

# Finding the number of permutations

*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  $\mathbf{x}_1 = (x_1, \dots, x_m)$  and  $\mathbf{y} = (y_1, \dots, y_n)$  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  $1 \leq j \leq n$ . 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 in the distributions against the alternative that the mean in the second group is larger than in the first group. This is equivalent to investigating how likely it is to observe  $s$  or a more extreme value in the direction of the alternative, i.e.  $P(S \leq s_{\text{obs}} | \mathbf{x}, \mathbf{y})$ .

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

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

$$P(S_Q = i) = \frac{\# \text{ permutations with sum equal to } i}{\# \text{ permutations selecting } m \text{ out of the } m + n \text{ values in } \mathbf{z}}, \quad 0 \leq i \leq \mathcal{S}$$

Let  $N(j, k)$  denote the vector of length  $\mathcal{S} + 1$ , which for  $0 \leq s \leq \mathcal{S}$  states how many of the possible  $\frac{(m+n-j)!}{j!}$  permutations yield a sum of  $s$ . 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):<sup>1</sup>

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

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

$$(s_0, \dots, s_w, 0, \dots, 0) \oplus i := (\underbrace{0, \dots, 0}_{i \text{ zeros}}, s_0, s_1, \dots, s_w, 0, \dots, 0)$$

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:

---

<sup>1</sup>... but without the fast Fourier transform to compute the sums

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

## 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
        N[,k] <- j*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 2 2 2 2 2
## 1  0 0 4 4 4 4
## 2  0 0 0 4 4 4
## 3  0 0 0 2 6 6
## 4  0 0 0 0 2 2
## 5  0 0 0 0 2 6
## 6  0 0 0 0 0 2
## 7  0 0 0 0 0 2
## 8  0 0 0 0 0 2
## 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 6  6  6  6
## 2  0 0 0  6  6  6
## 3  0 0 0 12 18 18
## 4  0 0 0  0 12 12
## 5  0 0 0  0 12 18
## 6  0 0 0  0  6 18
## 7  0 0 0  0  0 12
## 8  0 0 0  0  0 18
## 9  0 0 0  0  0  6
## 10 0 0 0  0  0  6

```

```

##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 6  6  6  6
## 2  0 0 0  6  6  6
## 3  0 0 0 12 18 18
## 4  0 0 0  0 12 12
## 5  0 0 0  0 12 18
## 6  0 0 0  0  6 18
## 7  0 0 0  0  0 12
## 8  0 0 0  0  0 18
## 9  0 0 0  0  0  6
## 10 0 0 0  0  0  6

```

```

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

```

```
## [1] 120
factorial(m+n)/factorial(m)

## [1] 120
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.