Mann-Whitney-Wilcoxon test: rationale

Assume that we have two samples $\mathbf{X} = (x_1, ..., x_{n1})$ and $\mathbf{Y} = (y_1, ..., y_{n2})$ and we want determine if they come from the same generating distribution, that is we want to test $H_0 : L_x = L_y$ using a nonparametric test, i.e. without assuming one parametric distribution, e.g. Normal, Binomial... Furthermore, assume that L_x and L_y are both continuous distributions and that x_i is independent of y_j for all i, j. One possible solution: the **Mann-Whitney-Wilcoxon** test otherwise known as the Wilcoxon's rank sum test.

Note: alternative versions of the Mann-Whitney-Wilcoxon test exists in case we specify the null hypothesis differently.

Let's uncover the mathematics behind this nonparametric test and its normal approximation. In a second time, we will perform the test and some simulations in the R language.

Basic notation and test statistic

Let $S = (x_1, ..., x_{n1}, y_1, ..., y_{n2})$ denote the pooled sample and $R = rank(S)$, that is the ordered rank of all observation. Let us denote $N = n_1 + n_2$. Finally, let us consider the r.v. R_X, which is the sum of the ranks of the observations from our first sample $x_1, ..., x_{n1}$ in the pooled sample S, defined as

$$R_X = \sum_{i=1}^{n1} R_i$$

The Mann-Whitney-Wilcoxon test statistic is usually given by

$$W = 2R_X - n_1(N + 1)$$

We are only concerned with testing $H_0 : L_x = L_y$.

Distribution of the test statistic and large-sample approximation

The distribution W depends on the sample sizes n_1 and n_2. If we do not have tables which gives us a critical value to test H_0, we can estimate it using simulations. If n_1 and n_2 are large enough, there is a possibility of a 'large-sample' approximation.

Asymptotic approximation: Using the CLT, we can prove that, for larges values of n_1 and n_2, W has the following distribution

$$W \sim N\left(0,\ n_1 n_2 \frac{N+1}{3}\right)$$

We can use the function $pnorm()$ in R to compute an approximate p-value.

Working example

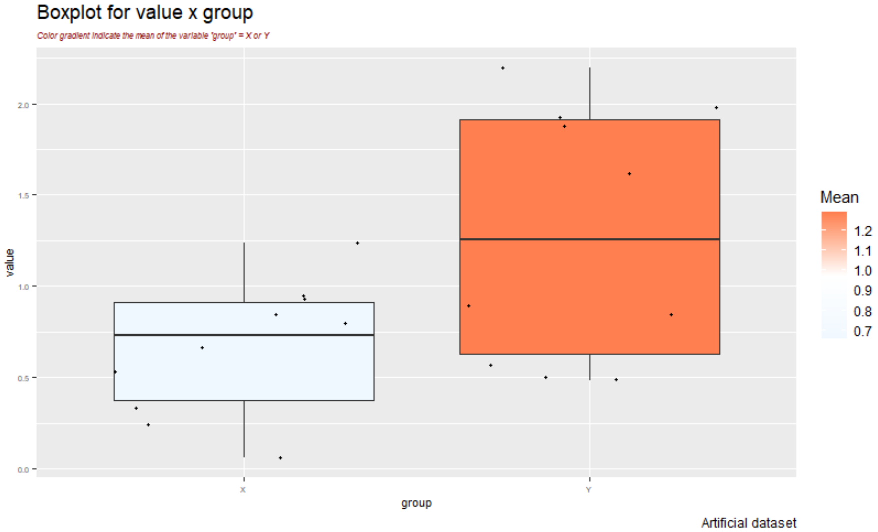
Example: The following values are uniform samples with $\mathbf{X} \sim U_{[0,2]}$ and $\mathbf{Y} \sim U_{[0.2,2.2]}$ of sample sizes $n_1 = n_2 = 10$

$$X = 0.93322789,\ 0.67038191,\ 0.32563512,\ 0.79224003,\ 0.06078346,\\ 0.24176974,\ 0.85233131,\ 1.23571576,\ 0.52641652,\ 0.95264774$$

$$Y = 1.9246380,\ 0.4977573,\ 0.5608602,\ 2.1985466,\ 1.8834824,\\ 0.4852981,\ 0.8896968,\ 1.9792509,\ 0.8396966,\ 1.6180317$$

Using the exact distribution of W, then the asymptotic approximation, what is the conclusion of the test at a level of significance of 5% testing $H_0 : L_x = L_y$, the data is coming from the same generating process or equivalently the temperatures are similar in both cities.

Visualizing the data



Mann-Whitney-Wilcoxon test using inbuilt function in R

```r
# 1. Mann-Whitney-Wilcoxon using inbuilt function in R

set.seed(2023)
X = runif(50, min = 0, max = 2)
Y = runif(50, min = 0.2, max = 2.2)

wilcox.test(X,Y)
# Wilcoxon rank sum exact test
#
# data:  X and Y
# W = 27, p-value = 0.08921
# alternative hypothesis: true location shift is not equal to 0
```

Conclusion: we can NOT reject H_0: 'both data come from the same generating distribution' at a significance level of 5% since for those samples, $p - value > 0.05$.

Asymptotic Mann-Whitney-Wilcoxon test in R

```r
# 2. Asymptotic Mann-Whitney-Wilcoxon using inbuilt function in R
set.seed(2023)
X = runif(10, min = 0, max = 2)
Y = runif(10, min = 0.2, max = 2.2)
S = c(X, Y)
R = rank(S)
Rx = sum(R[1:length(X)])
W = 2*Rx + (12*(N+1))
varW = 50*50*((N+1)/3)

pnorm(W, mean = 0, sd = sqrt(varW), lower.tail = FALSE)
# 0.0008313872

# Conclusion: using the asysmptotic approximation, we reject H0: Lx = Ly that the both
# data come from the same generating distribution.
```

Conclusion: we can reject H_0: 'both data come from the same generating distribution' at a significance level of 5% since for those samples, $p - value < 0.05$.

Expectation of the sum of ranks and sum or squared ranks

We know that, under $H_0 : L_x = L_y$, the R_i's are uniformly distributed on the set $\{1, ..., N\}$. When computing the expectation, it involves the sum of the first N integers, given by the formula $\frac{N(N+1)}{2}$. It can be proven by induction (appendix 1).

So the expectation of the ranks and the squared ranks are respectively given by

$E[R_i] = \frac{1}{N}\sum_{i=1}^{N} R_i = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$

$E[R_i^2] = \frac{1}{N}\sum_{i=1}^{N} R_i^2 = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} = \frac{(N+1)(2N+1)}{6}$

Variance of the ranks

By definition, the variance of the random variable R_i is given by

$$var(R_i) = E[R_i^2] - \big(E[R_i]\big)^2$$

Then replacing by the results that we derived on the previous slide, we get

$$var(R_i) = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12}$$

Since R_i is not independent from R_j, then the variance of the sum is equal to the sum of the variances plus a covariance term.

Covariance term among ranks

We first note that $\sum_{i=1}^{N} R_i$ is a constant and therefore its variance is 0.

$$\underbrace{\text{var}(\sum_{i=1}^{N} R_i)}_{0} = \underbrace{\sum_{i=1}^{N} var(R_i)}_{\frac{N(N^2-1)}{12}} + \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N-1} cov(R_i, R_j)}_{N(N-1)cov(R_i,R_j)} \qquad \text{for } i \neq j.$$

So we have that

$$-\frac{N(N^2-1)}{12} = N(N-1)cov(R_i, R_j)$$
$$-\frac{N(N-1)(N+1)}{12} = N(N-1)cov(R_i, R_j)$$
$$-\frac{(N+1)}{12} = cov(R_i, R_j)$$

Expectation of the r.v R_X

To compute $E[R_X]$, we have:

$$\begin{aligned} E[R_X] &= \sum_{i=1}^{n_1} E[R_i] \\ &= n_1 \frac{N+1}{2} \\ &= n_1 \frac{n_1+n_2+1}{2} \end{aligned}$$

Variance of the r.v R_X

To compute $var(R_X)$, we have:

$$var(R_X) = \sum_{i=1}^{n_1} var(R_i) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_1 - 1} cov(R_i, R_j)$$

where $var(R_i) = \frac{N^2-1}{12}$ and $cov(R_i, R_j) = -\frac{(N+1)}{12}$. By replacing these results in the previous equation, we get

$$\begin{aligned} var(R_X) &= \; n_1 \; \frac{N^2 - 1}{12} + n_1 \; (n_1 - 1)\Big(- \frac{N+1}{12}\Big) \\ &= \; \frac{n_1}{12} \; \Big[N^2 - 1 - (n_1 - 1) \; (N+1)\Big] \\ &= \; \frac{n_1 \; (N+1)}{12}\Big[N - 1 - (n_1 - 1)\Big] \\ &= \; \frac{n_1 \; n_2 \; (N+1)}{12} \end{aligned}$$

Distribution of the r.v R_X

So we have that

$$R_X \sim N\left(n_1 \frac{N+1}{2}, \frac{n_1 \; n_2 \; (N+1)}{12}\right)$$

This distribution is has obviously the same shape as the distribution of W since they both differ by the constant $n_1(N+1)$. See next slides.

Empirical distribution of R_X and W in R

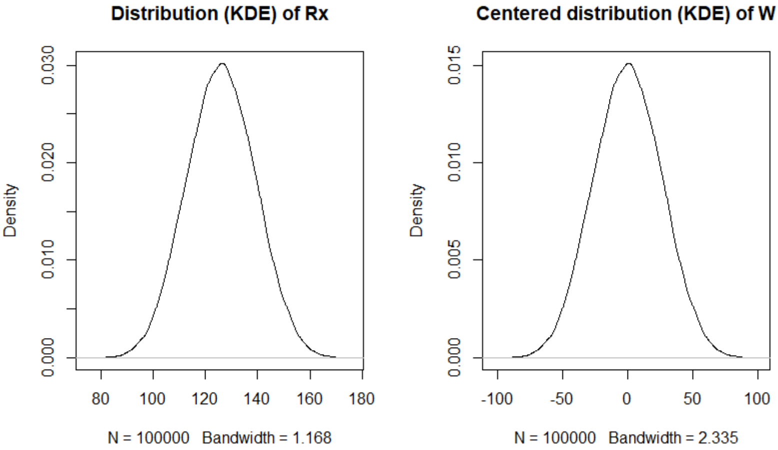
```r
# distribution of Rx and W under H0 using simulation
set.seed(2023)
n = 100000
X = matrix(rep(0, 12*n), nrow = n, ncol = 12)
Y = matrix(rep(0, 8*n), nrow = n, ncol = 8)
S = R = matrix(rep(0, (12+8)*n), nrow = n, ncol = 12+8)
Rx = W = numeric(n)

for(i in 1:n) {
  X[i,] = runif(12, min = 0, max = 2)
  Y[i,] = runif(8, min = 0, max = 2)
  S[i,] = c(X[i,], Y[i,])
  R[i,] = rank(S[i,])
  Rx[i] = sum(R[i, 1:12])
  W[i] = 2*Rx[i] + (12*(N+1))
}

# mean and variance of Rx and W
mean(Rx); var(Rx)
# [1] 126.048 about 12*(20+1)/2
# [1] 168.2862 about 12*8*(20+1)/12
mean(W); var(W)
# [1] 504.096 about (2*(12*(20+1)/2)) + 12*(20+1)
# [1] 673.145

# plots
par(mfrow = c(1,2))
plot(density(Rx), main = 'Distribution (KDE) of Rx')
plot(density(W - rep((2*(12*(20+1)/2)) + 12*(20+1), n)), main = 'Centered
      distribution (KDE) of W')
```

Density plots of R_X and W



References

Bagdonavičius V., Kruopis J., Nikulin M. S., Non-parametric Tests for Complete Data (2011), Wiley, ISBN 978-1-84821-269-5 (hardback)

The R Project for Statistical Computing: https://www.r-project.org/

course notes