The correlation

Guillaume Filion

September 26, 2011

1 The problem

In 1997, Robert Clarke and collaborators conducted a meta-analysis to determine the quantitative importance of dietary fatty acids to blood concentrations of cholesterol. They report 6 studies with the same protocol applied to 14 subjects each. Here is their data (the study is real but the data is fake):

Saturated fat in diet % total calories	Blood cholesterol mmol/L
25.7	5.9
5.1	5.9
13.0	4.8
20.5	5.8
7.0	6.3
27.9	6.1

In your `R` session, manually enter the data in 2 vectors of length 6 (`diet`, `blood`).

```
> diet <- c(25.7, 5.1, 13.0, 20.5, 7.0, 27.9);
> blood <- c(5.9, 5.9, 4.8, 5.8, 6.3, 6.1);
```

Exercise 1

Can we assume that the data is Gaussian? If not, can we assume that the response variable (blood cholesterol) is Gaussian given the diet? What is the null hypothesis?

Exercise 2

We would like to find a statistic that measures 'association' between the two variables. Can you suggest a score? What is the alternative hypothesis?

2 The test statistic

For this problem, `R` made our life particulary easy with the function `cor` which computes just our test statistic r.

Exercise 3

Plot the data in the (x, y) plane. Add the regression line with `abline(lm(blood ~ diet))`. Display the coefficients of the line by calling `lm(blood ~ diet)`. Also compute `cor(blood, diet) * sd(blood) / sd(diet)`, what do you observe?

Exercise 4

Verify by the formula, or with some examples at the terminal that r is invariant by translation and scaling **of any or both of the variables**. What does that mean for the resampling of r?

Exercise 5

Resample r, finish the test, estimate the p-value, compare with the results of `cor.test`.

3 Properties of the correlation

The coefficient of correlation has a few properties worth mentioning, mostly because **they can be very dangerous**.

Exercise 6

What is the maximum value that the coefficient of correlation can take? What is the minimum?

Exercise 7

What does that mean if two variables have a coefficient of correlation of 1 in absolute value?

Exercise 8

What does that mean if two variables have a coefficient of correlation of 0? Assign to `x` all the integer numbers between -5 and 5. Specify `y <- x^2`, plot the variables in the (x, y) plane and compute their correlation. Conclude.

Answer of Exercise 1

We can assume that given a certain diet, the distribution of blood cholesterol is Gaussian ('reaction norms' are often Gaussian). Like for the t test, the null hypothesis can be formulated as follows:

1. The blood cholesterol is sampled from a Gaussian distribution.
2. The parameters are unknown, but equal in all cases.
3. Sampling is IID.

Answer of Exercise 2

The coefficient of correlation r is a good score for the problem at hand. By definition it is the covariance of the two variables divided by he product of their standard deviations.

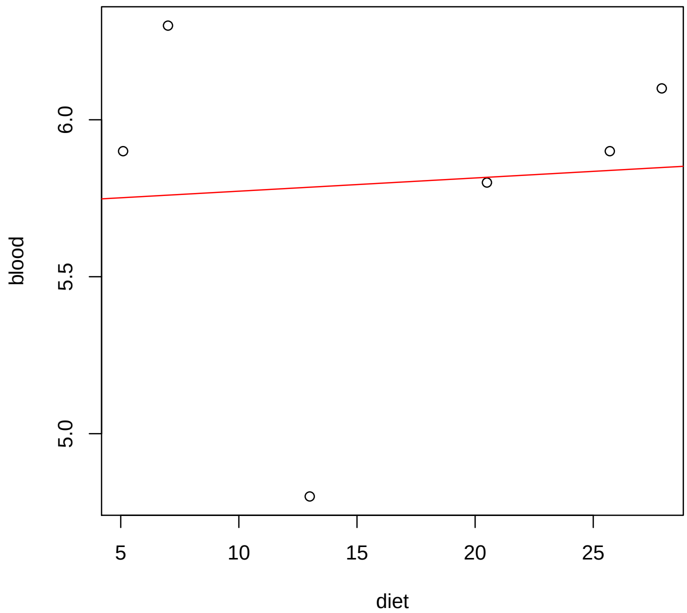
$$r = \frac{\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{6}(x_i - \bar{x})^2 \sum_{i=1}^{6}(y_i - \bar{y})^2}}.$$

You can check that this is the same as computing the covariance (`cov`) between two vectors and dividing by the product of their standard deviations (`sd`). The $n - 1$ terms cancel out in the numerator and the denominator.

In the alternative hypothesis, item 2 is replaced by 'The expected value of the response variable is a linear function of the independent variable (the diet here)'.

Answer of Exercise 3

```
> plot(diet, blood);
> abline(lm(blood ~ diet), col=2);
> lm(blood ~ diet);

Call:
lm(formula = blood ~ diet)

Coefficients:
(Intercept)         diet
   5.730375     0.004211

> cor(blood, diet)*sd(blood)/sd(diet);

[1] 0.004211178
```



The last term is the slope of the line. This is a property of the regression line, the slope is intimately linked to the correlation (by the formula used above).

Answer of Exercise 4

The proof is similar to that used for the effect size t. To verify it you can try different transformations.

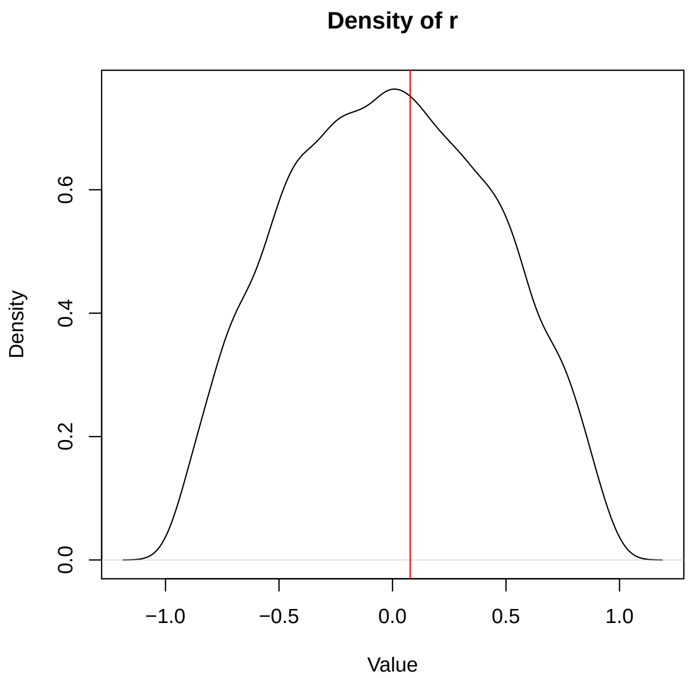
```
> r.obs <- cor(blood, diet);
> r.obs;

[1] 0.07770512

> cor(blood+1, diet);

[1] 0.07770512

> cor(pi*blood+1, diet);

[1] 0.07770512

> cor(pi*blood+1, diet+2);

[1] 0.07770512

> cor(pi*blood+1, -1*diet+2);

[1] -0.07770512
```

Note that the correlation changes sign but not value when one variable is multiplied by a negative number.

This means that we can resample r from standard Gaussian variables because all Gaussian variables have the same distribution of r under the null hypothesis.

Answer of Exercise 5

```
> r.smpl <- rep(NA, 10000);
> for (i in 1:10000) {
+   r.smpl[i] <- cor(rnorm(6), diet);
+ }
> plot(density(r.smpl), main="Density of r", xlab="Value");
> abline(v=r.obs, col=2);
> quantile(abs(r.smpl), probs=0.95);

      95%
0.8178391

> mean(abs(r.smpl) > r.obs);

[1] 0.8826

> cor.test(blood, diet);

        Pearson's product-moment correlation

data:  blood and diet
t = 0.1559, df = 4, p-value = 0.8837
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7832498  0.8365138
sample estimates:
       cor
0.07770512
```



Answer of Exercise 6

The coefficient of correlation is always between -1 and 1 (for real variables). This is a direct result of the Cauchy-Schwarz inequality, but it is rather difficult to prove if you don't know this result. You can verify by yourself by trying different inputs to `cor`. The highest you can find will be for `cor(blood, blood)` and the lowest for `cor(blood, -blood)`.

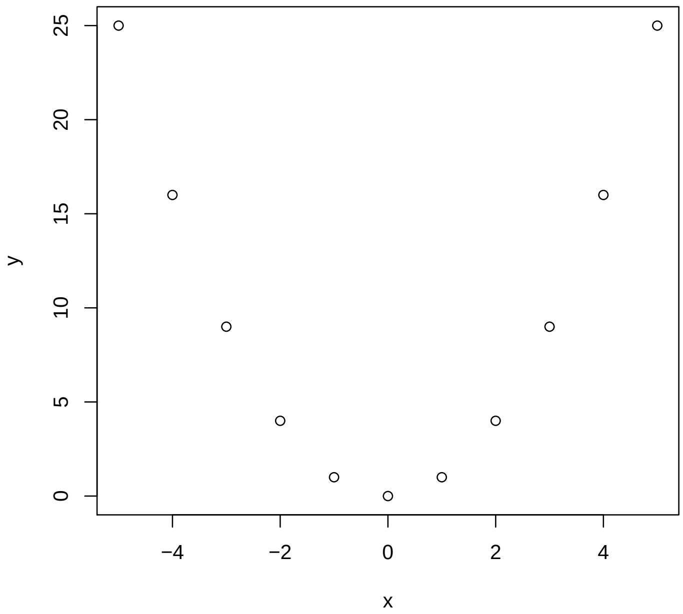
Answer of Exercise 7

This means that one of the variables is a linear transformation of the other, *i.e.* a scaling and translation of the other. You can see it with `cor(blood, 3*blood+4)` etc.

Answer of Exercise 8

```
> x <- -5:5;
> y <- x^2;
> plot(x,y);
> cor(x,y);
```

```
[1] 0
```



The coefficient of correlation is sometimes called the coefficient of **linear** correlation because it measures only linear trends between two variables. The coefficient of correlation should **never** be interpreted without a plot of y versus x, because many variables show a strong mutual dependency, but no linear correlation.