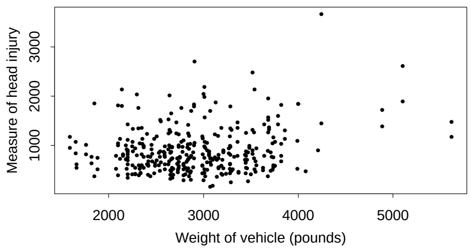
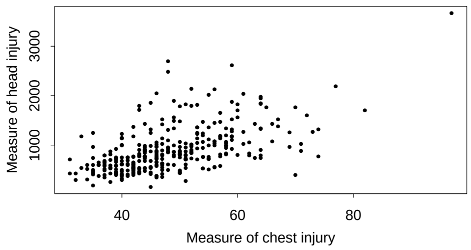
Scatterplots of Quantitative Variables

Relationships between two quantitative variables can be displayed by a *scatterplot*.

Scatterplots of variables from the crashtest data:

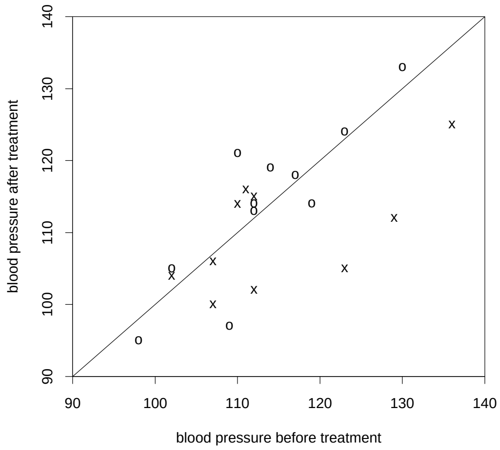




Marked Scatterplots

The relationships among two quantitative variables *and* a categorical variable can be shown by marking points in a scatterplot.

Here is the data on calcium and blood pressure, marked by who took the calcium (x) and who took the placebo (o):



Correlation of Two Variables

Some variables are *positively correlated*. When one is high, the other tends to be high too.

Example: height and weight of people.

Other variables are *negatively correlated*. When one is high, the other tends to be low.

Example: duration of lecture and fraction of class awake at end.

And some variables are *uncorrelated*. Whether one is high or low is not consistently related to with whether the other is high or low.

Example: age and size of nose (in adults).

Numerical Measure of Correlation

The *sample covariance* between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is defined to be

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

where $\bar{x}$ and $\bar{y}$ are the sample means of the x's and the y's.

The *sample correlation* between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ is

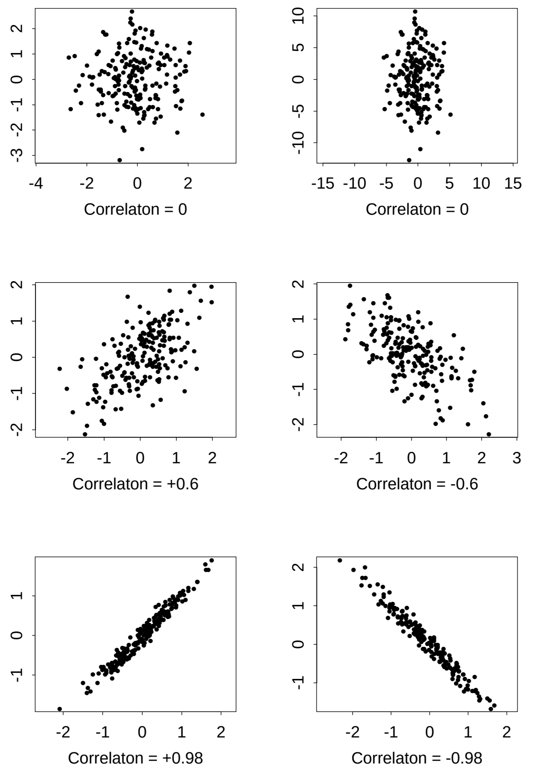
$$r \;=\; \frac{s_{xy}}{s_x\, s_y} \;=\; \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

where s_x and s_y are the sample standard deviations for the x's and y's.

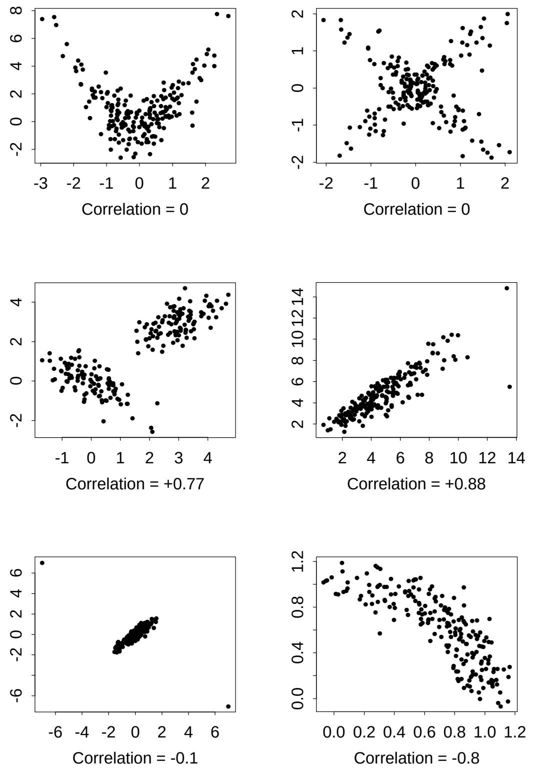
The correlation ranges from -1 to $+1$.

One can also talk about the correlation in the population (often denoted by ρ).

Correlation in Scatterplots (Normal)



Correlation in Scatterplots (Non-Normal)

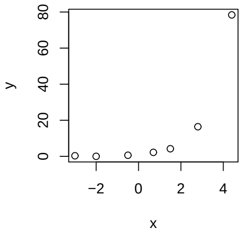


Spearman's Rank Correlation

Like the mean and standard deviation, correlation is not resistant to outliers. Ordinary ("Pearson's") correlation also measures only *linear* relationships.

Spearman's rank correlation can be used when these could be problems. It's the ordinary correlation between the *ranks* of the data:

x	rank(x)	y	rank(y)
2.8	6	16.4	6
1.5	5	4.2	5
-2.0	2	0.0	1
4.4	7	78.4	7
-0.5	3	0.6	3
-3.0	1	0.3	2
0.7	4	2.2	4




```
Pearson's correlation  = 0.756
Spearman's correlation = 0.964
```