Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 17: Correlation

Jinko Graham

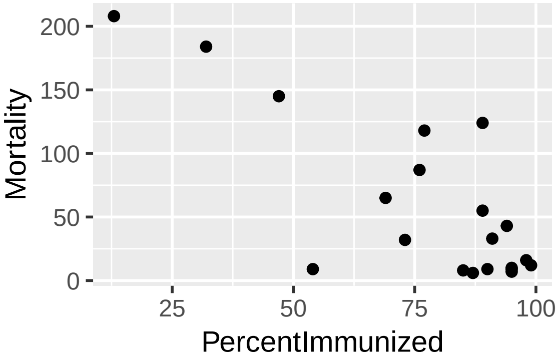
Example Data: Child Mortality by Country

- Introduce ideas by focusing on an example data set.
- Data on child mortality (number of deaths before age 5 years, per 1000 live births) and percentage of children who are immunized for diptheria, pertussis and tetanus (DPT) from a random sample of 20 countries (see Table 17.2 of text).

```
##        Nation PercentImmunized Mortality    Region
## 1     Bolivia               77       118 SouthAmer
## 2      Brazil               69        65 SouthAmer
## 3    Cambodia               32       184      Asia
## 4      Canada               85         8 NorthAmer
## 5       China               94        43      Asia
## 6 CzechRepub               99        12    Europe
```

Scatterplots

- For displaying a relationship between two quantitative variables.
- Each individual is represented by one point, comprised of a coordinate on the x-axis and a coordinate on the y-axis.

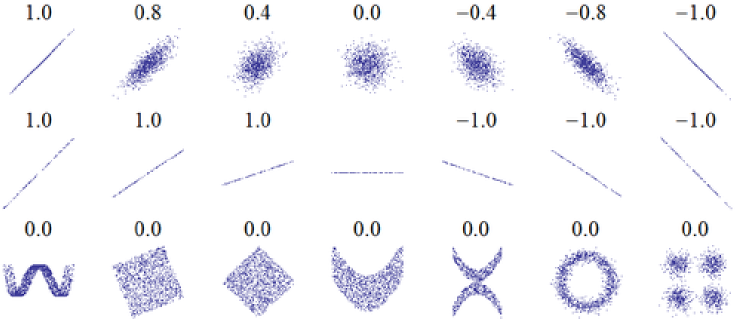


Interpreting Scatterplots

- The **direction** of the relationship – positive, negative, or no relationship
- The **form** of the relationship – patterns, such as linear or curved trends or even regional clusters in this particular example.
- The **strength** of the relationship – how tightly the data fall around the apparent trend
- The child mortality data give the impression of a fairly weak, negative linear relationship.
  - However, if we were to exclude the countries with immunization rates $< 50\%$ a pattern would not be obvious.

Overview of Correlation

- Correlation is a measure of the strength of a **linear** (as opposed to curved, circular etc.) relationship between two quantitative variables.
- The most commonly-used measure of correlation is the Pearson correlation coefficient.
- Example trends with their Pearson correlation coefficients:



Pearson Correlation Coefficient

- Suppose we have a simple random sample of data of size n, taken from some larger population.
- The Pearson correlation coefficient, r, is a measure of the strength and direction of a *linear* association between quantitative variables in the sample.
- Always between -1 and 1. Close to ± 1 suggests a strong linear relationship.
- Negative r suggests a negative association. Positive r, a positive association.
- The Pearson correlation coefficient is:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

(The formula can be simplified slightly – see page 400 of the text.)

Hypothesis Test of Correlation

- r estimates the *population* Pearson correlation, ρ, which we can think of as the Pearson correlation for the *entire population* of (large) size N:

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)$$

- The null hypothesis $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ can be tested with the test statistic

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

which, under H_0, has an approximate t-distribution with $n-2$ df.

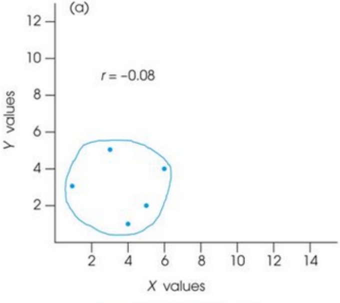
Testing correlation

- Sample correlation between `PercentImmunized` and `Mortality` is negative: $r = -0.791$.
- Let's test if the population correlation differs from zero
  - i.e. Test $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$.

```
##
##  Pearson's product-moment correlation
##
## data:  PercentImmunized and Mortality
## t = -5.4864, df = 18, p-value = 3.281e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9137250 -0.5362744
## sample estimates:
##        cor
## -0.7910654
```

- According to the test, there is strong statistical evidence that the population correlation is not 0.
- But could this be due to outlier countries, such as the three with low immunization rates?

- Pearson correlation coefficient is sensitive to outliers; e.g.:



From Gravetter and Wallnau, 8th Ed.

- An alternative measure that is less sensitive to outliers is the Spearman Rank-Correlation coefficient.

Spearman's Rank Correlation Coefficient

- To reduce the impact of outliers, replace the values with **ranks**.
- For a sample $x_1, x_2, \ldots, x_n$, let $x_{r1}, x_{r2}, \ldots, x_{rn}$ denote the ranks; E.G.
  - If x_1 is the 5th-largest value in the ordered list of x's, then its rank is 5; i.e., $x_{r1} = 5$.
  - If x_2 is the nth-largest value (i.e. smallest) in the ordered list of x's, then its rank is n; i.e. $x_{r2} = n$.
  - If x_3 is the 1st-largest value (i.e. largest) in the ordered list of the x's, then its rank is 1; i.e. $x_{r3} = 1$, etc.
- **Spearman's rank-correlation coefficient** is the Pearson correlation of the **ranks**:

$$r_s = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_{ri} - \overline{x}_r}{s_{rx}} \right) \left(\frac{y_{ri} - \overline{y}_r}{s_{ry}} \right), \quad \text{where}$$

  $\bar{x}_r$, $\bar{y}_r$ are sample means and s_{rx}, s_{ry} are sample SDs of ranks.
- The text (pg 405) shows that this formula can be simplified to depend only on the differences between ranks of the (x, y) pairs, but this is not our focus.

Hypothesis Test of Spearman's Correlation

- The *population* Spearman correlation coefficient, ρ_s, is the Spearman correlation for the entire population.
- To test $H_0 : \rho_s = 0$ vs. $H_a : \rho_s \neq 0$, about the presence of any rank-based correlation in the population, we can use the statistic

$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n-2)}} = r_s\sqrt{\frac{n-2}{1-r_s^2}},$$

  which has an approximate t-distribution with $n-2$ df under H_0.
  - The text considers this approximation to be reliable for $n \geq 10$.

Application to Child Mortality Data

```
## [1] -0.5431913
```

- The sample Spearman correlation coefficient of -0.54 is closer to zero than the sample Pearson correlation coefficient of -0.79.
- Let's test to see if there's any evidence that the population Spearman correlation coefficient differs from zero.
  - i.e., test $H_0 : \rho_s = 0$ vs. $H_a : \rho_s \neq 0$.
- We have a sample of $n = 20$ countries and so, according to the text, approximating the null distribution of the test statistic with t-distribution should be OK.

```
##
##  Spearman's rank correlation rho
##
## data:  PercentImmunized and Mortality
## S = 2052.4, p-value = 0.01332
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5431913
```

- There is statistical evidence that the population Spearman correlation differs from 0 (at level $\alpha = 0.05$).
  - `Mortality` and `PercentImmunized` appear to be negatively correlated, even when outlying countries are taken into account through a rank-based correlation test.