Stat 425 HW7 Solution

Fritz Scholz

1. Write a function `van.der.Waerden(x,y){...}` that computes for given samples `x` (length m) and `y` (length n) the exact p-value of the 2-sample van der Waerden rank test when testing the hypothesis H_0 : no difference between treatment and control against the alternative that y-values are stochastically larger than x values.

Do this by calculating the transformed-rank vector Z_N of length $N = m+n$ according to van der Waerden's scheme (you do this just once for any given pair of samples `x` and `y`), and then (using `combn`) compute all possible sums of n transformed-ranks taken from Z_N. Get as exact p-value the proportion of such transformed-rank sums that are $\geq$ to the observed transformed-rank sum for the given samples `x` and `y`.

Allow for ties to be present in such calculations. Apply the rank transformation to the midranks rather than averaging the tied transformed ranks. For example, when two observations are tied in rank positions 7 and 8, you take as transformed ranks $\Phi^{-1}(7.5/(N+1))$ in both positions and not $[\Phi^{-1}(7/(N+1))+\Phi^{-1}(8/(N+1))]/2$, although there probably is not much difference between the two schemes. This aspect should only affect how Z_N is calculated, the rest of `van.der.Waerden` should stay the same.

Since any such transformed-rank sum can be viewed as the sum of a random sample of size n from the finite population of N transformed-ranks use the normal approximation for such sums that was provided in the context of sampling from a finite population. Calculate the p-value using the normal approximation, but do not employ a continuity correction.

Have `van.der.Waerden` also produce a histrogram (using the option `probability=T` in the argument list of `hist`) with `breaks=seq(-10.1,10.1,.2)` for the exact null distribution and superimpose the approximating normal density. On this histogram mark the position of the observed value of the test statistic using `abline(v=...)` and annotate it with the exact and approximate p-values using `text(...)`.

Let `van.der.Waerden` output a named vector containing the exact p-value and the one obtained by normal approximation.

Compute the result for the following two samples and comment on the two p-values and on the symmetry or lack of symmetry in the histogram:

```
x1=c(87.8, 100.1, 94.4, 99.7, 88.9, 101.5, 96.3)
y1=c(106.1, 96.4, 108.8, 119.3, 102.2, 100, 107.4, 103.6)
```

To test it for the situation with ties repeat this for the following two samples

```
x2=c(87.8, 100.1, 94.4, 99.7, 88.9, 101.8, 96.3)
y2=c(106.1, 96.4, 108.8, 119.3, 101.8, 100.1, 107.4, 103.6)
```

Give the results, plots and comments in each case, and provide the code of `van.der.Waerden`.

For extra credit: Amend the code of `van.der.Waerden` (in the spirit of problem 2 of HW 3) so that it can handle larger sample sizes m and n (where full enumeration is no longer feasible) via simulation, using `Nsim` simulated sums of size n taken from Z_N instead of the fully enumerated set of such sums. Here you would want to include `Nsim=100000` in your argument list and switch to simulation whenever the full enumeration leads to more than `Nsim` cases. When simulation is invoked your histogram will be of those `Nsim` simulated sums. As test case use

```
x3=c(102.5, 100.3, 96.8, 90, 96.4, 94.6, 100.9, 100.6, 94.6, 94.3)
y3=c(100.3, 104.5, 102.1, 97.7, 106.6, 105.3, 111.7, 97.4, 98.5, 110.2)
```

Provide the plot and commentary, and amended code instead of the code without this capability.

Solution: The code of `van.der.Waerden` with the simulation option included follows.

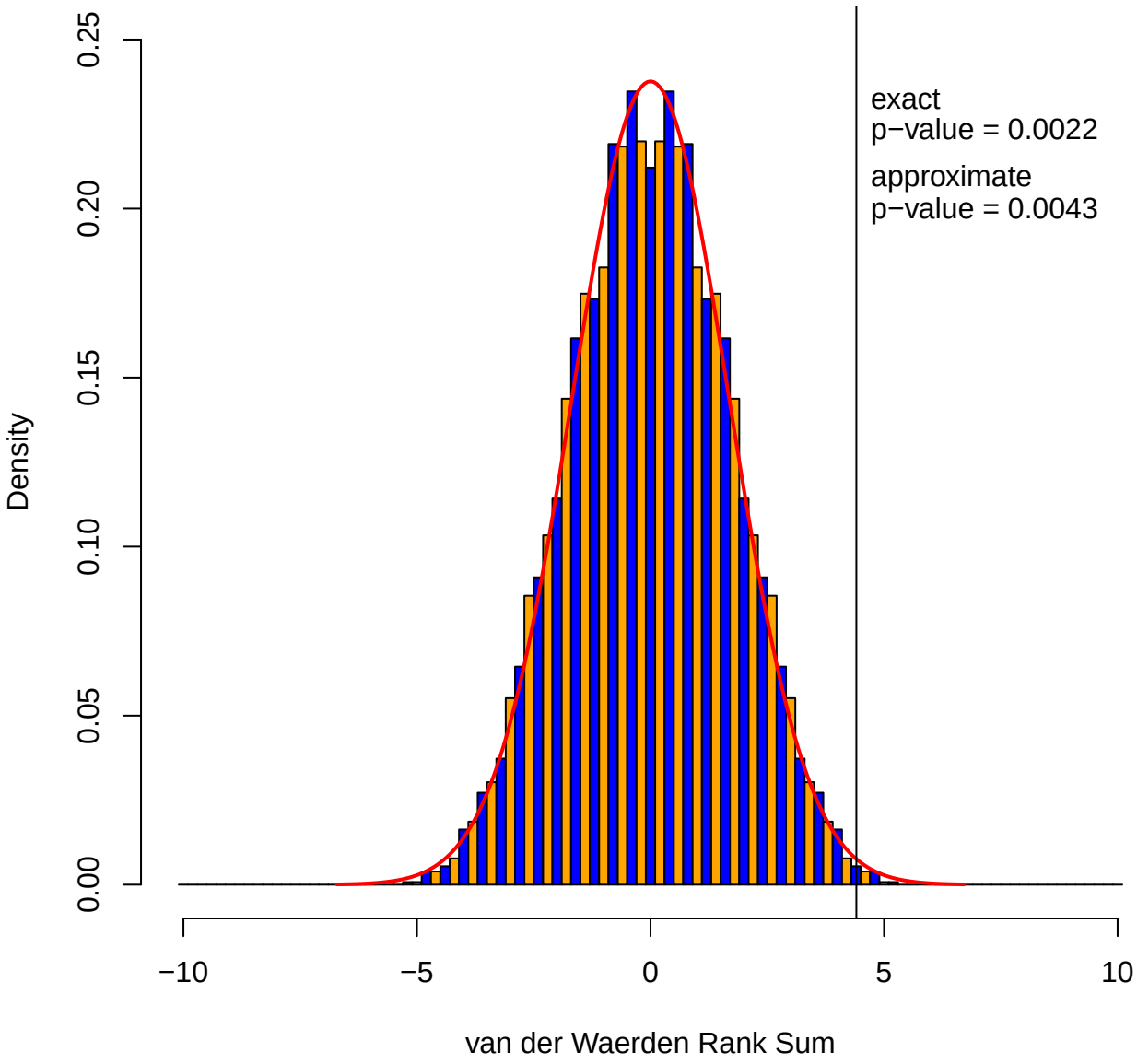
```
van.der.Waerden=function(x,y,Nsim=100000,PDF=F){
if(PDF==T) pdf(file="vanderWaerden.pdf",width=7)
m=length(x)
n=length(y)
rxy=rank(c(y,x))
N=m+n
M=choose(N,n)
flag=F
scores=qnorm(rxy/(N+1))
vdW.obs=sum(scores[1:n])
if(M <= Nsim){
  out=combn(scores,n,FUN=sum)}else{
  flag=T
  out=rep(0,Nsim)
  for(i in 1:Nsim){
    out[i]=sum(sample(scores,n,replace=F))
  }
}
p.val=mean(out>=vdW.obs)
mu=n*mean(scores)
sig=sqrt(n*var(scores)*(N-n)/N)
p.val.norm=1-pnorm((vdW.obs-mu)/sig)
```

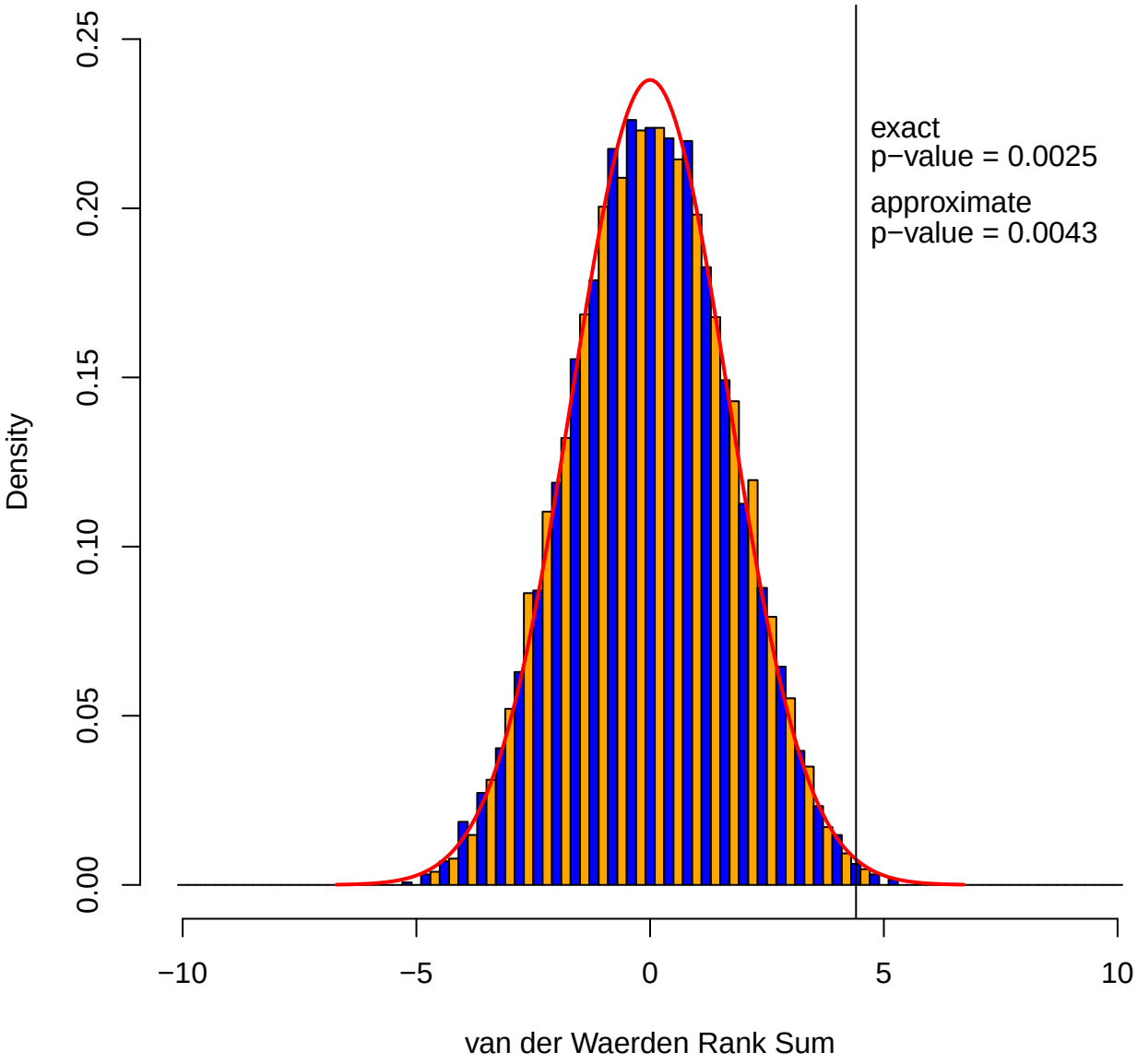
```
xx=c(p.val,p.val.norm)
if(flag==F){
   names(xx)=c("p.val.exact","p.val.appr")}else{
   names(xx)=c("p.val.sim","p.val.appr")
}
hist.out=hist(out,breaks=seq(-10.1,10.1,.2),main="",
    xlab="van der Waerden Rank Sum",probability=T,
    col=c("blue","orange"),ylim=c(0,.25))
x=seq(mu-4*sig,mu+4*sig,length.out=201)
z=dnorm(x,mu,sig)
lines(x,z,lwd=2,col="red")
delta=max(out)-min(out)
abline(v=vdW.obs)
if(flag==F){
  text(vdW.obs+.03*delta,max(hist.out$density)*.99,"exact",adj=0)}else{
  text(vdW.obs+.03*delta,max(hist.out$density)*.99,"simulated",adj=0)
}
text(vdW.obs+.03*delta,max(hist.out$density)*.95,
   paste("p-value =",round(p.val,4)),adj=0)
text(vdW.obs+.03*delta,max(hist.out$density)*.89,"approximate",adj=0)
text(vdW.obs+.03*delta,max(hist.out$density)*.85,
   paste("p-value =",round(p.val.norm,4)),adj=0)
if(PDF==T) dev.off()
xx
}
```

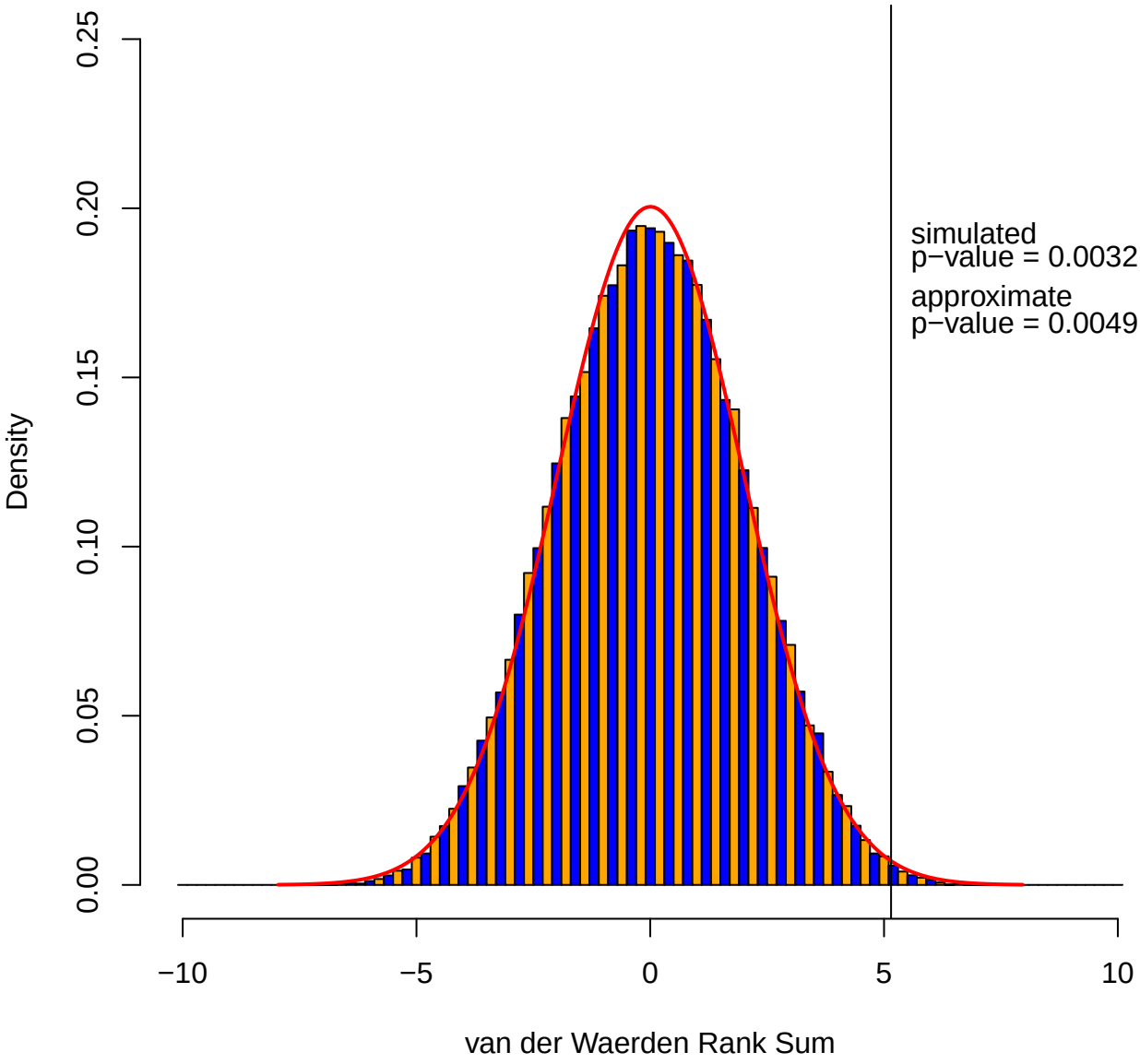
The runs for the three sample sets were as follows. The corresponding histograms follow.

```
> van.der.Waerden(x1,y1,PDF=T)
p.val.exact  p.val.appr
0.002175602 0.004326541
> van.der.Waerden(x2,y2,PDF=T)
p.val.exact  p.val.appr
0.002486402 0.004300139
> van.der.Waerden(x3,y3,PDF=T)
  p.val.sim  p.val.appr
0.003170000 0.004854839
```

The p-value approximations show high relative error compared to the exact (or simulated) p-values. However, because they are small in either case the hypothesis would typically be rejected.
The histogram shows symmetry in the first case but not in the other two situation. One can prove symmetry when the score vector Z_N is symmetric, which is the case when there are no ties. However, when there are ties the score vector Z_N most often is no longer symmetric unless the ties appear in a symmetric pattern around $(N+1)/2$.


exact
p−value = 0.0022
approximate
p−value = 0.0043
Density
0.00
0.05
0.10
0.15
0.20
0.25
−10
−5
0
5
10
van der Waerden Rank Sum


Density
0.00
0.05
0.10
0.15
0.20
0.25
−10
−5
0
5
10
van der Waerden Rank Sum
exact
p−value = 0.0025
approximate
p−value = 0.0043


Density
0.00
0.05
0.10
0.15
0.20
0.25
−10
−5
0
5
10
van der Waerden Rank Sum
simulated
p−value = 0.0032
approximate
p−value = 0.0049