

# Stat 324: Lecture 21

## Correlation Analysis

Moo K. Chung  
mchung@stat.wisc.edu

April 21, 2005

1. The *sample correlation coefficient*  $r$  is defined as  $r = S_{xy} / \sqrt{S_{xx}S_{yy}}$ . In the previous lecture, we showed that  $0 \leq r^2 \leq 1$  so  $-1 \leq r \leq 1$ .

2. *Example 1.* Is there any linear relationship between the mathematical achievement test scores ( $x_i$ ) and calculus grades ( $y_i$ ) for 10 college freshmen?

$x \parallel 39, 43, 21, 64, 57, 47, 28, 75, 34, 52$

$y \parallel 65, 78, 52, 82, 92, 89, 73, 98, 56, 75$

```
> cor(x, y)
[1] 0.8397859
```

3. *Example 2.* Suppose that data  $(x_j, y_j)$  satisfy a nonlinear relationship  $y_j = \sqrt{10^2 - x_j^2}$ . Since this is a highly nonlinear relationship, we should expect less correlation.

```
> x <- -10:10
> y <- sqrt(10^2 - x^2)
> cor(x, y)
[1] 0
> summary(lm(y~x))
Coefficients: Estimate
(Intercept) .... 7.392e+00
x .... 4.997e-18
Multiple R-Squared: 9.504e-32
```

For this data set,  $r = 0$ .

4.  $r = \pm 1$  if and only if paired observations  $(x_j, y_j)$  are on the straight line. This can be seen from the fact that  $r^2 = 1$  iff.  $SSE = \sum_{j=1}^n (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)^2 = 0$ . This is equivalent to all points  $(x_j, y_j)$  passing through a line  $y = \hat{\beta}_0 + \hat{\beta}_1 x$ . The correlation can be used to measure if the paired observations show linear relationship.

5. *Population correlation coefficient* is defined as

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

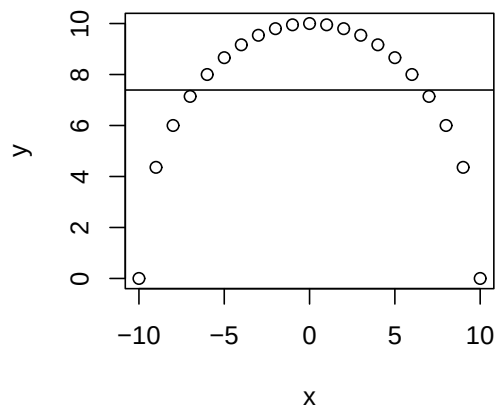


Figure 1: Complete lack of linear fit.

. The sample correlation is an estimate for the population correlation.

*Example 3.* Is there any relationship between the number of times you miss submitting homeworks ( $x_i$ ) and the final grade ( $y_i$ )?

$x \parallel 03, 04, 01, 02, 01, 00, 00, 07, 08, 05$

$y \parallel 70, 75, 80, 75, 85, 95, 90, 70, 50, 60$

```
> cor(x, y)
[1] -0.8938321
```

6. Assume  $(X, Y)$  is distributed as a *bivariate normal* whose joint density is given by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[ -\left( \frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \frac{x-\mu_1}{\sigma_1} \frac{y-\mu_2}{\sigma_2} - \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] / (2(1-\rho^2)).$$

The marginal distributions are  $X \sim N(\mu_1, \sigma_1^2)$ ,  $Y \sim N(\mu_2, \sigma_2^2)$ . Also  $\rho(X, Y) = \rho$ . If  $\rho = 0$ , then the joint density can be factored into the

product of marginal distributions showing the independence of  $X$  and  $Y$ .

It can be shown that  $Y|X = x$  is normally distributed with parameters  $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$ , where  $\beta_1 = \rho \frac{\sigma_2}{\sigma_1}$  and  $\beta_0 = \mu_2 - \rho \mu_1 \frac{\sigma_2}{\sigma_1}$ , and  $\mathbb{V}(Y|X = x) = (1 - \rho^2) \sigma_2^2$ . This implies that  $Y$  given  $X = x$  can be modeled using the linear regression.

7. Testing  $H_0 : \rho = 0$  vs.  $H_1 : \rho \neq 0$  is equivalent to  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$  in bivariate normal data. So the appropriate test statistic for testing  $H_0 : \rho = 0$  is  $T = (\hat{\beta}_1 - \beta_1) / S_{\hat{\beta}_1} \sim t_{n-2}$ . Under  $H_0$ , the test statistic is equivalent to  $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ .
8. Revisiting *Example 3*. Test for the dependence of the mathematical achievement test scores ( $x_i$ ) and calculus grades ( $y_i$ ) for 10 college freshmen assuming  $(x_i, y_i)$  are from bivariate normal.  
 $x \parallel 39, 43, 21, 64, 57, 47, 28, 75, 34, 52$   
 $y \parallel 65, 78, 52, 82, 92, 89, 73, 98, 56, 75$

*Solution.* There are two ways to solve this problem. The first approach might be much simpler conceptually. The only thing you need is to compute the correlation coefficient and follow the solution of Exercise 12.67.

```
> r<-cor(x,y)
[1] 0.8397859
> r*sqrt(10-2)/sqrt(1-r^2)
[1] 4.375015
> qt(0.025,8)
[1] -2.306004
```

An alternate method would be to perform a simple linear regression analysis.

```
> summary(lm(y~x))
Coefficients: Estimate
              x         4.4
```

9. *Example 4*. Exercise 12.67. A sample of 10,000  $(x_i, y_i)$  pairs resulted in the sample correlation coefficient  $r = 0.022$ . Test  $H_0 : \rho = 0$  vs.  $H_1 : \rho \neq 0$  at level  $= 0.05$ . Also test if  $x_i$  and  $y_i$  are independent.

```
> r<-0.022
> r*sqrt(10000-2)/sqrt(1-r^2)
[1] 2.200313
> qnorm(0.025)
[1] -1.959964
```

Since  $t = 2.2$  we reject  $H_0$ .

10. In general, test  $H_0 : \rho = \rho_0$  vs.  $H_1 : \rho \neq \rho_0$  requires the Fisher transformation. It can be shown that  $\frac{1}{2} \ln \frac{1+r}{1-r} \sim N(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3})$ . So an appropriate test statistic for testing  $H_0 : \rho = \rho_0$  is

$$Z = \frac{\frac{1}{2} \ln \frac{1+r}{1-r} - \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}}{(n-3)^{-1/2}}.$$

and under  $H_0$ ,  $Z \sim N(0, 1)$ . Constructing CI is based on the Fisher transform.

**Assigned problems.** Exercise 12.60, 12.64. HW 6 due April 28th. Solve the assigned problems in Lecture 19-22.