# Finding the number of permuations

*Michael Höhle, Stockholm University*

*2019-06-01*

## Abstract

We write up a difference equations to compute the distribution under the null hypothesis for an exact permutation test for the means in two independent groups.

## The number of permutations

Let $\boldsymbol{x}_1 = (x_1, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be two samples consisting of integer values, i.e. $x_i \in \mathbb{N}_0$ for $1 \leq i \leq m$ and $y_j \in \mathbb{N}_0$ for $1 \leq j \leq m$. We compute $s_{\text{obs}} = \sum_{i=1}^{m} x_i$ as the observed sum in the first sample and want to know how extreme this value is under the null hypothesis that the mean does not differ between the two distributions against the alternative that the mean in the second group is larger than in the first group (aka. Fisher-Pitman test?). This is equivalent to investigating how likely it is to observe $s_{\text{obs}}$ or a more extreme value in the direction of the alternative, i.e. $P(S \leq s_{\text{obs}} | \boldsymbol{x}, \boldsymbol{y})$, where $S$ is a random variable denoting the value of the sum in the first sample under the permutation distribution.

The computational approach is as follows: Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_{n+m})$ be the ordered vector of the $n + m$ values from the two samples, i.e. the vector is such such that $z_1 = \min(\boldsymbol{x}, \boldsymbol{y})$ is the smallest value in the two samples and $z_{n+m} = \max(\boldsymbol{x}, \boldsymbol{y})$ is the largest value and for all $1 \leq i < j \leq n + m$ we have $z_i \leq z_j$ for $i < j$. WLOG we can assume that $z_1 = 0$.

Let the possible states of the integer sum be $\Omega = \{0, 1, \ldots, \mathcal{S}\}$, which we shall denote the support of the sum. We know that the maximal value the sum can obtain when using the values from $\boldsymbol{z}$ is $\mathcal{S} = z_{n+1} + \ldots + z_{m+n}$, i.e. the sum of the $m$ largest values. Assume we let the random variable $\boldsymbol{Q} = (Q_1, \ldots, Q_m)$ denote a randomly picked permutation where we pick $m$ values from $\boldsymbol{z}$ and let $S = \sum_{i=1}^{m} Q_i$ be the corresponding sum. We compute the PMF of $S$, which has support on $\Omega$ as follows:

$$P(S = s) = \frac{\#\text{ permutations selecting } m \text{ out of the } m+n \text{ values in } \boldsymbol{z} \text{ s.t. sum is equal to } s}{\#\text{ permutations selecting } m \text{ out of the } m+n \text{ values in } \boldsymbol{z}},$$

where $0 \leq s \leq \mathcal{S}$. From basic combinatorics we know that the denominator of the above is equal to

$$_mC_{m+n} = \frac{(m+n)!}{m!}.$$

To find the numerator, let $\boldsymbol{N}(j, k)$ denote the vector of length $\mathcal{S} + 1$, which for $0 \leq s \leq \mathcal{S}$ states how many combinations are possible when selecting $j$, $1 \leq j \leq k$, elements out of $z_1, \ldots, z_k$ such that the sum of the elements is $s$. Note that this can be computed using the following difference equation, which is inspired by the computation of Theorem 1 in Pagano and Tritchler (1983)[1] and made explicit in Zimmermann (1985)[2] :

$$\boldsymbol{N}(j, k) = j \cdot (\boldsymbol{N}(j-1, k-1) \oplus z_k) + \boldsymbol{N}(j, k-1),$$

[1] ... but without the fast Fourier transform to compute the sums

[2] In the Zimmermann (1985) paper the equation is given for combinations where order does not matter. In the present note we count the number of combinations where order matters. The resulting p-values are the same, though.

where the $\oplus$ operator for the integer $0 \leq i \leq z_{n+m}$ is defined as follows:

$$(s_0, \ldots, s_w, 0, \ldots, 0) \oplus i := (\underbrace{0, \ldots, 0}_{izeros}, s_0, s_1, \ldots, s_w, 0, \ldots, 0)$$

In other words, the $\oplus$ operator shifts the elements of its left argument by $i$ positions forward. Note: The input argument is always such that there are enough zeros at the end such that this shift just reduces the number of zeroes at the end, but does not lead to any non-zero entries being shifted outside the vector. Example:

$$(2, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)' \oplus 3 := (0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0)'$$

The desired value of the numerator is then found as $N(m, m+n)$. For a direct recursive equation without the shift operator see Zimmermann (1985).

## In code

```
x <- c(0,3,0) # first sample
y <- c(1,2,5) # second sample
m <- length(x) #length of the 1st sample (x)
n <- length(y) #length of the 2nd sample (y)
z <- sort(c(x,y)) #vector of the combined first and second sample (ordered!)
S <- sum(tail(z,n=m)) #largest possible value of the sum in the permutations of z

##Vector and matrix containing just zeroes
zero_vec <- rep(0, S+1)
zero_matrix <- matrix(NA, S+1, m+n,dimnames=list(0:S,1:(m+n)))

##helper function
shift <- function(v, i) c(rep(0,i),v[1:(length(v)-i)])

##Compute N(i,j) by loops
for (j in 1:m) {
  N <- zero_matrix
  for (k in 1:(m+n)) {
    if (j == 1) {
      ##Base setup - we count what we observed in the first k positions of z
      N[,k] <- table(factor(z[1:k], levels=0:S))
    } else {
      if (j>k) {
        ##Nothing to do in this case
        N[,k] <- zero_vec
      } else {
        ##Compute using the difference equation
        ##When order matters
        N[,k] <- j*shift(Njm1[,k-1],z[k]) + N[,k-1]
        ##When order does not matter as in Zimmermann (1985)
        N[,k] <- shift(Njm1[,k-1],z[k]) + N[,k-1]
      }
    }
```

```
  }

  Njm1 <- N
  cat("\n\nj=",j, "   k=",k,"\n")
  print(N)
}
```

```
##
##
## j= 1    k= 6
##    1 2 3 4 5 6
## 0  1 2 2 2 2 2
## 1  0 0 1 1 1 1
## 2  0 0 0 1 1 1
## 3  0 0 0 0 1 1
## 4  0 0 0 0 0 0
## 5  0 0 0 0 0 1
## 6  0 0 0 0 0 0
## 7  0 0 0 0 0 0
## 8  0 0 0 0 0 0
## 9  0 0 0 0 0 0
## 10 0 0 0 0 0 0
##
##
## j= 2    k= 6
##    1 2 3 4 5 6
## 0  0 1 1 1 1 1
## 1  0 0 2 2 2 2
## 2  0 0 0 2 2 2
## 3  0 0 0 1 3 3
## 4  0 0 0 0 1 1
## 5  0 0 0 0 1 3
## 6  0 0 0 0 0 1
## 7  0 0 0 0 0 1
## 8  0 0 0 0 0 1
## 9  0 0 0 0 0 0
## 10 0 0 0 0 0 0
##
##
## j= 3    k= 6
##    1 2 3 4 5 6
## 0  0 0 0 0 0 0
## 1  0 0 1 1 1 1
## 2  0 0 0 1 1 1
## 3  0 0 0 2 3 3
## 4  0 0 0 0 2 2
## 5  0 0 0 0 2 3
## 6  0 0 0 0 1 3
## 7  0 0 0 0 0 2
## 8  0 0 0 0 0 3
## 9  0 0 0 0 0 1
## 10 0 0 0 0 0 1
```

```
##Show result, this is N(3,6)
print(N)
```

```
##    1 2 3 4 5 6
## 0  0 0 0 0 0 0
## 1  0 0 1 1 1 1
## 2  0 0 0 1 1 1
## 3  0 0 0 2 3 3
## 4  0 0 0 0 2 2
## 5  0 0 0 0 2 3
## 6  0 0 0 0 1 3
## 7  0 0 0 0 0 2
## 8  0 0 0 0 0 3
## 9  0 0 0 0 0 1
## 10 0 0 0 0 0 1
```

```
##Sanity checks
sum(N[,m+n])
```

```
## [1] 20
```

```
##Number of permutations when order matters
factorial(m+n)/factorial(m)
```

```
## [1] 120
```

```
##Number of permutations when order does not matter
choose(m+n, m)
```

```
## [1] 20
```

```
##Resulting PMF
pmf <- N[,m+n] / sum(N[,m+n])
##More extreme observations, i.e. this is the p-value
(p_value <- sum(pmf[(sum(x)+1):(S+1)]))
```

```
## [1] 0.9
```

We compare the result with performing all permutations manually:

```
##Compute all permutations manually
perms <- combinat::combn(m+n, m=m)
dist <- apply(perms, 2, function(i) sum(z[i]))
(p_value_exact <- mean(dist>=sum(x)))
```

```
## [1] 0.9
```

```
##
stopifnot(isTRUE(all.equal(p_value, p_value_exact)))
```

# Literature


Pagano, Marcello, and David Tritchler. 1983. "On Obtaining Permutation Distributions in Polynomial Time." *Journal of the American Statistical Association* 78 (382). Taylor & Francis: 435–40. doi:10.1080/01621459.1983.10477990.

Zimmermann, H. 1985. "Exact Calculation of Permutational Distributions for Two Independent Samples." *Biometrical Journal* 4: 431–34. doi:10.1002/bimj.4710270414.