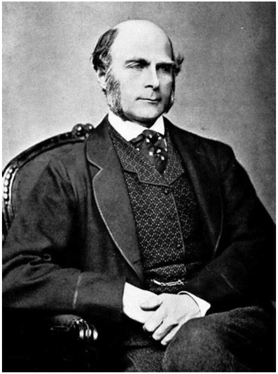
生物统计学

第十一章 相关分析

云南大学 生命科学学院

會澤百家 至公天下

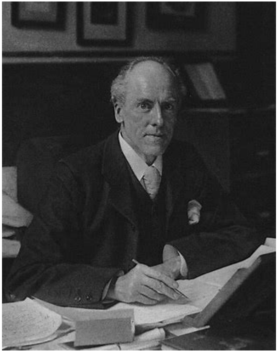
1. 相关性和相关系数的由来
2. 线性相关分析
3. 秩相关分析
4. 相关分析的注意事项
5. 回归系数与相关系数的关系



Francis Galton (1822-1911)

相关(correlation) 一词最早出在 Galton 于 1889[a]发表的一篇文章，当时写作 co-relation。

[a] Galton F. 1889. I. Co-relations and their measurement, chiefly from anthropometric data. Proc. R. Soc. Lond., 45:135–145. 1888 年 12 月 5 日收稿，1889 年 1 月 1 日发表。



Karl Pearson (1857-1936)

两个服从正态分布的随机变量 X 和 Y，且 $X \sim N(0, \sigma_1^2)$ 和 $Y \sim N(0, \sigma_2^2)$。二维平面上的所有点 $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ 就是两个随机变量构成的总体中的 N 个观测值。

Pearson 从该总体的概率密度（又称为联合密度函数）入手，得

$$f_{XY}(x,y) = (\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}})^N \times \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{\sum x_i^2}{\sigma_1^2} - 2\rho\frac{\sum x_i y_i}{\sigma_1\sigma_2} + \frac{\sum y_i^2}{\sigma_2^2}\right]} \quad (11.1)$$

其中 ρ 为总体相关系数。

$$f_{XY}(x,y) = \left(\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\right)^N \times \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{\sum x_i^2}{\sigma_1^2} - 2\rho\frac{\sum x_i y_i}{\sigma_1\sigma_2} + \frac{\sum y_i^2}{\sigma_2^2}\right]} \quad (11.1)$$

其中 ρ 为总体相关系数。

$$\rho = \frac{\sum x_i y_i}{N\sigma_x\sigma_y} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \quad (11.3)$$

将随机变量 X 和 Y 泛化到更一般的情形，即它们分别服从正态分布 $N(\mu_x, \sigma_1^2)$ 和 $N(\mu_y, \sigma_2^2)$，

则有总体相关系数的计算公式为

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} \tag{11.5}$$

将随机变量 X 和 Y 泛化到更一般的情形，即它们分别服从正态分布 $N(\mu_x, \sigma_1^2)$ 和 $N(\mu_y, \sigma_2^2)$，

则有总体相关系数的计算公式为

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} \tag{11.5}$$

当用样本统计量来估计总体参数时，有样本相关系数的计算公式

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{11.6}$$

在回归分析中，曾记 $SP_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$，$SS_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$，$SS_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$。

所以相关系数 r 的公式可改写为

$$r = \frac{SP_{xy}}{\sqrt{SS_x} \times \sqrt{SS_y}} \tag{11.10}$$

在回归分析中，曾记 $SP_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$，$SS_x=\sum_{i=1}^{n}(x_i-\overline{x})^2$，$SS_y=\sum_{i=1}^{n}(y_i-\overline{y})^2$。

所以相关系数 r 的公式可改写为

$$r=\frac{SP_{xy}}{\sqrt{SS_x}\times\sqrt{SS_y}} \tag{11.10}$$

又因回归平方和 $SS_r=b^2SS_x=\frac{SP_{xy}^2}{SS_x}$，所以相关系数 r 又可以表达为

$$r=\sqrt{\frac{SS_r}{SS_y}} \tag{11.11}$$

定义 (11.1)

设随机变量 X 和 Y，若

$$\mathrm{Cov}(X, Y) = E\Big[(X - EX)(Y - EY)\Big] \tag{1}$$

存在，则称其为随机变量 X 和 Y 的协方差(covariance)，记作 $\mathrm{Cov}(X, Y)$。

$$
\begin{aligned}
r &= \frac{SP_{xy}}{\sqrt{SS_x} \times \sqrt{SS_y}} = \frac{\frac{SP_{xy}}{n-1}}{\sqrt{\frac{SS_x}{n-1}} \times \sqrt{\frac{SS_y}{n-1}}} \\
&= \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)} \times \sqrt{\mathrm{Var}(y)}}
\end{aligned}
\tag{11.13}
$$

因样本相关系数 r 的期望等于总体相关系数 ρ，所以检验统计量可写为

$$\frac{r-\rho}{s_r} \tag{11.14}$$

其中 $s_r=\sqrt{\frac{1-r^2}{n-2}}$。代入上式，得检验统计量

$$\frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} \tag{11.15}$$

零假设 $H_0 : \rho = 0$ 成立时，检验统计量服从 $n-2$ 的 t 分布。即

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t(n-2) \tag{11.16}$$

例 (11.1)

对例题 10.1 中的数据（`nitrogenGrass` 数据集），进行 Pearson 相关分析。

```
> cor(x = nitrogenGrass$N, y = nitrogenGrass$DW)
[1] 0.943256
> cor.test(x = nitrogenGrass$N, y = nitrogenGrass$DW,
method = "pearson")

        Pearson's product-moment correlation

data:  nitrogenGrass$N and nitrogenGrass$DW
t = 6.3517, df = 5, p-value = 0.001429
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6565943 0.9918071
sample estimates:
     cor
0.943256
```

例 (11.2)

对例题 10.1 中的数据（**nitrogenGrass** 数据集），利用 F 检验法检验相关性的显著性。

```
> summary(lm(DW ~ N, data = nitrogenGrass))
...

Residual standard error: 1.346 on 5 degrees of freedom
Multiple R-squared:  0.8897, Adjusted R-squared:  0.8677
F-statistic: 40.34 on 1 and 5 DF,  p-value: 0.001429
```

对于非正态分布的数据资料要进行相关分析，解决问题的思路是将变量 x 和 y 先转变成秩统计量，然后计算秩相关系数 (coefficient of rank correlation) 以表示秩相关的性质及其相关程度。

常用的秩相关分析方法包括**Spearman** 秩相关和**Kendall** 秩相关。

Spearman 秩相关系数(Spearman's rank correlation coefficient，记作 r_s)，是英国心理学家 Charles E. Spearman 在 1904 年提出的一种非参数秩统计量，用于衡量两个变量之间的相关强度。

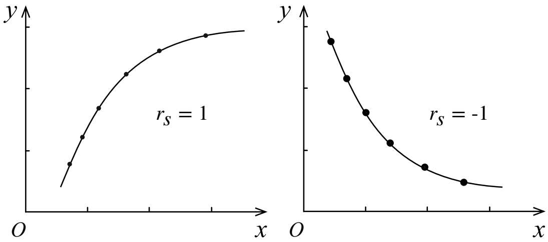


图 **11.1** 非线性关系的 Spearman 相关系数

秩统计量转换

假设有 3 个数据对 $(2,-3),(5,4),(-6,1)$，即 x 有 3 个观测值：$2,5,-6$，y 有 3 个观测值：$-3,4,1$。

秩统计量转换

假设有 3 个数据对 $(2,-3),(5,4),(-6,1)$，即 x 有 3 个观测值：$2,5,-6$，y 有 3 个观测值：$-3,4,1$。对 x 的 3 个观测值，按从小到大排序得

$$-6,2,5$$

用各观测值的位置编号替换原序列中的观测值，得秩统计量

$$R_x=\{2,3,1\}$$

秩统计量转换

假设有 3 个数据对 $(2,-3),(5,4),(-6,1)$，即 x 有 3 个观测值：$2,5,-6$，y 有 3 个观测值：$-3,4,1$。对 x 的 3 个观测值，按从小到大排序得

$$-6,2,5$$

用各观测值的位置编号替换原序列中的观测值，得秩统计量

$$R_x=\{2,3,1\}$$

对 y 作相同的处理，得

$$R_y=\{1,3,2\}$$

计算两秩统计量中相同位置上的秩差 d，即

$$d = R_x - R_y = 1, 0, -1$$

代入公式

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{11.24}$$

即可计算**Spearman** 秩相关系数。

例 (11.5)

对例题 10.1 中的数据（`nitrogenGrass` 数据集），进行 Spearman 秩相关分析。

```
> cor.test(x = nitrogenGrass$N, y = nitrogenGrass$DW,
method = "spearman")

    Spearman's rank correlation rho

data:  nitrogenGrass$N and nitrogenGrass$DW
S = 1.2434e-14, p-value = 0.0003968
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

Kendall 秩相关系数是由英国统计学家 Maurice G. Kendall 于 1938 年提出的一种相关程度的度量方法。

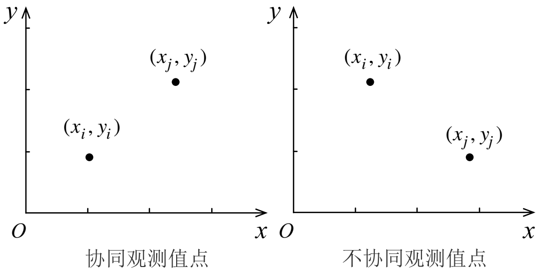


图 **11.3** 数据的协同关系示意图

当 $(x_j - x_i)(y_j - y_i) > 0$ 则两观测值点协同；当 $(x_j - x_i)(y_j - y_i) < 0$ 则两观测值点不协同。

Kendall 秩相关系数计算公式如下:

$$\tau = \frac{N_c - N_d}{\mathrm{C}_n^2} = \frac{N_c - N_d}{\frac{n(n-1)}{2}} \tag{11.32}$$

例 (11.7)

对例题 10.1 中的数据（`nitrogenGrass` 数据集），进行 Kendall 秩相关分析。

```
> cor.test(x = nitrogenGrass$N, y = nitrogenGrass$DW,
method = "kendall")

    Kendall's rank correlation tau

data:  nitrogenGrass$N and nitrogenGrass$DW
T = 21, p-value = 0.0003968
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
  1
```

11.4 相关分析的注意事项

- 数据分析实践中应特别注意 Pearson 相关系数要求数据服从正态分布，对离群值和非线性关系较为敏感。
- 相关系数应进行显著性检验。
- 为保证分析结果的可靠性变量的观测值应尽可能的多。

11.4 相关分析的注意事项

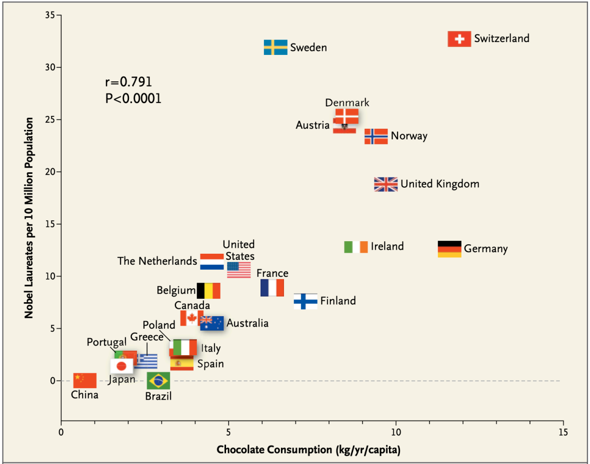


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

引自： Messerli FH. 2012. Chocolate Consumption, Cognitive Function, and Nobel Laureates. N. Engl. J. Med., 367(16):1562–1564.

$$
\begin{aligned}
b_{y|x} \times b_{x|y} &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} \times \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(y-\overline{y})^2} \\
&= \left(\frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}}\right)^2 \\
&= r^2
\end{aligned} \tag{11.40}
$$

本章小结

1. 相关性和相关系数的由来
2. 线性相关分析
   - Pearson 相关系数
   - 相关系数的显著性检验
   - 相关系数的区间估计
3. 秩相关分析
   - Spearman 秩相关系数
   - Kendall 秩相关系数
4. 相关分析的注意事项
5. 回归系数与相关系数的关系