Tortoise and Hare

Juntao Zhang

10/03/2021

Tortoise and Hare Racing Problem

After the famous fast tortoise and slow hare race, the team of 10 hares and the team of 10 tortoise had a rematch. Their finishing times are given in the dataset race.csv, in which the first column records team hares' finishing times, and the second column records team tortoise's finishing times.

Hypotheses

We are interested in testing whether the true mean finishing time is the same for team tortoise and team hare.

Let x's be the finishing times of hares and y's be the finishing times of the tortoises. Our null hypothesis is $\bar{x} = \bar{y}$ and the alternative hypothesis is $\bar{x} \neq \bar{y}$.

Using a one-sided alternative we can only test the effect size in one direction: $\bar{x} >$ or $< \bar{y}$. The effect is divided in two directions for the two-sided test. For the same size level, a one-sided alternative would have a larger power and larger size than the two-tailed test in one direction.

```
#read data
data <-read.csv("race.csv")
#calculate for the sample mean
h = mean(data$Hare)
t = mean(data$Tortoise)
#calculate the difference in sample means
h-t
```

```
## [1] -5.045642
```

The difference in sample mean is -5.05.

Same Variance Assumption

If we assume that the variance of the finishing time distributions for the two teams,then:

Let N_1 be the number of observations in group hare and $N2$ be the number of observations in group tortoise.

We know that the pooled variance (the variance representing both the variance of team hare and team tortoise based on the weighted average of the variances of both teams) is:

$S_p{}^2 = \frac{S^2_{hare}(N_1-1)+S^2_{tortoise}(N_2-1)}{N_1+N_2-2}$

Then:

$$\begin{aligned}
Var\left(\bar{X}_1 - \bar{X}_2\right) &= Var(\bar{X}_1) + Var(\bar{X}_2) - 2Cov(\bar{X}_1\bar{X}_2) \\
&= Var(\bar{X}_1) + Var(\bar{X}_2) \\
&= \frac{S_p^{\,2}}{N_1} + \frac{S_p^{\,2}}{N_2} \\
&= S_p^{\,2}(\frac{1}{N_1} + \frac{1}{N_2}) \\
&= \frac{S_{hare}^2(N_1 - 1) + S_{tortoise}^2(N_2 - 1)}{N_1 + N_2 - 2}(\frac{1}{N_1} + \frac{1}{N_2})
\end{aligned}$$

Since observations are independent, the means of the group are also independent. Thus the covariance of the means is zero. Due to assumption of equal variance of the two groups, the variance of means can be estimated using the pooled variance.

```
#calculate variance of the difference of two means using the equation above

n=nrow(data)
s_h = var(data$Hare)
s_t = var(data$Tortoise)
var_ht = (s_h*(n-1)+s_t*(n-1))/(n+n-2)*(1/n+1/n)
var_ht
```

```
## [1] 82.62257
```

$Var(\overline{X_1\text{-}X_2}) = 82.62$

T-test

```
t_sta=(h-t)/(var_ht^0.5)
t_sta
```

```
## [1] -0.5550947
```

```
2*pt(t_sta,n+n-2)
```

```
## [1] 0.5856628
```

The degree of freedom is 18 and the t-statistic is -0.555 with a p-value of 0.59.

```
qt(p=0.05/2,df=18)
```

```
## [1] -2.100922
```

At the level of 5% significance, the rejection region is {t>2.10} or {t<-2.10}. So the t score of -0.555 is not in the rejection region. We can conclude that we do not reject the null hypothesis of $\bar{x} = \bar{y}$.

Normality Assumption

The t test assumes normally distributed random variables thus we need to test for normality using the Shapiro Wilk test.

```
shapiro.test(data$Hare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Hare
## W = 0.52331, p-value = 6.83e-06
```

```
shapiro.test(data$Tortoise)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Tortoise
## W = 0.7704, p-value = 0.006324
```

The p-values are small, so both groups' observations are not normally distributed.

Equal Variance Assumption

Another assumption of the T-test is the equal variance of the two groups. We can use Levene's test to test for equal variance. Note that, Levene's test is less sensitive to the normality of the variables.

```
library(car)
```

```
## Loading required package: carData
```

```
#create new data frame
d2<-data.frame(group=rep(c("H","T"),each=10),time=c(data$Hare,data$Tortoise))

leveneTest(time ~ group,data=d2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1   0.613 0.4438
##       18
```

The p-value of the Levene's test is larger than 0.05, so we can conclude that we do not reject the hypothesis of equal variance of the two groups.

```
boxplot(time ~ group,
  data = d2,
  main = "Finishing Time Distribution by Group",
  xlab = "Group",
  ylab = "Finishing Time",
  names=c("Hare","Tortoise"),
  col = "steelblue",
  border = "black")
```

Finishing Time Distribution by Group
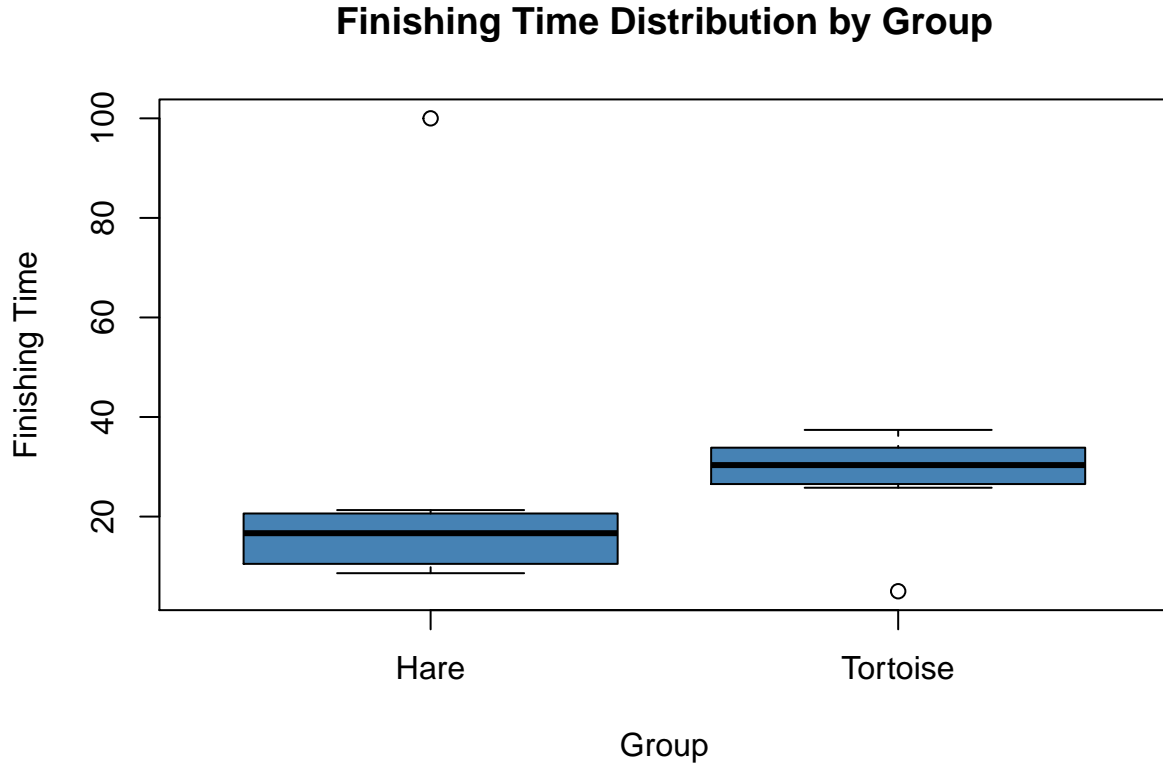
In addition to the tests, we can examine the box plots of the two groups. The lengths of the two boxes look similar and there is a clear difference between the means of two groups.

We can conclude that the two sample t-test is not appropriate for this problem because the sample size is relatively small and we cannot invoke the central limit theorem. The observations are not normally distributed.

The Mann-Whitney U-test

If the two teams are about the same in finishing times, then we would expect the number of hares passing the number of tortoises to be roughly the same as the number of tortoise passing the number of hares. In probability terms, $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$ should hold. Therefore, it is of interest to test the hypotheses:

$$H0 : P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$$
$$H1 : P(X_{hare} < X_{tortoise}) \neq P(X_{tortoise} < X_{hare})$$

U Statistic

The Mann-Whitney U-test may be used to test the above hypotheses. This test is based on calculating U-statistics that look at all pair-wise comparisons between members of the two teams and summarizes the total number of wins for one of the teams.

$$U_{hare} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{hare,i} < X_{tortoise,j})$$

$$U_{tortoise} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{tortoise,j} < X_{hare,i})$$

```
#Calculate the U-statistic
#for Hare:
u_h=0
for (i in 1:length(data$Hare)){
  u_h = u_h+sum(as.numeric(data$Hare[i] < data$Tortoise))
}
u_h
```

```
## [1] 81
```

```
#for Tortoise:
u_t = 0
for (i in 1:length(data$Tortoise)){
  u_t = u_t+sum(as.numeric(data$Tortoise[i] < data$Hare))
}
u_t
```

```
## [1] 19
```

So $U_{hare} = 81$ and $U_{tortoise} = 19$

Null Hypothesis of U Test

Under the null hypothesis, $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$ $P(X_{hare} < X_{tortoise}) + P(X_{tortoise} < X_{hare}) = 1$ So, $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare}) = 0.5$

$$\begin{aligned} E(U_{hare}) &= E(\sum_{i}^{10} \sum_{j}^{10} I) \\ &= \sum_{i}^{10} \sum_{j}^{10} E(I) \\ &= \sum_{i}^{10} \sum_{j}^{10} 0 \cdot P(X_{hare,i} > X_{tortoise,j}) + 1 \cdot P(X_{hare,i} < X_{tortoise,j}) \\ &= \sum_{i}^{10} \sum_{j}^{10} P(X_{hare,i} < X_{tortoise,j}) \\ &= 100 P(X_{hare} < X_{tortoise}) \\ &= 100 \cdot 0.5 \\ &= 50 \end{aligned}$$

Similarly, $E(U_{tortoise}) = 100 P(X_{tortoise} < X_{hare}) = 50$

The expected value for U-statistic of both groups is 50.

Z-test of U Statistic

```
#calculate the z-statistics
#for hare:
z_h = (u_h-50)/(10*10*(10+10+1)/12)^0.5
z_h
```

```
## [1] 2.34338
```

```
#p-value:
2*pnorm(-abs(z_h))
```

```
## [1] 0.01910992
```

```
#for tortoise:
z_t = (u_t-50)/(10*10*(10+10+1)/12)^0.5
z_t
```

```
## [1] -2.34338
```

```
#p-value
2*pnorm(-abs(z_t))
```

```
## [1] 0.01910992
```

For hare, the z-statistic is 2.34 and the p-value is 0.019. For tortoise, the z-statistic is -2.34 and the p-value is 0.019.

At a 0.05 significance level, the null hypothesis is rejected. So there is a difference between the two teams in terms of finishing time.

Alternatively, we can use the wilcox.text() function in R to verify these results.

```
wilcox.test(data$Hare,data$Tortoise,exact=F,correct =F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data$Hare and data$Tortoise
## W = 19, p-value = 0.01911
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(data$Tortoise,data$Hare,exact=F,correct =F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data$Tortoise and data$Hare
## W = 81, p-value = 0.01911
## alternative hypothesis: true location shift is not equal to 0
```

```r
wh=wilcox.test(data$Tortoise,data$Hare,exact=F,correct =F)$statistic
wt=wilcox.test(data$Hare,data$Tortoise,exact=F,correct =F)$statistic
```

The results are the same for the calculation and the r function.

Permutation Test

Permutation or randomization based tests are an alternative way to test these types of hypotheses. As with all other hypothesis tests, we must compute the sampling distribution for the test statistic under the null hypothesis. One way to construct this sampling distribution is to consider that when the null hypothesis is true, switching the group labels of the team members for the two teams should not affect the distributions of the expected outcomes or test statistics. Therefore, we can generate the null distribution by permuting the group labels for a large number of times, and computing any test statistic for each permutation.

Generate 3000 permuted datasets:

```r
#create an empty list and put all observations into one vector
obs = c(data$Hare,data$Tortoise)
d=list()
#sample 3000 times from the vector (re-arrangement of the observations) and put into the list
for (i in 1:3000){
  s <- sample(obs,20,replace=F)
  d[[i]] <-data.frame(Hare=s[1:10],Tortoise=s[11:20])
}
```

```r
#define functions
#for mean difference:
diff <-function(x){(mean(x$Hare)-mean(x$Tortoise))}

#for std error of the mean difference
sd_mean<-function(x){((var(x$Hare)+var(x$Tortoise))*0.1)^0.5}

#for t statistic:
t_stat <-function(x){diff(x)/sd_mean(x)}

#for u statistic:
u_hare <-function(x){u_h=0
for (i in 1:length(x$Hare)){
  u_h = u_h+sum(as.numeric(x$Hare[i] < x$Tortoise))}
return(u_h)}

u_tortoise<-function(x){u_t = 0
for (i in 1:length(x$Tortoise)){
  u_t = u_t+sum(as.numeric(x$Tortoise[i] < x$Hare))
}
return(u_t)}

#z-statistic
z_stath <- function(x){(u_hare(x)-50)/(175^0.5)}
z_statt <- function(x){(u_tortoise(x)-50)/(175^0.5)}

#Wilcox's rank sum statistics
```

```r
w_t <-function(x){wilcox.test(x$Hare,x$Tortoise,exact=F,correct =F)$statistic}
w_h<-function(x){wilcox.test(x$Tortoise,x$Hare,exact=F,correct =F)$statistic}
```

```r
#calculate the required statistics for each permuted dataset
#difference of mean

dm <- sapply(d,diff)

#t-statistic
dt<-sapply(d,t_stat)

#u-statistic
#for team hare:
ush<- sapply(d,u_hare)
#for team tortoise:
ust<-sapply(d,u_tortoise)

#z-statistic
#for team hare:
zsh<-sapply(d,z_stath)
#for team tortoise:
zst<-sapply(d,z_statt)

#wilcox's rank sum statistic
#for team hare:
wsh<-sapply(d,w_h)
#for team tortoise:
wst<-sapply(d,w_t)
```

```r
#create histograms for the statistics calculated
#add vertical line for p-value calculation
#calculate p-value for each statistic
par(mfrow=c(1,2))
#difference of mean
hist(dm,main="Histogram of difference in group mean",xlab="Difference in group mean")
abline(v=h-t,col="blue")

#t-statistic
hist(dt,main="Histogram of t-statistic",xlab="t-statistic")
abline(v=t_sta,col="blue")
```

Histogram of difference in group mean

Difference in group mean

Histogram of t–statistic
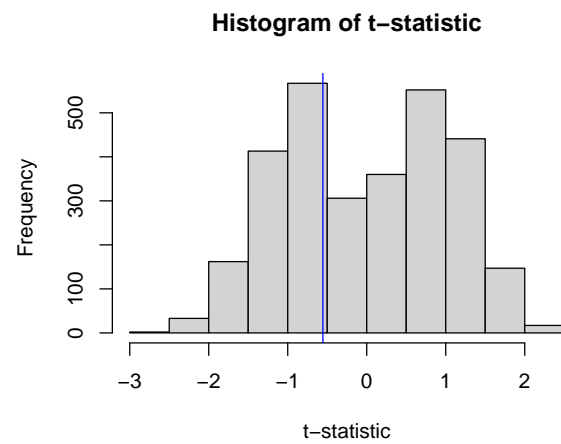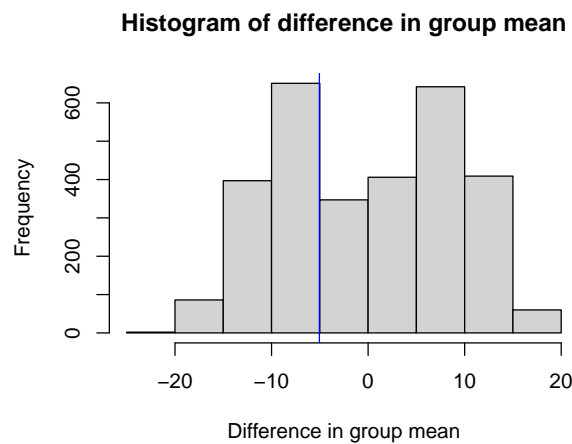
t–statistic

```r
#u-statistic
par(mfrow=c(1,2))
#for team hare:
hist(ush,main="Histogram of u-statistic(hare)",xlab="u-statistic")
abline(v=u_h,col="blue")
#for team tortoise:
hist(ust,main="Histogram of u-statistic(tortoise)",xlab="u-statistic")
abline(v=u_t,col="blue")
```

Histogram of u–statistic(hare)

u–statistic

## Histogram of u–statistic(tortoise)
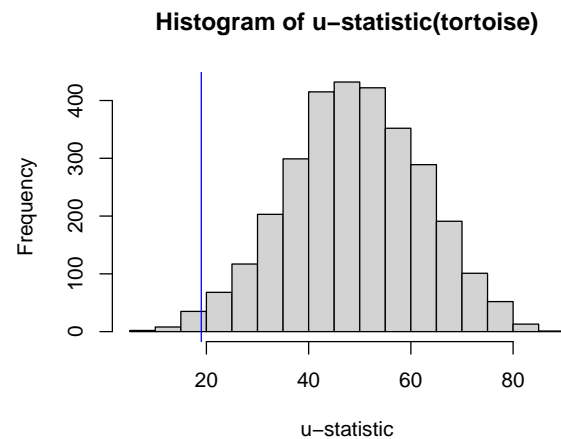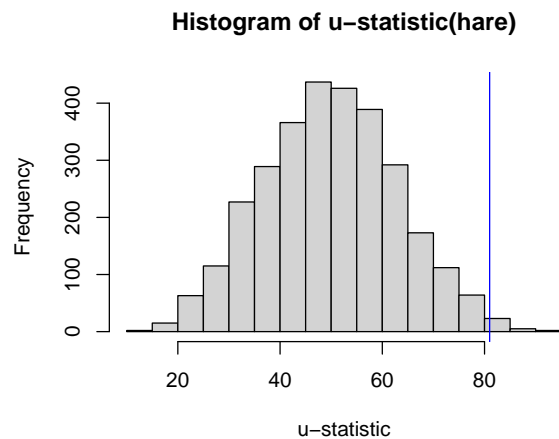
u–statistic

```r
#z-statistic
par(mfrow=c(1,2))
#for team hare:
hist(zsh,main="Histogram of z-statistic(hare)",xlab="z-statistic")
abline(v=z_h,col="blue")
#for team tortoise:
hist(zsh,main="Histogram of z-statistic(tortoise)",xlab="z-statistic")
abline(v=z_t,col="blue")
```

Histogram of z–statistic(hare)

Histogram of z–statistic(tortoise)
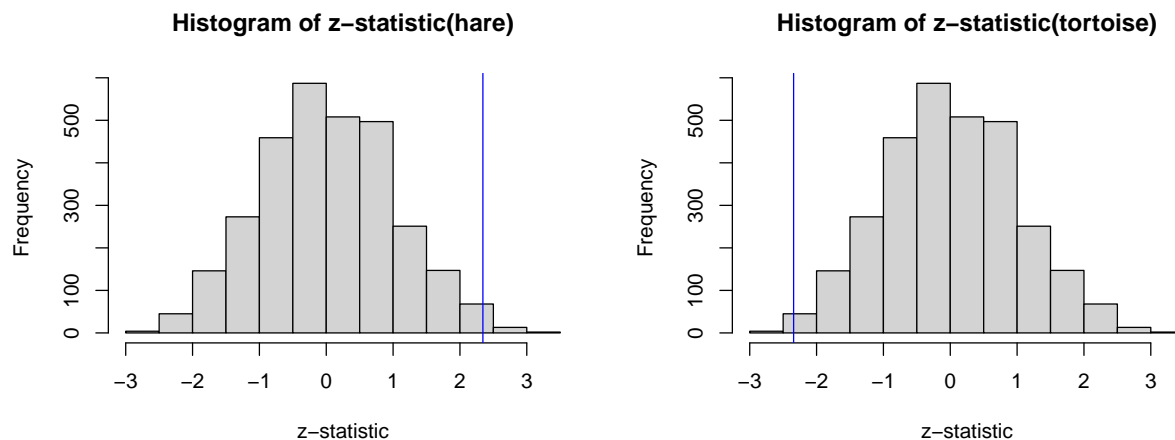

```
#wilcox's rank sum statistic
par(mfrow=c(1,2))
#for team hare:
hist(wsh,main="Histogram of W-statistic(hare)",xlab="w-statistic")
abline(v=wh,col="blue")
#for team tortoise:
hist(wst,main="Histogram of W-statistic(tortoise)",xlab="w-statistic")
abline(v=wt,col="blue")
```

Histogram of W–statistic(hare)

Histogram of W–statistic(tortoise)
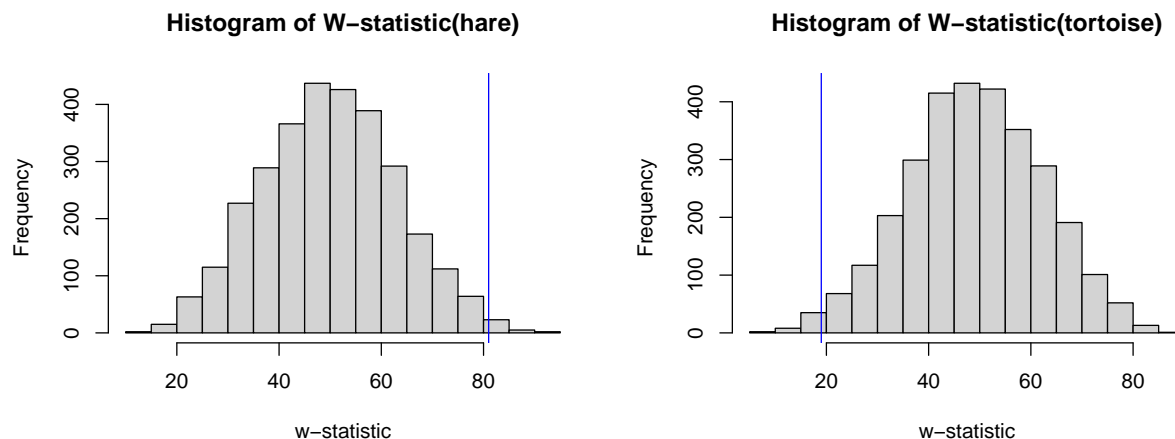
The distributions of the statistics calculated are the same as sampling distribution under the null hypothesis. Under null hypothesis, the two groups are the same in terms of finishing time. There is no significant difference between the two groups, so it does not matter how we label the data points. Thus the sampling distributions of statistics of the permuted datasets should show the same distribution as the sampling distribution under the null hypothesis.

All histograms are roughly symmetric. The difference in mean and the t-statistic histograms have similar shape, but they do not show a t-distribution shape. The u,z,and w-statistics histograms have a bell shape.

I expected the mean value of difference in mean,t-statistic and z-statistic to be 0. The mean value of u-statistic and w-statistic to be 50. The mean value of these statistics should be the same as what the null hypothesis suggested.

The p-values can be calculated by using the observed statistics from the original race data set and applying the central limited theorem (since we have 3000 samples for each statistic). The vertical lines added to the plot are the values of the observed data.

```
#Calculate p-values for the sampling distributions
#for difference in mean:
2*pnorm(-abs((h-t-mean(dm))/sd(dm)))
```

```
## [1] 0.5786716
```

```
#for t-statistic:
2*pnorm(-abs((t_sta-mean(dt))/sd(dt)))
```

```
## [1] 0.5960947
```

```
#for u-statistic:
2*pnorm(-abs((u_h-mean(ush))/sd(ush)))
```

```
## [1] 0.01979277
```

```
2*pnorm(-abs((u_t-mean(ust))/sd(ust)))
```

```
## [1] 0.01979277
```

```
#for z-statistic:
2*pnorm(-abs((z_h-mean(zsh))/sd(zsh)))
```

```
## [1] 0.01979277
```

```
2*pnorm(-abs((z_t-mean(zst))/sd(zst)))
```

```
## [1] 0.01979277
```

```
#for wilcox's rank sum statistic
2*pnorm(-abs((wh-mean(wsh))/sd(wsh)))
```

```
##          W
## 0.01979277
```

```
2*pnorm(-abs((wt-mean(wst))/sd(wst)))
```

```
##          W
## 0.01979277
```

From the p-values calculated, results are similar to the results from the previous tests. The difference in mean and t-statistic do not reject the null hypothesis. The U-statistic, Z-statistic and W-statistic reject the null hypothesis.

Summary

Under the same null hypothesis of hare and tortoise having the same mean finishing time of the race, the two sample t-test and the Mann-Whitney test yield different results. The two sample t-test does not reject the null hypothesis and the Mann-Whitney test reject the null hypothesis.

However, there are several underlying assumptions for the two sample t-test. It requires the two groups of data points be drawn independently from normal distributions with the same variance. After testing for the assumptions, both Hare and Tortoise finishing times have the variance but are not normally distributed. Since the normality assumption is not satisfied, the test results are not validated.

The Mann-Whitney test, on the other hand, does not assume the data follow any particular distribution(Rice,435). It is a non-parametric method. The calculated u-statistics and the Wilcoxon test in R yield the same results. They both reject the null hypothesis that hares and tortoises are the same in terms of finishing time. The result is more reliable than the two sample t-test because this method does not depend on an assumption of normality.

The randomization based tests yield the same results as the previous tests. The difference in mean statistics and the t statistic still shows no significant evidence for the alternative hypothesis. The U statistics, Z statistics and W statistics shows significant evidence for the alternative hypothesis.

When the sample size is relatively small and there is no prior knowledge on the distribution of the observations, choosing the Mann-Whitney test is safer. However, the Mann-Whitney test is not sensitive to outliers since the actual values of the observations are replaced by their ranks (Rice,439). When the assumptions of the two-sample t tests are confirmed and the sample size is relatively larger,the two sample t tests are more powerful.

The final conclusion of the hypothesis test is that the finishing time of hares and tortoises is different.

Reference

Rice, John. "Mathematical Statistics and Data Analysis." Mathematical Statistics and Data Analysis. Brantford, Ontario: W. Ross MacDonald School Resource Services Library, 2015. 421-42. Print.