

Stat 312: Lecture 22

Correlation coefficient.

Moo K. Chung
mchung@stat.wisc.edu

Dec 9, 2004

1. *Correlation coefficient* between two random variables X and Y is defined as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The *sample correlation coefficient* r between two samples x_i and y_j is defined as $r = S_{xy} / \sqrt{S_{xx}S_{yy}}$. *Example* Is there any linear relationship between the mathematical achievement test scores (x_i) and calculus grades (y_i) for 10 college freshmen?

x ||39, 43, 21, 64, 57, 47, 28, 75, 34, 52

y ||65, 78, 52, 82, 92, 89, 73, 98, 56, 75

$x <- c(39, 43, 21, 64, 57, 47, 28, 75, 34, 52)$

$y <- c(65, 78, 52, 82, 92, 89, 73, 98, 56, 75)$

$> \text{cor}(x, y)$

[1] 0.8397859

We are interested in testing $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. In order to do this, we need to know the distribution of the sample correlation coefficient r under the null assumption.

2. Suppose the joint probability density function f of X and Y is the *bivariate normal* such that $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $\rho(X, Y) = \rho$. In this situation, it can be shown that

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x,$$

where $\beta_1 = \rho\sigma_Y/\sigma_X$. This is the regression model we studied before. Note that $\rho = 0$ if and only if $\beta_1 = 0$. Hence testing

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

is equivalent to

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

So the appropriate test statistic is

$$T = (\hat{\beta}_1 - \beta_1) / S_{\hat{\beta}_1} = \beta_1 / S_{\hat{\beta}_1} \sim t_{n-2}$$

under the null assumption. It can be shown that the test statistic is equivalent to $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Note. In bivariate normal data, $\rho = 0$ if and only if X and Y are independent. So testing for independence is equivalent to testing $\rho = 0$ in this situation.

```
> summary(lm(y~x))
```

```
Call: lm(formula = y ~ x)
```

```
Coefficients:
```

```
... t value Pr(>|t|)
```

```
(Intercept) ... 4.794 0.00137
```

```
...
```

```
x 0.7656 0.1750 4.375 0.00236
```

```
... Multiple R-Squared: 0.7052
```

```
...
```

3. *Fisher z-transformation*. Given sample correlation r ,

$$z = \tanh^{-1} r = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \sim N\left(\rho, \frac{1}{\sqrt{n-3}}\right).$$

This is asymptotic results for large sample.

```
>fisher.z<-function(r){atanh(r)}
```

```
>cor.test(x,y,"two.sided","pearson",
,conf.level=0.95)
```

```
Pearson's product-moment correlation dat
x and y t=4.375, df =8, p-value = 0.0023
...
```

```
95 percent confidence interval:
```

```
0.4459632 0.9611846
```

```
sample estimates:
```

```
cor
```

```
0.8397859
```

R-package `cor.test` with `pearson` option is based on the exact distribution of a test statistic using t_{n-2} distribution.