Nonparametric Statistics Notes

Chapter 5: Some Methods Based on Ranks

Jesse Crawford

Department of Mathematics
Tarleton State University

Outline

1. Section 5.1: Two Independent Samples
2. Section 5.2: Several Independent Samples
3. Section 5.7: Matched Pairs
4. Section 5.4: Measures of Rank Correlation
5. Section 5.8: Several Related Samples

The Mann-Whitney-Wilcoxon Test

- Assumptions
  - Random sample $X_1, \ldots, X_n$ from Population 1
  - Random sample $Y_1, \ldots, Y_m$ from Population 2
  - Samples are statistically independent of each other
  - Measurement scale is at least ordinal
- H_0 : The probability distributions for populations 1 and 2 are identical.
- Rank all observations
  - Rank $1 =$ smallest
  - Rank $n + m =$ largest.
  - $R(X_i) =$ rank of X_i
  - $R(Y_j) =$ rank of Y_j
- Test statistic: $T = \sum_{i=1}^{n} R(X_i)$

The Mann-Whitney-Wilcoxon Test

- Test statistic: $T = \sum_{i=1}^{n} R(X_i)$

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

- Exact distribution of T can be found.
- Asymptotically, $T_1 \approx N(0,1)$
- Alternative test statistic: $W = \sum_{i=1}^{n} R(X_i) - \frac{n(n+1)}{2}$

Outline

1. Section 5.1: Two Independent Samples
2. Section 5.2: Several Independent Samples
3. Section 5.7: Matched Pairs
4. Section 5.4: Measures of Rank Correlation
5. Section 5.8: Several Related Samples

The Kruskal-Wallis Test

- Assumptions
  - Random sample $X_{11}, \ldots, X_{1n_1}$ from Population 1
  - Random sample $X_{21}, \ldots, X_{2n_1}$ from Population 2

$$\vdots$$

  - Random sample $X_{k1}, \ldots, X_{kn_k}$ from Population k
  - Samples are statistically independent of each other
  - Measurement scale is at least ordinal
- H_0 : The probability distributions for all k populations are identical.

Outline

1. Section 5.1: Two Independent Samples
2. Section 5.2: Several Independent Samples
3. Section 5.7: Matched Pairs
4. Section 5.4: Measures of Rank Correlation
5. Section 5.8: Several Related Samples

The Wilcoxon Signed Ranks Test

- Random sample of pairs: $(X_1, Y_1), \ldots, (X_{n'}, Y_{n'})$.
- $\mathrm{H}_0 : E(X_i) = E(Y_i)$
- Discard ties $(X_i = Y_i)$.
- Now we have: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- $D_i = Y_i - X_i$
- Assumptions:
  - Distribution of each D_i is symmetric
  - The D_i's are statistically independent and have the same mean.
- Rank all pairs from 1 to n based on absolute differences $|D_i|$.
  - $R_i =$ Rank assigned to $|D_i|$, if $D_i > 0$
  - $R_i = -$(Rank assigned to $|D_i|$), if $D_i < 0$
- Test statistic: $T^+ = \sum (R_i \mid D_i > 0)$

Outline

1. Section 5.1: Two Independent Samples
2. Section 5.2: Several Independent Samples
3. Section 5.7: Matched Pairs
4. Section 5.4: Measures of Rank Correlation
5. Section 5.8: Several Related Samples

Pearson's Correlation Coefficient

Definition

- Consider a random sample of ordered pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Pearson's correlation coefficient is

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

- $-1 \leq r \leq 1$
- Values of r near 1 suggest a strong positive relationship.
- Values of r near -1 suggest a strong negative relationship.
- Values of r near 0 suggest a weak or nonlinear relationship.

Pearson's Correlation Coefficient

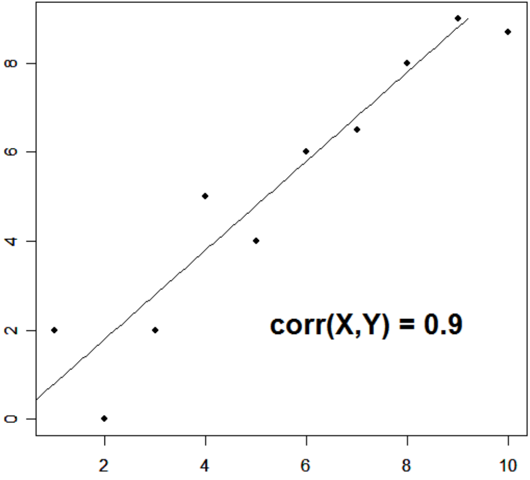
Definition

- Consider a random sample of ordered pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Pearson's correlation coefficient is

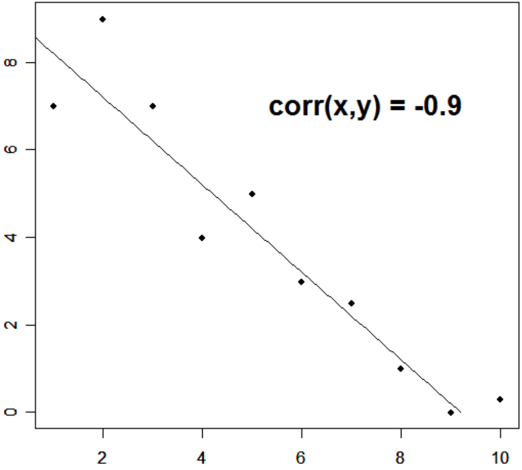
$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2)}}.$$

- $-1 \leq r \leq 1$
- Values of r near 1 suggest a strong positive relationship.
- Values of r near -1 suggest a strong negative relationship.
- Values of r near 0 suggest a weak or nonlinear relationship.

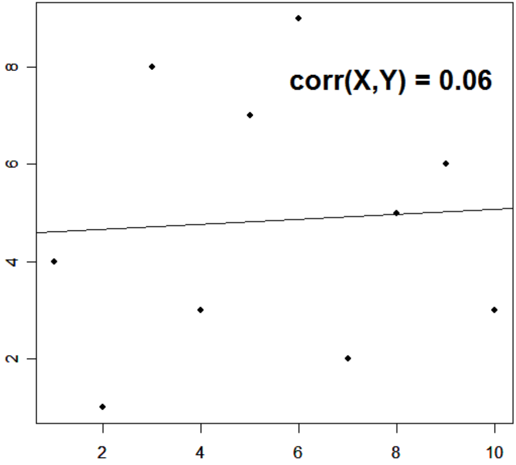
Strong Positive Correlation



Strong Negative Correlation



Virtually No Correlation



Spearman's Correlation Coefficient

Definition

- Consider a random sample of ordered pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- $R(X_i)$ is the rank of X_i among $X_1, \ldots, X_n$.
- $R(Y_i)$ is the rank of Y_i among $Y_1, \ldots, Y_n$.
- Spearman's correlation coefficient $=$ Pearson's coefficient applied to the ranks.

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n(\frac{n+1}{2})^2}{\sqrt{\left(\sum_{i=1}^{n} R(X_i)^2 - n(\frac{n+1}{2})^2\right)\left(\sum_{i=1}^{n} R(Y_i)^2 - n(\frac{n+1}{2})^2\right)}}.$$

- $-1 \leq \rho \leq 1$
- Values of ρ near 1 suggest a strong positive relationship.
- Values of ρ near -1 suggest a strong negative relationship.
- Values of ρ near 0 suggest a weak or nonlinear relationship.

Outline

1. Section 5.1: Two Independent Samples
2. Section 5.2: Several Independent Samples
3. Section 5.7: Matched Pairs
4. Section 5.4: Measures of Rank Correlation
5. Section 5.8: Several Related Samples

The Quade Test for Related Observations

	Treatments			
Block	1	2	$\cdots$	k
1	X_{11}	X_{12}	$\cdots$	X_{1k}
2	X_{21}	X_{22}	$\cdots$	X_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
b	X_{b1}	X_{b2}	$\cdots$	X_{bk}

- **Assumptions**
  - The blocks $1, \ldots, b$ are statistically independent (between blocks, not necessarily within blocks).
  - Within each block i, the variables $X_{i1}, \ldots, X_{ik}$ can be ranked, and it is possible to compute a range between the maximum and minimum values (this holds for quantitative variables).
- **Testing problem:**

$$\mathrm{H}_0 : \text{For each block } i, \text{all rankings of } X_{i1}, \ldots, X_{ik} \text{ are equilikely}$$

$$\mathrm{H}_0 : \text{For every subject, all treatments are equally effective for that subject.}$$

The Quade Test for Related Observations

- **Test Statistic**

$$T_1 = \frac{(b-1)B_1}{A_1 - B_1}$$

- **Null distribution**

$$T_1 \approx F[k-1, (b-1)(k-1)]$$

- **Post-hoc Pairwise Tests.** If the above test rejects the null hypothesis, we do post-hoc tests. There is a statistically significant difference between treatments i and j if

$$|S_i - S_j| > t_{\alpha/2,(b-1)(k-1)} \left[\frac{2b(A_1 - B_1)}{(b-1)(k-1)} \right]^{\frac{1}{2}}$$

Note that these post-hoc tests use the *same* α as the overall Quade Test, with *no* p-value adjustment.

- **Ranks** $R(X_{ij})$ = rank of X_{ij} within block i.
- **Block Ranges**

$$\text{Range(Block } i) = \max_j X_{ij} - \min_j X_{ij}$$

$$Q_i = \text{rank of block } i$$

- **Standardized Ranks**

$$S_{ij} = Q_i \left[R(X_{ij}) - \frac{k+1}{2} \right]$$

$$S_j = \sum_{i=1}^{b} S_{ij}$$

- **Sums of Squares**

$$A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} S_{ij}^2 \qquad \text{(Total SS)}$$

$$B_1 = \frac{1}{b} \sum_{j=1}^{k} S_j^2 \qquad \text{(Treatment SS)}$$

Example

The following table shows the number of bottles sold for five brands of lotion in seven stores. Test the null hypothesis that, within each store, all brands of lotion are equally likely to be purchased.

	Lotion Brand				
Store	A	B	C	D	E
1	5	4	7	10	12
2	1	3	1	0	2
3	16	12	22	22	35
4	5	4	3	5	4
5	10	9	7	13	10
6	19	18	28	37	58
7	10	7	6	8	7

Quade Test for Machine Learning

	ML Regression Model			
Subjects	1	2	$\cdots$	k
1	$\|e_{11}\|$	$\|e_{12}\|$	$\cdots$	$\|e_{1k}\|$
2	$\|e_{21}\|$	$\|e_{22}\|$	$\cdots$	$\|e_{2k}\|$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	$\|e_{n1}\|$	$\|e_{n2}\|$	$\cdots$	$\|e_{nk}\|$

- $e_{ij} = y_i - \hat{y}_{ij} =$ ith residual from jth model
- Test uses absolute values, $|e_{ij}|$