Introduction to bivariate analysis

- When one measurement is made on each observation, univariate analysis is applied.

  If more than one measurement is made on each observation, multivariate analysis is applied.

  In this section, we focus on bivariate analysis, where exactly two measurements are made on each observation.

  The two measurements will be called X and Y. Since X and Y are obtained for each observation, the data for one observation is the pair (X, Y).

- Bivariate data can be stored in a table with two columns:

	X	Y
Obs. 1	2	1
Obs. 2	4	4
Obs. 3	3	1
Obs. 4	7	5
Obs. 5	5	6
Obs. 6	2	1
Obs. 7	4	4
Obs. 8	3	1
Obs. 9	7	5
Obs. 10	5	6

- Some examples:
  - Height (X) and weight (Y) are measured for each individual in a sample.
  - Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
  - A cell culture is treated with varying concentrations of a drug, and the growth rate (X) and drug concentration (Y) are recorded for each trial.
  - Temperature (X) and precipitation (Y) are measured on a given day at a set of weather stations.

- Be clear about the difference between bivariate data and two sample data. In two sample data, the X and Y values are not paired, and there aren't necessarily the same number of X and Y values.

  Two-sample data:

  Sample 1: 3,2,5,1,3,4,2,3
  Sample 2: 4,4,3,6,5

- A bivariate simple random sample (SRS) can be written

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

Each observation is a pair of values, for example (X_3, Y_3) is the third observation.

In a bivariate SRS, the observations are independent of each other, but the two measurements within an observation may not be (taller individuals tend to be heavier, profitable companies tend to have higher stock market valuations, etc.).

- The distribution of X and the distribution of Y can be considered individually using univariate methods. That is, we can analyze

$$X_1, X_2, \ldots, X_n$$

or

$$Y_1, Y_2, \ldots, Y_n$$

using CDF's, densities, quantile functions, etc. Any property that described the behavior of the X_i values alone or the Y_i values alone is called marginal property.

For example the ECDF $\hat{F}_X(t)$ of X, the quantile function $\hat{Q}_Y(p)$ of Y, the sample standard deviation of $\hat{\sigma}_Y$ of Y, and the sample mean $\bar{X}$ of X are all marginal properties.

- The most interesting questions relating to bivariate data deal with X and Y simultaneously.

  These questions are investigated using properties that describe X and Y simultaneously.

  Such properties are called joint properties. For example the mean of $X - Y$, the IQR of X/Y, and the average of all X_i such that the corresponding Y_i is negative are all joint properties.

- A complete summary of the statistical properties of (X, Y) is given by the joint distribution.

- If the sample space is finite, the joint distribution is represented in a table, where the X sample space corresponds to the rows, and the Y sample space corresponds to the columns. For example, if we flip two coins, the joint distribution is

	H	T
H	1/4	1/4
T	1/4	1/4.

The marginal distributions can always be obtained from the joint distribution by summing the rows (to get the marginal X distribution), or by summing the columns (to get the marginal Y distribution). For this example, the marginal X and Y distributions are both $\{H \to 1/2, T \to 1/2\}$.

- For another example, suppose we flip a fair coin three times, let X be the number of heads in the first and second flips, and let Y be the number of heads in the second and third flips. These are the possible outcomes:

HHH HTH HTT TTH
HHT THH THT TTT.

The joint distribution is:

	0	1	2
0	1/8	1/8	0
1	1/8	1/4	1/8
2	0	1/8	1/8

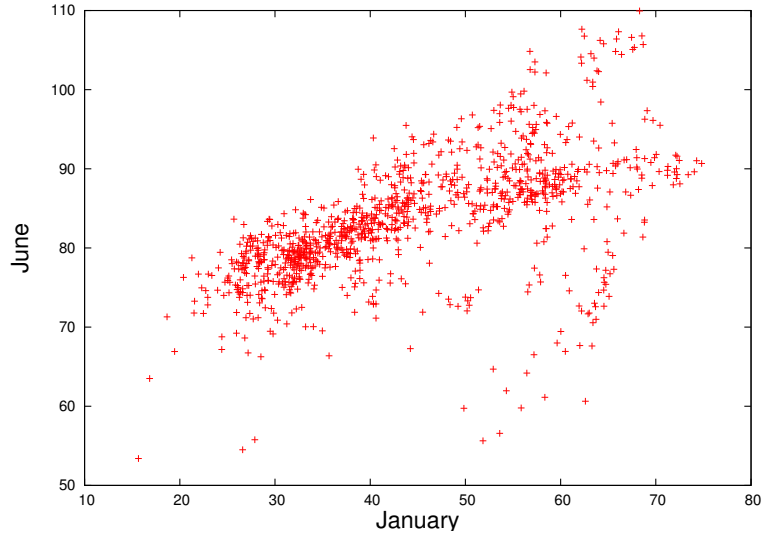
The marginal X and Y distributions are both $\{0 \to 1/4, 1 \to 1/2, 2 \to 1/4\}$.

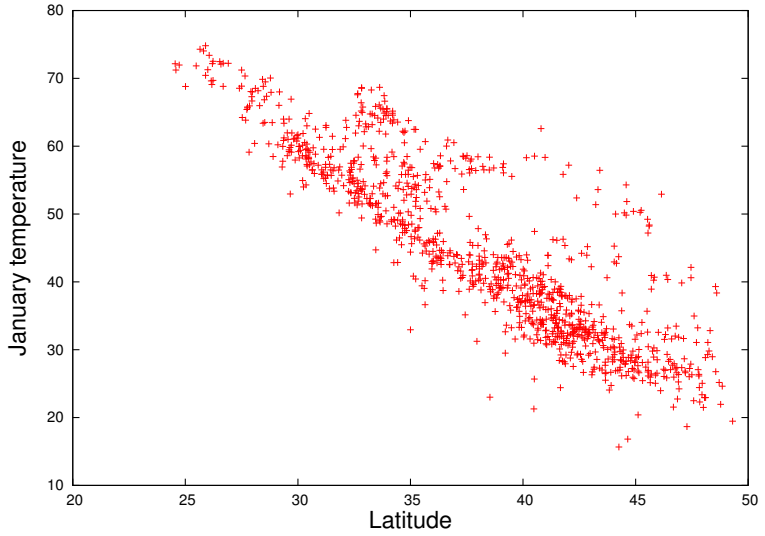
- An important fact is that two different joint distributions can have the same X and Y marginal distributions. In other words, the joint distribution is not determined completely by the marginal distributions, so information is lost if we summarize a bivariate distribution using only the two marginal distributions. The following two joint distributions have the same marginal distributions:

	0	1
0	2/5	1/5
1	1/10	3/10

	0	1
0	3/10	3/10
1	1/5	1/5

- The most important graphical summary of bivariate data is the scatterplot. This is simply a plot of the points (X_i, Y_i) in the plane. The following figures show scatterplots of June maximum temperatures against January maximum temperatures, and of January maximum temperatures against latitude.





- A key feature in a scatterplot is the association, or trend between X and Y.

  Higher January temperatures tend to be paired with higher June temperatures, so these two values have a positive association.

  Higher latitudes tend to be paired with lower January temperature decreases, so these values have a negative association.

  If higher X values are paired with low or with high Y values equally often, there is no association.

- Do not draw causal implications from statements about associations, unless your data come from a randomized experiment.

  Just because January and June temperatures increase together does not mean that January temperatures cause June temperatures to increase (or vice versa).

  The only certain way to sort out causality is to move beyond statistical analysis and talk about mechanisms.

- In general, if X and Y have an association, then

  (i) X could cause Y to change

  (ii) Y could cause X to change

  (iii) a third unmeasured (perhaps unknown) variable Z could cause both X and Y to change.

  Unless your data come from a randomized experiment, statistical analysis alone is not capable of answering questions about causality.

- For the association between January and July temperatures, we can try to propose some simple mechanisms:

  Possible mechanism for (i): warmer or cooler air masses in January persist in the atmosphere until July, causing similar effects on the July temperature.

  Possible mechanism for (ii): None, it is impossible for one event to cause another event that preceded it in time.

  Possible mechanism (iii): If Z is latitude, then latitude influences temperature because it determines the amount of atmosphere that solar energy must traverse to reach a particular point on the Earth's surface.

  Case (iii) is the correct one.

- Suppose we would like to numerically quantify the trend in a bivariate scatterplot.

  The most common means of doing this is the correlation coefficient (sometimes called *Pearson's correlation coefficient*):

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

  The numerator

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(n-1)$$

  is called the covariance.

- The correlation coefficient r is a function of the data, so it really should be called the sample correlation coefficient.

  The (sample) correlation coefficient r estimates the population correlation coefficient ρ.

- If either the X_i or the Y_i values are constant (i.e. all have the same value), then one of the sample standard deviations is zero, and therefore the correlation coefficient is not defined.

- Both the sample and population correlation coefficients always fall between -1 and 1.

  If $r = 1$ then the X_i, Y_i pairs fall exactly on a line with positive slope.

  If $r = -1$ then the X_i, Y_i pairs fall exactly on a line with negative slope.

  If r is strictly between -1 and 1, then the X_i, Y_i points do not fall exactly on any line.

- Consider one term in the correlation coefficient:

$$(X_i - \bar{X})(Y_i - \bar{Y}).$$

If X_i and Y_i both fall on the same side of their respective means,

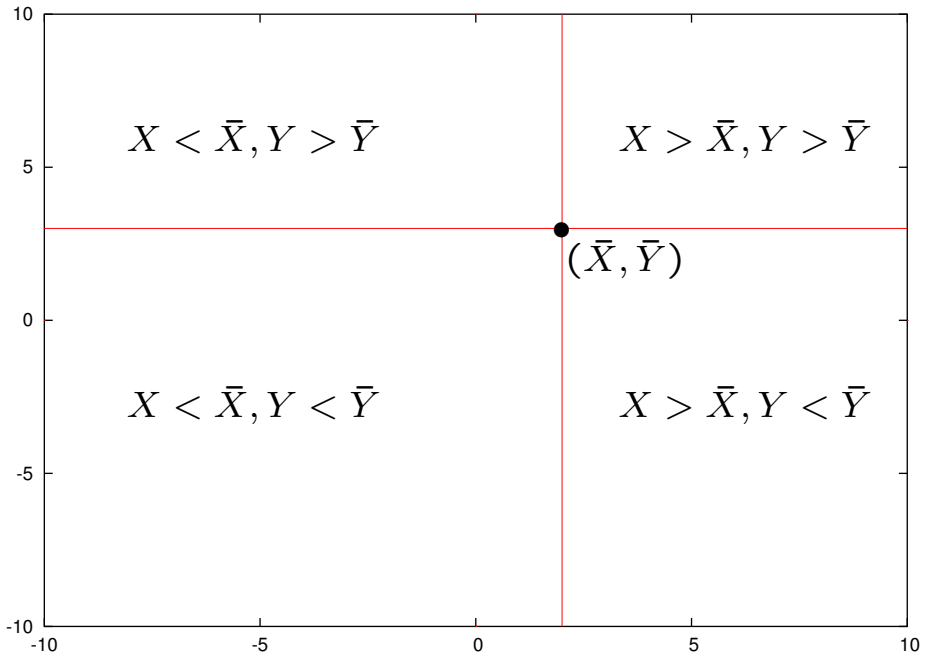
$$X_i > \bar{X} \text{ and } Y_i > \bar{Y} \qquad \text{or} \qquad X_i < \bar{X} \text{ and } Y_i < \bar{Y}$$

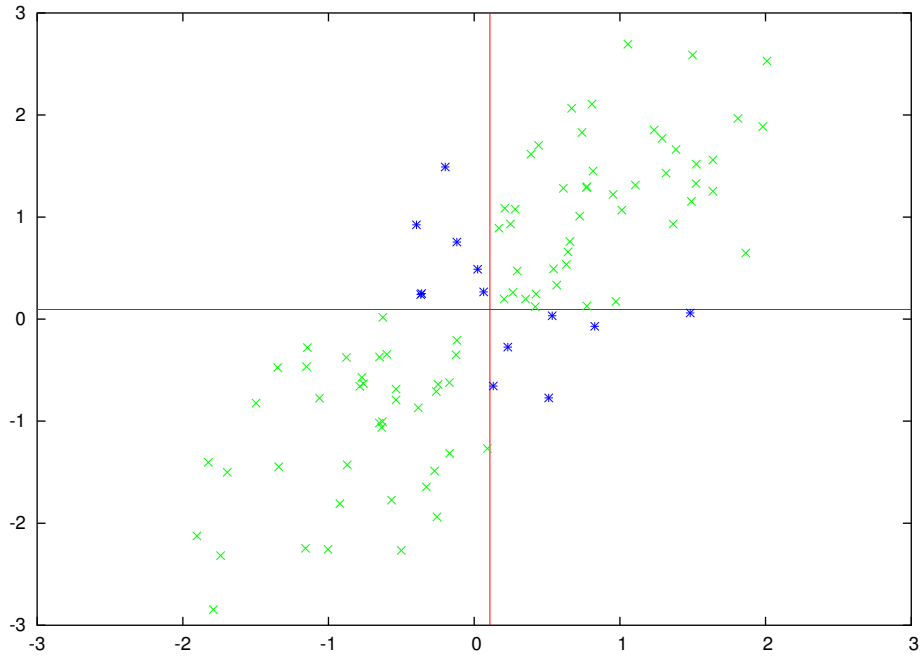
then this term is positive. If X_i and Y_i fall on opposite sides of their respective means,

$$X_i > \bar{X} \text{ and } Y_i < \bar{Y} \qquad \text{or} \qquad X_i < \bar{X} \text{ and } Y_i > \bar{Y}$$

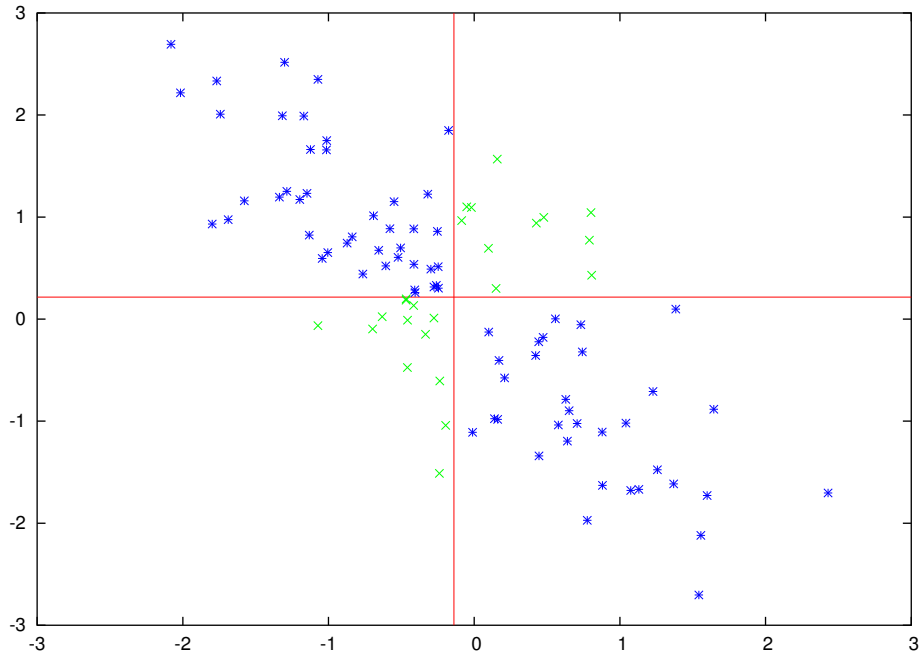
then this term is negative.

So $r > 0$ if X_i and Y_i tend to fall on the same side of their means together. If they tend to fall on opposite sides of their means, then r is negative.


10
5
0
-5
-10
-10
-5
0
5
10
$X < \bar{X}, Y > \bar{Y}$
$X > \bar{X}, Y > \bar{Y}$
$(\bar{X}, \bar{Y})$
$X < \bar{X}, Y < \bar{Y}$
$X > \bar{X}, Y < \bar{Y}$



The green points contribute positively to r, the blue points contribute negatively to r. In this case the result will be $r > 0$.



The green points contribute positively to r, the blue points contribute negatively to r. In this case the result will be $r < 0$.

- Summary of the interpretation of the correlation coefficient:
  - Positive values of r indicate a positive linear association (i.e. large X_i and large Y_i values tend to occur together, small X_i and small Y_i values tend to occur together).
  - Negative values of r indicate a negative linear association (i.e. large X_i values tend to occur with small Y_i values, small X_i values tend to occur with large X_i values).
  - Values of r close to zero indicate no linear association (i.e. large X_i are equally likely to occur with large or small Y_i values).

- Suppose we calculate $\bar{X}$ for $X_1, X_2, \ldots, X_n$. Construct two new data sets:

$$Y_i = X_i + b$$

$$Z_i = cX_i$$

Then $\bar{Y} = \bar{X} + b$ and $\bar{Z} = c\bar{X}$.

The Y_i are a translation of the X_i.

The Z_i are a scaling of the X_i.

It follows that $Y_i - \bar{Y} = X_i - \bar{X}$ and $Z_i - \bar{Z} = c(X_i - \bar{X})$.

- From the previous slide, if we are calculating the sample covariance

$$C = \sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(n-1),$$

it follows that if we translate the X_i or the Y_i, C does not change.

If we scale the X_i by a and the Y_i by b, then C is changed to abC.

- Suppose we calculate $\hat{\sigma}_X$ for $X_1, X_2, \ldots, X_n$. Construct two new data sets:

$$Y_i = X_i + b$$

$$Z_i = cX_i$$

Then $\hat{\sigma}_Y = \hat{\sigma}_X$ and $\hat{\sigma}_Z = |c|\hat{\sigma}_X$.

- From the previous two slides, it follows that the sample correlation coefficient is not affected by translation.

  If we scale the X_i by a and the Y_i by b, then the sample covariance gets scaled by $|ab|$, $\widehat{\sigma}_X$ is scaled by $|a|$, and $\widehat{\sigma}_Y$ is scaled by $|b|$.

  This correlation r is scaled by $ab/|ab|$, which is the sign of ab: $\text{sgn}(ab)$.

- Four key properties of covariance and correlation are:

$$\text{cor}(X, X) = 1$$

$$\text{cov}(X, X) = \text{var}(X)$$

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)$$

- *More on the paired two sample test.*

  If paired data X_i, Y_i are observed and we are interested in testing whether the X mean and the Y mean differ, the paired and unpaired test statistics are

$$\sqrt{n}\frac{\bar{Y} - \bar{X}}{\hat{\sigma}_D} \qquad \text{and} \qquad \frac{\bar{Y} - \bar{X}}{\hat{\sigma}_{XY}}.$$

  Using the properties given above,

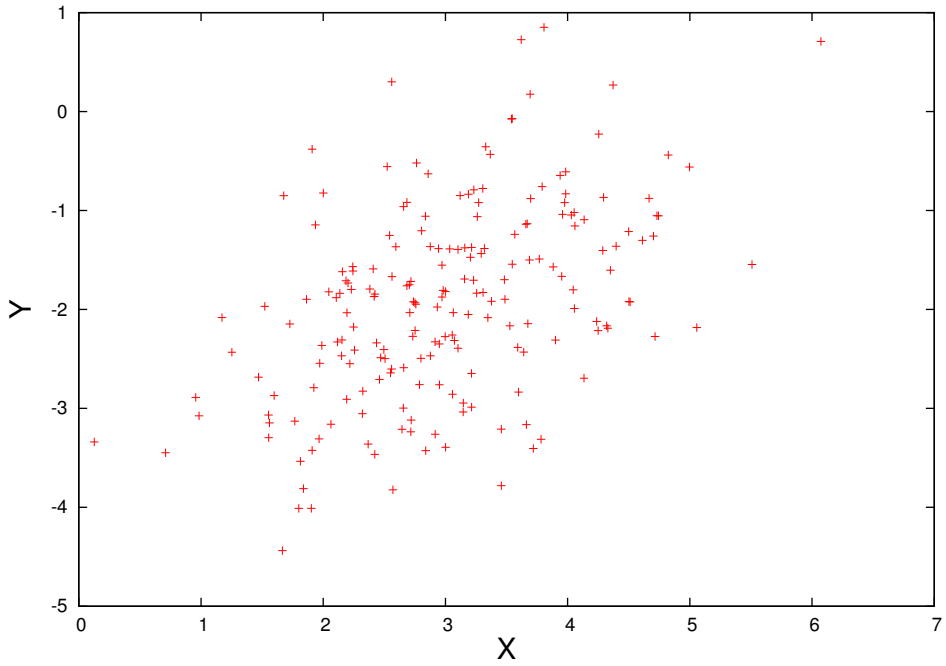
$$\text{var}(D) = \text{cov}(X - Y, X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)$$

  If $\text{cov}(X, Y) > 0$ then $\hat{\sigma}_D < \hat{\sigma}_{XY}$, so the paired test statistic will be larger and hence more significant.

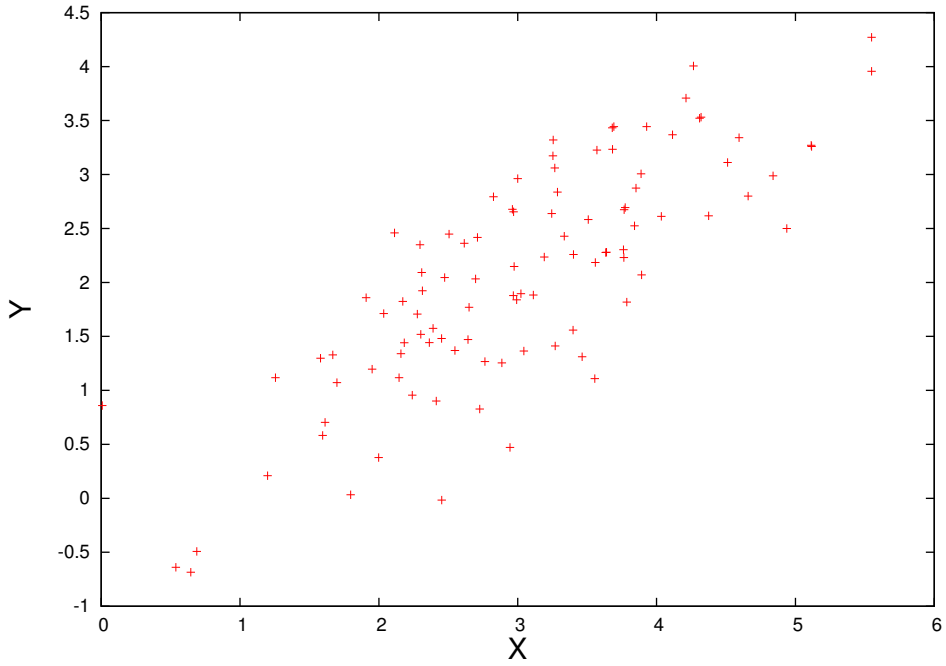
  If $\text{cov}(X, Y) < 0$ then $\hat{\sigma}_D > \hat{\sigma}_{XY}$, so the paired test statistic will be less significant.

- In the paired two sample test, the covariance will be generally be positive, so the paired test statistic gives a more favorable result.

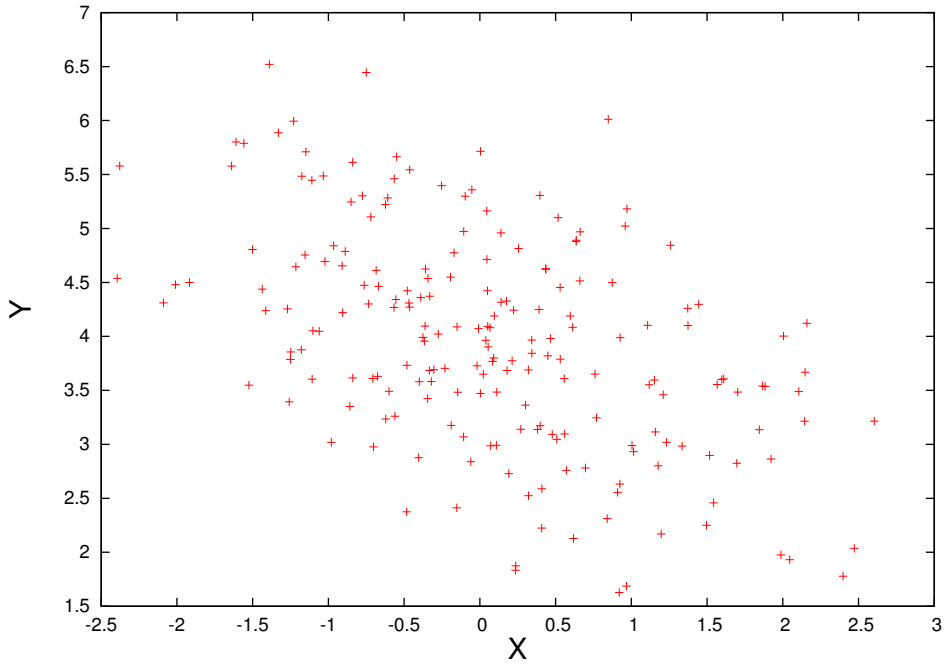
  For example, consider the typical "before and after" situation. Suppose a certain cancer drug kills 30% of the cells in every patient's tumor. Patients with larger than average tumors before treatment will still have larger than average tumors after treatment.



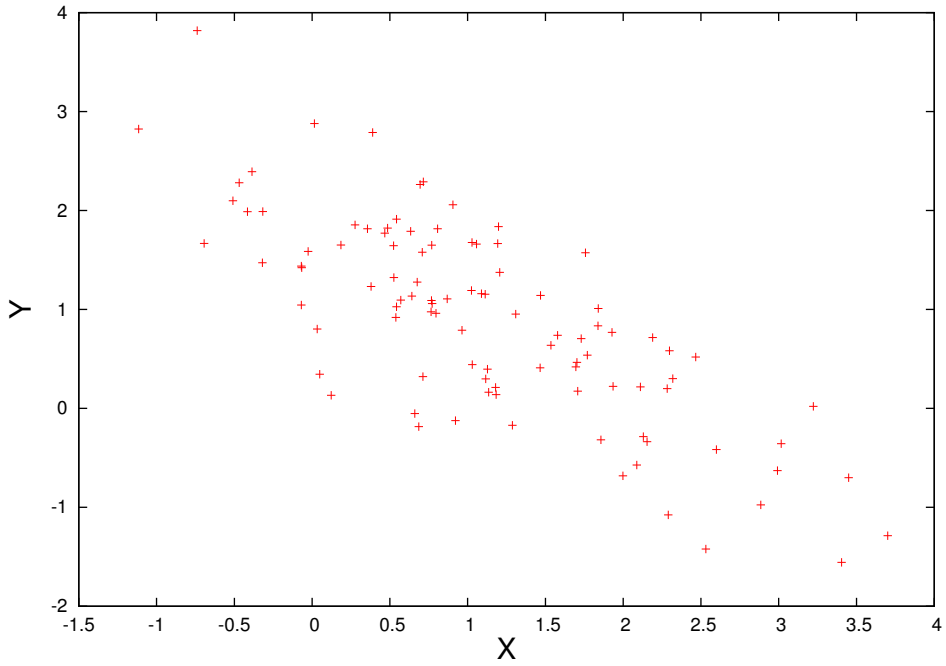
Scatterplot of 200 points with $\rho = .5$



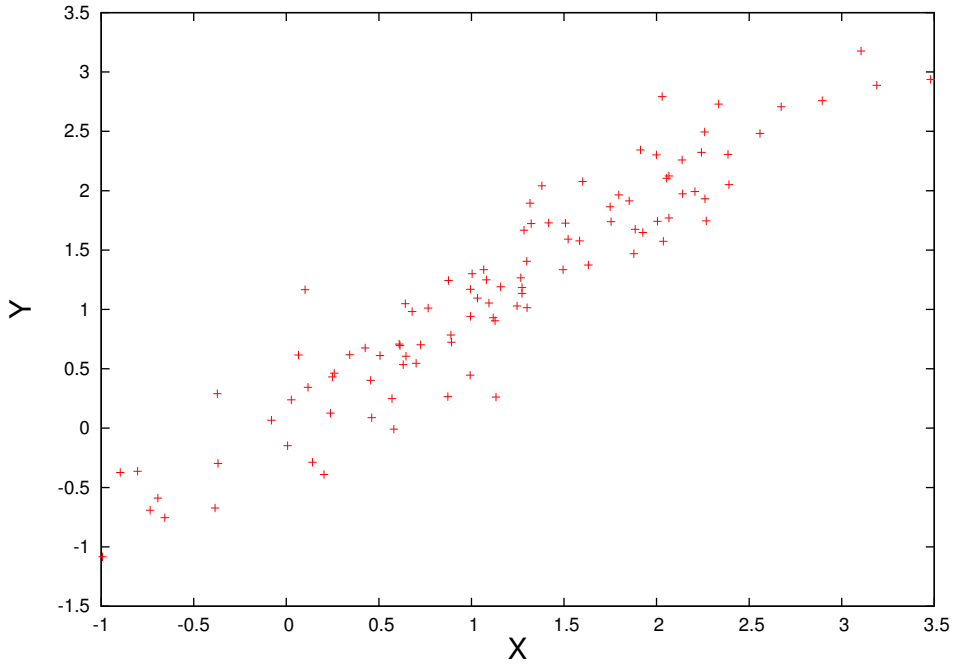
Scatterplot of 100 points with $\rho = .8$



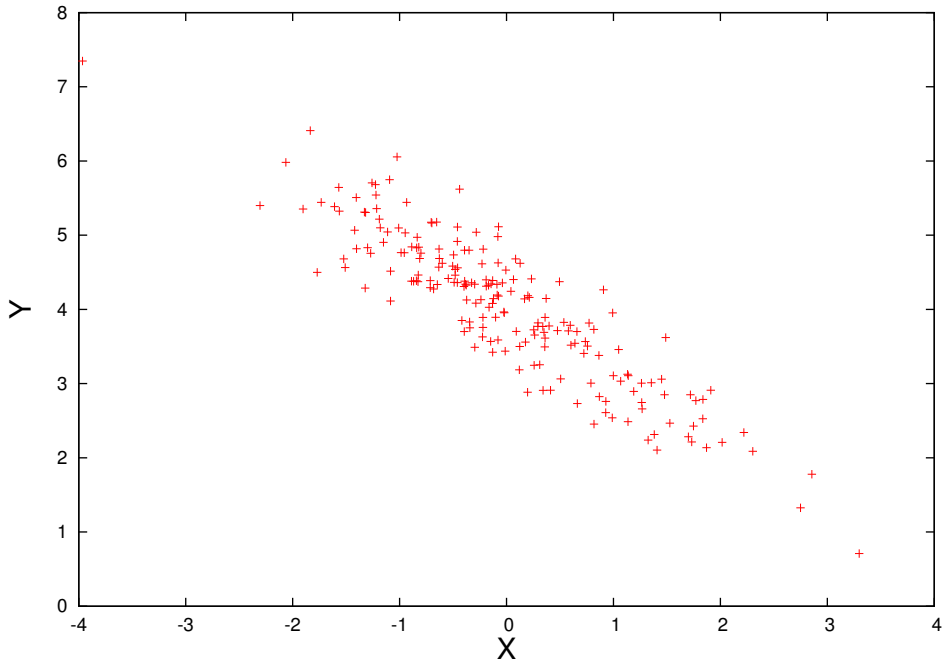
Scatterplot of 200 points with $\rho = -.5$



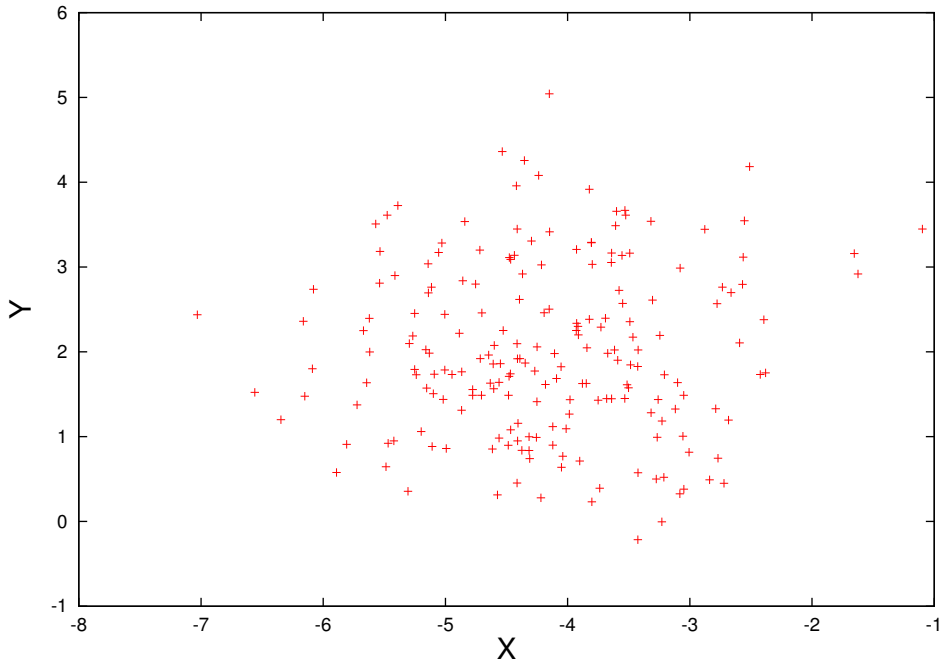
Scatterplot of 100 points with $\rho = -.8$



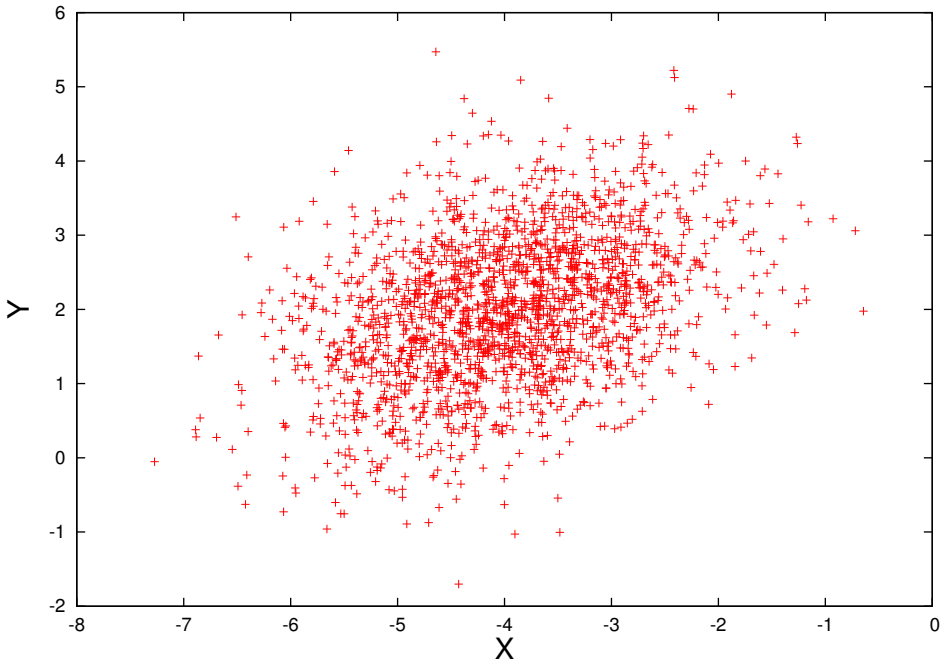
Scatterplot of 100 points with $\rho = .95$



Scatterplot of 200 points with $\rho = -.9$

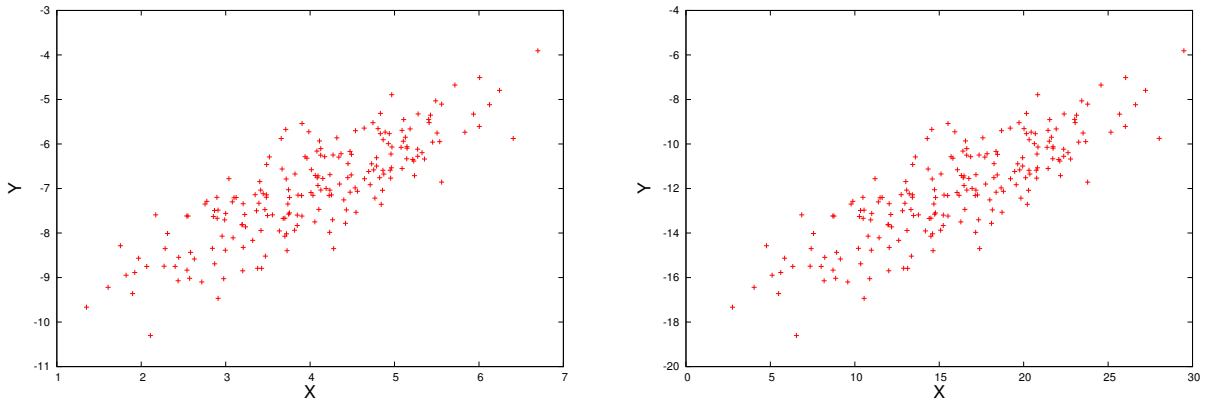


Scatterplot of 100 points with $\rho = 0$



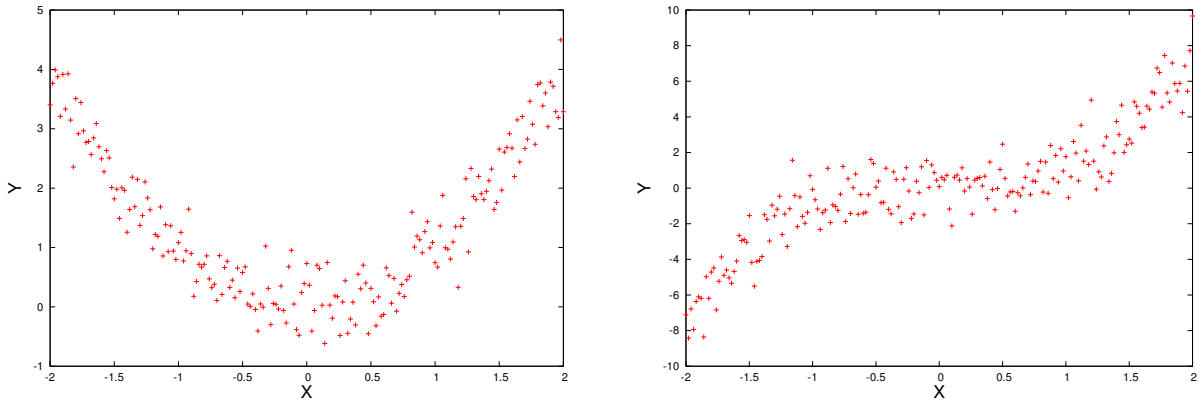
Scatterplot of 2000 points with $\rho = .4$

- Correlation has nothing to do with the marginal mean or variance of X or Y. The following two scatterplots show identical correlations with very different means and variances.



Two sets of values with $\rho = .8$,
but different marginal means and variances

- Correlation detects only the linear trend between X and Y. The left plot below is quadratic and the correlation is nearly 0. However a nonlinear trend that is approximately linear, like the cubic relationship on the right (which is approximately linear between -2 and 2) still shows sizable correlation (around .85).

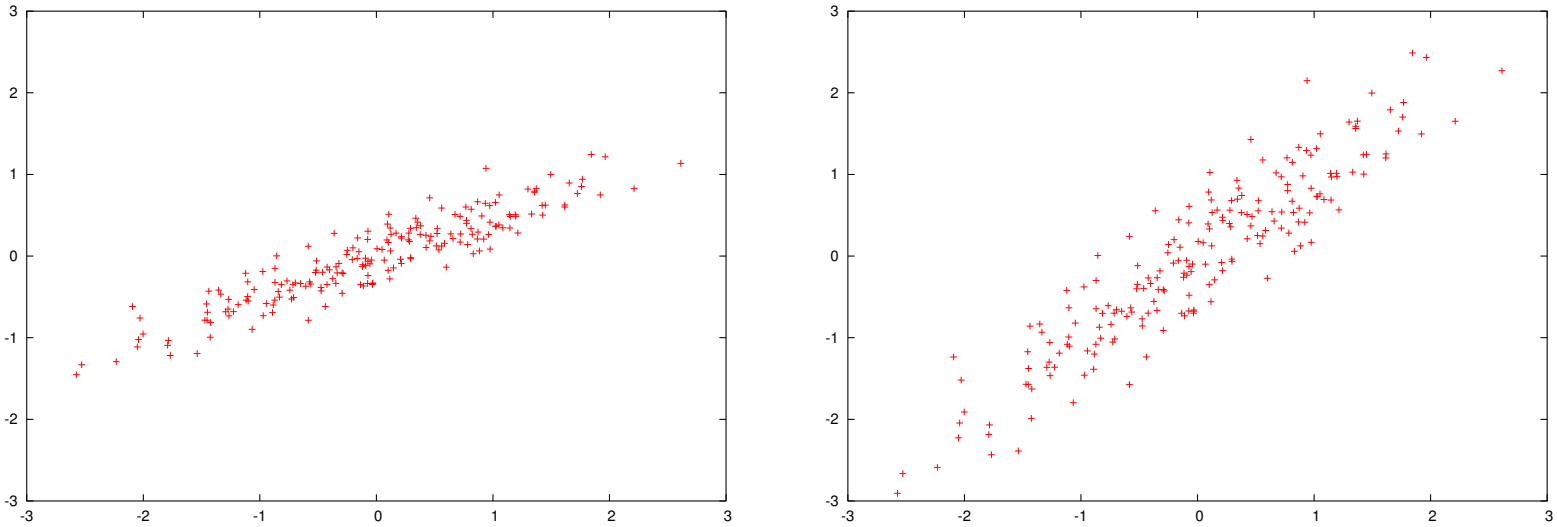


Two nonlinear associations.

- Correlation is not the same as slope.

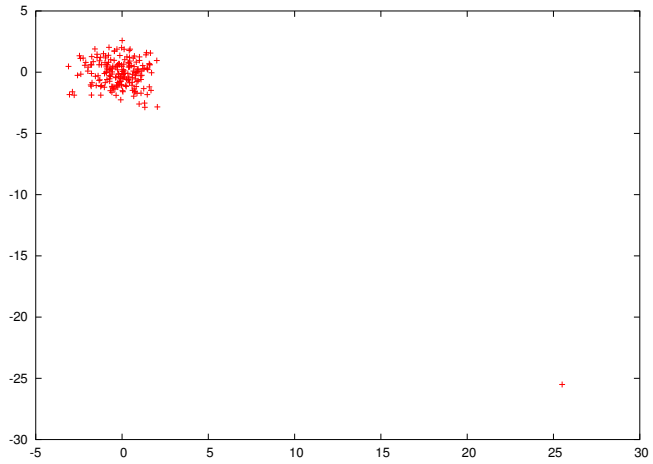
  Correlation measures the strength of a linear trend and is not affected by scaling.

  Scaling each Y_i by c scales the slope of Y on X by a factor of c. Scaling each X_i by c scales the slope by a factor of $1/c$.



Two bivariate data sets with the equal correlations but different slopes.

- A single outlying observation can have a substantial effect on the correlation coefficient. The following scatterplot shows a bivariate data set in which a single point produces a correlation of around $-.75$. The correlation would be around .01 if the point were removed.



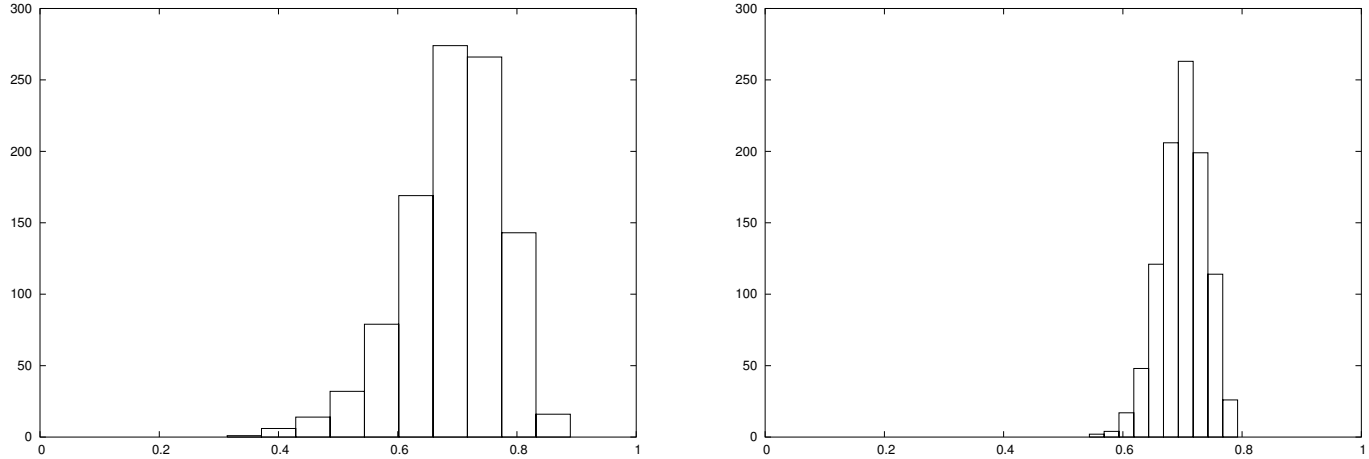
A correlation of $\approx -.75$ produced by a single outlier.

Statistical inference with correlation coefficients

- The sample correlation coefficient is a function of the data, which are random.

  Therefore the sample correlation coefficient is a random variable, it has a sampling distribution.

  The variation of the sampling distribution depends on the sample size (greater sample size $\leftrightarrow$ less variation).



Correlation coefficients for 1000 bivariate data sets for which the population correlation coefficient is .7. The sample sizes are 40 (left) and 200 (right).

- If n is large then Fisher's Z transform can be used. Recall that t is the sample correlation coefficient and ρ is the population correlation coefficient. Let

$$r^* = \log\left(\frac{1+r}{1-r}\right)/2$$

and let

$$\rho^* = \log\left(\frac{1+\rho}{1-\rho}\right)/2,$$

Then r^* is approximately normal with mean ρ^* and variance $1/(n-3)$.

- *Example:* If $n = 30$ and $r = 0.4$ then $r^* \approx 0.42$.

  If we are testing the null hypothesis $\rho = 0$, then $\rho^* = 0$, so $\sqrt{n-3}r^*$ is the standardization of r^*.

  Therefore the one-sided p-value is

$$P(r^* > 0.42) = P(Z > 0.42 \cdot \sqrt{27}),$$

  where Z is standard normal.

  The approximate one sided p-value is .015.

- Suppose we would like to determine how large r must be to give p-value α (for the null hypothesis $\rho = 0$ and the right tailed alternative). First we answer the question in terms of r^*:

$$P(r^* > t) = \alpha$$

$$P(\sqrt{n-3}r^* > \sqrt{n-3}t) = \alpha$$

$$t = Q(1-\alpha)/\sqrt{n-3},$$

where Q is the standard normal quantile function.

Next we solve $r^* = t$ for r, giving

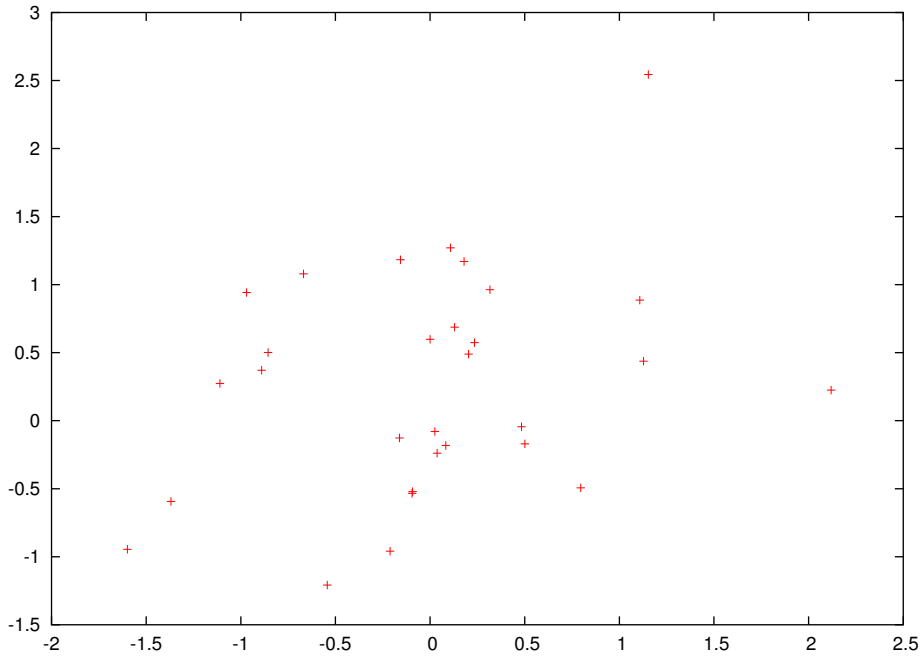
$$\log\left(\frac{1+r}{1-r}\right) = 2t$$

$$\frac{1+r}{1-r} = \exp(2t)$$

$$r = \frac{\exp(2t)-1}{\exp(2t)+1}.$$

To get a p-value smaller than .05 under a one sided test with $n = 30$ we need $r^* > 1.64/\sqrt{27} \approx 0.32$, so we need $r > 0.31$.

It is surprising how visually weak a significant trend can be. The following 30 points have correlation 0.31, and hence have p-value smaller than 0.05 for the right tailed test of $\rho = 0$:



- The Fisher Z transform can be used to construct confidence intervals for the correlation coefficient. To get a 95% CI, start with

$$P(-1.96 \leq \sqrt{n-3}(r^* - \rho^*) \leq 1.96) = 0.95,$$

and work it into this expression

$$P(r^* - 1.96/\sqrt{n-3} \leq \rho^* \leq r^* + 1.96/\sqrt{n-3}) = 0.95,$$

Next we need to invert the Fisher transform to change ρ^* into ρ.

Solve

$$\rho^* = \log\left(\frac{1+\rho}{1-\rho}\right)/2$$

for ρ, yielding

$$\rho = \frac{\exp(2\rho^*) - 1}{\exp(2\rho^*) + 1}.$$

Now apply the function

$$f(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1}$$

to both sides above, yielding

$$P(f(r^* - 1.96/\sqrt{n-3}) \leq \rho \leq f(r^* + 1.96/\sqrt{n-3})) = 0.95.$$

So

$$f(r^* - 1.96/\sqrt{n-3}), f(r^* + 1.96/\sqrt{n-3})$$

is a 95% CI.

For example, if we observe $r = 0.4$ when $n = 30$, the CI is $(.02, .65)$.

- *Comparison of two correlations.*

  Suppose we observe two bivariate samples, where r_1 is computed from (X_1, Y_1), ..., (X_m, Y_m) and r_2 is computed from (X'_1, Y'_1), ..., (X'_n, Y'_n).

  We wish to test the null hypothesis $\rho_1 = \rho_2$. We will base our inference on the statistic

$$D = (\sqrt{m-3}r_1^* - \sqrt{n-3}r_2^*)/\sqrt{2},$$

  which is approximately normal with mean $\rho_1^* - \rho_2^*$ and standard deviation 1.

  For example, if $r_1 = 0.4$ with $n = 30$ and $r_2 = 0.2$ with $m = 20$, then $r_1^* = 0.42$, $r_2^* = 0.20$, and $D = 0.96$. The one sided p-value is 0.17, so there is no strong evidence of a difference between ρ_1 and ρ_2.

Conditional mean and variance function

- The mean of a random variable Y may be written EY, where the E symbol stands for expectation, or expected value.

- Now suppose that we observe bivariate data (X, Y), and we select from the population all points where $X = 3$.

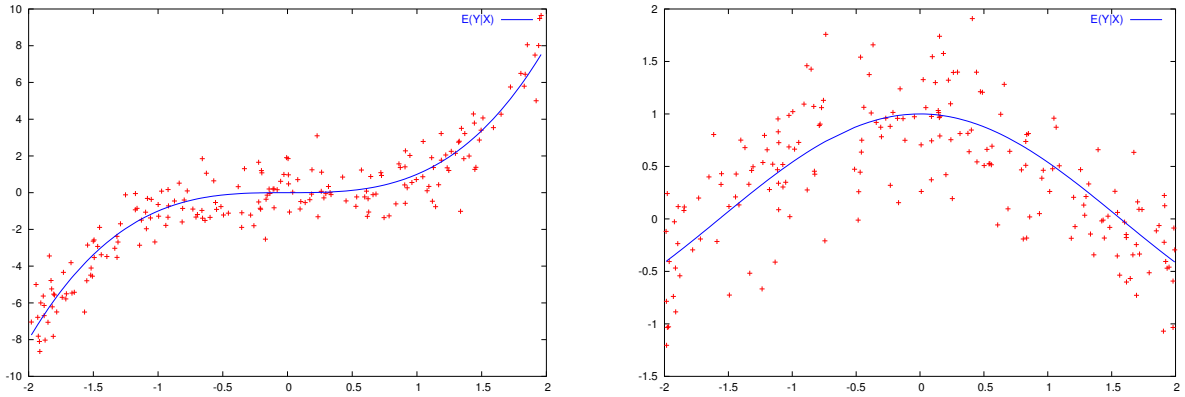
  If we then take the average of all Y values that are paired with these X values, we would obtain the conditional mean of Y given $X = 3$, or the conditional expectation of Y given $X = 3$, written $E(Y|X = 3)$.

- More generally, for any value x in the X sample space we can average all values of Y that are paired with $X = x$, yielding $E(Y|X = x)$.

  Viewed as a function of x, this is called the conditional mean function, the conditional expectation function, or the regression function.

  A crucial point is that $E(Y|X = x)$ is a function of x, but not of Y – the "Y" is part of the notation to remind you what is being averaged.

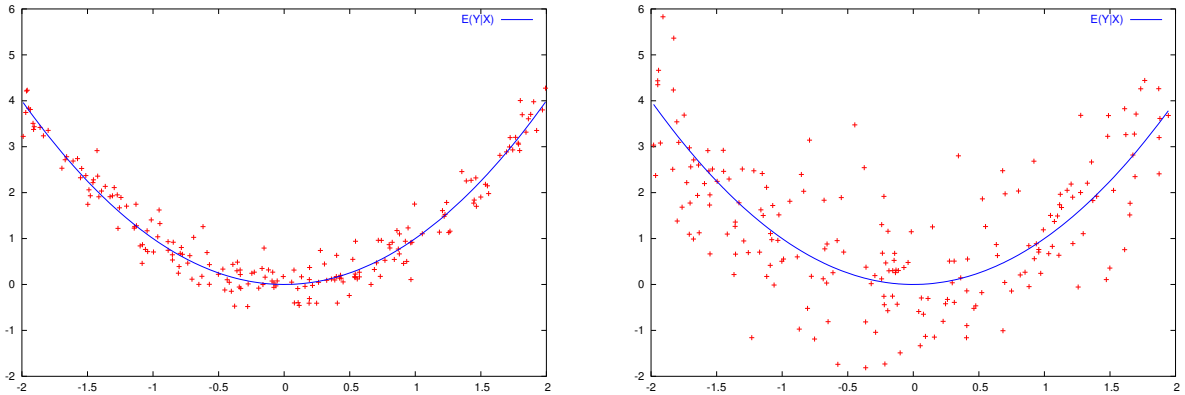
- Viewed graphically, the regression function $E(Y|X = x)$ is approximately equal to the average of the Y coordinates for points that fall in a narrow vertical band around x.



Left: a bivariate data set whose regression function is $E(Y|X = x) = x^3$. Right: A bivariate data set whose regression function is $E(Y|X = x) = \cos(x)$.

- The regression function is not equivalent to the joint distribution.

  The following plots show two distinct joint distributions with the same regression function.



Two bivariate data sets whose regression function is $E(Y|X = x) = x^2$.

- A critical piece of information that the regression function ignores is the level of variability around the conditional mean.

  For example, in the previous plots, the variability is much lower on the left than on the right.

  To summarize the variability in a bivariate data set, we use the conditional standard deviation function, which we will write $\sigma(Y|X = x)$.

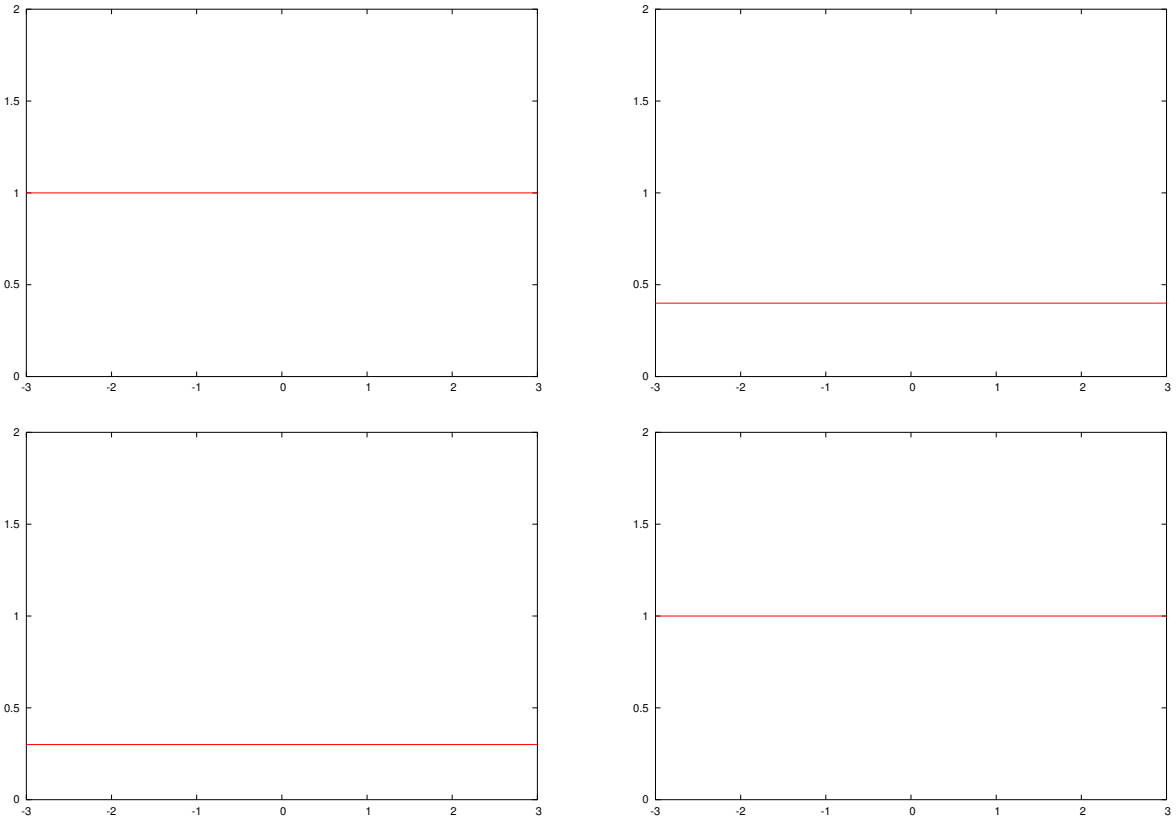
  We will also refer to the conditional variance function:

$$\text{var}(Y|X = x) = \sigma(Y|X = x)^2.$$

- The value of $\sigma(Y|X = x)$ is the standard deviation of all Y values that are paired with $X = x$.

  Graphically, this is equal to the standard deviation of the points falling in a narrow vertical band around x. In the examples shown above, $\sigma(Y|X = x)$ is a constant function of x.

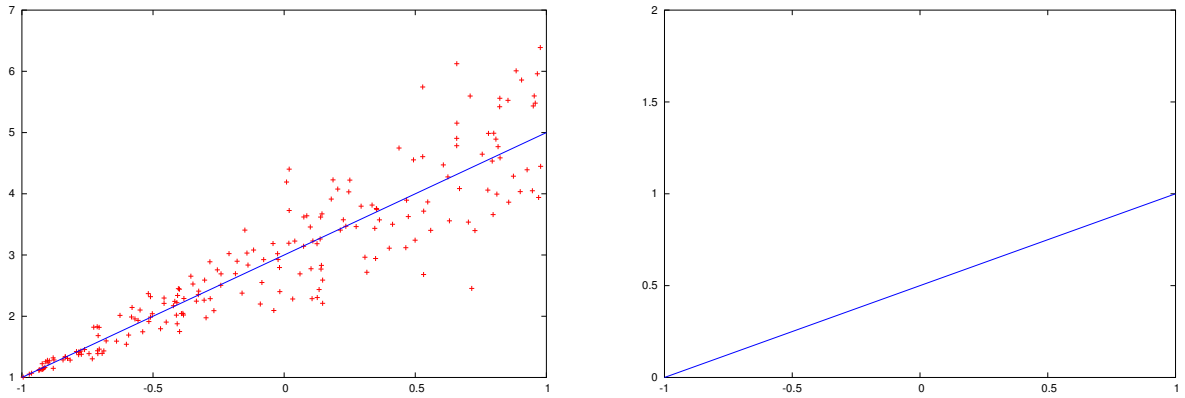
- Plots of $\sigma(Y|X = x)$ for these examples are shown below.



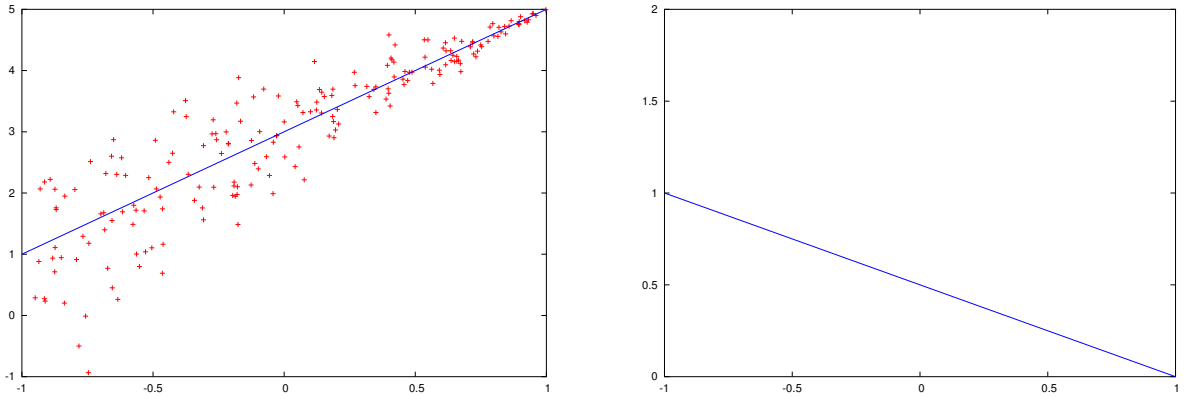
Plots of $\sigma(Y|X = x)$ for the four examples above.

- The above examples are called homoscedastic, meaning the standard deviation does not depend on x (i.e., it is a constant function of x).

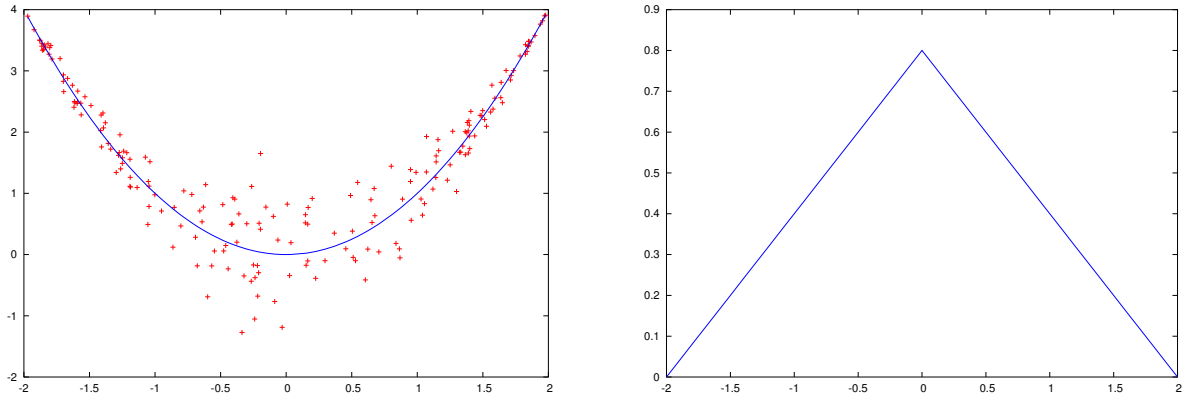
  The following examples show heteroscedastic bivariate distributions. The regression function and a scatterplot are shown on the left, the conditional standard deviation function is shown on the right.



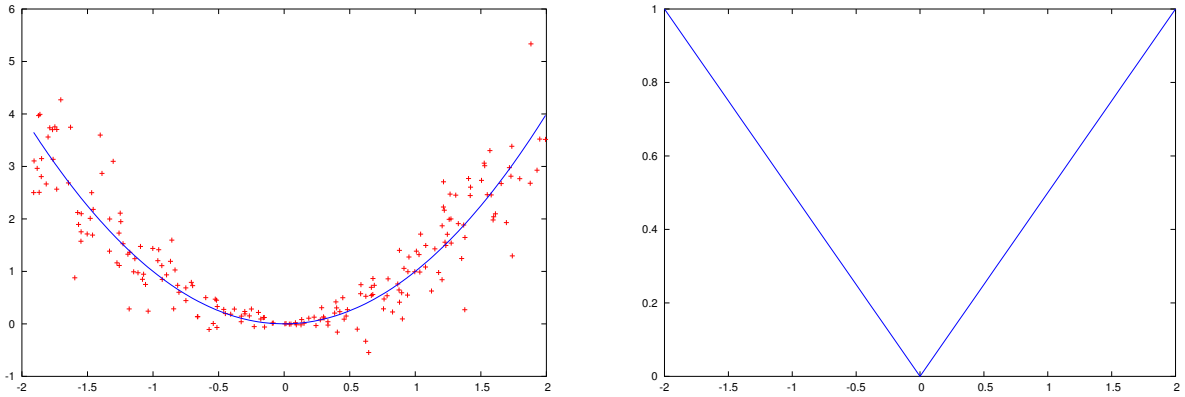
A bivariate data set with $E(Y|X = x) = 3 + 2x$ and $\sigma(Y|X = x) = (x + 1)/2$. Left: A scatterplot of the data, and the regression function. Right: the conditional standard deviation function.



A bivariate data set with $E(Y|X = x) = 3 + 2x$ and $\sigma(Y|X = x) = (1 - x)/2$. Left: A scatterplot of the data, and the regression function. Right: the conditional standard deviation function.



A bivariate data set with $E(Y|X = x) = x^2$ and $\sigma(Y|X = x) = .4|2 - x|$. Left: A scatterplot of the data, and the regression function. Right: the conditional standard deviation function.



A bivariate data set with $E(Y|X = x) = x^2$ and $\sigma(Y|X = x) = |x|/2$. Left: A scatterplot of the data, and the regression function. Right: the conditional standard deviation function.