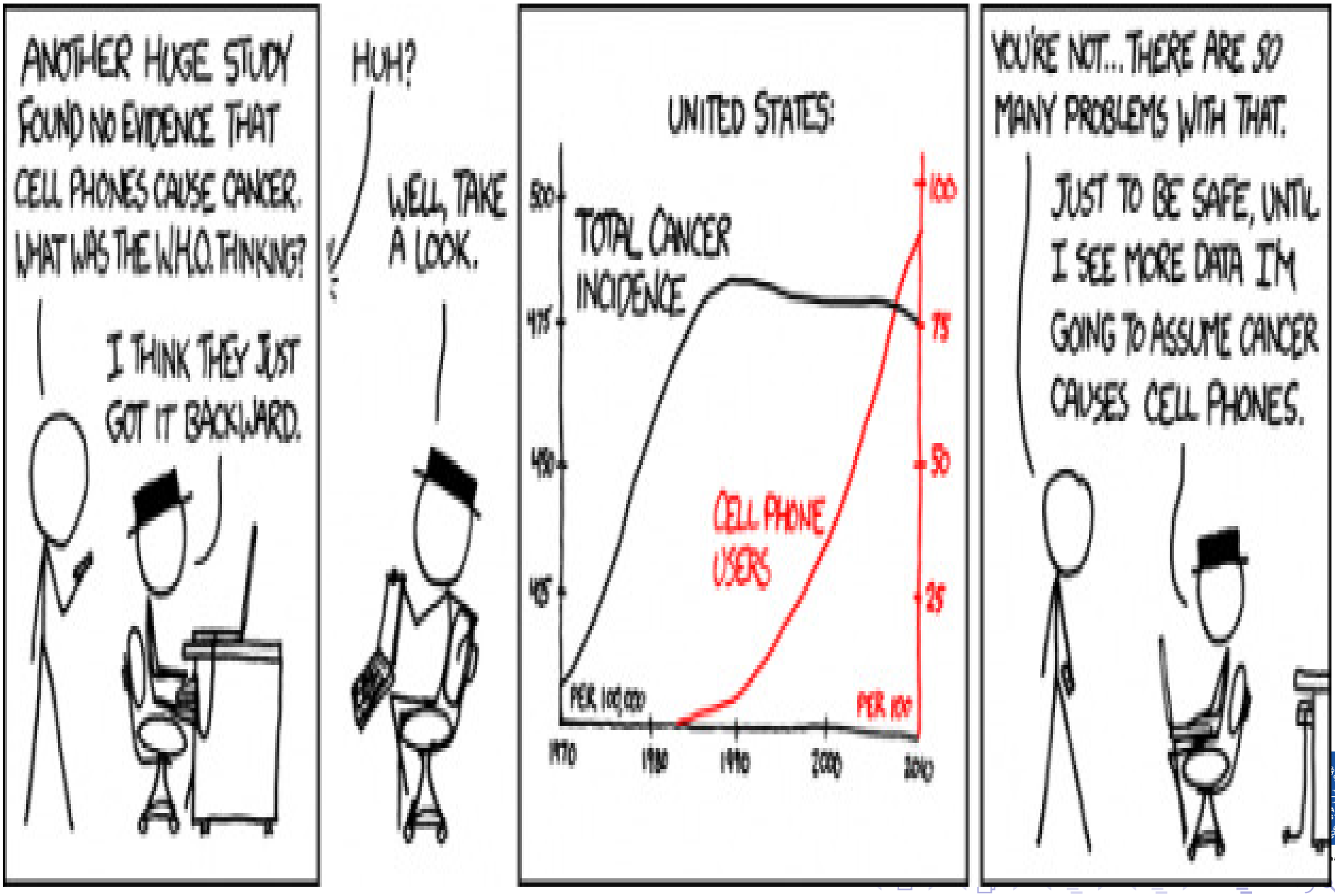
KTH Matematik

But **randomized controlled trials** showed that HRT caused a small but **statistically significant** ($\leftarrow$ förklaras senare) increase in risk of CHD. Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups, with better-than-average diet and exercise regimens.

The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socioeconomic status), rather than a non-confounded cause and effect, as had been supposed.

KTH VETENSKAP OCH KONST
KTH Matematik

Correlation and Causation



Lite till om linjär regression

Regression: what is in the name ?

Sir Francis Galton, 1822-1911, bought family records that contained the heights of 205 sets of parents and their adult children. If the parents were short their children were slightly taller, on the other hand, if the parents were tall then the children were slightly shorter. This lead Galton to invent the word regression. *Regression was defined as the process of returning to the mean*. In the experiment the smallest parents had offspring who were bigger and closer to the mean. The largest parents had offspring who were smaller and once again closer to the mean.
During Galton's studies and experiments he invented words such as eugenics and regression. Galton first thought that breeding two smart people would produce an even smarter person. He also thought that breeding two tall people would produce an even taller person.

KTH Matematik

Return to the mean

If $-1 < r < 1$, then (c.f. visas på övningarna)

$$\frac{\widehat{y} - \bar{y}}{s_y} = r\frac{(x - \bar{x})}{s_x}$$

that the predicted standardized value $\widehat{y}$ of y is closer to its mean $\bar{y}$ than the standardized value of x is to its mean $\bar{x}$. Thus the data points $(x_i, y_i)_{i=1}^n$ display regression toward the mean (as found and formulated by Francis Galton).

KTH Matematik

Return to the mean (4)

The idea of regression toward the mean resolved an important difficulty of Darwinian selection:

> *if offspring were always identical to parents, then evolution by natural selection was not possible. But, on the other hand, there was also intergenerational stability, as all experience under fairly constant environmental conditions showed that the range of variability on short timescales, as between two generations, was essentially constant. (Stephen Stigler)*

KTH Matematik

Regression line: terminology

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x \tag{1}$$

is called the **regression line**. Here x is called the **independent variable** or **predictor variable** or **explanatory variable** and $\widehat{y}$ is called the **dependent variable** or **response variable**. a is called the **intercept** and b is the **slope**. We will find later $\widehat{\alpha}$ and $\widehat{\beta}$ as sample estimates of *population intercept* (**population: förklaras senare**) α and *population slope* β (**population: förklaras senare**).

KTH Matematik

Galton and his students: biometrics

$$y = \widehat{\alpha} + \widehat{\beta} x \tag{2}$$

The predictor/explanatory variable

$$x = \frac{\text{mother's height} + \text{father's height}}{2}.$$

The dependent variable or response variable y is the height of the child. The slope β measures heritability and the intercept α is like an average of environmental effects.

KTH Matematik

Examples

- x = systolic reading, y = diastolic blood pressure
- x = cholesterol level, y = weight
- x = tree circumference, y = tree height
- e.t.c.

KTH VETENSKAP OCH KONST
KTH Matematik

Regression: marginal change

$$\widehat{y} = \alpha + \beta x \tag{3}$$

When x changes by one unit we get the $\widehat{y}^{\dagger} = \alpha + \beta(x+1)$. We have the **marginal change** $\triangle y$

$$\triangle y = \widehat{y}^{\dagger} - \widehat{y} = \beta$$

Hence the slope β represents the change in response, when the explanatory variable is changed by one unit.

KTH VETENSKAP OCH KONST
KTH Matematik

Residuals

The vertical distances e_i from y_i to regression line at x_i,

$$e_i \stackrel{\text{def}}{=} y_i - \widehat{y} = y_i - \widehat{\alpha} - \widehat{\beta} x_i$$

when a and b are computed by the formulae above, are called **residuals**.

$$\underline{\text{Residual} = \text{observed y - predicted y}}$$

$$\sum_{i=1}^{n} e_i^2.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Residuals

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

its minimum value is

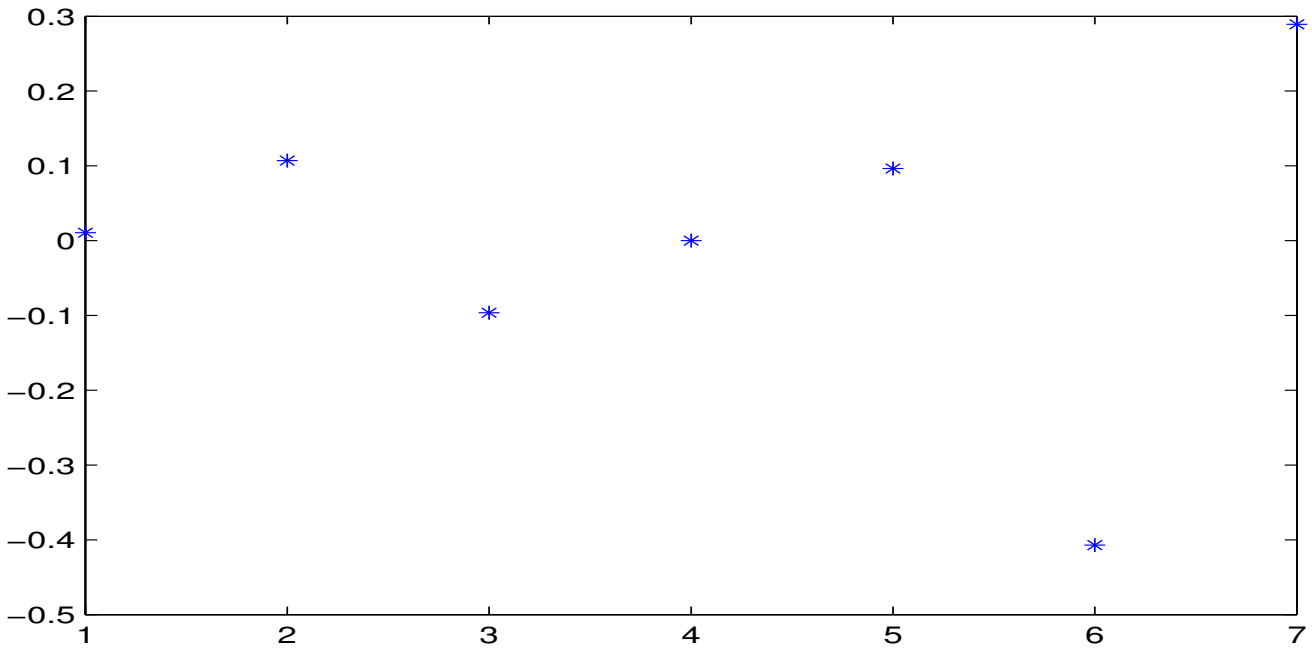
$$Q_0 = \sum_{i=1}^{n} e_i^2.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Residual plot

Definition

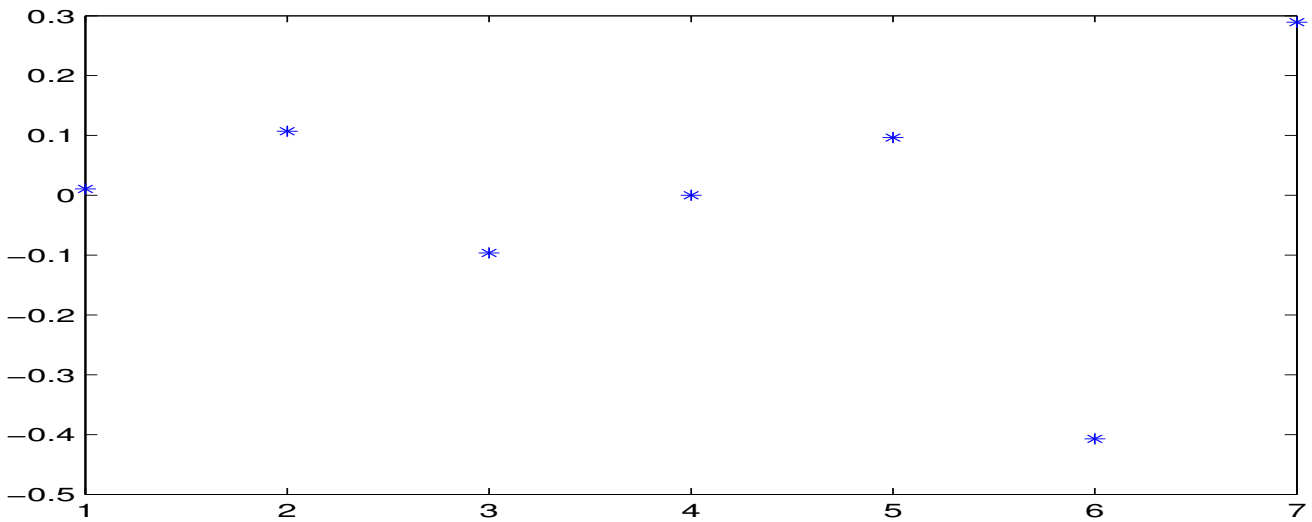
A **residual plot** *is a scatterplot of the pairs* (x_i, ε_i) $i = 1, 2, \ldots n$.



KTH Matematik

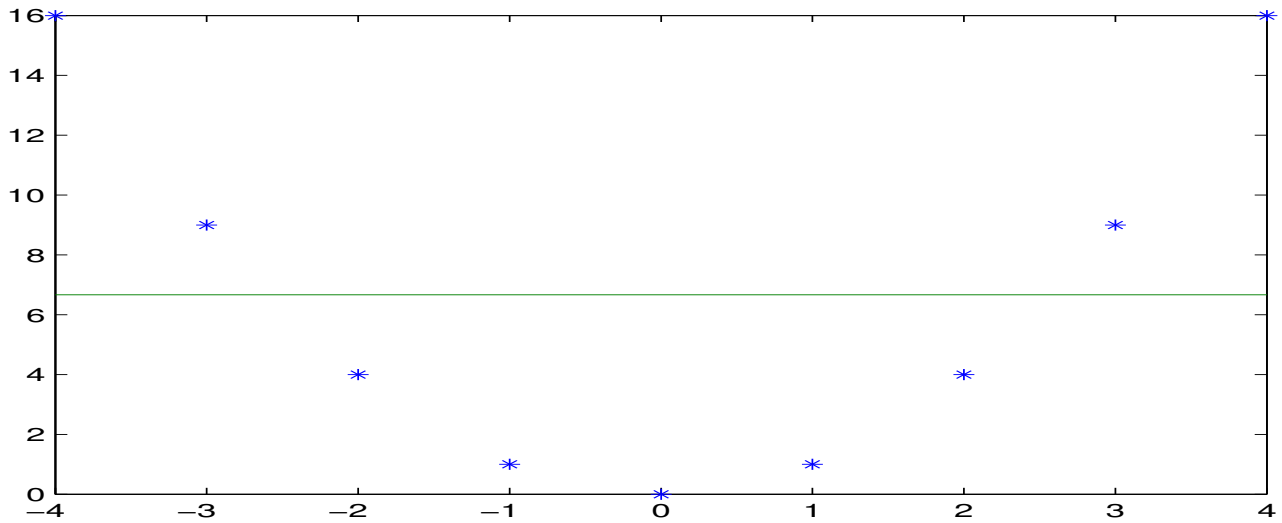
Residual plot

If the residual plot reveals no pattern, the regression equation is a good representation of the association between the two variables. If the residual plot reveals some systematic pattern, the regression equation is not a good representation of the association between the two variables. Here is the residual plot for the data in the example.



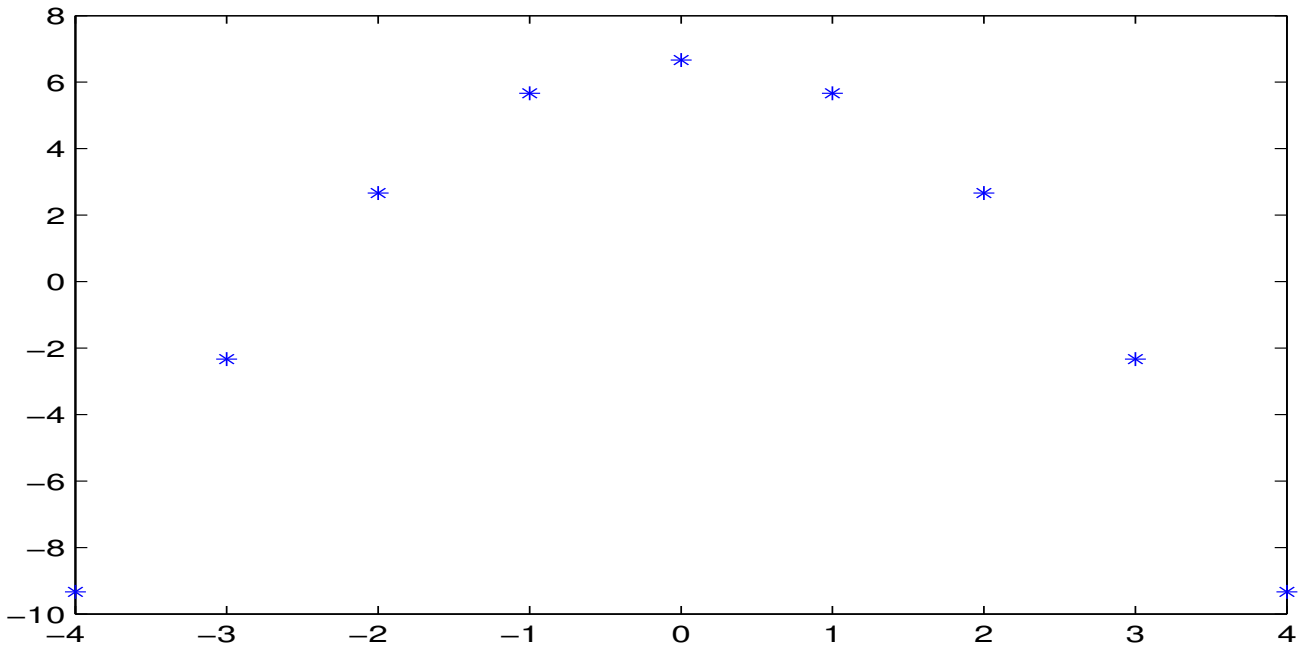
KTH VETENSKAP OCH KONST
KTH Matematik

Regression for $y = x^2$



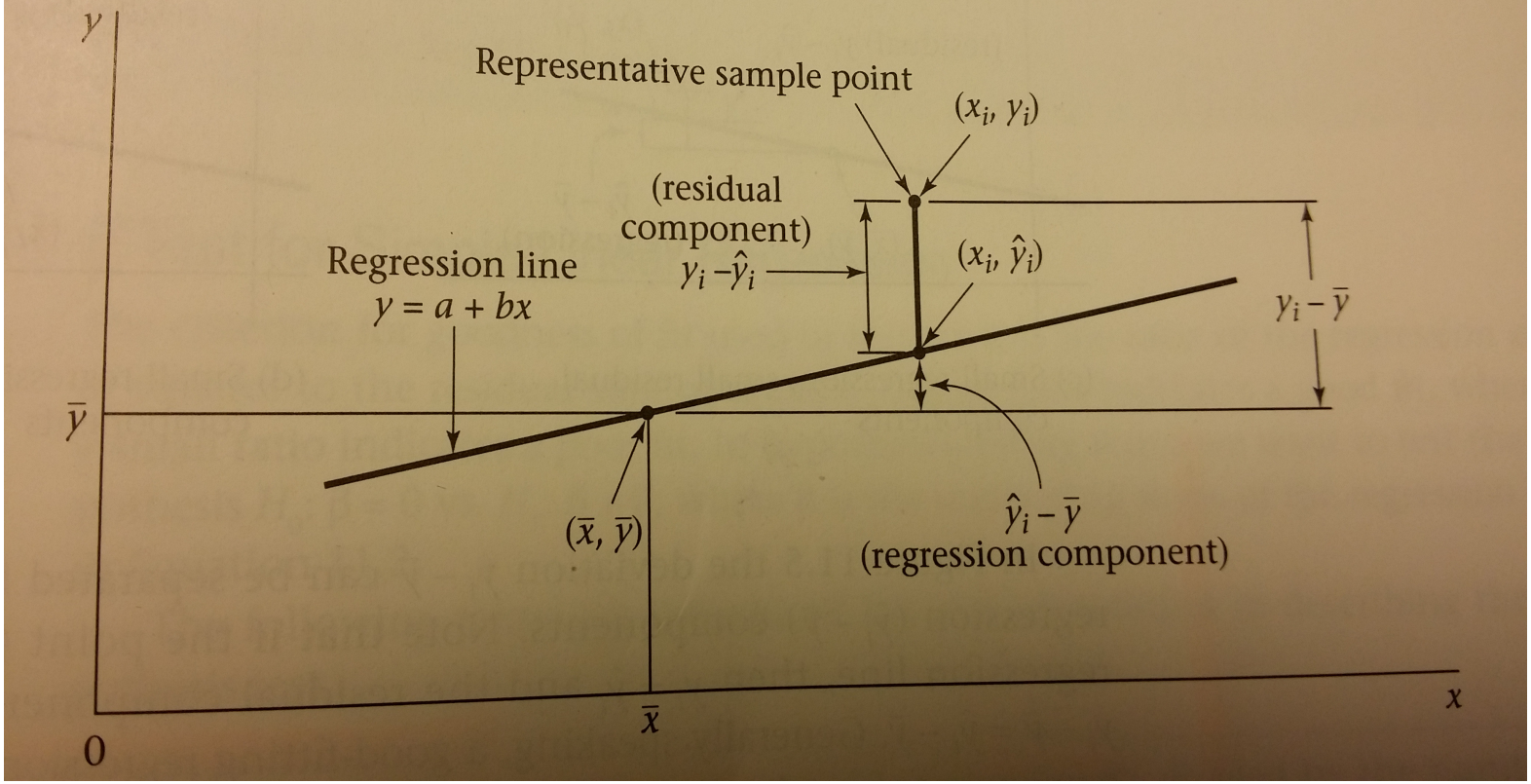
KTH VETENSKAP OCH KONST
KTH Matematik

Regression for $y = x^2$ and the residual plot



regression line in Figure 11.5.

Goodness of fit of a regression line



Goodness of fit & Coefficient of determination

For any (x_i, y_i) the **regression component** of that pair about the regression line is $\widehat{y}_i - \bar{y}$.
A good-fitting regression line will have regression components large in absolute value relative to the residuals.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \textbf{Total Sum of Squares, Total SS}$$

$$\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \leftrightarrow \textbf{Regression Sum of Squares, Reg SS}$$

KTH Matematik

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2,$$

i.e.,

Total SS = Reg SS + Residual Sum of Squares

KTH VETENSKAP OCH KONST
KTH Matematik

For any (x_i, y_i) the **explained deviation** of that pair about the regression line is $\widehat{y}_i - \bar{y}$.
For any (x_i, y_i) the residual or the **unexplained deviation** of that pair about the regression line is $e_i = y_i - \widehat{y}_i$.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \textbf{Total Variation}$$

$$\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \leftrightarrow \quad \textbf{Explained Variation}$$

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \leftrightarrow \quad \textbf{Unexplained Variation}$$

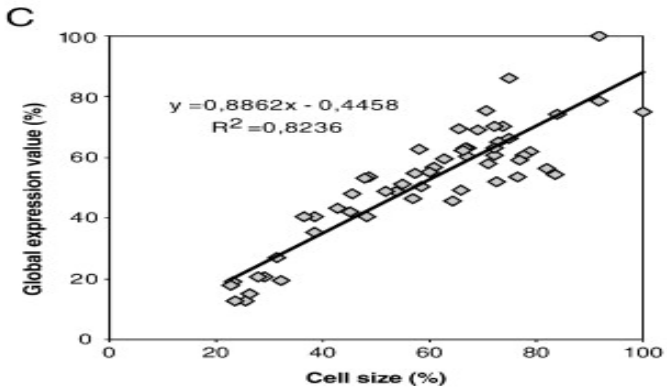
KTH Matematik

The **coefficient of determination** R^2 is the amount of variation in y that is explained by the regression line.

$$R^2 \overset{\text{def}}{=} \frac{\textbf{Explained Variation}}{\textbf{Total Variation}}$$

KTH Matematik

Emma Lundberg et.al.: *The correlation between cellular size and protein expression levels -Normalization for global protein profiling.* **Journal of Proteomics**, 71, 2008, 448−460



Decomposition of the Total Sum of Squares in Regression and Residual Components

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

Summan $\sum_{i=1}^{n} (y_i - \bar{y})^2$ är given av data: den påverkas inte av regressionslinjen. Således: om $\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$ är 'stor', så måste $\sum_{i=1}^{n} e_i^2$ vara liten, och vice versa.

KTH Matematik

Kvantiler, lådagram, utliggare, histogram

KTH Matematik

För en datamängd $\{x_1, \ldots, x_n\}$ är $x_{(k)}$ dess **k:te ordningsstatistika** om

$$x_{(k)} = \text{ k:te minsta värdet i } \{x_1, \ldots, x_n\}$$

Härmed är $x_{(1)}$ det minsta värdet i $\{x_1, \ldots, x_n\}$ och $x_{(n)}$ det största värdet i $\{x_1, \ldots, x_n\}$. Med andra ord, har vi den **ordnade datamängden**

$$\{x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}\}$$

Till exempel, $\{x_1 = 53, x_2 = 41, x_3 = 6.5, x_4 = 65, x_5 = 19\}$ så är ordningsstatistikorna

$$\{x_{(1)} = 6.5, x_{(2)} = 19, x_{(3)} = 41, x_{(4)} = 53, x_{(5)} = 65\}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Kvantiler, percentiler och kvartiler

Vi har ett data set med n värden som i storleksordning är $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

1. För varje tal p som är av formen $p = \frac{i-0.5}{n}$, där i är ett heltal i $\{1, \ldots, n\}$, är p-**kvantilen** lika med $x_{(i)}$.
2. För tal p mellan $\frac{0.5}{n}$ och $\frac{n-0.5}{n}$, som inte är lika med $\frac{i-0.5}{n}$, fås p-**kvantilen** för detta data set genom (linjär) interpolation mellan två tal av formen $\frac{i-0.5}{n}$ sådana att p ligger mellan dem.

I båda fallen betecknas p-**kvantilen** med $Q(p)$. Olika programpaket utnyttjar var sin interpolationsmetod i fall 2..

KTH VETENSKAP OCH KONST
KTH Matematik

En **percentil** är det värde nedanför vilken en viss procent av observationerna i datat ligger. Så är till exempel den tjugonde percentilen P_{20} det värde som delar observationsvärden så att 20 procent av dem är mindre än P_{20} och 80 procent är större.
De percentiler som delar in materialet i fyra delar, P_{25} d.v.s. $Q(0.25)$(även betecknad med Q_1), $P_{0.50}$, d.v.s. $Q(0.50)$(även betecknad med Q_2) och P_{75} d.v.s. $Q(0.75)$(även betecknad som Q_3)), kallas **kvartiler**.

KTH Matematik

Interquartile range, (på sv. kvartilavstånd) (IQR) is defined as

$$IQR \stackrel{\text{def}}{=} Q(0.75) - Q(0.25) = Q_3 - Q_1.$$

and is implemented in every boxplot, see next.

KTH Matematik

KTH Matematik

Lådagram eller låddiagram eller boxplot

är ett diagram där ett statistiskt material åskådliggörs i form av en låda, som anger den mittersta hälften av datamaterialet. Diagrammet sammanfattar datamaterialet med hjälp av fem värden: medianen, $Q(0.25)$ och $Q(0.75)$ samt minimum och maximum. Eventuella extrema värden betraktas som utliggare (outliers) och markeras med egna symboler, t.ex. $+$.

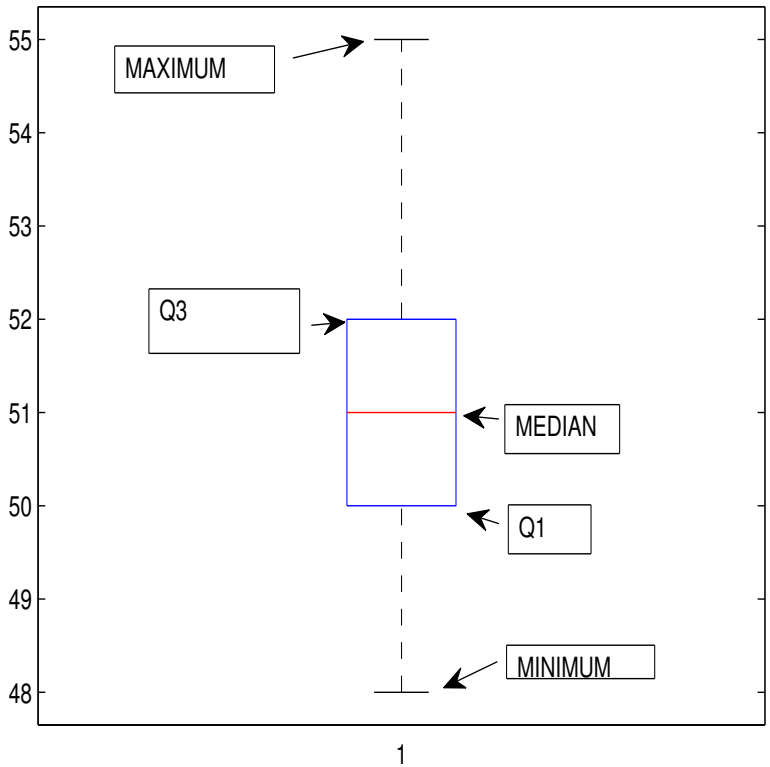
KTH Matematik

Lådan begränsas nedåt av $Q(0.25) = Q_1$ och uppåt av $Q(0.75) = Q_3$. Medianen ritas med ett streck genom lådan. Värden som ligger längre ifrån boxen än 1.5 gånger kvartilavståndet IQR som utliggare och är markerade med $+$. Värden som ligger mer än 3 gånger kvartilavståndet IQR från lådan betraktas som avlägsna utliggare, och betecknas i Matlab med $+$. De strecken som går ut från boxen dras till det lägsta värdet och det högsta bland de värden som inte är utliggare.

KTH Matematik

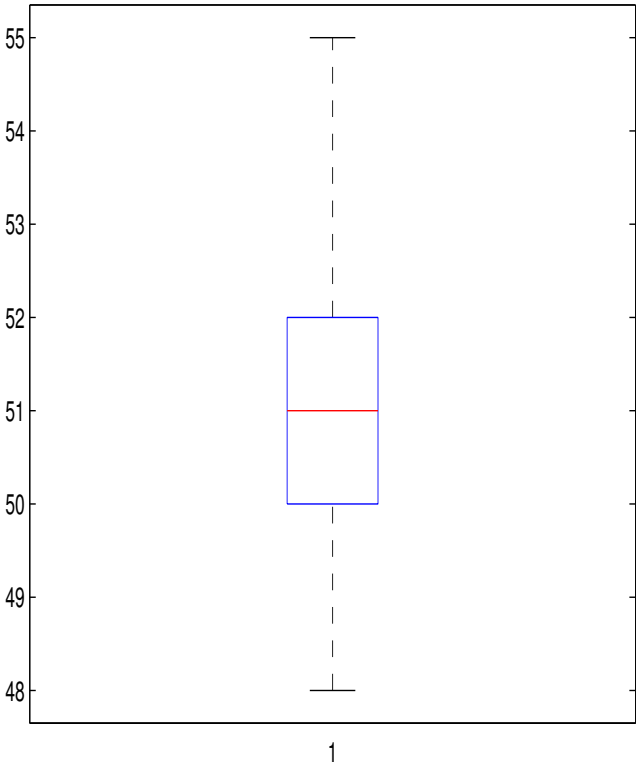
BOXPLOT

first quartile Q_1= separates the bottom 25% of the sorted values, the median = midpoint, 50 % below, the third quartile Q_3 = separates the bottom 75% of the **sorted** *values*



KTH Matematik

Lådagrammet för tändsticksdata: `boxplot(x)` i Matlab



51	52	49	51	52	51	53
52	48	52	50	53	49	50
51	53	51	52	50	51	53
53	55	50	49	53	50	51
51	52	48	53	50	49	51

KTH VETENSKAP OCH KONST
KTH Matematik

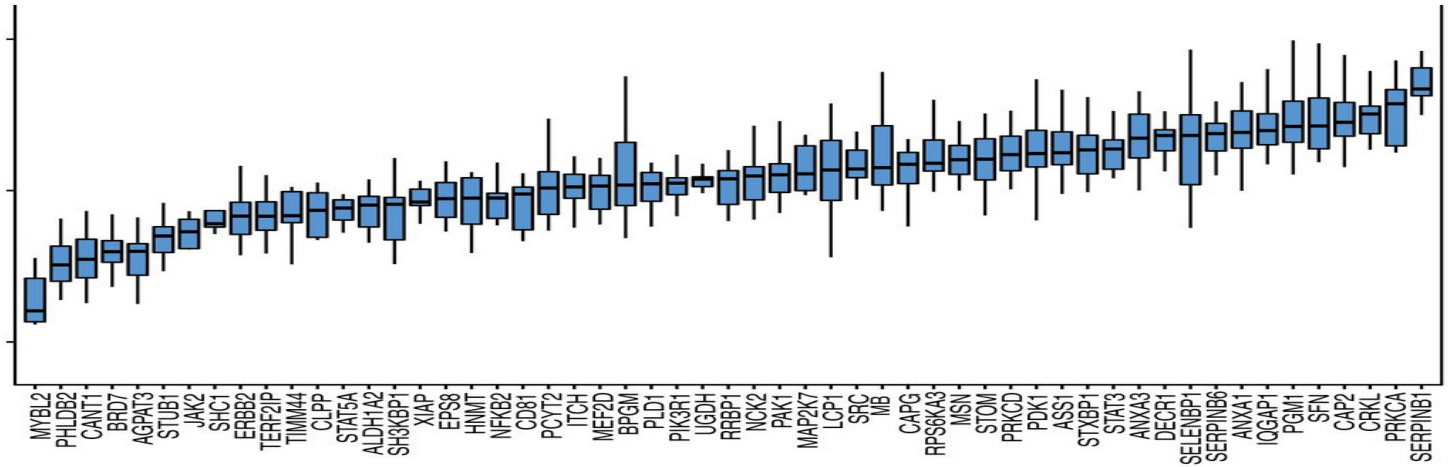
Edfors, Fredrik and Danielsson, Frida and Hallström, Björn M and Käll, Lukas and Lundberg, Emma and Pontén, Fredrik and Forsström, Björn and Uhlén, Mathias: *Gene-specific correlation of RNA and protein levels in human cells and tissues*, **Molecular Systems Biology**, 12, 883, 2016.

An important issue for molecular biology is to establish whether transcript levels of a given gene can be used as proxies for the corresponding protein levels. Here, we have developed a targeted proteomics approach for a set of human nonsecreted proteins based on parallel reaction monitoring to measure, at steadystate conditions, absolute protein copy numbers across human tissues and cell lines and compared these levels with the corresponding mRNA levels using transcriptomics.

KTH Matematik

The study shows that the transcript and protein levels do not correlate well unless a genespecific RNAtoprotein (RTP) conversion factor independent of the tissue type is introduced, thus significantly enhancing the predictability of protein copy numbers from RNA levels. The results show that the RTP ratio varies **significantly** with a few hundred copies per mRNA molecule for some genes to several hundred thousands of protein copies per mRNA molecule for others. In conclusion, our data suggest that transcriptome analysis can be used as a tool to predict the protein copy numbers per cell, thus forming an attractive link between the field of genomics and proteomics.

KTH VETENSKAP OCH KONST
KTH Matematik



KTH Matematik

Boxplottarna ovan: The genespecific RNAtoprotein correlation factors are shown for all the 55 genes with a boxplot showing the average correlation factor for each gene and the variation observed in the nine cell lines and 11 tissues. All the values for each of the cell lines and tissues. Horizontal lines = median. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). Length of the whiskers as multiple of IQR = 1.5.

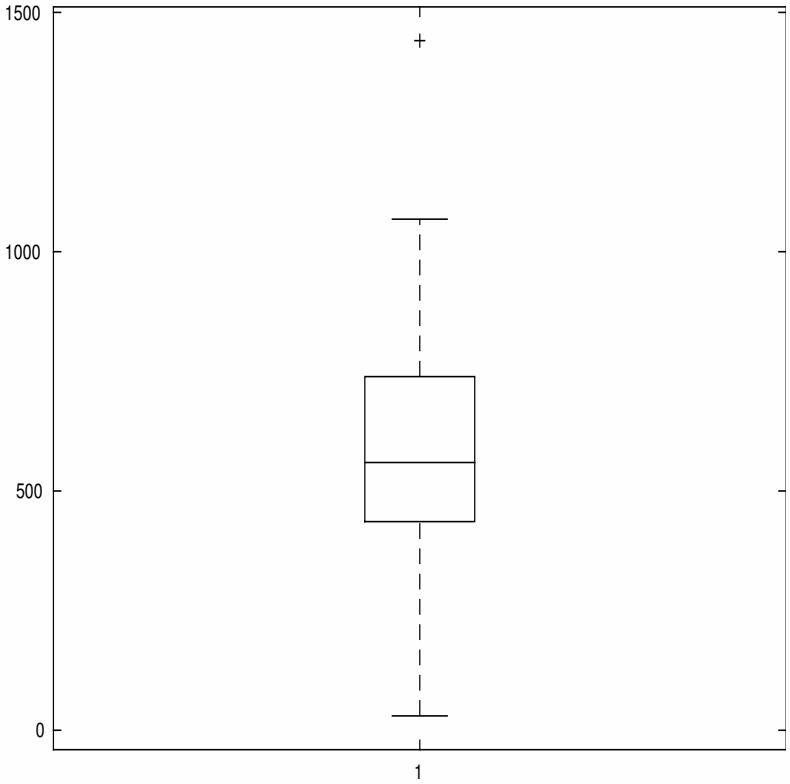
KTH Matematik

An **outlier** (utliggare) is an observation that lies an abnormal distance from other values in a random sample from a population.
To detect abnormality we define

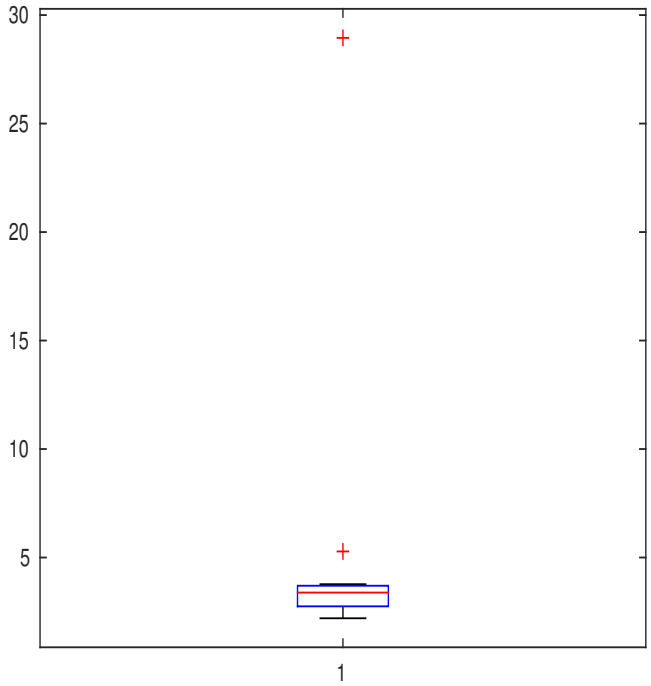
- **lower inner fence**: $\stackrel{\text{def}}{=} Q_1 - 1.5 \cdot IQR$,
- **upper inner fence**: $\stackrel{\text{def}}{=} Q_3 + 1.5 \cdot IQR$,
- **lower outer fence**: $\stackrel{\text{def}}{=} Q_1 - 3 \cdot IQR$,
- **upper outer fence**: $\stackrel{\text{def}}{=} Q_3 + 3 \cdot IQR$.

A data point beyond an inner fence on either side is considered a **mild outlier**. A data point beyond an outer fence is considered an **extreme outlier**.

KTH VETENSKAP OCH KONST
KTH Matematik



KTH Matematik


30
25
20
15
10
5
1

KTH Matematik

Michel Adanson, French botanist, 1727−1806

Adanson's book *Familles des plantes* (1763) described his classification system for plants, which was much opposed by Carolus Linnaeus. Adansons classification was based on all anatomical characters. *Adanson also* **introduced the use of statistical methods in botanical classification**. Adanson's system was superseded by the Linnaean system.

KTH Matematik

Ett **histogram** är en grafisk framställning av frekvenser eller relativa frekvenser för en kvantitativ (kontinuerlig) datamängd. Konstruktion av ett histogram kallas 'data binning" och görs enligt följande (1.-3.):

1. intervallet mellan det största och minsta värdet av datapunkter uppdelas i ett antal intervall (bin), vanligtvis är dessa lika långa och icke-överlappande.
2. räkna antalet datapunkter som ligger i varje intervall.
3. rita en rektangel ovanför varje intervall. Rektangelns höjd är proportionell mot frekvensen, d.v.s. mot antalet datapunkter i intervallet. Rektangeln kan även normaliseras, så att rektangelns höjd är proportionell mot den relativa frekvensen av antalet datapunkter i intervallet.

KTH Matematik

Matlab har automatiserat stegen 1.-3. ovan för konstruktion av histogram:

- **N = hist(X)** bins the elements of X into 10 equally spaced containers and returns the number of elements in each container.
- **hist(X)** bins the elements of X into 10 equally spaced containers and produces a histogram bar plot.
- **hist(Y,M),** where M is a scalar, uses M bins.

Den kvarstående frågan är tydligen att välja M, antalet intervall. Det finns ingen universell lösning. Experimentering behövs vanligtvis.

KTH VETENSKAP OCH KONST
KTH Matematik

Vi bestämmer först $h =$ bredden av intervallet (=längden av bin). Sedan beräknar vi $M =$ antalet bins eller intervall enligt formeln:

$$M = \left\lceil \frac{\max X - \min X}{h} \right\rceil .$$

där $\lceil a \rceil$ avrundar talet a uppåt till närmaste heltalet större än eller lika med a, i Matlab `ceil`. Det återstår att välja h. Här har statistikerna härlett/hittat på många regler.

KTH VETENSKAP OCH KONST
KTH Matematik

Freedman−Diaconis regel för val av bredden h (IQR= interquartile range, kvartilavstånd), se ovan,

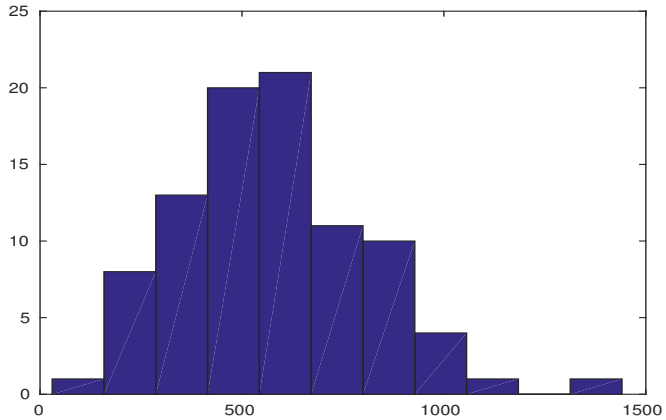
$$h = 2\frac{\mathrm{IQR}(X)}{n^{1/3}}$$

Om funktionen `ceil`

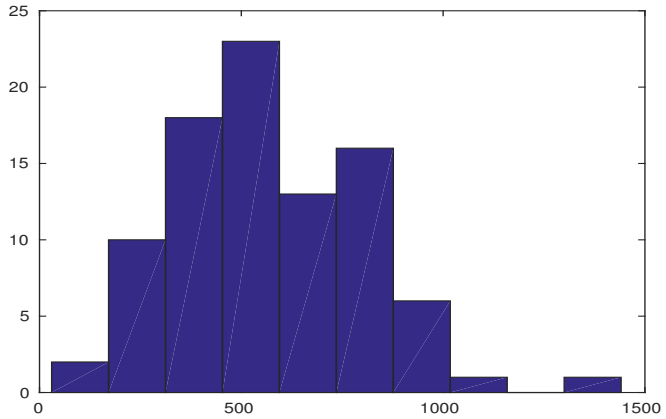
Funktionen $\lceil a \rceil$ avrundar talet a uppåt till **närmaste heltalet** **större än eller lika med** a. Till exempel:

$$\texttt{ceil(1)} = 1, \quad \texttt{ceil(1)}\ (1.5) = 2, \quad \texttt{ceil(7.25)} = 8.$$

Freedman−Diaconis regel ger för en datamängd ($\bar{x} = 576$, median$= 559$, skewness $= 0.54$, trimodal) $M = 11$



Default i Matlab med M=10, ger för samma data



Vilket av dessa två sätt att välja M d.v.s. antalet bin är bättre för detta data och varför?

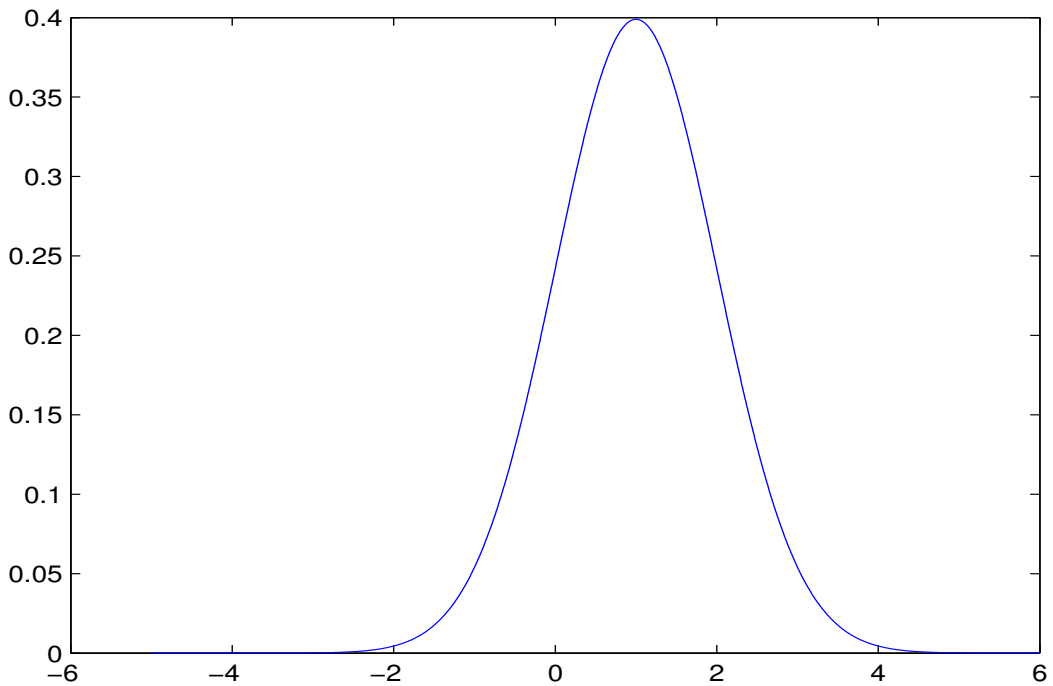
KTH Matematik

Klockkurvan i en sedel

KTH Matematik

Låt $\sigma > 0$ och betrakta (den s.k.) klockkurvan eller normalkurvan

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



$f(x)$ med $\mu = \sigma = 1$ i figuren.

KTH Matematik

Statistisk teori: sannolikhetskurva eller sannolikhetstäthet

Ett **histogram** är även en skattning av en sannolikhetskurva för en kvantitativ (kontinuerlig) datamängd.

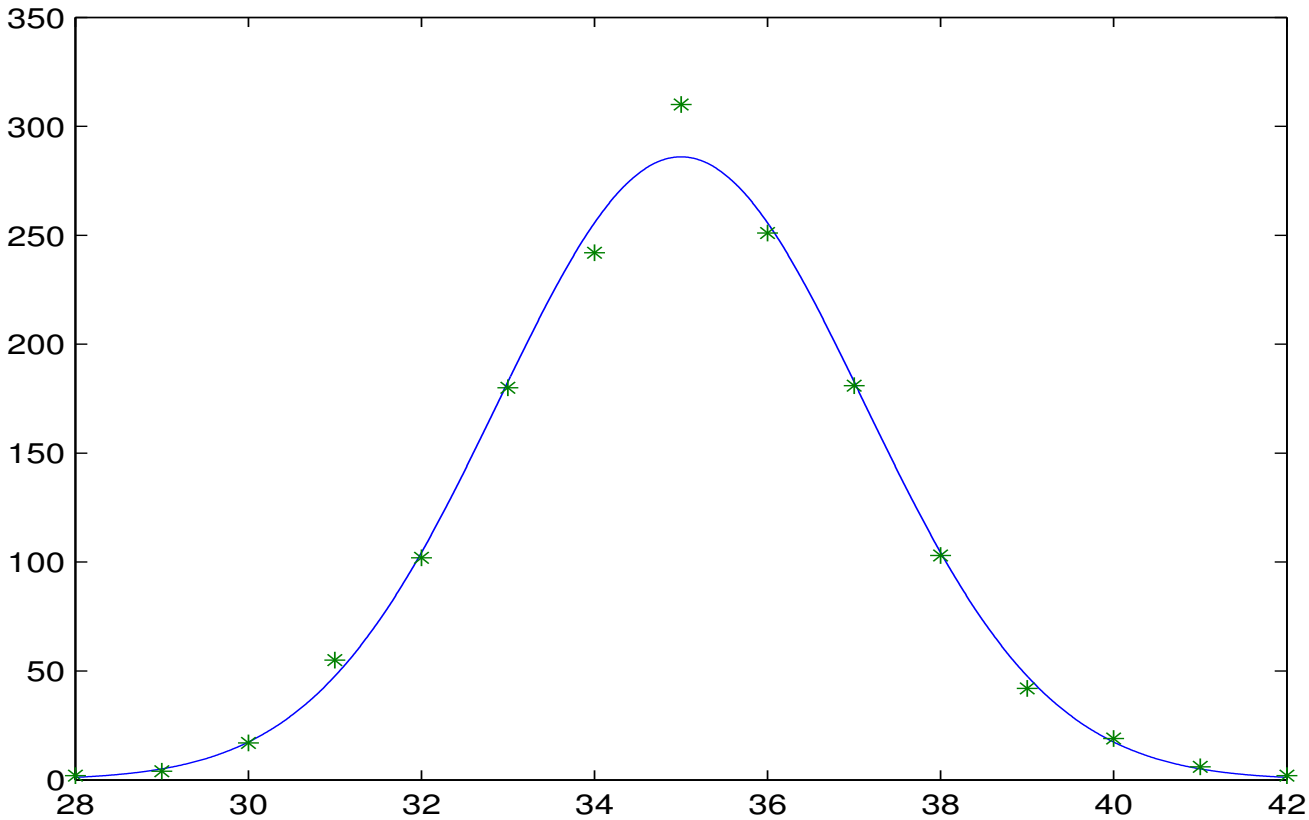
KTH Matematik

Nedan har vi **grupperade** data, en studie av Adam Quetelet från 1800-talet, och ger frekvenserna på omfånget av bröstkorg i cm för 1516 belgiska soldater.

Klass	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Frekvens	2	4	17	55	102	180	242	310	251	181	103	42	19	6	2

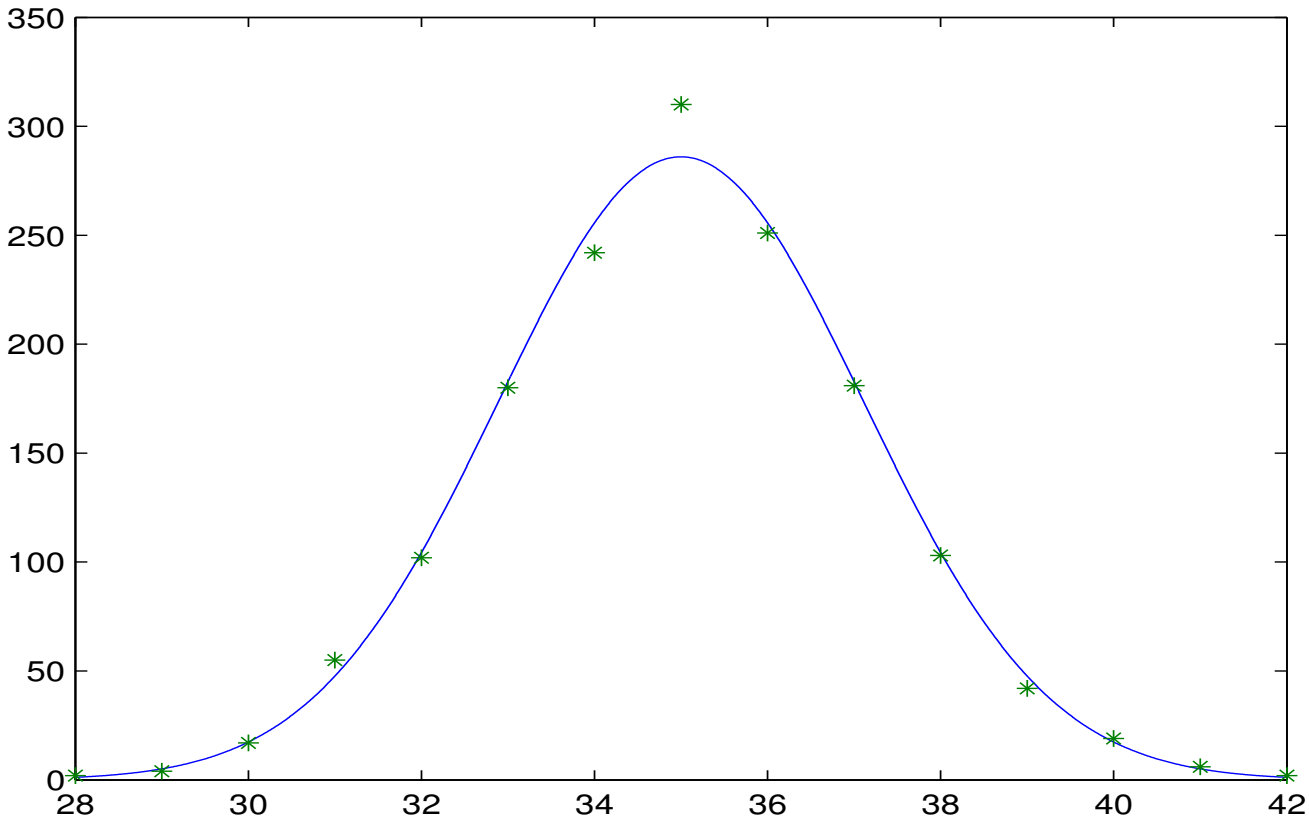
KTH Matematik

Ett stolpdiagram mot en skalad normalkurva/klockurva:



KTH VETENSKAP OCH KONST
KTH Matematik

Klockurvan som **matematisk och statistisk modell**.



KTH VETENSKAP OCH KONST
KTH Matematik