Hypothesis Testing Project

Frequentist Inference

Liuyang Xu

03/10/2022

Libraries and Data

```
race <- read.csv("race.csv")
```

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. Your task is to follow the instructions in each of the questions to make statistical inference about the tortoise and hare race problem.

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

(a) Specify the appropriate null and alternative hypotheses for the problem of interest using a two-sided alternative hypothesis. Comment on the implications regarding size and power of using a two-sided versus a one-sided alternative in this case. **Answer:** Using a two-sided alternative hypothesis, the *null hypotheses* for the problem of interest is that "the true means finishing time are the same for team tortoise and team hare" and the *alternative hypotheses* for the problem of interest is that "the true means finishing time are not the same for team tortoise and team hare".
If using a one-sided alternative hypothesis, the *null hypotheses* would remain the same. But the *alternative hypotheses* can be "the true mean finishing time for team tortoise is larger than team hare", or can also be "the true mean finishing time for team tortoise is larger than team hare".
Since the relationship between size and power is $\alpha = size = \max_{\theta \to \theta_0} \beta(\theta)$ and power is $\beta(\theta) = P_\theta \{(X_1, ..., X_n) \in C\}$, once we set the same size(α), compared to the two-sided alternative, we would gain power in the direction of the alternative hypotheses when using the one-sided alternative.

(b) Assume the finishing times of all racers are independent from each other, find the difference in sample mean in finishing times for the two teams, $\overline{X}_{hare} - \overline{X}_{tortoise}$, by calculating this quantity in `R`

```
Sample_Mean_Diff <- mean(race$Hare) - mean(race$Tortoise)
```

Answer: The difference in sample mean in finishing times for the two teams, $\overline{X}_{hare} - \overline{X}_{tortoise} = -5.045642$

(c) If we assume that the variance of the finishing time distributions for the two teams are equal, show that the variance of the sampling distribution of $\overline{X}_{hare} - \overline{X}_{tortoise}$ can be estimated as follows,

$$Var(\overline{X}_1 - \overline{X}_2) = \left(\frac{S^2_{hare}(N_1 - 1) + S^2_{tortoise}(N_2 - 1)}{N_1 + N_2 - 2} \right) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

by relating $Var(\overline{X}_1 - \overline{X}_2)$ to the quantities $Var(\overline{X}_1)$ and $Var(\overline{X}_2)$. Make sure to justify each step in your derivation. Note that S^2_{hare} and $S^2_{tortoise}$ are the sample variance of the two teams, N_1 and N_2 are the numbers of hares and tortoises. Note that you may look up and use the standard definition of the

pooled variance. **Answer:**
According to the variance of the sum of two random variables property, we have

$$Var(\bar{X_1} - \bar{X_2}) = Var(\bar{X_1}) + Var(\bar{X_2}) - 2Cov(\bar{X_1}, \bar{X_2})$$

Since X_1 and X_2 are independent, we have $Cov(\bar{X_1}, \bar{X_2}) = 0$, so $Var(\bar{X_1} - \bar{X_2}) = Var(\bar{X_1}) + Var(\bar{X_2})$. then we have

$$Var(\bar{X_1} - \bar{X_2})$$
$$= Var(\frac{1}{N_1}\sum_{i=1}^{N_1} X_{1i}) + Var(\frac{1}{N_2}\sum_{j=1}^{N_2} X_{2j})$$
$$= \frac{1}{N_1}Var(\sum_{i=1}^{N_1} X_{1i}) + \frac{1}{N_2}Var(\sum_{j=1}^{N_2} X_{2j})$$

Since the X_{1i} are independent and X_{2j} are independent, we have

$$Var(\bar{X_1} - \bar{X_2})$$
$$= (\frac{1}{N_1})^2\sum_{i=1}^{N_1} Var(X_i) + (\frac{1}{N_2})^2\sum_{j=1}^{N_2} Var(X_j)$$
$$= (\frac{1}{N_1})^2 \times N_1 \times Var(X_1) + (\frac{1}{N_2})^2 \times N_2 \times Var(X_2)$$
$$= \frac{1}{N_1}Var(X_1) + \frac{1}{N_2}Var(X_2)$$

According to the assumption in the question, $Var(X_1) = Var(X_2)$. Let $Var(X)$ be the pooled variance of X_1 and X_2, so we have

$$Var(\bar{X_1} - \bar{X_2})$$
$$= (\frac{1}{N_1} + \frac{1}{N_2})Var(X)$$

According to wikipedia, the pooled variance can be estimated in the following way

$$Var(X) = \frac{(N_1 - 1)S_2^2 + (N_2 - 1)S_1^2}{N1 + N2 - 2}$$

So we can have

$$Var(\overline{X}_1 - \overline{X}_2) = \left(\frac{S^2_{hare}(N_1 - 1) + S^2_{tortoise}(N_2 - 1)}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)$$

(d) Calculate the above quantity $Var(\overline{X}_1 - \overline{X}_2)$ in `R`

```
hare.s2 <- var(race$Hare) # the sample variance of the hare team
tortoise.s2 <- var(race$Tortoise)# the sample variance of the tortoise team

N1 <- length(race$Hare) # the number of hares
N2 <- length(race$Tortoise) # the number of tortoises

# the variance of the sampling distribution of sample mean difference
Sampling_Dis_Var <- (hare.s2*(N1-1) + tortoise.s2*(N2-1))/(N1+N2-2)*(1/N1+1/N2)
```

Answer: The above quantity $Var(\overline{X}_1 - \overline{X}_2) = 82.62257$

(e) Use the independent two-sample t-test to test the hypotheses in (a). The test statistic for such a problem is given as follows,

$$t = \frac{\overline{X}_{hare} - \overline{X}_{tortoise}}{\text{std.err}(\overline{X}_{hare} - \overline{X}_{tortoise})}$$

i. Under the assumption that the difference in sample mean is normally distributed, this test statistic follows a t distribution with degrees of freedom $N_1 + N_2 - 2$ under the null hypothesis. Calculate the test statistic and the p-value for this dataset.

```
t <- Sample_Mean_Diff / (sqrt(Sampling_Dis_Var)) # test statistic

dof <- N1 + N2 - 2 # degrees of freedom

p <- 2 * pt(q = t, df = dof) # p-value
```

Answer: The test statistic is -0.5550947.
The p-value is 0.5856628.

ii. Setting the level of test at 5%, report the rejection region for this problem, and report your conclusion of this hypothesis test.

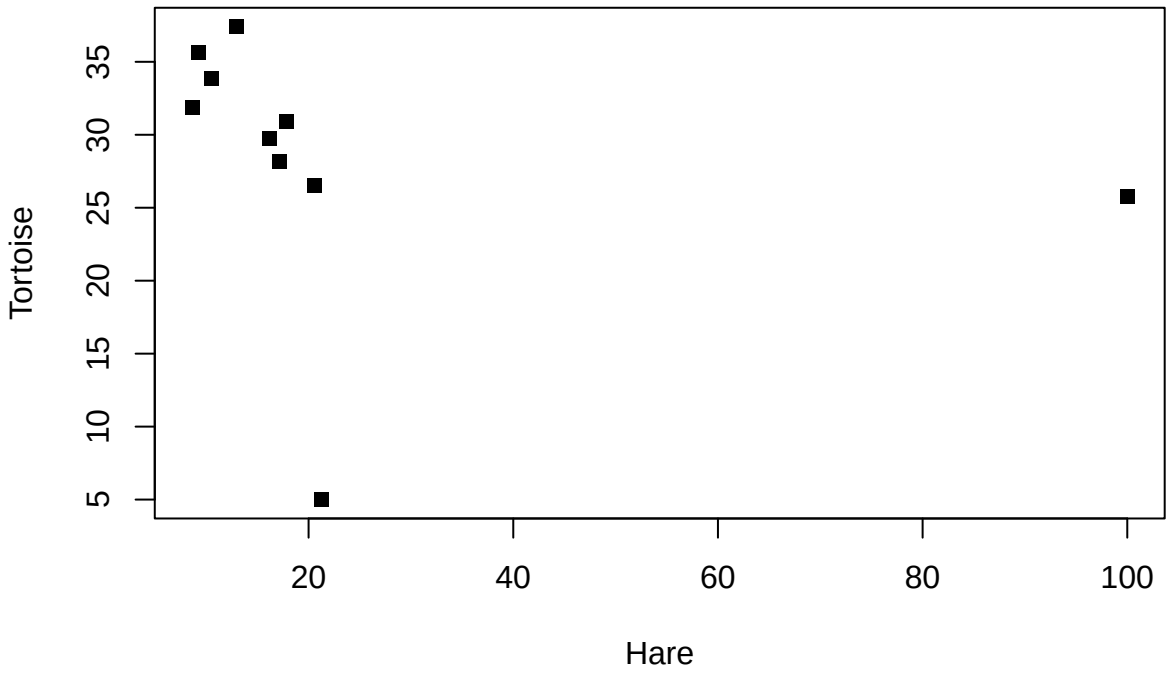
```
qt(p=.05/2, df=dof, lower.tail=FALSE) # critical_value
```

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

Answer: The degrees of freedom is 18 and the the level of this test is at 5%, the rejection region is $t \geqslant 2.100922$ and $t \leqslant -2.100922$.
We can not reject the null hypothesis because the test statistic doesn't fall into the rejection region and the p-value is larger than the level 5%.

iii. Is the two-sample t-test used in this problem appropriate? Justify your answer by checking the assumptions of the test you just performed. You may use pre-coded hypothesis test functions for this question.

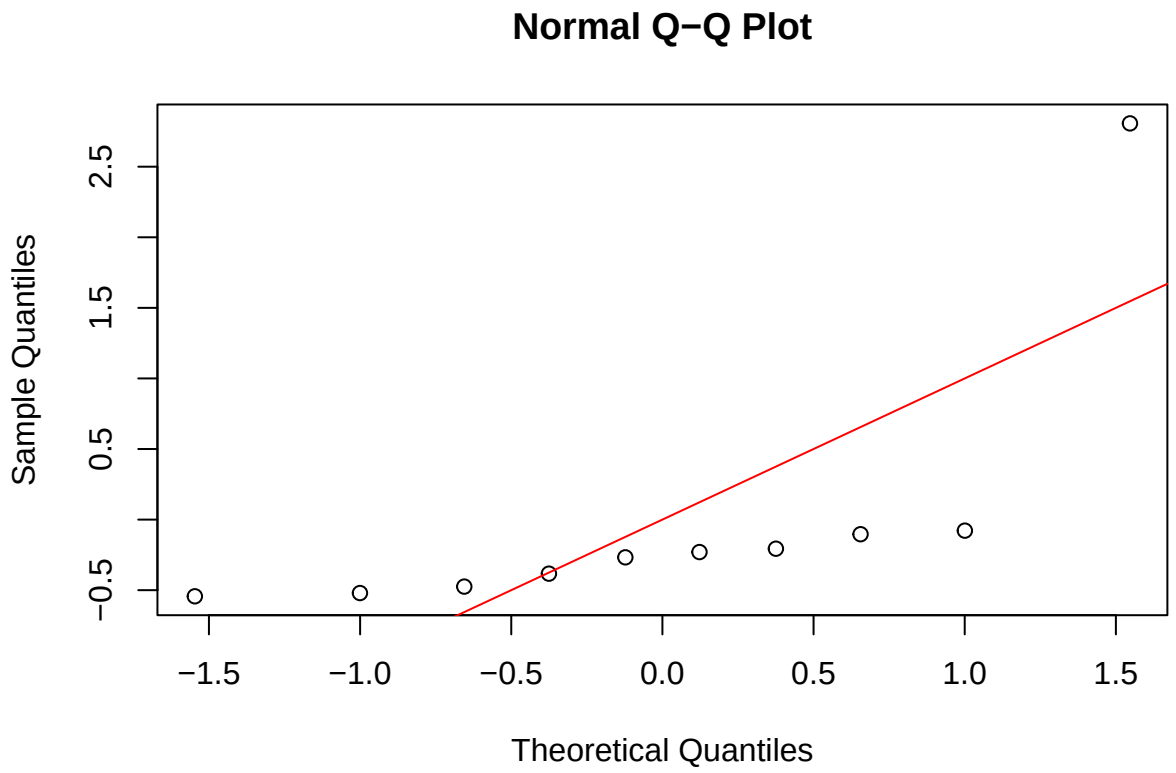
```
# Plot the two against each other
plot(race$Hare, race$Tortoise, pch = 15, xlab = "Hare", ylab = "Tortoise")
```



```r
# Default on cor.test is Pearson's correlation
cor.test(race$Hare, race$Tortoise)
```

```
##
##  Pearson's product-moment correlation
##
## data:  race$Hare and race$Tortoise
## t = -0.63136, df = 8, p-value = 0.5454
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7452574  0.4772303
## sample estimates:
##        cor
## -0.2178569
```

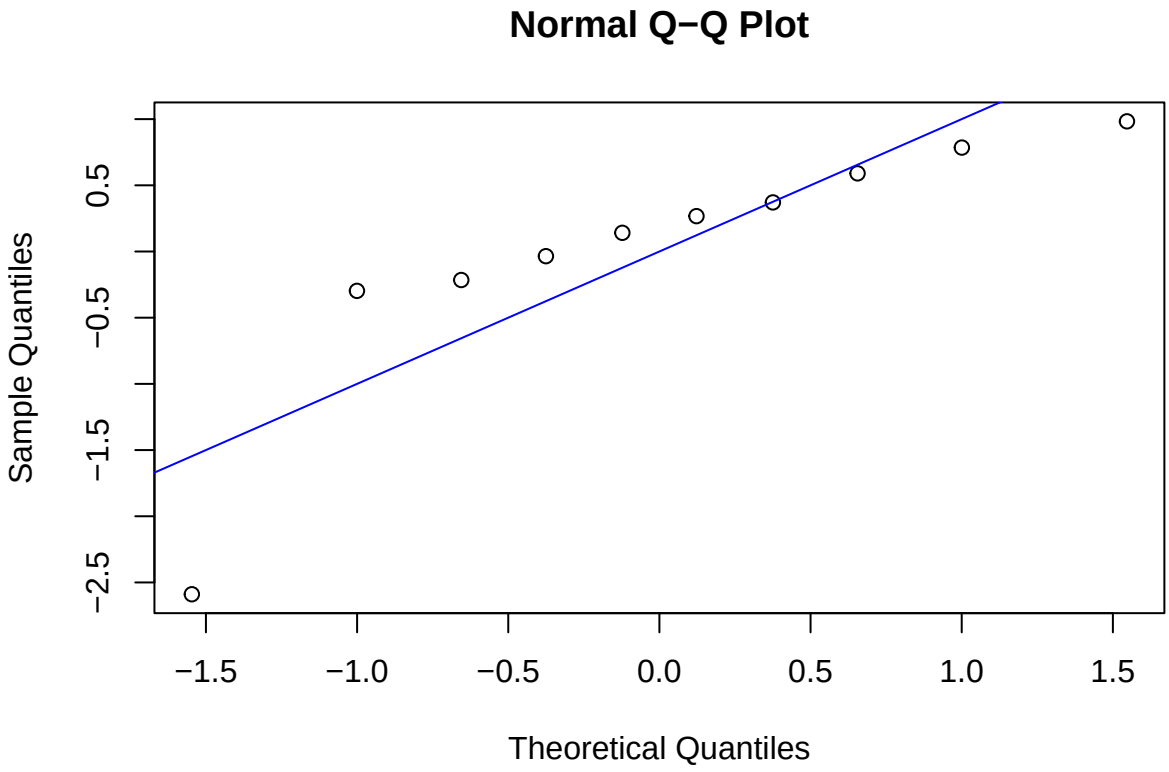
```r
# Check for normality
qqnorm(scale(race$Hare))
abline(0,1, col="red")
```




```r
shapiro.test(race$Hare)
```

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

```r
# Check for normality
qqnorm(scale(race$Tortoise))
abline(0,1, col="blue")
```




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

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

```
# check if they have the same variance
var.test(race$Hare, race$Tortoise)
```

```
##
##  F test to compare two variances
##
## data:  race$Hare and race$Tortoise
## F = 9.039, num df = 9, denom df = 9, p-value = 0.00306
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   2.245169 36.391105
## sample estimates:
## ratio of variances
##           9.039036
```

Answer: When the two data group values are independent, are randomly sampled from two normal populations and the two independent groups have equal variances, we can use the two-sample t-test.

When plotting the two against each other, it seems the data are independent. The p-value is 0.5454 in Pearson's product-moment correlation also indicates that the independency holds.

When using Q-Q plot and the Shapiro-Wilk normality test on both 2 groups, the points in the Q-Q plot are not around the y=x line and the p-values are 6.83e-06 and 0.006324 which all lead to conclusion that these two groups of data dont come from the normal distribution.

When using the F test to compare two variances, the p-value is 0.00306 which reject the null hypothesis that true ratio of variances is equal to 1. SO the two variances are no the same.

The assumption of normality and the equal variance dont hold so the two-sample t-test used in this problem isn't appropriate.

2. Let's consider a different 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:

$$H_0 : P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$$
$$H_1 : P(X_{hare} < X_{tortoise}) \neq P(X_{tortoise} < X_{hare})$$

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.

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

where $I(\cdot)$ is an indicator function. For example, if $X_{hare,1} = 0.5$ and $X_{tortoise,1} = 0.6$, then $I(X_{hare,1} < X_{tortoise,1}) = 1$ and $I(X_{tortoise,1} < X_{hare,1}) = 0$. Note that in order to win a race, your finishing time must be shorter than your opponent. For the sake of simplicity, assume that the times are recorded with fine-grained precision so that there are no exact ties.

(a) Calculate the U-statistic for each of the teams in `R`. Do this by hand, not by using any formulas that relate U to a Wilcoxon's statistic.

```
U.hare <- 0
U.tortoise <- 0

for (i in 1:nrow(race)){
  for (j in 1:nrow(race)){
    ifelse(race$Hare[i] < race$Tortoise[j],
           U.hare <- U.hare + 1,
           U.tortoise <- U.tortoise + 1)
  }
}
```

Answer: The U-statistic for team hare is 81.
The U-statistic for team tortoise is 19.

(b) Under the null hypothesis, given that there are 10 members on each team, what is the expected value of the U-statistic for each team? Explain how you arrived at this answer. You can either show some mathematical derivations or explain it in heuristic terms. **Answer:** The expected values of the U-statistic for team hare and team tortoise are both 50.
Since we assume that the times are recorded with fine-grained precision so that there are no exact ties, the event $X_{hare} < X_{tortoise}$ and $X_{hare} > X_{tortoise}$ are mutually exclusive. Therefore $P(X_{hare} < X_{tortoise}) + P(X_{tortoise} < X_{hare}) = 1$. Under the null hypothesis $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$, we can have $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare}) = 0.5$
So given that there are 10 members on each team, each hare in the team hare would smaller than $10 * 0.5 = 5$ of the tortoise in the team tortoise, and vice versa. So in total it would be $5 * 10 = 50$

(c) Under the null hypothesis, when the sample size is large enough, the U-statistic is approximately normally distributed. The mean for this distribution, μ_{U_0}, was calculated in the previous part (2.b).

The standard deviation for this normal distribution is $\sigma_{U_0} = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$. Therefore, we may use the following test statistic:

$$Z = \frac{U - \mu_{U_0}}{\sigma_{U_0}} \tag{1}$$

i. Calculate the z-statistics for the Mann-Whitney U-test and report the appropriate p-values.

```
n1 <- length(race$Hare)
n2 <- length(race$Tortoise)

mu_U0 <- 50
sigma_U0 <- sqrt(n1*n2*(n1+n2+1)/12)

z.hare <- (U.hare - mu_U0)/sigma_U0
z.tortoise <- (U.tortoise - mu_U0)/sigma_U0

pval.z.hare <- pnorm(abs(z.hare), lower.tail = F)*2
pval.z.tortoise <- pnorm(abs(z.tortoise), lower.tail = F)*2
```

Answer: The z-statistics for hare is 2.34338 and the p-value is 0.01910992. The z-statistics for tortoise is -2.34338 and the p-value is 0.01910992.

ii. Report your conclusion for this hypothesis test at the $\alpha = 0.05$ significance level. **Answer:** The p-value is 0.01910992 which is smaller than 0.05. So I reject the null hypothesis which means that the conclusion is $P(X_{hare} < X_{tortoise}) \neq P(X_{tortoise} < X_{hare})$

iii. Mann-Whitney U test is sometimes referred to as a version of the Wilcox rank sum test. Use `wilcox.test` function in `R` to test the same hypothesis and compare your results. Set options `exact=F, correct=F` when running your `wilcox.test` function.

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

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

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

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

Answer: The p-values are the same comparing to my previous results above.

3. 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. Note that permutations should not be done in a pairwise fashion (that is, do not switch across rows, switch across the whole dataset).

(a) Generate 3000 permuted datasets as described above.

```
Itr <- 3000
permuted_dataset.total <- array(NA, dim=c(nrow(race),2,Itr),dimnames = list(c(1:10),c("Hare","Tortoise"))

observed_data = c(race$Hare,race$Tortoise)
for (i in 1:Itr){
  permuted_dataset.total[,,i] <- matrix(sample(observed_data), nrow = nrow(race), ncol = 2)
}
```

(b) For each permuted dataset, calculate:

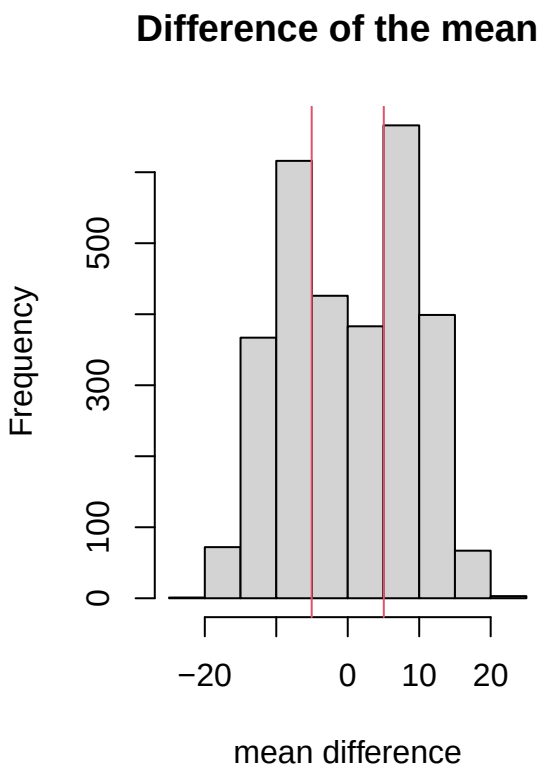
- $\overline{X}_{hare} - \overline{X}_{tortoise}$
- The t statistic as in equation (1)
- U_{hare} and $U_{tortoise}$
- The Z statistics as in equation (2)
- Wilcox's rank sum statistics for team Hare(W_{hare}) and for team Tortoise($W_{tortoise}$)

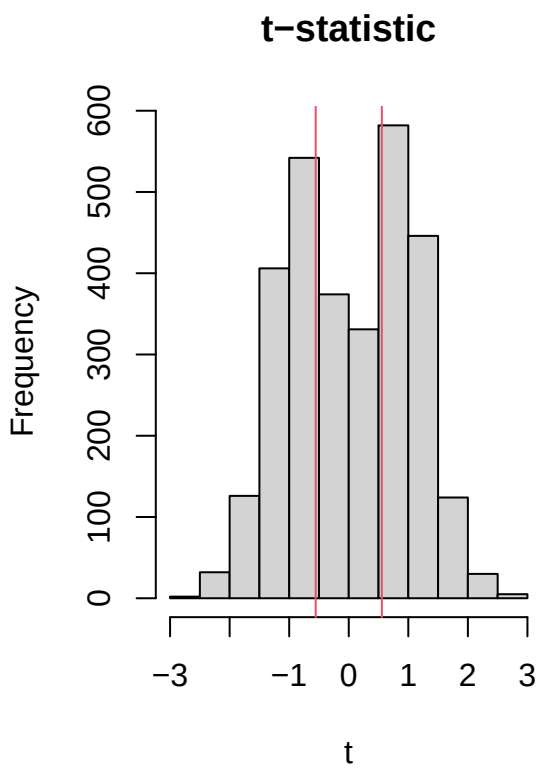
```
Sample_Mean_Diff.sampling <- rep(NA, Itr)

t.sampling <- rep(NA, Itr)

U.hare.sampling <- rep(0, Itr)
U.tortoise.sampling <- rep(0, Itr)

z.hare.sampling <- rep(NA, Itr)
z.tortoise.sampling <- rep(NA, Itr)

W.hare.sampling <- rep(NA, Itr)
W.tortoise.sampling <- rep(NA, Itr)

for (i in 1:Itr){
  pm_dat <- as.data.frame(permuted_dataset.total[,,i]) # convert the subset data to data frame

  N1 <- length(pm_dat$Hare)
  N2 <- length(pm_dat$Tortoise)

  # X_bar(hare) - X_bar(tortoise)
  Sample_Mean_Diff.sampling[i] <- mean(pm_dat$Hare) - mean(pm_dat$Tortoise)

  # The t statistic as in equation (1)
  t.sampling[i] <- (mean(pm_dat$Hare) - mean(pm_dat$Tortoise))/
    (sqrt((var(pm_dat$Hare)*(N1-1) + var(pm_dat$Tortoise)*(N2-1))/(N1+N2-2)*(1/N1+1/N2)))

  # U.hare and U.tortoise
  for (j in 1:nrow(pm_dat)){
    for (k in 1:nrow(pm_dat)){
    ifelse(pm_dat$Hare[j] < pm_dat$Tortoise[k],
           U.hare.sampling[i] <- U.hare.sampling[i] + 1,
           U.tortoise.sampling[i] <- U.tortoise.sampling[i] + 1)
      }
  }

  # The Z statistics as in equation (2)
  z.hare.sampling[i] <- (U.hare.sampling[i] - mu_U0)/sqrt(N1*N2*(N1+N2+1)/12)
  z.tortoise.sampling[i] <- (U.tortoise.sampling[i] - mu_U0)/sqrt(N1*N2*(N1+N2+1)/12)
```

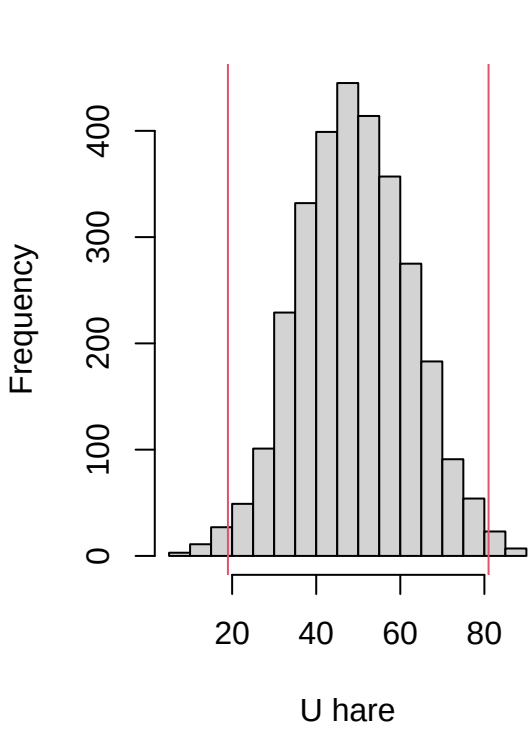
```
  # Wilcox's rank sum statistics for team Hare(W_hare) and for team Tortoise(W_tortoise)
  W.hare.sampling[i] <-  sum(rank(pm_dat)[1:N1])
  W.tortoise.sampling[i] <-  sum(rank(pm_dat)[(N1+1):(N1+N2)])
}
```

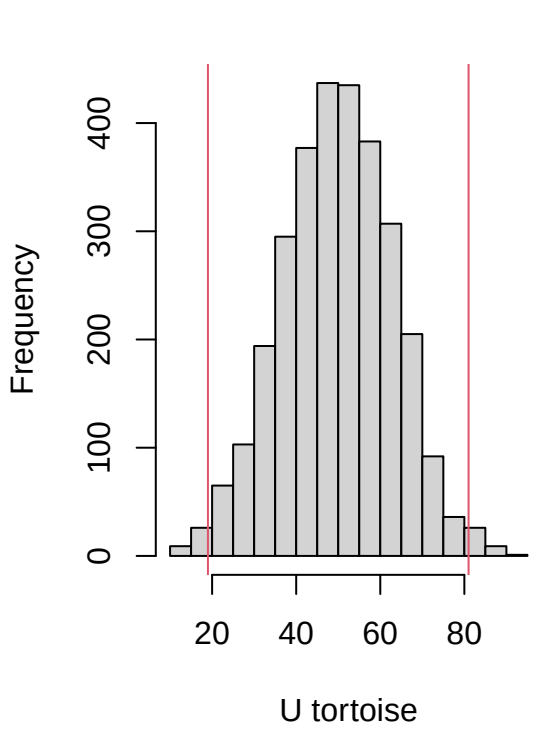
(c) For each of the quantities you calculated in (3.b), use a histogram to display its distribution.

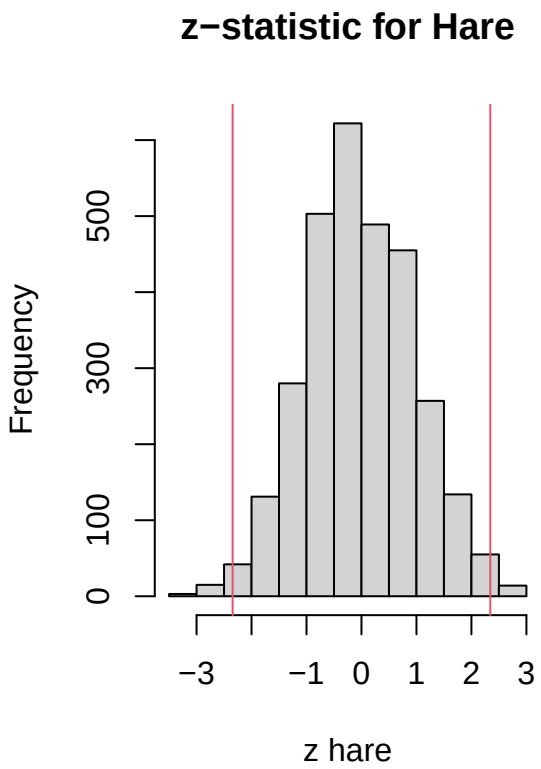
```
# Wilcox's rank sum statistics for team Hare(W_hare) and for team Tortoise(W_tortoise)
W.hare <-  sum(rank(race)[1:N1])
W.tortoise <-  sum(rank(race)[(N1+1):(N1+N2)])

quantities.sampling <- cbind(Sample_Mean_Diff.sampling, t.sampling,
                    U.hare.sampling, U.tortoise.sampling,
                    z.hare.sampling, z.tortoise.sampling,
                    W.hare.sampling, W.tortoise.sampling)

quantities.sample <- cbind(Sample_Mean_Diff, t,
                    U.hare, U.tortoise,
                    z.hare, z.tortoise,
                    W.hare, W.tortoise)

main_names <- c("Difference of the means", "t-statistic",
                "U-statistic for Hare", "U-statistic for Tortoise",
                "z-statistic for Hare", "z-statistic for Tortoise",
                "rank sum statistics for Hare", "rank sum statistics for Tortoise")

xlab_names <- c("mean difference", "t",
                "U hare", "U tortoise",
                "z hare", "z tortoise",
                "W hare", "W tortoise")

p_value <- t(rep(NA, 8))
colnames(p_value) <- c("Sample_Mean_Diff", "t",
                       "U.hare", "U.tortoise",
                       "z.hare", "z.tortoise",
                       "W.hare", "W.tortoise")

lines <- rbind(quantities.sample, -quantities.sample)
lines[2,3] <- U.tortoise
lines[2,4] <- U.hare
lines[2,7] <- W.tortoise
lines[2,8] <- W.hare

par(mfrow=c(1, 2))
for (i in 1:8){
  hist(quantities.sampling[,i], main = main_names[i], xlab = xlab_names[i])

  abline(v = lines[1,i], col = 2)
  abline(v = lines[2,i], col = 2)
}
```

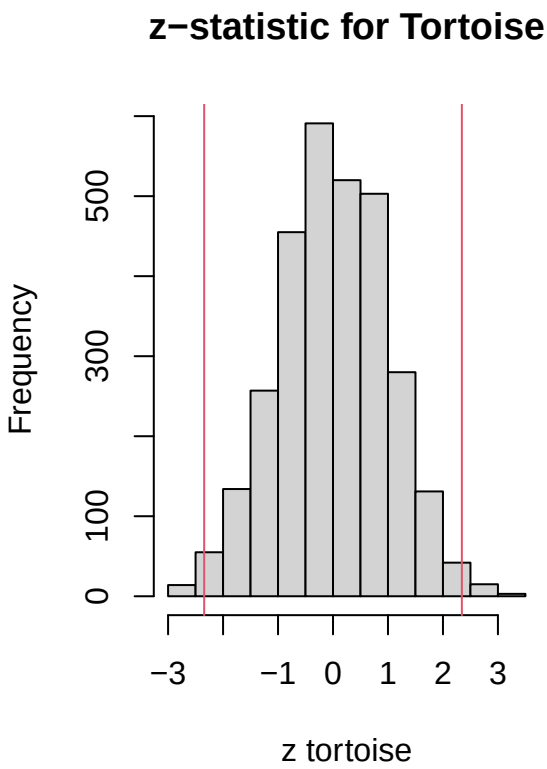

Difference of the means
Frequency
0
100
300
500
−20
0
10
20
mean difference

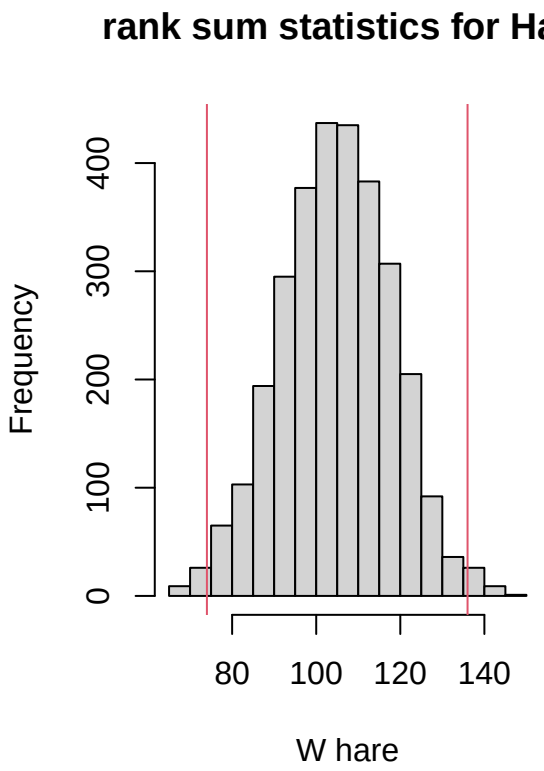

t−statistic
Frequency
0
100
200
300
400
500
600
−3
−1
0
1
2
3
t

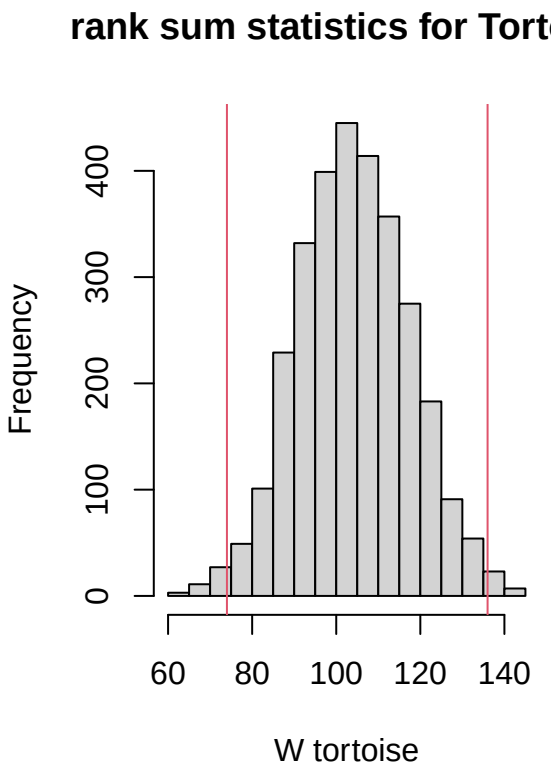

U−statistic for Hare
Frequency
0
100
200
300
400
20
40
60
80
U hare


U−statistic for Tortoise
Frequency
0
100
200
300
400
20
40
60
80
U tortoise









- Explain how these distributions are related to the sampling distribution of the test statistics. Comment on the similarities and/or differences of these distributions you observe. **Answer:** We generated 3000 permuted datasets which is quite large, so these distributions would approximate to the sampling distribution. The similarity is that all of these distributions have the shape of normal distributions. The difference is that mean difference and t statistics have two peaks in the graph while the others

have only one peak.

- What do you expect the mean value of each of the sampling distributions to be? **Answer:** I expect the mean value of the mean difference to be 0, the mean value of the t-statistic to be 0, the mean value of the U-statistics for both teams to be 50, the mean value of the z-statistics for both teams to be 0, the mean value of the Wilcox's rank sum statistics for both teams to be 105.
- How would you obtain a p-value from each of the distribution? Add vertical line(s) to the histogram to illustrate the p-value calculation. I would obtain the p-value based on the definition. I would record the number of the test statistic data as the race data I observed or more extreme(in the direction of the alternitive) under the null hypothesis and diveded by the iterations.
- For each of the quantities, test the null hypothesis using the p-values you just calculated. To calculate the p-value for a two-sided test, assume that the sampling distribution is symmetric.

```
(p_value.Sample_Mean_Diff <-2 * sum(Sample_Mean_Diff >= Sample_Mean_Diff.sampling)/Itr)
```

```
## [1] 0.7006667
```

```
(p_value.t <-2 * sum(t >= t.sampling)/Itr)
```

```
## [1] 0.7006667
```

```
(p_value.U.hare <-2 * sum(U.hare <= U.hare.sampling)/Itr)
```

```
## [1] 0.02
```

```
(p_value.U.tortoise <-2 * sum(U.tortoise >= U.tortoise.sampling)/Itr)
```

```
## [1] 0.02
```

```
(p_value.z.hare <-2 * sum(z.hare <= z.hare.sampling)/Itr)
```

```
## [1] 0.02
```

```
(p_value.z.tortoise <-2 * sum(z.tortoise >= z.tortoise.sampling)/Itr)
```

```
## [1] 0.02
```

```
(p_value.W.hare <-2 * sum(W.hare >= W.hare.sampling)/Itr)
```

```
## [1] 0.02
```

```
(p_value.W.tortoise <-2 * sum(W.tortoise <= W.tortoise.sampling)/Itr)
```

```
## [1] 0.02
```

Answer:
At the $\alpha = 0.05$ significance level, the p-value for mean difference and t-statistics are larger than 0.05, which failed to reject the null hypothesis.
The p-value for U-statistic, z-statistic, Wilcox's rank sum statistics for both teams are smaller than 0.05, which reject the null hypothesis.

4. Summarize your findings from the first three questions by comparing your results across different tests. In which situations would you prefer one of these tests over another? Broadly comment on the pro's and con's of each of these approaches. What is the final conclusion to your hypothesis test?
   Answer:
   Set the significance level as $\alpha = 0.05$, by using the t-test in the first question, the p-value we obtained is larger than 0.05 so we can't reject the null hypothesis. In the second question, the p-value for z-statistic and Wilcox rank sum test are all smaller than 0.05 so we can reject the null hypothesis. The p-value of difference of mean and t-statistic are larger than 0.05 while the others are smaller than 0.05.

When the sample size is large and the data may come from a normal distribution and have independency, I'd prefer the t-test because it's simple to use and easy to interpret. The conclusion based on the t-test can also applied to the entire population. However, when the sample size is small, the central limit theorem will not hold and the t-test result wouldn't be accurate.
So if we are not comfortable choosing a parametric family and the sample size is small, the non-parametric methods like Wilcox's rank sum in question 2 has the benifits. It also holds the fewest assumptions which means there are less chance to get the wrong assumption. But it may not be straightforward. Tied values in the data can be problematic.
The permutation method in question three can reduce the effect of sample sizes but it might cost more time and recources.
The final conclusion to your hypothesis test is that we reject the null hypothesis and the two teams are not the same.