Dec 12, 2022

# Theoretical Hypothesis Testing

## Week 9: Tests on correlation coefficients

# Introduction

A correlation coefficient is used to **measure the relationship** between two random variables.

It shows **magnitude** and **direction** of the relationship.

It takes the values between -1 and 1. The value of -1 indicates the perfect negative, and 1 indicates the perfect positive relationship. The value of 0 shows there is no relationship between these variables.

# Notations

As in any hypothesis testing procedure, **we are also dealing with the populations' correlation coefficient in here**. Thus **we use the sample to test any hypothesis** about the population parameters.

In the tests on correlation coefficient, the following notations are used:

- $\rho$ : population's correlation coefficient
- $r$ : sample's correlation coefficient

# Correlation coefficients

There are the three most commonly used correlation coefficients:

1. Pearson correlation coefficient ($r_{XY}^{P}$)
2. Spearman correlation coefficient ($r_{XY}^{S}$)
3. Kendall correlation coefficient ($r_{XY}^{K}$)

# Pearson’s correlation coefficient

# Pearson's correlation coefficient

Pearson's correlation coefficient is commonly represented by $r_{xy}^P$ and referred as the sample correlation coefficient. Given data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ consisting of $n$ observations, $r_{xy}^P$ is defined as:

$$r_{xy}^P = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $n$ is sample size, $x_i, y_i$ are the individual sample points indexed with $i$, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

# Test on Pearson's correlation coefficient

This test is used to test that the Pearson's correlation coefficient of population $\rho$ differs from a specific value $\rho_0$.

Assumptions:

- Data are measured on an interval or ratio scale.
- The relationship between $X$ and $Y$ is linear.
- Data points $(x_i, y_i)$ are randomly sampled from normal distribution.

Hypotheses:

- $H_0 : \rho = \rho_0$ vs. $H_A : \rho \neq \rho_0$
- $H_0 : \rho \leq \rho_0$ vs. $H_A : \rho > \rho_0$
- $H_0 : \rho \geq \rho_0$ vs. $H_A : \rho < \rho_0$

# Test on Pearson's correlation coefficient

Test statistic:

$$T = r_{xy}^{P} \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^{P}}} \sim t_{n-2}$$

# Application

Test the correlation between the mpg (mileage per gallon) and qsec (fastest time to travel 1/4 mile from standstill in seconds) of mtcars dataset whether is significant. Do not forget the follow the steps required in any hypothesis testing.

```
> shapiro.test(mtcars$mpg)

Shapiro-Wilk normality test

data:  mtcars$mpg
W = 0.94756, p-value = 0.1229


> shapiro.test(mtcars$qsec)

Shapiro-Wilk normality test

data:  mtcars$qsec
W = 0.97325, p-value = 0.5935


> cor.test(mtcars$mpg, mtcars$qsec, method = "pearson")

Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$qsec
t = 2.5252, df = 30, p-value = 0.01708
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08195487 0.66961864
sample estimates:
     cor
0.418684
```

# Spearman’s correlation coefficient

# Spearman's correlation coefficient

Spearman's correlation coefficient is commonly represented by $r_{xy}^S$ and referred as the sample correlation coefficient. Given data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ consisting of $n$ observations, $r_{xy}^S$ is defined as:

$$r_{xy}^S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where $n$ is sample size, $d_i = R(x_i) - R(y_i)$ is the difference between the two ranks of each observations.

# Test on Spearman's correlation coefficient

This test is used to test that the Spearman's correlation coefficient of population $\rho$ differs from a specific value $\rho_0$.

Assumptions:

- Data are measured at least on an ordinal scale.
- The relationship between $X$ and $Y$ is monotonic.

Hypotheses:

- $H_0 : \rho = \rho_0$ vs. $H_A : \rho \neq \rho_0$
- $H_0 : \rho \leq \rho_0$ vs. $H_A : \rho > \rho_0$
- $H_0 : \rho \geq \rho_0$ vs. $H_A : \rho < \rho_0$

# Test on Spearman's correlation coefficient

Test statistic:

$T_S = r_{xy}^S \sqrt{(n-1)} \sim N(0,1)$ for large sample size.

# Application

Test the correlation between the mpg (mileage per gallon) and disp (the volume of the engine) of mtcars dataset whether is significant. Do not forget the follow the steps required in any hypothesis testing.

```
> shapiro.test(mtcars$mpg)

Shapiro-Wilk normality test

data:  mtcars$mpg
W = 0.94756, p-value = 0.1229


> shapiro.test(mtcars$disp)

Shapiro-Wilk normality test

data:  mtcars$disp
W = 0.92001, p-value = 0.02081


> cor.test(mtcars$mpg, mtcars$disp, method = "spearman")

Spearman's rank correlation rho

data:  mtcars$mpg and mtcars$disp
S = 10415, p-value = 6.37e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.9088824
```

# Kendall’s correlation coefficient

# Kendall's correlation coefficient

Kendall's correlation coefficient is commonly represented by $r_{xy}^K$ and referred as the sample correlation coefficient. Given the ordered data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, then for each $y_i$, count the number of $y_j > y_i$ as concordant pairs, and the number of $y_j < y_i$ as discordant paris, $r_{xy}^K$ is defined as:

$$r_{xy}^K = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where $n$ is sample size, $n_c$ is the number of concordant pairs, and $n_d$ is the number of discordant pairs.

# Test on Kendall's correlation coefficient

This test is used to test that the Kendall's correlation coefficient of population $\rho$ differs from a specific value $\rho_0$.

Assumptions:

- Data are measured at least on an ordinal scale.
- The relationship between $X$ and $Y$ is monotonic.

Hypotheses:

- $H_0 : \rho = \rho_0$ vs. $H_A : \rho \neq \rho_0$
- $H_0 : \rho \leq \rho_0$ vs. $H_A : \rho > \rho_0$
- $H_0 : \rho \geq \rho_0$ vs. $H_A : \rho < \rho_0$

# Test on Spearman's correlation coefficient

Test statistic:

$$T_K = \frac{r_{xy}^K}{\sqrt{\frac{n(n+1)(2n+5)}{18}}} \sim N(0,1)$$

# Application

Test the correlation between the mpg (mileage per gallon) and disp (the volume of the engine) of mtcars dataset whether is significant. Do not forget the follow the steps required in any hypothesis testing.

```
> shapiro.test(mtcars$mpg)

Shapiro-Wilk normality test

data:  mtcars$mpg
W = 0.94756, p-value = 0.1229


> shapiro.test(mtcars$disp)

Shapiro-Wilk normality test

data:  mtcars$disp
W = 0.92001, p-value = 0.02081


> cor.test(mtcars$mpg, mtcars$disp, method = "kendall")

Kendall's rank correlation tau

data:  mtcars$mpg and mtcars$disp
z = -6.1083, p-value = 1.007e-09
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.7681311
```

# Comparison of the CCs

# Comparison of the CCs

**Pearson vs Spearman and Kendall's CC**

- Non-parametric correlations are less powerful because they use less information in calculations. For example, in the calculation of Pearson's CC uses the information about the mean and the deviation from the mean, while non-parametric correlations (Spearman and Kendall) use only the ordinal information of the observations.
- In the case of non-parametric correlation, there is no assumption that the distribution of X and Y should ve normal distribution.

**Spearman vs. Kendall's CC**

- Kendall is more robust and efficient than Spearman that means Kendall is preferred when there are small samples or some outliers.
- Kendall correlation has $O(n^2)$ computation complexity* comparing with Spearman is $O(nlogn)$.
- The value of Spearman CC is generally higher than Kendall's.

*computation complexity is the amount of resources (mostly about the computation time and memory) required to run it.

# Applications

```
> shapiro.test(mtcars$mpg)

Shapiro-Wilk normality test
data:  mtcars$mpg
W = 0.94756, p-value = 0.1229

> shapiro.test(mtcars$disp)

Shapiro-Wilk normality test
data:  mtcars$disp
W = 0.92001, p-value = 0.02081

> cor.test(mtcars$mpg, mtcars$disp, method = "pearson")

Pearson's product-moment correlation
data:  mtcars$mpg and mtcars$disp
t = -8.7472, df = 30, p-value = 9.38e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9233594 -0.7081376
sample estimates:
       cor
-0.8475514

> cor.test(mtcars$mpg, mtcars$disp, method = "spearman")

Spearman's rank correlation rho
data:  mtcars$mpg and mtcars$disp
S = 10415, p-value = 6.37e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.9088824

> cor.test(mtcars$mpg, mtcars$disp, method = "kendall")

Kendall's rank correlation tau
data:  mtcars$mpg and mtcars$disp
z = -6.1083, p-value = 1.007e-09
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.7681311
```

```
> shapiro.test(mtcars$mpg)

Shapiro-Wilk normality test
data:  mtcars$mpg
W = 0.94756, p-value = 0.1229

> shapiro.test(mtcars$qsec)

Shapiro-Wilk normality test
data:  mtcars$qsec
W = 0.97325, p-value = 0.5935

> cor.test(mtcars$mpg, mtcars$qsec, method = "pearson")

Pearson's product-moment correlation
data:  mtcars$mpg and mtcars$qsec
t = 2.5252, df = 30, p-value = 0.01708
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08195487 0.66961864
sample estimates:
     cor
0.418684

> cor.test(mtcars$mpg, mtcars$qsec, method = "spearman")

Spearman's rank correlation rho
data:  mtcars$mpg and mtcars$qsec
S = 2908.4, p-value = 0.007056
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4669358

> cor.test(mtcars$mpg, mtcars$qsec, method = "kendall")

Kendall's rank correlation tau
data:  mtcars$mpg and mtcars$qsec
z = 2.5165, p-value = 0.01185
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3153652
```