Class Notes: STAT 501
Nonparametrics & Log-Linear Models
Kendall Correlation Coefficient, Spearman Corrlation Coefficient, Cohen’s Kappa

Da Kuang

University of Pennsylvania

Contents

1 Kendall Correlation Coefficient

1.1 Background

1.1.1 Independence

Two random variables X and Y are independent if and only if, for any two sets of real numbers, A and B,

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B)$$

Equivalently, X and Y are independent if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

Independent means that knowing the value of one does not change the distribution of the other.

Random variables that are not independent are said to be dependent.

In practice, we usually decide the dependence by common sense and prior knowledge, instead of the definition. Then the definition and properties are applied to the model.

1.1.2 Correlation

If X and Y are jointly distributed random variables with expectations μ_X and μ_Y, respectively. The covariance of X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)].$$

If X and Y are independent, then $\mathbf{Cov}(X, Y) = 0$. But the converse is not necessary to be true. There could be non-linear relationship between X and Y.

Suppose $\mathbf{Var}(X)$ and $\mathbf{Var}(Y)$ are both non-zero. The correlation of X and Y is

$$\mathbf{Corr}(X,Y) \equiv \rho = \frac{\mathbf{Cov}(Y,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}$$

1.1.3 Pearson Correlation Coefficient

The definition of correlation is about random variables. Now suppose we have a data with sample size n. Then the sample correlation is actually the **Pearson Correlation Coefficient**

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The Example R code is as follows. Note that for Person correlation, the hypothesis test is based on t-distribution.

```
x=c(44.4,45.9, 41.9,53.3,44.7,44.1,50.7,45.2,60.1)
y=c(2.6,3.1,2.5, 5.0,3.6,4.0,5.2,2.8, 3.8)
n=length(x)
cor(x,y)
cor(x,y,method="pearson")
# [1] 0.5711816
mx=mean(y); my=mean(y)
sum((x-mx)*(y-my))/(sd(x)*sd(y)*(n-1))
# [1] 0.5711816
cor.test(x,y,method="pearson",alternative="greater")
#  Pearson’s product-moment correlation

# data: x and y
# t = 1.8411, df = 7, p-value = 0.05409
# alternative hypothesis: true correlation is greater than 0
# 95 percent confidence interval:
# -0.02223023 1.00000000
# sample estimates:
# cor
# 0.5711816
cor.test(x,y,method="pearson",alternative="less")
# Pearson’s product-moment correlation

# data: x and y
# t = 1.8411, df = 7, p-value = 0.9459
# alternative hypothesis: true correlation is less than 0
# 95 percent confidence interval:
# -1.0000000 0.8669786
# sample estimates:
# cor
# 0.5711816
cor.test(x,y,method="pearson",alternative="two.sided")
# Pearson’s product-moment correlation

# data: x and y
# t = 1.8411, df = 7, p-value = 0.1082
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
# -0.1497426 0.8955795
# sample estimates:
# cor
# 0.5711816
pr=cor(x,y); stat=pr*sqrt((n-2)/(1-pr^2))
c(stat , pt(stat,n-2,lower.tail=F) )
# [1] 1.84108264 0.05408653
```

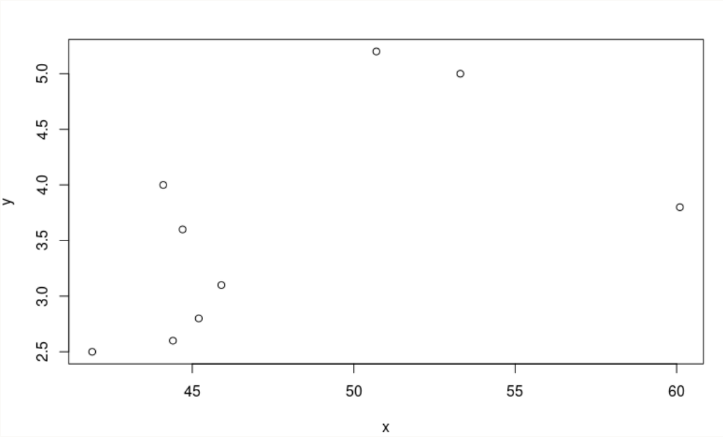

Figure 1: Dot Plot in Peason Example

1.2 Assumption

We suppose that the Data is a random sample from a bivariate population with n subjects,

$$(X_1, Y_1), \cdots, (X_n, Y_n).$$

We look for the statistical relationship between the two variables in the bivariate structure.

- Whether the two variables are independent.
- If not independent, what are the type and degree of dependency.

Assume $(X_1, Y_1), \cdots, (X_n, Y_n)$ are a random sample from a continuous bivariate population. That is (X_i, Y_i) iid from a continuous bivariate distribution.

1.3 Kendall Correlation Coefficient

Suppose we have

- $F_{X,Y}$: the joint distribution function of the (X, Y) pairs.
- F_X: the marginal distribution function of X.
- F_Y: the marginal distribution function of Y.

The null hypothesis is

$$H_0 : F_{X,Y}(x, y) = F_X(x)F_Y(y), \text{ for all } (x, y) \text{ pairs.}$$

Kendall population correlation coefficient

$$\begin{aligned}\tau =& 2\mathbf{P}[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1\\ =& \mathbf{P}[(Y_2 - Y_1)(X_2 - X_1) > 0] - \mathbf{P}[(Y_2 - Y_1)(X_2 - X_1) < 0]\end{aligned}$$

If $\tau > 0$, then it is more likely that $\{X_2 > X_1$ and $Y_2 > Y_1\}$ or $\{X_2 < X_1$ and $Y_2 < Y_1\}$ occurs than either of the complementary events $\{X_2 > X_1$ and $Y_2 < Y_1\}$ or $\{X_2 < X_1$ and $Y_2 > Y_1\}$.

Thus, if $\tau > 0$, it is more likely that the change from X_1 to X_2 has the same (rather than opposite) sign as that from Y_1 to Y_2. It is reasonable to interpret this type of relationship between X and Y as indicative of a positive association (as measured by τ).

Similarly, $\tau < 0$ may reasonably be interpreted as indicative of a negative association (as measured by τ) between X and Y.

Note that

$$\mathbf{P}[(Y_2 - Y_1)(X_2 - X_1) > 0] = \mathbf{P}(X_2 > X_1, Y_2 > Y_1) + \mathbf{P}(X_2 < X_1, Y_2 < Y_1)$$

If X and Y are independent, we have $\tau = 0$, since

$$\begin{aligned}
&\mathbf{P}(X_2 > X_1, Y_2 > Y_1)\\
=&\mathbf{P}(X_2 > X_1)\mathbf{P}(Y_2 > Y_1)\\
=&\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}\\
&\mathbf{P}(X_2 < X_1, Y_2 < Y_1)\\
=&\mathbf{P}(X_2 < X_1)\mathbf{P}(Y_2 < Y_1)\\
=&\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}
\end{aligned}$$

Note that $\tau = 0$ does not necessarily imply that X and Y are independent.

1.4 Kendall Sample Correlation Statistics K

Given $i \le i \le j \le n$, for each pair of bivariate object,

$$Q[(X_i, Y_i), (X_j, Y_j)] = \begin{cases} 1, \text{ if } (Y_j - Y_i)(X_j - X_i) > 0 \\ -1, \text{ if } (Y_j - Y_i)(X_j - X_i) < 0. \end{cases}$$

Define sign statistics based on the $n(n-1)/2$ paired objects.

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i), (X_j, Y_j)]$$

1.5 Hypothesis Test

The null hypothesis is $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, which means for all (x,y) pairs, $\tau = 0$. Let $\widehat{\tau}$ be the average of K. Then

$$\widehat{\tau} = \frac{1}{\frac{n(n-1)}{2}} K$$

The alternative hypothesis are

- $H_1 : \tau > 0$. Reject H_0 if $\widehat{\tau} \geq k_\alpha$
- $H_1 : \tau < 0$. Reject H_0 if $\widehat{\tau} \leq k_\alpha$
- $H_1 : \tau \neq 0$. Reject H_0 if $|\widehat{\tau}| \geq k_\alpha$

Under H_0, we have

- $\mathbf{E}K = 0$
- $\mathbf{Var}K = \frac{n(n-1)(2n+5)}{18}$
- $\frac{K}{\sqrt{\mathbf{Var}K}} \to N(0,1)$

The example code is as follows. Note that Kendall correlation is more conservative since it only check the signs in comparisons. Also, even though we got the significance during the normal approximation test, the size is only 9 so that the result is not reliable.

```
cor(x,y,method="kendal")
# [1] 0.4444444
cor.test(x,y,method = "kendall",
        alternative="g")
# Kendall’s rank correlation tau

# data: x and y
# T = 26, p-value = 0.05972
# alternative hypothesis: true tau is greater than 0
# sample estimates:
```

```
# tau
# 0.4444444
n2=n*(n-1)/2
oy=outer(y,y,"-"); z=oy[lower.tri(oy)]
ox=outer(x,x,"-"); w=ox[lower.tri(ox)]
Tstat=sum(z*w>0)
Tstat
# [1] 26
K=2*Tstat-n2 # K=T- [n(n-1)/2-T] ;
K/n2
# [1] 0.4444444
sdk=sqrt(2*n2*(2*n+5)/18)
K/sdk
# [1] 1.668115
pnorm(K/sdk,lower.tail=F)
# [1] 0.04764642
cor.test(x,y,method="kendall",exact=F,
        alternative="g")
# Kendall's rank correlation tau

# data: x and y
# z = 1.6681, p-value = 0.04765
# alternative hypothesis: true tau is greater than 0
# sample estimates:
# tau
# 0.4444444
```

2 Spearman Correlation Coefficient