

In this discussion, we look at confidence intervals in the Wilcoxon Rank-Sum Test for one sample. The data comes from an article by B. Flaconneche *et. al.*, "Permeability, Diffusion and Solubility of Gases," in *Oil and Gas Science and Technology*, 2001, Navidi, which is quoted by Navidi, "Statistics for Engineers and Scientists," 2nd ed., MacGraw Hill, New York, 2008, problem 5.6.12. The study measured the effect of temperature on the gas transport coefficients in semicrystalline polymers. The permeability coefficient (in  $10^{-6} \text{ cm}^3 \text{ (STP)/cm s MPa}$ ) of  $\text{CO}_2$  was measured for extruded medium density polyethylene at both  $60^\circ \text{ C}$  and  $61^\circ \text{ C}$ . Find a 2-sided CI on  $\mu_X - \mu_Y$ .

We assume that the independent samples  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  where  $m \leq n$  come from a continuous distributions with pdf's  $f(x - \mu_X)$  and  $f(y - \mu_Y)$  which have the same shape and spread, but have different means  $\mu_X$  and  $\mu_Y$ . The null and alternative hypotheses are

$$\begin{aligned} \mathcal{H}_0 : \mu_X - \mu_Y &= \Delta_0; \\ \mathcal{H}_a : \mu_X - \mu_Y &\neq \Delta_0. \end{aligned}$$

The test statistic  $W$  is defined as follows. We replace  $X_i$  by  $X_i - \Delta_0$  so that we may take  $\Delta_0 = 0$  for simplicity. We order the combined values  $X_i$  and  $Y_j$  from lowest to highest. We rank them 1 to  $m+n$ .  $W$  is the sum of the ranks corresponding to  $X_i$  terms. The null hypothesis is rejected at the significance level  $\alpha$  if  $W \geq c$  or  $W \leq \frac{n(n+1)}{2} - c$  where  $2P(W \geq c) = \alpha$ . The CI is facilitated by another interpretation of  $W$ .

**Theorem 1.** Assume  $\Delta_0 = 0$  and that there are no ties among the combined  $X_i$  and  $Y_j$  values. Let  $N$  equal the number of pairs  $(i, j)$  such that  $X_i - X_j \geq 0$ . Then

$$W = N + \frac{m(m+1)}{2}.$$

*Proof.* Observe that the number of such pairs is unchanged if the  $X_i$ 's are rearranged among themselves and the  $Y_j$ 's are rearranged among themselves. Let  $X_{(i)}$  denote the sorting from lowest to highest. Let  $s_i$  denote the rank of  $X_{(i)}$  in the union of  $X_i$ 's and  $Y_j$ 's. Similarly, let  $t_j$  denote the rank of  $Y_{(j)}$  in the union.

Then for each  $i$ , since  $X_{(1)} \dots, X_{(i)}$  are no larger than  $X_{(i)}$  and  $X_{(i+1)}, \dots, X_{(m)}$  are larger, it follows that exactly  $i$  of the  $X$ 's are at most  $X_{(i)}$ . Because  $X_{(i)}$  is the  $s_i$ th smallest in the union, there must be  $s_i - i$  of the  $Y_j$ 's smaller than  $X_{(i)}$ . It follows that

$$\sum_{j=1}^n \chi_{\{X_{(i)} - Y_{(j)} \geq 0\}} = s_i - i$$

where  $\chi_S$  is the indicator function: it is one if  $S$  is true and zero otherwise. Summing over  $i$  gives the result

$$N = \sum_{i=1}^m \sum_{j=1}^n \chi_{\{X_{(i)} - Y_{(j)} \geq 0\}} = \sum_{i=1}^m s_i - \sum_{i=1}^m i = W - \frac{m(m+1)}{2}. \quad \square$$

The confidence interval associated to the Rank-Sum test of  $\mathcal{H}_0$  is motivated by the theorem. We consider the set of  $mn$  values

$$\mathcal{A} = \{X_i - Y_j : 1 \leq i \leq m, 1 \leq j \leq n\}$$

If there are no ties, then  $N$  is the number of elements of  $\mathcal{A}$  such that  $X_i \geq Y_j$ . We sort  $\mathcal{A}$  into  $\{A_{(1)} \leq A_{(2)} \leq \dots \leq A_{(mn)}\}$ . The one and two-sided critical values for  $W$  and  $N$  may be

computed as follows. Assuming  $\mathcal{H}_0$ , the each combination of  $m$  ranks taken  $m + n$  at a time is equally likely. Thus the distribution of  $W$  is obtained by looking at the histogram of values

$$\sum_{i=1}^m s_i$$

where  $(s_1 < \dots < s_m)$  is a combination of  $\{1 < 2 < \dots < m + n\}$  taken  $m$  at a time, each of which has an equal chance  $\binom{m+n}{m}^{-1}$  of occurring. They range from  $\frac{m(m+1)}{2}$  to  $\frac{m(m+2n+1)}{2}$ .

The pmf  $p(w)$  of  $W$  is symmetrical about the mean  $\mu_W = \frac{m(m+n+1)}{2}$ . The one-sided upper critical value is  $P(W \geq c_1) = \sum_{i \geq c_1} p(i) = \alpha$ , which is tabulated in Table A14. We indicate a computation of the  $c_1$  critical values in the program.

The two-sided critical value  $c$  satisfies  $P(N \geq c) = P(N \leq mn - c) = \frac{\alpha}{2}$ . These numbers are tabulated in Table A16 of the text. Note that  $c + \frac{m(m+1)}{2}$  would be the corresponding critical value of  $W$ . We show how to compute the values of these tables in the code, albeit inefficiently.

Then the two-sided confidence interval for  $\mu_X - \mu_Y$  at the level  $\alpha$  where  $c$  is the two-sided critical value that satisfies  $P(N \geq c) = \frac{\alpha}{2}$  is given by

$$(A_{(mn+1-c)}, A_{(c)}).$$

If the sample sizes are large enough, (rule of thumb both  $m > 8$  and  $n > 8$ ) then the normal approximation to  $p(w)$  may be used. the critical value for  $N$  is given by

$$c \approx \frac{mn}{2} + z_{\frac{\alpha}{2}} \sqrt{\frac{mn(m+n+1)}{12}}.$$

## Data Set Used in this Analysis :

---

```
# Math 3080-1                Gas Transport Data                April 20, 2014
# Treibergs
#
# From Navidi, "Statistics for Engineers and Scientists," 2nd ed., MacGraw
# Hill, New York, problem 5.6.12, who quotes the article by B. Flaconneche
# et. al., "Permeability, Diffusion and Solubility of Gases,"
# in "Oil and Gas Science and Technology," 2001. The study measured the
# effect of temperature on the gas transport coefficients in
# semicrystalline polymers. The permeability coefficient
# (in  $10^{-6}$  cm3 (STP) / cm s MPa) of CO2 was measured for extruded
# medium density polyethylene at both 60 C and 61 C. Find a 2-sided CI on
#  $\mu_x - \mu_y$ .
Temp  Perm
C61  58
C61  60
C61  66
C61  66
C61  68
C61  61
C61  60
C60  54
C60  51
C60  61
C60  67
C60  57
C60  69
C60  60
C60  60
C60  63
C60  62
```

---

### R Session:

---

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or
```

'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]

```
> tt=read.table("M3082DataGasTransport.txt",header=T)
> attach(tt)
> tt
  Temp Perm
1  C61  58
2  C61  60
3  C61  66
4  C61  66
5  C61  68
6  C61  61
7  C61  60
8  C60  54
9  C60  51
10 C60  61
11 C60  67
12 C60  57
13 C60  69
14 C60  60
15 C60  60
16 C60  63
17 C60  62
> x=Perm[Temp=="C61"]
> y=Perm[Temp=="C60"]
> m=length(x); m; n=length(y); n
[1] 7
[1] 10
> ##### SIDE-BY-SIDE HISTOGRAM OF x AND y #####
> hx=hist(x,breaks=seq(50,70,2.5))
> hy=hist(y,breaks=seq(50,70,2.5))
> mx=t(cbind(hx$density,hy$density))
> colo=c(rainbow(10,alpha=.7)[4],rainbow(10,alpha=.5)[7])

> colnames(mx)= c("50-","52.5-","55-","57.5-","60-","62.5-","65-","67.5-")
> barplot(mx,beside=T,col=colo,main="CO2 Gas Permeability",
  legend.text=c("61 C","60 C"),space=c(0,.5))
>
```

```

> ##### CRITICAL VALUES OF RANK-SUM STATISTIC W #####
> ##### TABLE OF ALL COMBINATIONS OF RANKS #####
> m=3; n=4
> M=combn(m+n,m); M
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]  1   1   1   1   1   1   1   1   1   1   1   1   1
[2,]  2   2   2   2   2   3   3   3   3   4   4   4   5
[3,]  3   4   5   6   7   4   5   6   7   5   6   7   6
  [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
[1,]  1   1   2   2   2   2   2   2   2   2   2   2
[2,]  5   6   3   3   3   3   4   4   4   5   5   6
[3,]  7   7   4   5   6   7   5   6   7   6   7   7
  [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35]
[1,]  3   3   3   3   3   3   4   4   4   5
[2,]  4   4   4   5   5   6   5   5   6   6
[3,]  5   6   7   6   7   7   6   7   7   7

> ##### ALL POSSIBLE RTANK-SUMS #####
> margin.table(M,2)
 [1]  6  7  8  9 10  8  9 10 11 10 11 12 12 13 14  9 10 11 12 11 12 13 13 14 15
[26] 12 13 14 14 15 16 15 16 17 18
> ##### FREQUENCIES OF W #####
> table(margin.table(M,2))

  6  7  8  9 10 11 12 13 14 15 16 17 18
1  1  2  3  4  4  5  4  4  3  2  1  1
> ##### PMF OF W #####
> p=table(margin.table(M,2))/choose(m+n,m);sum(p)
[1] 1
> p

      6      7      8      9      10      11      12
0.02857143 0.02857143 0.05714286 0.08571429 0.11428571 0.11428571 0.14285714
      13      14      15      16      17      18
0.11428571 0.11428571 0.08571429 0.05714286 0.02857143 0.02857143

> ##### TAIL PROBABILITIES #####
> cp=cumsum(p);cp
> names(cp)=seq(m*(m+n+n+1)/2,m*(m+1)/2,-1)
> ##### ONE-TAILED CRIT. VALUES c1 FOR W AS IN TABLE A14 #####
> cp
      18      17      16      15      14      13      12
0.02857143 0.05714286 0.11428571 0.20000000 0.31428571 0.42857143 0.57142857
      11      10      9      8      7      6
0.68571429 0.80000000 0.88571429 0.94285714 0.97142857 1.00000000
> # table A14
>

```

```

> ##### TWO-TAILED CRITICAL VALUES FOR N AND W #####
> m=7; n=9
> M=combn(m+n,m)
> p=table(margin.table(M,2))/choose(m+n,m);sum(p)
[1] 1
> p

      28      29      30      31      32      33
8.741259e-05 8.741259e-05 1.748252e-04 2.622378e-04 4.370629e-04 6.118881e-04
      34      35      36      37      38      39
9.615385e-04 1.311189e-03 1.835664e-03 2.447552e-03 3.234266e-03 4.108392e-03
      40      41      42      43      44      45
5.332168e-03 6.555944e-03 8.129371e-03 9.790210e-03 1.171329e-02 1.372378e-02
      46      47      48      49      50      51
1.608392e-02 1.835664e-02 2.089161e-02 2.342657e-02 2.596154e-02 2.840909e-02
      52      53      54      55      56      57
3.094406e-02 3.312937e-02 3.522727e-02 3.706294e-02 3.854895e-02 3.968531e-02
      58      59      60      61      62      63
4.055944e-02 4.090909e-02 4.090909e-02 4.055944e-02 3.968531e-02 3.854895e-02
      64      65      66      67      68      69
3.706294e-02 3.522727e-02 3.312937e-02 3.094406e-02 2.840909e-02 2.596154e-02
      70      71      72      73      74      75
2.342657e-02 2.089161e-02 1.835664e-02 1.608392e-02 1.372378e-02 1.171329e-02
      76      77      78      79      80      81
9.790210e-03 8.129371e-03 6.555944e-03 5.332168e-03 4.108392e-03 3.234266e-03
      82      83      84      85      86      87
2.447552e-03 1.835664e-03 1.311189e-03 9.615385e-04 6.118881e-04 4.370629e-04
      88      89      90      91
2.622378e-04 1.748252e-04 8.741259e-05 8.741259e-05

> ### TABLE OF CRITICAL VALUES FOR N = NO. (I,J) ST X[I] \GE Y[J] ###
> cp2=1-2*cumsum(p)
> names(cp2)=seq(m*n,1,-1)
> cp2[1:(m*n/2)]
> ##### TABLE A16 FOR m = 7 AND n = 9 #####
      63      62      61      60      59      58      57
0.99982517 0.99965035 0.99930070 0.99877622 0.99790210 0.99667832 0.99475524
      56      55      54      53      52      51      50
0.99213287 0.98846154 0.98356643 0.97709790 0.96888112 0.95821678 0.94510490
      49      48      47      46      45      44      43
0.92884615 0.90926573 0.88583916 0.85839161 0.82622378 0.78951049 0.74772727
      42      41      40      39      38      37      36
0.70087413 0.64895105 0.59213287 0.53024476 0.46398601 0.39353147 0.31940559
      35      34      33
0.24230769 0.16293706 0.08181818
> # table A16
>

```

```

> ##### TABLE A16 FOR GAS TRANSPORT DATA #####
> m=length(x); m; n=length(y); n
[1] 7
[1] 10
> M=combn(m+n,m)
> p=table(margin.table(M,2))/choose(m+n,m);sum(p)
[1] 1
> cp2=1-2*cumsum(p)
> names(cp2)=seq(m*n,1,-1)
> cp2[1:(m*n/2)]
      70      69      68      67      66      65      64
0.99989716 0.99979432 0.99958865 0.99928013 0.99876594 0.99804607 0.99691485
      63      62      61      60      59      58      57
0.99537227 0.99321267 0.99033320 0.98642534 0.98148910 0.97501028 0.96698889
      56      55      54      53      52      51      50
0.95691074 0.94467297 0.92976142 0.91217606 0.89119704 0.86692719 0.83874949
      49      48      47      46      45      44      43
0.80676676 0.77046483 0.73015220 0.68521185 0.63615796 0.58268202 0.52529823
      42      41      40      39      38      37      36
0.46380090 0.39911559 0.33093377 0.26038667 0.18747429 0.11322501 0.03774167
>
>
> ##### CANNED WILCOXON RANK-SUM TEST #####
> wilcox.test(x,y,conf.int=T)

```

Wilcoxon rank sum test with continuity correction

```

data: x and y
W = 41.5, p-value = 0.5553
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.999974  7.000008
sample estimates:
difference in location
      2.183133

```

Warning messages:

- 1: In wilcox.test.default(x, y, conf.int = T) :  
cannot compute exact p-value with ties
- 2: In wilcox.test.default(x, y, conf.int = T) :  
cannot compute exact confidence intervals with ties

```

> ##### GENERATE ALL DIFFERENCES X[I]-Y[J] #####
> A=outer(x,y,"-")
> A
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    4    7   -3   -9    1  -11   -2   -2   -5   -4
[2,]    6    9   -1   -7    3   -9    0    0   -3   -2
[3,]   12   15    5   -1    9   -3    6    6    3    4
[4,]   12   15    5   -1    9   -3    6    6    3    4
[5,]   14   17    7    1   11   -1    8    8    5    6
[6,]    7   10    0   -6    4   -8    1    1   -2   -1
[7,]    6    9   -1   -7    3   -9    0    0   -3   -2
> A=sort(as.vector(outer(x,y,"-")))
> A
 [1] -11 -9 -9 -9 -8 -7 -7 -6 -5 -4 -3 -3 -3 -3 -3 -2 -2 -2
[19] -2 -2 -1 -1 -1 -1 -1 -1 -1  0  0  0  0  0  1  1  1  1  3
[37]  3  3  3  4  4  4  4  5  5  5  6  6  6  6  6  6  6  7
[55]  7  7  8  8  9  9  9  9 10 11 12 12 14 15 15 17
> ##### 2-SIDED ALPHA=.05 CI FOR muX - muY #####
> alpha=.05
> c=56
> c( A[m*n+1-c],A[c])
[1] -3  7
>
> ##### 2-SIDED ALPHA=.01 CI FOR muX - muY #####
> alpha=.01
> c=61
> c( A[m*n+1-c],A[c])
[1] -4  9
> ##### CANNED 2-SIDED ALPHA = .01 CI #####
> wilcox.test(x,y,conf.int=T, conf.level=.99)

Wilcoxon rank sum test with continuity correction

data:  x and y
W = 41.5, p-value = 0.5553
alternative hypothesis: true location shift is not equal to 0
99 percent confidence interval:
 -5.000058  9.000009
sample estimates:
difference in location
      2.183133

Warning messages:
1: In wilcox.test.default(x, y, conf.int = T, conf.level = 0.99) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x, y, conf.int = T, conf.level = 0.99) :
  cannot compute exact confidence intervals with ties

> ### NORMAL APPROXIMATION OF ALPHA=.05 2-SIDED CRIT. VAL. C ###
> alpha=.05
> zalpha2=qnorm(alpha/2,lower.tail=F); zalpha2
[1] 1.959964

```



```
> m*n/2+zalpha2*sqrt(m*n*(m+n+1)/12)
[1] 55.08365
> c=55
> c( A[m*n+1-c],A[c])
[1] -2 7
>
```

### CO2 Gas Permeability

