

Nonparametric Hypothesis Tests

Consider data X_1, \dots, X_m iid with distribution F , and data Y_1, \dots, Y_n iid with distribution G . We are interested in the following test:

$$H_o : F = G$$

$$H_a : F \neq G$$

One approach for conducting this test is to replace the data realizations with their ranks. The rank of a data realization is its index in the ascending-order-sorted list of all realizations from F and G . Let r_F denote the sum of the ranks of data realized from distribution F , and let r_G denote the sum of ranks of data realized from distribution G .

Assuming H_o is true, we can explicitly compute the probability mass function from which r_F was realized and then decide to reject H_o if this realization is too extreme.

Example

Let $m = 3$ and $n = 2$ and say that we realized data:

$$\{x_1, x_2, x_3\} = \{11.2, 12.1, 10.4\}$$

$$\{y_1, y_2\} = \{18.5, 14.4\}$$

The ascending-order-sorted list of all realizations from F and G is

$$\{x_3, x_1, x_2, y_2, y_1\}$$

Now replace the data with the ranks:

$$\{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\} = \{2, 3, 1\}$$

$$\{\tilde{y}_1, \tilde{y}_2\} = \{5, 4\}$$

Now compute the rank sums:

$$r_F = 2 + 3 + 1 = 6$$

$$r_G = 5 + 4 = 9$$

Do these values r_F and r_G allow us to reject H_o ?

Computing the Rank Sum Distribution

Assuming H_o is true, all possible rank sums are equally likely. Also note that:

$$r_F = (m + n)(m + n + 1)/2 - r_G$$

thus we need only compute the probability mass function from which r_F was realized, let R denote the random variable that had the realization r_F . We can find the pmf of R by enumerating all of the $\binom{m+n}{m}$ possible rank sums. This computation can be done in R for the example when $m = 3$ and $n = 2$ we have:

```
m=3
n=2
rF = 6
R = NULL
k=1
for ( i.1 in 1:(m+n-2) )
{
  for ( i.2 in (i.1+1):(m+n-1))
  {
    for ( i.3 in (i.2+1):(m+n) )
    {
      R[k] = i.1+i.2+i.3
      k=k+1
    }
  }
}

## Probability that R = rF
sum(R == 6)/length(R)

## P(R <= 6)+P(R >= 12)
sum(R <= 6)/length(R) + sum(R >= 12)/length(R)
```

Note that the smallest possible value of R is 6 and the largest possible value for R is 12. For the two-sided test, we should reject H_o if $P(R \leq 6) + P(R \geq 12) \leq 0.05$. We thus cannot reject H_o because $P(R \leq 6) + P(R \geq 12) = 0.2$.

Suppose we have independent samples from two populations: X_1, \dots, X_m are from population A, and Y_1, \dots, Y_n are from population B. Let W_i be the number of the Y 's that are smaller than X_i . Define:

$$W = \sum_{i=1}^m W_i$$

$$W1 = \sum_{i=1}^m \text{Rank}(X_i) - m(m+1)/2$$

It is a fact that $W = W1$. To exemplify this fact let's generate $m = 5, n = 7$ realizations from population A and B respectively, and compute W and $W1$

```
m = 5
```

```
n = 7
```

```
X = rnorm(m)
```

```
Y = rnorm(n)
```

```
## Compute W
```

```
W = 0
```

```
for (i in 1:m)
```

```
{
```

```
  W = W + sum(Y < X[i])
```

```
}
```

```
## Compute W1=Ranksum(X) - m(m+1)/2
```

```
Z = c(X, Y)
```

```
R = rank(Z)
```

```
RX = R[1:m]
```

```
W1 = sum(RX) - m*(m+1)/2
```

```
c(W, W1)
```

In the homework, you will show by simulation that when all of the data are iid, in other words the null hypothesis is true, then:

- $EW = mn/2$
- $\text{var } W = mn(m+n+1)/12$
- For m, n sufficiently large: $W_s = \frac{W - EW}{\text{SD } W} \approx Z \sim N(0, 1)$

```
## Compute the proportion of times
```

```
## we reject Ho, if it is actually true
```

```
m = 25
```

```
n = 15

mu = m*n/2
s = sqrt(m*n*(m+n+1)/12)

nrep = 1e4
WS = NULL

for (k in 1:nrep)
{
  ## Generate Data under H0
  X = rexp(m)
  Y = rexp(n)

  Z = c(X, Y)
  R = rank(Z)
  RX = R[1:m]
  W = sum(RX) - m*(m+1)/2

  WS[k] = (W - mu) / s
}
## Reject using alpha = 0.05
mean(abs(WS) > qnorm(0.975))
```