Lecture 6

Instructor: Prof. Bowen Gang *Scribes: Yize Wang, Jingyi Zhou*

6.1 Review: Kendall's Tau

Last class we discussed Kendall's τ as a non-parametric measure of associations. It is defined as

$$P_c - P_d$$

where

$$\begin{aligned} P_c &= P\left[(X_i - X_j)(Y_i - Y_j) > 0\right] \ i.e., \ concordance \\ P_d &= P\left[(X_i - X_j)(Y_i - Y_j) < 0\right] \ i.e., \ discordance \end{aligned}$$

6.2 More on Kendall's Tau

6.2.1 Estimate Kendall's Tau

Given $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, how to estimate Kendall's τ? Consider

$$A_{ij} = sgn\,(X_j - X_i)\, sgn\,(Y_j - Y_i)$$

where

$$sgn(u) = \begin{cases} 1, \ if \ u > 0 \\ -1, \ if \ u < 0 \\ 0, \ if \ u = 0 \end{cases}$$

Note that here A_{ij} is a discrete random variable and its p.m.f. is given by

$$f_{A_{ij}}(a_{ij}) = \begin{cases} P_c & , \ a_{ij} = 1 \\ P_d & , \ a_{ij} = -1 \\ 1 - P_c - P_d & , \ a_{ij} = 0 \end{cases}$$

Then the expectation is

$$\begin{aligned} \mathbb{E}\left[A_{ij}\right] &= P_c \cdot 1 + P_d \cdot (-1) + (1 - P_c - P_d) \cdot 0 \\ &= P_c - P_d = \tau \end{aligned}$$

Therefore, we can calculate the mean of all A_{ij}'s and that would be an unbiased estimator for τ. In other words, we can estimate τ using

$$T = \frac{\sum\sum_{1 \leqslant i < j \leqslant n} A_{ij}}{\binom{n}{2}}$$

We call this Kendall's sample τ coefficient. We have another observation: recall Pearson sample correlation is given by

$$\frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left[\sum \left(X_i - \bar{X}\right)^2\right]^{\frac{1}{2}} \left[\sum \left(Y_i - \bar{Y}\right)^2\right]^{\frac{1}{2}}}$$

We can let

$$\begin{aligned} u_{ij} &= sgn\left(X_j - X_i\right) \\ v_{ij} &= sgn\left(Y_j - Y_i\right) \end{aligned}$$

Then we have

$$\begin{aligned} A_{ij} &= sgn\left(X_j - X_i\right) sgn\left(Y_j - Y_i\right) \\ &= u_{ij} v_{ij} \end{aligned}$$

Assume $X_i \neq X_j$ and $Y_i \neq Y_j$, $\forall i, j$, then it is also immediate that

$$\sum_{i=1}^{n} \sum_{i=1}^{n} u_{ij}^2 = \sum_{i=1}^{n} \sum_{i=1}^{n} v_{ij}^2 = n(n-1)$$

where when $i = j$, $u_{ij} = v_{ij} = 0$. Finally we get

$$\begin{aligned} T = \frac{\sum\sum_{1 \leqslant i < j \leqslant n} A_{ij}}{\binom{n}{2}} &= \frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}}{\frac{n(n-1)}{2}} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} v_{ij}}{n(n-1)} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} v_{ij}}{\left[\left(\sum\sum u_{ij}^2\right)\left(\sum\sum v_{ij}^2\right)\right]^{\frac{1}{2}}} \end{aligned}$$

which is quite similar with the expression of Pearson sample correlation. The main idea is: previously $X_i - \bar{X}$ and $Y_i - \bar{Y}$ depend on the distributions of X and Y. So we change them to the sign function and do similar things to the denominator. This is how we transform parametric stuff to non-parametric stuff. Somehow we suppressed the magnitude.

6.2.2 Asymptotic Null distribution of Tau

Here null distribution will be based on $H_0 : X, Y$ are independent. In parametric setting, we might calculate the correlation between X and Y, and when it is too small, we may reject the null. Here in non-parametric setting, we can take the Kendall's τ coefficient of X and Y and when it differs significantly from zero, we will have strong evidence to say that X and Y are not independent. For the null distribution when X and Y are independent, we have

$$Z = \frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} T \xrightarrow{D} N(0,1)$$

Accordingly we can do hypothesis testing and calculate the p-value.

6.3 Spearman Correlation

Spearman Correlation is another measure of association. It is even more straightforward than Kendall's τ. Recall Pearson again:

$$\hat{\rho} = \frac{\sum\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left[\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2\right]^{\frac{1}{2}}}$$

In Kendall's τ we replace the nominator with sign functions. Another way that we can make it distribution free is to use ranks as follows:

$$\begin{aligned} R_i = &\ \textit{rank of } X_i \textit{ among the } X \textit{ sample} \\ S_i = &\ \textit{rank of } Y_i \textit{ among the } Y \textit{ sample} \end{aligned}$$

and put them into $\hat{\rho}$, then it becomes

$$\hat{\rho}^S = \frac{\sum_{i=1}^n \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\left[\sum \left(R_i - \bar{R}\right)^2 \sum \left(S_i - \bar{S}\right)^2\right]^{\frac{1}{2}}}$$

Note that $\bar{R} = \bar{S} = \frac{n+1}{2}$ and $\sum \left(R_i - \bar{R}\right)^2 = \left(1 - \frac{n+1}{2}\right)^2 + \left(2 - \frac{n+1}{2}\right)^2 + \cdots = \frac{n(n^2-1)}{12} = \sum \left(S_i - \bar{S}\right)^2$, we can simplify the expression of $\hat{\rho}^S$.

Define

$$D_i = R_i - S_i = \left(R_i - \bar{R}\right) - \left(S_i - \bar{S}\right)$$

then

$$\sum D_i^2 = \sum \left(R_i - \bar{R}\right)^2 + \sum \left(S_i - \bar{S}\right)^2 - 2\sum \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)$$

After rearranging the terms, we get

$$\sum \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right) = \frac{n(n^2-1)}{12} - \frac{1}{2}\sum D_i^2$$

So,

$$\begin{aligned} \hat{\rho}^S &= \frac{\sum_{i=1}^n \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum \left(R_i - \bar{R}\right)^2 \sum \left(S_i - \bar{S}\right)^2}} \\ &= \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}\sum D_i^2}{\frac{n(n^2-1)}{12}} \\ &= 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)} \end{aligned}$$

where Spearman Correlation also satisfies all the 7 requirements discussed last class. The corresponding null distribution of Spearman correlation is given by

$$\frac{\sqrt{n-2}}{1-R^2} \to t_{n-2}$$

where $R =$ *Spearman Correlation*

6.4 Another Measure of Association of Two Samples

We also use the idea of converting something to ranks. Let s_i denote the rank of Y observation which is paired with the i-th smallest X observation. Then the random sample of pairs of ranks are $(1, s_1), (2, s_2), (3, s_3), ...$ For example, if we have 3 pairs

$$(1.1, 2.1), (1.5, 2.0), (1.3, 3.0)$$

then we can transform it into

$$(1, s_1 = 2), (3, s_3 = 1), (2, s_2 = 3)$$

where the first entries are the ranks of X_i's in X observations. Let $\xi_i = \mathbb{E}\left[U_{(i)}\right]$ where $U_{(i)}$ is the i-th order statistic in a sample of n from $N(0, 1)$, which is similar to the Terry-Hoeffding normal score test. Consider $(\xi_1, \xi_{s_1}), (\xi_2, \xi_{s_2}), ..., (\xi_n, \xi_{s_n})$ Then we can have the following measure:

$$R = \frac{\sum_{i=1}^{n} \xi_i \xi_{s_i}}{\sum_{i=1}^{n} \xi_i^2}$$

So here instead of replacing $X_i - \bar{X}$ with ranks, we replace it with the i-th order statistic from $N(0, 1)$

6.5 Measures of Association in Multiple Classifications

Previously we talked about association in two classes. Here our basic setup is as follows

$$X_{ij} = \mu + \underbrace{\beta_i}_{row\ effect} + \underbrace{\theta_j}_{column\ effect} + \underbrace{\epsilon_{ij}}_{noise}$$

To be more specific, here we have a table to show the idea.

	Treatment				
block	1	2	3	$\cdots$	J
1	X_{11}	X_{12}	X_{13}	$\cdots$	X_{1J}
2	X_{21}	X_{22}	X_{23}	$\cdots$	X_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	X_{I1}	X_{I2}	X_{I3}	$\cdots$	X_{IJ}

For example, you can imagine that for each row, we totally have I plants and for each column, we have J fertilizers. Then X_{ij} denotes the growth of i-th plant with j-th fertilizer.

Another example is that we have J athletes and I referees. Then β_i is the effect of referees/plants and θ_j is the effect of athletes/fertilizers. Here μ is just a shift in order to have all β_i, θ_j and ϵ_{ij} mean zero.

The goal here is to test the column effect. To be more specific, our null hypothesis is given by

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J$$

The alternative is that at least one θ_j is different from other θ_j's.

6.5.1 Friedman's Two Way ANOVA

We have been familiar with the common trick of transforming X_{ij}'s into ranks. Similar here, let R_{ij} be the rank of treatment j when observed in block. Then our table becomes

	Treatment					
block	1	2	3	$\cdots$	n	**Total**
1	R_{11}	R_{12}	R_{13}	$\cdots$	R_{1n}	$\frac{n(n+1)}{2}$
2	R_{21}	R_{22}	R_{23}	$\cdots$	R_{2n}	$\frac{n(n+1)}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	R_{k1}	R_{k2}	R_{k3}	$\cdots$	R_{kn}	$\frac{n(n+1)}{2}$
Total	R_1	R_2	R_3	$\cdots$	R_n	$\frac{k(n+1)n}{2}$

Our test statistic to consider is then given by

$$S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2$$

Here we know that the average of R_j is $\frac{k(n+1)}{2}$ and if every/most R_j is close to $\frac{k(n+1)}{2}$, then we will have strong evidence to say, for example, the scores for each athlete given by each referee is random. Otherwise, we will reject the null that there are significant column effects. The S statistic looks similar to Chi-square test statistic. So unsurprisingly the asymptotic distribution is given by

$$Q = \frac{12S}{kn(n+1)} \to \chi^2_{n-1}$$

Note that here our test is one-sided and only when Q or S is too big, we will reject the null hypothesis. Here is a testing procedure and next we will discuss some real 'measures' of association or correlation which is similar to Kendall's τ.

6.5.2 The Coefficient of Concordance for k Sets of Rankings of n Objects

6.5.2.1 Definition of the Kendall's Coefficient of Concordance

Think about is similar as before: we have k judges and n objects/athletes. We want a single measure of overall association. Idealistically, if there is perfect agreement, say, the rankings of athlete 1, 2 and 3 are respectively 3, 1 and 2, we will then have

	Treatment					
block	1	2	3	$\cdots$	n	**Total**
1	3	1	2	$\cdots$	R_{1n}	$\frac{n(n+1)}{2}$
2	3	1	2	$\cdots$	R_{2n}	$\frac{n(n+1)}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	3	1	2	$\cdots$	R_{kn}	$\frac{n(n+1)}{2}$
Total	$R_1 = 3k$	$R_2 = k$	$R_3 = 2k$	$\cdots$	R_n	$\frac{k(n+1)n}{2}$

Then the set of column total would be

$$\{R_1, R_2, ..., R_n\} = \{k, 2k, ..., nk\}$$

Under perfect agreement,

$$S_{\text{perfect agreement}} = \sum_{j=1}^{n} \left[jk - \frac{k(n+1)}{2} \right]^2 = k^2 n \frac{n^2-1}{12}$$

The actual agreement you observe is

$$S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2$$

We claim that the maximum value of S is $k^2 n^2 \frac{n^2-1}{12}$ and this should be intuitive: the maximum value of S exists when there is perfect agreement. Therefore, by dividing the test statistic by its possible maximum value, we have

$$0 \leqslant \frac{12S}{k^2 n(n^2-1)} \leqslant 1$$

This is a measure of overall concordance: Kendall's Coefficient of Concordance.

$$W = \frac{12S}{k^2 n(n^2-1)}$$

6.5.2.2 Another Interpretation of Kendall's Coefficient of Concordance

Here is another interpretation of W: W is related to the average of rank correlation. The rank correlation is calculated by

$$r_{i,m} = \frac{12}{n(n^2-1)} \sum_{j=1}^{n} \left(r_{ij} - \frac{n+1}{2} \right) \left(r_{mj} - \frac{n+1}{2} \right), \forall i \neq m$$

where essentially $\left(r_{ij} - \frac{n+1}{2}\right)$ is analogous to $\left(X - \bar{X}\right)$. Here $r_{i,m}$ can be viewed as the agreement level between the i-th judge and the m-th judge. Then the average rank correlation is

$$r_{av} = \frac{\sum \sum_{1 \leqslant i < m \leqslant k} r_{i,m}}{\binom{k}{2}}$$

Finally one can show that

$$W = r_{av} + \frac{1 - r_{av}}{k}$$

So $W = 1$ when $r_{av} = 1$.

6.5.2.3 How to Estimate the True Preferential Ordering

Suppose we have rejected the null hypothesis that $\theta_1, ..., \theta_n$ are not all equal from the test statistic $S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2$. Then the next question we are interested in would be, say, which athletes perform better? For the most naive way, we can rank them according to their column sums. However, why this kind of estimation is sensible or reasonable?

In fact, this estimation is the best in the sense that if the coefficient of rank correlation is calculated between this estimated ranking and each of the k observed rankings, the average of these k correlation coefficients is a

maximum. The proof is as follows. Let $r_{e1}, r_{e2}, ..., r_{en}$ be any estimate of the true preferential ordering. r_{ej} is the estimated rank of object j. Let $R_{e,i}$ be the rank correlation coefficient between this estimated ranking and the ranking of observer(judge) i. Then the average rank correlation is given by

$$\sum_{i=1}^{k} \frac{R_{e,i}}{k} = \frac{12 \sum_{i=1}^{k} \sum_{j=1}^{n} \left(r_{e_j} - \mu\right)\left(r_{ij} - \mu\right)}{kn(n^2-1)} = 12 \sum_{j=1}^{n} \frac{\left(r_{ej} - \mu\right)\left(r_j - k\mu\right)}{kn(n^2-1)}$$

where μ is the average rank $\frac{n+1}{2}$ and r_j is the column sum of the j-th column. You can further simplify it as

$$\frac{12 \sum r_{ej} r_j}{kn(n^2-1)} - \frac{3(n+1)}{n-1}$$

In order to maximize the above function, we only need to maximize the first term: $\sum r_{ej} r_j$, which is maximized when $r_{ej} \propto r_j$.

This estimate is also the best in the least-square sense. If r_{ej} is the estimated rank of object j and the estimate is true, the measure of error will be

$$\underbrace{r_j}_{actually\ observe} \quad -k \quad \underbrace{r_{ej}}_{true\ ranking}$$

where r_j will fluctuate because of different judges and the actual performance of certain athlete on certain day. You can think of the overall error as a sum of square

$$\begin{aligned} \sum_{j=1}^{k} \left[r_j - k r_{ej}\right]^2 &= \underbrace{\sum_{j=1}^{n} r_j^2 + k^2 \sum_{j=1}^{n} r_{ej}^2}_{constant} - 2k \sum_{j=1}^{n} r_j r_{ej} \\ &= C - 2k \sum_{j=1}^{n} r_j r_{ej} \end{aligned}$$

Then in order to minimize the error, we need to maximize $\sum_{j=1}^{n} r_j r_{ej}$, which is maximized when r_j and r_{ej} are perfectly correlated.

6.5.3 Kendall's Tau for Partial Correlation

6.5.3.1 Definition

Partial correlation coefficients measure association in the conditional probability distribution of two variables given one or more other variables. For example, we have three variables: age, height and weight. Obviously weight will not have impact on age or height. However, age will have impact on height and weight, and height will have impact on weight. The age variable will have influence on weight either directly or through height variable. Therefore, by conditional on height, we can measure the direct impact of age on weight.

Assume we are given m independent triplets,

$$(X_i, Y_i, Z_i),\ i = 1, ..., m$$

which are from a trivariate population. Also we assume that the marginal distribution of X, Y and Z are continuous. Define

$$\begin{aligned} u_{ij} &= sgn\left(X_j - X_i\right) \\ v_{ij} &= sgn\left(Y_j - Y_i\right) \\ w_{ij} &= sgn\left(Z_j - Z_i\right) \end{aligned}$$

which are somehow similar to the definition of Kendall's tau. Let

$$n(u,v,w) = \ \#\ of\ values\ of\ i,j\ s.t.\ u_{ij}=u, v_{ij}=v, w_{ij}=w$$

Also let

$$\begin{aligned}
X_{11} &= n(1,1,1)\ i.e.,\ X,Y,Z\ concordant\\
X_{22} &= n(-1,-1,1)\ i.e.,\ X,Y\ concordant,\ Z\ discordant\\
X_{12} &= n(-1,1,1)\ i.e.,\ X,Y\ discordant,\ Y,Z\ concordant\\
X_{21} &= n(1,-1,1)\ i.e.,\ X,Z\ concordant,\ Y\ discordant
\end{aligned}$$

Then we have the following table

	ranking X		
ranking Y	concordant with Z	discordant with Z	**Total**
concordant with Z	X_{11}	X_{12}	$X_{1\cdot}$
discordant with Z	X_{21}	X_{22}	$X_{2\cdot}$
Total	$X_{\cdot 1}$	$X_{\cdot 2}$	$N=\binom{m}{2}$

The partial correlation between X and Y when Z is held constant is then defined as

$$T_{XY,Z} = \frac{X_{11}X_{22} - X_{12}X_{21}}{(X_{\cdot 1}X_{\cdot 2}X_{1\cdot}X_{2\cdot})^{\frac{1}{2}}}$$

where X_{11} and X_{22} correspond to situations where X and Y are concordant, X_{12} and X_{21} correspond to situations where X and Y are discordant. You can also show that

$$-1 \leqslant T_{XY,Z} \leqslant 1$$

Recall Kendall's tau has something to do with the Pearson correlation. Here is also the case:

$$\begin{aligned}
\binom{m}{2}\underbrace{T_{XY}}_{Kendall's\ \tau} &= \sum\sum A_{ij} = \underbrace{(X_{11}+X_{22})}_{concordance} - \underbrace{(X_{12}+X_{21})}_{discordance}\\
\binom{m}{2}T_{XZ} &= (X_{11}+X_{21}) - (X_{22}+X_{12})\\
\binom{m}{2}T_{YZ} &= (X_{11}+X_{12}) - (X_{22}+X_{21})
\end{aligned}$$

where

$$\binom{m}{2} = X_{11}+X_{12}+X_{21}+X_{22} = N$$

Then we have

$$\begin{aligned}
1-T_{XZ}^2 &= \frac{4\,(X_{11}+X_{21})\,(X_{12}+X_{22})}{N^2} = \frac{4X_{\cdot 1}X_{\cdot 2}}{N^2}\\
1-T_{YZ}^2 &= \frac{4\,(X_{11}+X_{12})\,(X_{22}+X_{21})}{N^2} = \frac{4X_{1\cdot}X_{2\cdot}}{N^2}\\
N^2T_{XY} &= \underbrace{[(X_{11}+X_{22})-(X_{12}+X_{21})]}_{N\cdot T_{XY}}\underbrace{[(X_{11}+X_{12})+(X_{21}+X_{22})]}_{N}
\end{aligned}$$

We have the following fact

$$N^2 \left(T_{XY} - T_{XZ}T_{YZ}\right) = 4\left(X_{11}X_{22} - X_{12}X_{21}\right)$$

Therefore, finally we get

$$\begin{aligned} T_{XY,Z} &= \frac{X_{11}X_{22} - X_{12}X_{21}}{\left(X_{\cdot 1}X_{\cdot 2}X_{1\cdot}X_{2\cdot}\right)^{\frac{1}{2}}} \\ &= \frac{T_{XY} - T_{XZ}T_{YZ}}{\left[\left(1 - T_{XZ}^2\right)\left(1 - T_{YZ}^2\right)\right]^{\frac{1}{2}}} \end{aligned}$$

Remember partial correlation is given by

$$\rho_{XY,Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

which is entirely parallel with our $T_{XY,Z}$. Note that this result is analogous to regression of Z on X and Y respectively.

6.5.3.2 Interpretation of Partial Correlation

Say we are given (X_i, Y_i, Z_i) for $i = 1, ..., n$ and we want to regress X on Z, i.e., use Z to estimate X. Then we will have an error term e_{Xi}. Similar we have E_{Yi}. The sample partial correlation is then given by

$$\frac{N\sum_{i=1}^N e_{Xi}e_{Yi} - \sum_{i=1}^N e_{Xi}\sum_{i=1}^N e_{Yi}}{\sqrt{N\sum_{i=1}^N e_{Xi}^2 - \left(\sum_{i=1}^N e_{Xi}\right)^2}\sqrt{N\sum_{i=1}^N e_{Yi}^2 - \left(\sum_{i=1}^N e_{Yi}\right)^2}}$$

where the nominator looks quite similar to $Cov\left(e_X, e_Y\right) = \mathbb{E}\left[e_X e_Y\right] - \mathbb{E}\left[e_X\right]\mathbb{E}\left[e_Y\right]$ and the denominator looks quite similar to variances. For example, if e_X and e_Y both tend to be negative, we can conclude that we tend to overestimate both X and Y by using Z. Therefore, there is still some left correlation between X and Y conditional on Z.

Example. *Suppose* $\mathrm{Var}(\epsilon_X) = \mathrm{Var}(\epsilon_Y) = 1$ *and*

$$\begin{cases} X &= Z + \epsilon_X \\ Y &= X + Z + \epsilon_Y \end{cases}$$

Then $Y = 2Z + \epsilon_X + \epsilon_Y$. *We can immediately get*

$$\mathrm{corr}(\epsilon_X, \epsilon_X + \epsilon_Y) = \frac{\mathrm{cov}(\epsilon_X, \epsilon_X + \epsilon_Y)}{\sqrt{\mathrm{Var}(\epsilon_X)}\sqrt{\mathrm{Var}(\epsilon_X + \epsilon_Y)}} = \frac{1+0}{1\cdot\sqrt{2}} = \frac{1}{\sqrt{2}}$$

which is larger than the following situation:

$$\begin{cases} X &= Z + \epsilon_X \\ Y &= 2X + Z + \epsilon_Y = 3Z + 2\epsilon_X + \epsilon_Y \end{cases} \Rightarrow \mathrm{corr}(\epsilon_X, 2\epsilon_X + \epsilon_Y) = \frac{2+0}{\sqrt{1}\sqrt{5}} = \frac{2}{\sqrt{5}}$$