Chapter 6: Correlation

Contents

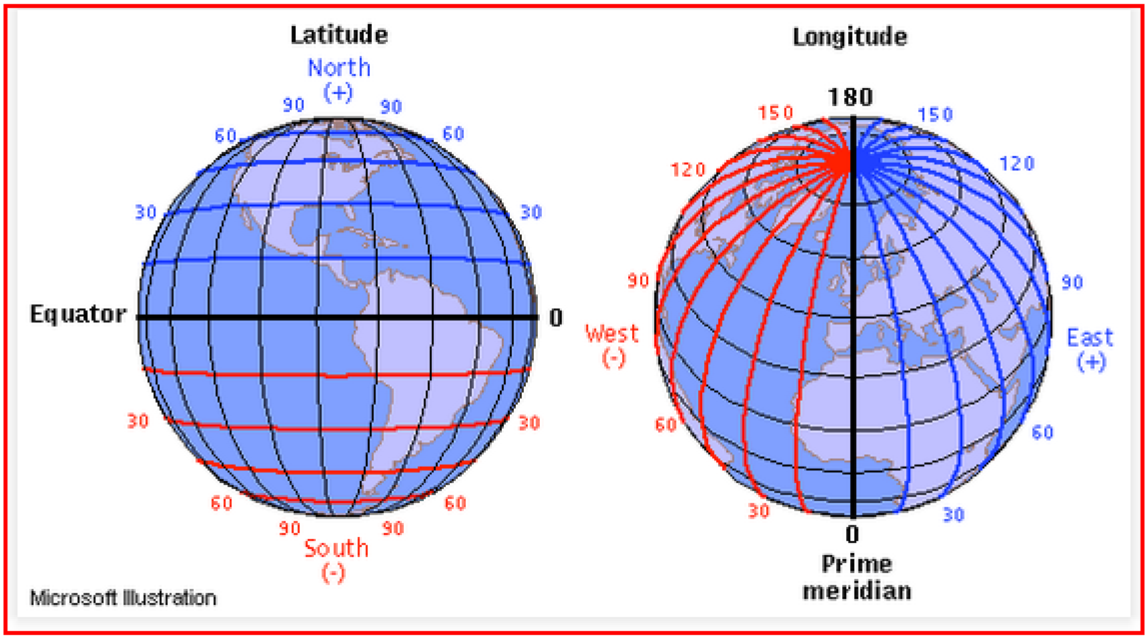
1 Starting example: Temperature data

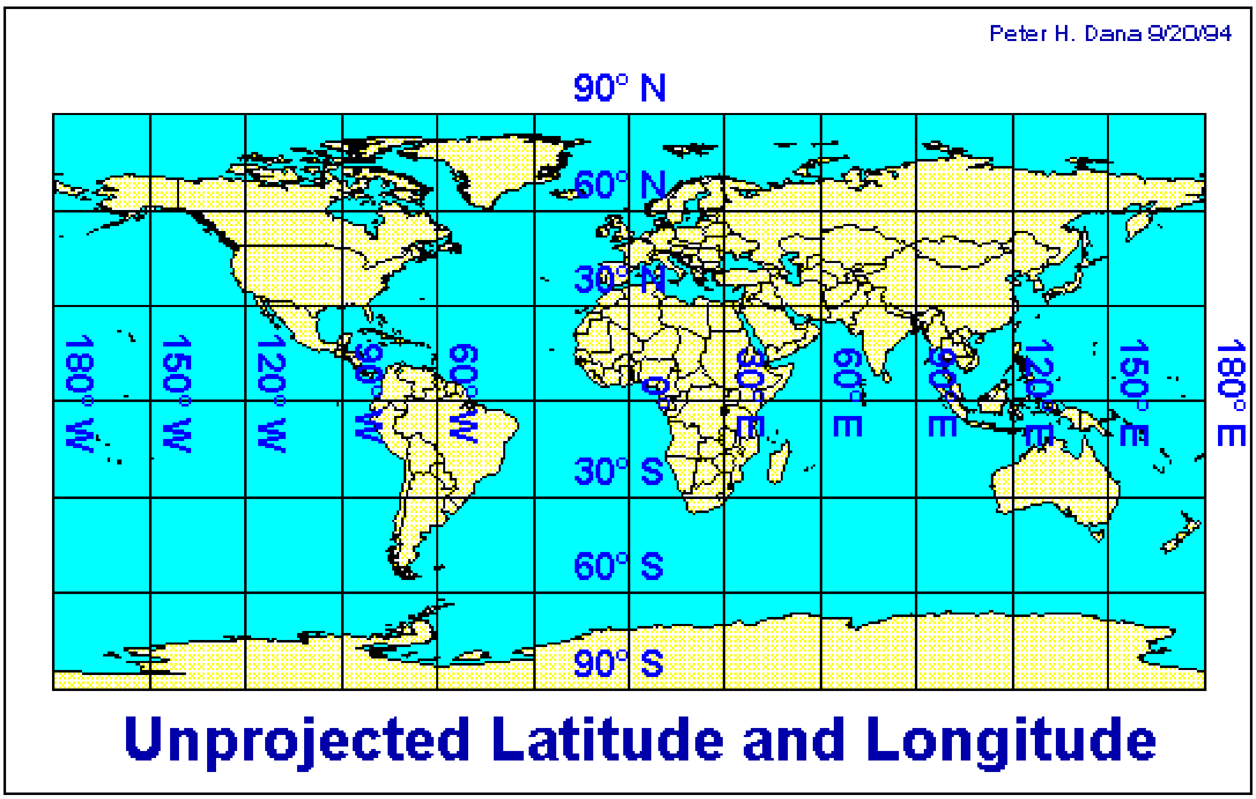
Example *Temperature*
Import the data set *temp_warm.txt* as `temperature`.

We are interested in the relationship between annual temperature (*annual*) on the one hand and *Latitude* and *Longitude* on the other hand.

```
head(temperature[15:19])
```

```
##   Amplitude Latitude Longitude  Area annual
## 1      14.6     52.2       4.5  West    9.9
## 2      18.3     37.6      23.5 South   17.8
## 3      18.5     52.3      13.2  West    9.1
## 4      14.4     50.5       4.2  West   10.3
## 5      23.1     47.3      19.0  East   10.9
## 6      17.5     55.4      12.3 North    7.8
```

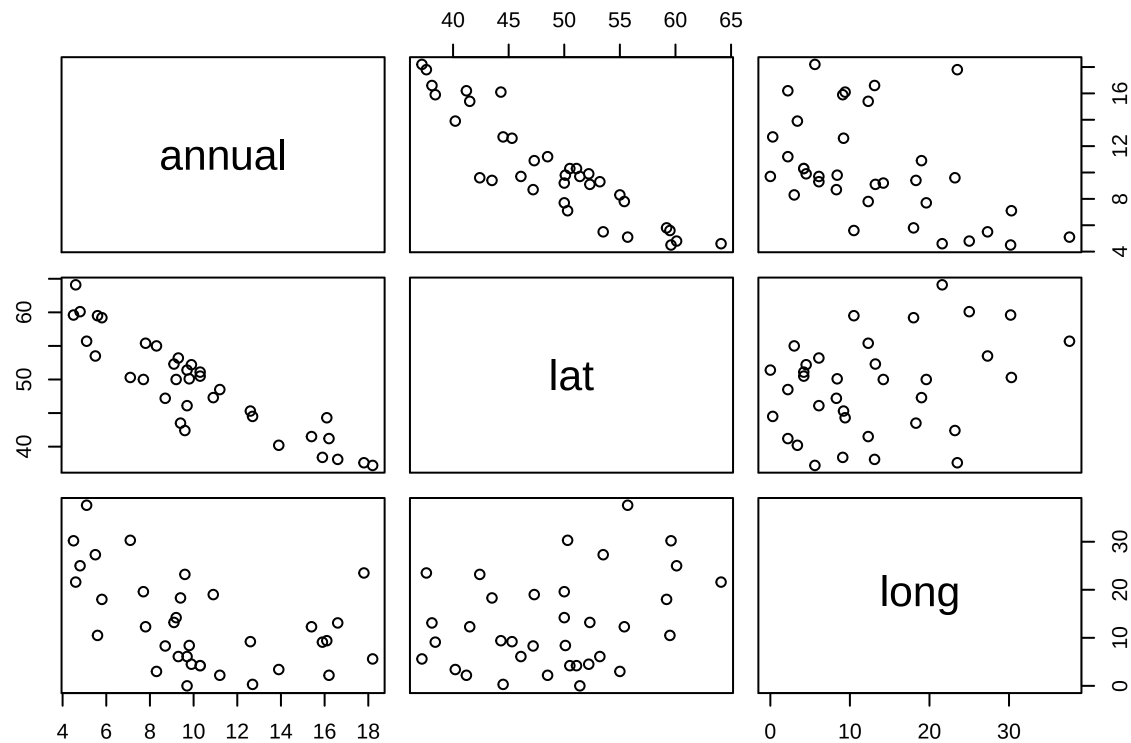




- We want to know if there is a linear relationship between annual temperature (*annual*) and *Latitude* and between annual temperature (*annual*) and *Longitude*.
- If such a relationship exists, how can we express it? How can we model this?
- And if we can model it, can we also use it for prediction?

We start by simply making a scatter plot, which visualizes the relationship between annual temperature (*annual*) and the variables *Latitude* and *Longitude*.

```
annual <- temperature$annual
lat <- temperature$Latitude
long <- temperature$Longitude
combine <- data.frame(annual, lat, long)
pairs(combine)
```



- There seems to be a decreasing trend between annual temperature and latitude.
- There seems to be a trend between annual temperature and longitude, but this is not so clear.

A first possibility to express a linear relationship between continuous variables, is to use the **correlation coefficient**.

2 Pearson correlation coefficient

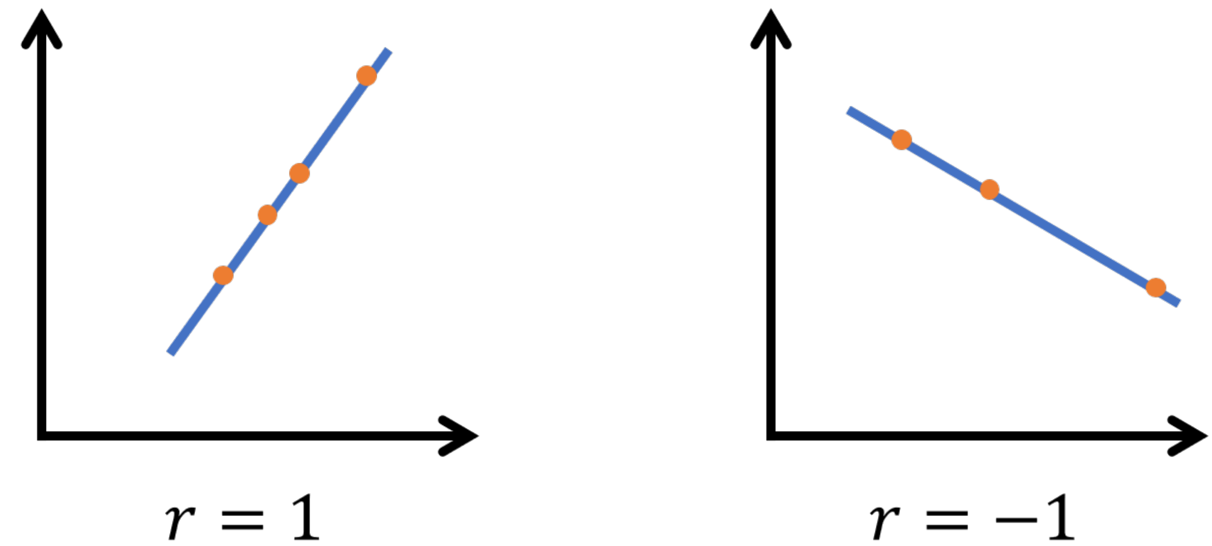
2.1 Definition

The **Pearson correlation coefficient** is expressed by

$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_X} \cdot \frac{Y_i - \overline{Y}}{S_Y} \right)$.

It gives an **indication if there is a *linear* relationship** between two continuous variables X and Y and it expresses how strong this linear relationship is. n is the total number of observations and S is the sample standard deviation.

r always has a value between -1 and $+1$.
The correlation coefficient r takes the value -1 or 1 if the pairs (x, y) are on a straight line: $+1$ if it is an increasing straight line and -1 if the line decreases.



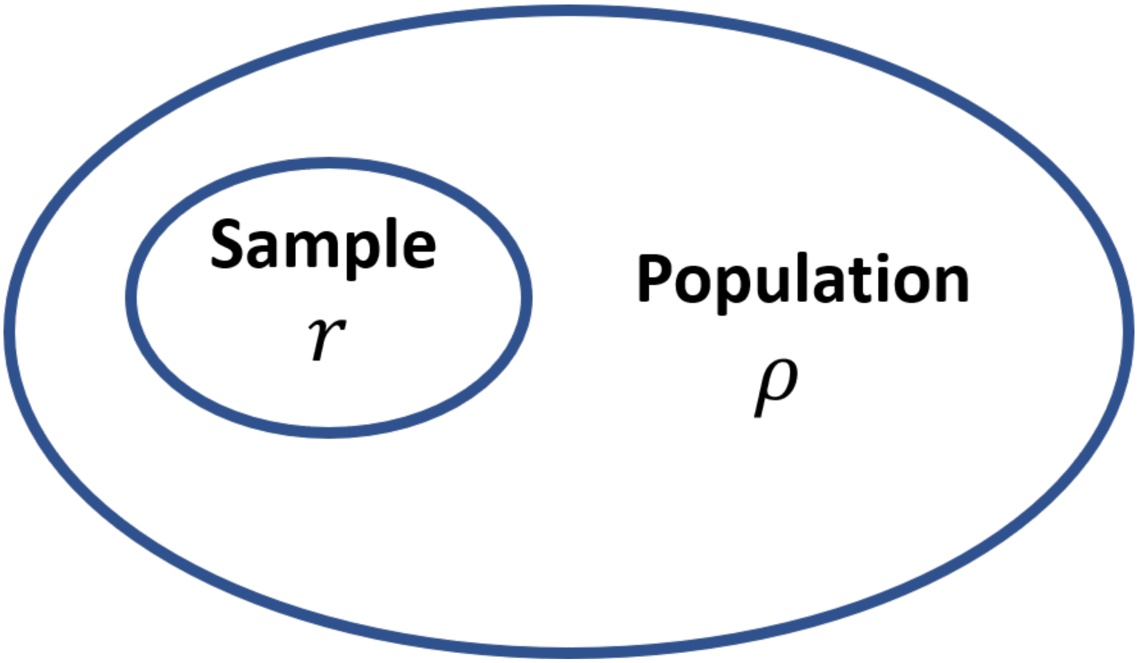
Some extra graphs:

2.2 Properties

- $-1 \leq r \leq 1$
- $r < 0$: negative linear trend between the x_i and y_i
- $r > 0$: positive linear trend between the x_i and y_i

- $r = -1$: points (x_i, y_i) are on a decreasing straight line
- $r = +1$: points (x_i, y_i) are on an increasing straight line
- $r = 0$ does not mean that there is no relation between x_i and y_i. It means that there is no *LINEAR* relation between x_i and y_i.

2.3 Population correlation coefficient



The **sample correlation coefficient** r is an estimate of the **population correlation coefficient** ρ, which is expressed by

$\rho = E\left(\frac{X-\mu_X}{\sigma_X} \cdot \frac{Y-\mu_Y}{\sigma_Y}\right)$

Hypothesis test:

Step 1
Often, we are interested in the following hypothesis about ρ:
$H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$

Step 2
We usually take $\alpha = 0.05$.

Step 3
We use the following test statistic:

$T = \sqrt{n-2} \cdot \frac{r}{\sqrt{1-r^2}} \approx t_{n-2}$

This test statistic T follows a t-distribution (also called student distribution) with $n-2$ degrees of freedom if H_0 holds and if X and Y are normally distributed.

Step 4
Compute the test statistic and the $p-value$.

Step 5
Based on the corresponding $p-value$, we formulate the appropriate conclusion.

2.4 Correlation coefficients in R

Let's compute the correlation coefficients in R.

In R

```
cor(combine)
```

```
##            annual        lat       long
## annual  1.0000000 -0.9027853 -0.4769927
## lat    -0.9027853  1.0000000  0.3154657
## long   -0.4769927  0.3154657  1.0000000
```

```
cor.test(annual, lat)
```

```
##
##  Pearson's product-moment correlation
##
## data:  annual and lat
## t = -12.058, df = 33, p-value = 1.226e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9501724 -0.8146160
## sample estimates:
##        cor
## -0.9027853
```

```
cor.test(annual, long)
```

```
##
##  Pearson's product-moment correlation
##
## data:  annual and long
## t = -3.1176, df = 33, p-value = 0.003765
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6991112 -0.1709140
## sample estimates:
##        cor
## -0.4769927
```

Remark:
In case the distribution of the variables is not normal, then Spearman correlations should be used.

3 Spearman correlation coefficient

The **Spearman correlation coefficient** r_S is a non-parametric alternative to the Pearson correlation coefficient which should be used when the distribution of the variables is not normal. The Spearman rank correlation coefficient (r_S) is an ordinary correlation coefficient based on the ranks (r_{ij}) of the data. It is expressed by
$r_S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{r_{X_i}-\bar{r}_X}{S_{r_X}} \cdot \frac{r_{Y_i}-\bar{r}_Y}{S_{r_Y}}\right)$

Example: ***Temperature***

1. Check the normality of the variables

```
shapiro.test(annual)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  annual
## W = 0.93475, p-value = 0.03879
```

Normality is rejected for the variable *annual*.

2. Compute Spearman correlations

```
cor(combine, method = "spearman")
```

```
##            annual        lat       long
## annual  1.0000000 -0.8668488 -0.5276026
## lat    -0.8668488  1.0000000  0.2670403
## long   -0.5276026  0.2670403  1.0000000
```

```
cor.test(annual, lat, method = "spearman")
```

```
## Warning in cor.test.default(annual, lat, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  annual and lat
## S = 13329, p-value = 1.669e-11
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.8668488
```

```
cor.test(annual, long, method = "spearman")
```

```
## Warning in cor.test.default(annual, long, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  annual and long
## S = 10907, p-value = 0.001126
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5276026
```