Stat 425 HW2 Solutions

Fritz Scholz

Chapter 1, Section 4, Problem 46.:
To test the effectiveness of vitamin B_1 in stimulating growth in mushrooms, vitamin B_1 was applied to 13 mushrooms selected at random from a group of 24, while the remaining 11 did not receive this treatment. The weights of the mushrooms at the end of the period of observation were[1] (in milligrams)

Controls:	18	14.5	13.5	12.5	23	24	21	17	18.5	9.5	14		
Treated:	27	34	20.5	29.5	20	28	20	26.5	22	24.5	34	35.5	19

1. Give the sorted vector of midranks for all observations.
```
[1] 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.5 10.5 12.0 13.0 14.0 15.0
[16] 16.0 17.0 18.0 19.0 20.0 21.0 22.5 22.5 24.0
```

2. What is the observed value of W_s^*?
```
19+22.5+12+21+10.5+20+10.5+18+14+17+22.5+24+9 = 220
```

3. Use the normal approximation to find the significance probability of these results.
Since we expect stimulated growth (higher weight) under the treatment, the p-value is $P_{H_0}(W_s^* \geq 220)$.
With $E_{H_0}(W_s^*) = 162.5$ and $\text{var}_{H_0}(W_s^*) = 297.6576$ the normal approximation gives

$$P_{H_0}(W_s^* \geq 220) \approx 1 - \Phi\left(\frac{220 - 162.5}{\sqrt{297.6576}}\right)$$

$$= 1 - \Phi(3.332801) = 1 - \texttt{pnorm}(3.332801) = 0.0004298825$$

4. For the exact null distribution of W_s^* how many midrank sums need to be computed?
`choose(24,13)=2496144`, almost 2.5 million.

5. Compute the exact p-value for the observed value of W_s^* (this may take a few minutes).
Proportion of all 2496144 possible sums $W_s^* \geq 220$ =`0.0001842842`

6. Compute an estimated p-value based on $N_{\text{sim}} = 100,000$ simulations. (execute `set.seed(35)` just prior to running the simulation). This may take a few minutes.
Proportion of simulated $W_s^* \geq 220$ =`0.00018`. Your answer would be `.00012` if you set up your midrank vector in your code as `Z=rank(c(control,treatment))` rather than `Z=rank(c(treatment,control))` as I did. Think about why you would get different results even though you start out with the same seed. So much for having all students getting the same result. It had me puzzled for some time.

[1] From Linder, *Statistische Methoden*, 2d ed., Birkhäuser, Basel, 1951, p. 91. Original data from Schopfer and Blumer, "Zur Wirkstoffphysiologie von Trichophyton album Sab.," *Ber. Schweiz. Botan Ges.* **53**:409–456 (1943).

7. Discuss the merits of the three p-value calculations, in terms of accuracy, computation time, and general usability.
   The normal approximation for the p-value is off by a factor of 2.33. However, it would have been judged significant for most practical purposes. The simulation comes as close as it can get for $N_{\text{sim}} = 100,000$. That may be an accident, but other simulation results would not be far off. Try different `set.seed(...)` prior to simulation. While the calculation of the exact p-value will encounter its limit of computability for slightly larger problems, the same is not true for the simulation approach. As m and n get larger the normal approximation should get better, but we don't know how much.

The code for the two functions that get the above answers is given below.

```
Problem46=function(){
treatment=c(27, 34 ,20.5, 29.5, 20, 28, 20, 26.5, 22, 24.5, 34, 35.5, 19)
control=c(18 ,14.5, 13.5, 12.5, 23, 24, 21, 17, 18.5, 9.5, 14)
Z=rank(c(treatment,control))
n=length(treatment)
m=length(control)
N=m+n
Ws.star=sum(Z[1:n])
mean.Ws=mean(Z)*n
var.Ws=var(Z)*n*(N-n)/N
combs=choose(m+n,n)
out=combn(Z,n,FUN=sum)
pval=mean(out>=Ws.star)
pval.norm=1-pnorm((Ws.star-mean.Ws)/sqrt(var.Ws))
out=list(m=m,n=n,Z.sort=sort(Z),combs=combs,Ws.star=Ws.star,
mean.Ws=mean.Ws,var.Ws=var.Ws,pval=pval,pval.norm=pval.norm)
out
}

Problem46.sim=function(Nsim=100000){
treatment=c(27, 34 ,20.5, 29.5, 20, 28, 20, 26.5, 22, 24.5, 34, 35.5, 19)
control=c(18 ,14.5, 13.5, 12.5, 23, 24, 21, 17, 18.5, 9.5, 14)
Z=rank(c(treatment,control))
n=length(treatment)
m=length(control)
Ws.star=sum(Z[1:n])
out=rep(0,Nsim) # This gives a much faster simulation than out=NULL, see HW3.
set.seed(35)
for(i in 1:Nsim){
out[i]=sum(sample(Z,n,replace=F))
}
pval.sim=mean(out>=Ws.star)
out=list(Nsim,Ws.star,pval.sim)
names(out)=c("Nsim","Ws.star","pval.sim")
out}
```

Chapter 1, Section 4, Problem 49.:
In the context of Prob. 42, suppose that $m = n = 10$ and the data are given in the following table:

	Very Poor	Poor	Indifferent	Good	Very Good
Control	2	2	5	1	0
Treatment	0	2	4	3	1

1. Find the observed value of W_s^*.
   W_s^* =`2*4.5+4*11+3*17.5+20=125.5`

2. Give the exact p-value for this observed value of W_s^*.
   Based on the problem description high values of W_s^* are significant. Thus the p-value is $P_{H_0}(W_s^* \geq 125.5) = 11922/184756 = 0.06452835$.

3. Find the normal approximation for this p-value.
   We have $E_{H_0}(W_s^*) = 105$ and $\text{var}_{H_0}(W_s^*) = 156.4474$ and thus

   $$P_{H_0}(W_s^* \geq 125.5) \approx 1 - \texttt{pnorm}((125.5 - 105)/\texttt{sqrt}(156.4474)) = 0.05061027.$$

   not very close to `0.06452835`.

4. Find the critical value c giving the significance level closest to .01, using the exact distribution vector developed in 2. Here it helps to look at the sorted unique values of the exact distribution vector (use `unique(...)`) and compute some of the corresponding upper tail probabilities.

```
 > out49=Problem49() # see below for code of Problem49
 > sort(unique(out49$out))
 [1]  65.0  71.5  74.0  74.5  78.0  80.5  81.0  83.5  84.0  84.5  87.0  87.5
[13]  90.0  90.5  91.0  93.0  93.5  94.0  96.5  97.0  97.5  99.5 100.0 100.5
[25] 103.0 103.5 104.0 106.0 106.5 107.0 109.5 110.0 110.5 112.5 113.0 113.5
[37] 116.0 116.5 117.0 119.0 119.5 120.0 122.5 123.0 125.5 126.0 126.5 129.0
[49] 129.5 132.0 135.5 136.0 138.5 145.0
> mean(out49$out>=135.5)
[1] 0.007047132
> mean(out49$out>=132)
[1] 0.02185585
```

   shows that c= `135.5` gives tail probability closest to `.01`.

5. Compare this $P(W_s^* \geqslant c)$ (closest to .01) with its normal approximation.
   `1-pnorm((135.5-105)/sqrt(156.4474)) = 0.007374992`, pretty close to `0.007047132`.

6. Use the normal approximation of the above c corresponding to .01, rounding it to the nearest multiple of .5 (Due to midranks W_s^* takes only values that are multiples of .5).
   The normal approximation gives

   $$P_{H_0}(W_s^* \geq c) \approx 1 - \Phi\left(\frac{c - 105}{\sqrt{(156.4474)}}\right) = .01 \implies \texttt{qnorm}(.99) = (\texttt{c} - 105)/\texttt{sqrt}(156.4474)$$

$$c = \texttt{qnorm}(.99) * \texttt{sqrt}(156.4474) + 105 = 134.0977$$

which rounds to $c =$`134`, which is a bit off from `135.5`. Of course, the latter had tail probability `.007`. Note that there is no possible value between `132` and `135.5`.

The code for `Problem49` follows:

```
Problem49=function(){
y=c(0,2,4,3,1); x=c(2,2,5,1,0)
d=x+y; n=sum(y); m=sum(x); ell=length(y); N=m+n
midrank=cumsum(d)-(d-1)/2
midrank.vec=NULL
for(j in 1:ell){
midrank.vec=c(midrank.vec,rep(midrank[j],d[j]))}
out=combn(midrank.vec,n,FUN=sum)
out=sort(out)
Ws.star=sum(y*midrank)
pval=mean(out>=Ws.star)
mean.Ws=n*mean(midrank.vec)
var.Ws=var(midrank.vec)*n*(N-n)/N
pval.norm=1-pnorm((Ws.star-mean.Ws)/sqrt(var.Ws))
list(out=out,Ws.star=Ws.star,mean.Ws=mean.Ws,var.Ws=var.Ws,
        pval=pval,pval.norm=pval.norm)
}
```