Assignment_1

Ri Liu (rl4508)

2/20/2022

Tortoise and Hare Racing 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:

Null hypothesis H_0: The true mean finishing time is the same between team tortoise and team hare. ($\mu_{tortoise} = \mu_{hare}$)

Alternative hypothesis H_a: The true mean finishing time is different between team tortoise and team hare. ($\mu_{tortoise} \neq \mu_{hare}$)

We use a two-sided test as we don't want to presuppose which animal must be faster. Comparing with an upper one-sided alternative having the probability of rejecting H_0 when H_0 is true on the upper side, the size will be doubled if we want the same probability of rejecting H_0 when H_0 is true on both the lower and the upper side.

When the mean difference is mild, the power of a one-sided alternative will be a little bit higher given the direction of the alternative is decided correctly. However, when the mean difference is extreme to some extent, the power between two-sided and one-sided will approximate the same.

(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, $\hat{X}_{hare} - \hat{X}_{tortoise}$, by calculating this quantity in R.

```
race = read.csv('race.csv')
```

```
# the difference in sample mean in finishing times for the two teams:
mean_diff_observe <- mean(race$Hare) - mean(race$Tortoise)
paste('The difference in sample mean in finishing times for the two teams: ', mean_diff_observe)
```

```
## [1] "The difference in sample mean in finishing times for the two teams:  -5.0456417317"
```

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 mean difference can be estimated as follows,

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

by relating $Var(\bar{X}_1 - \bar{X}_2)$ to the quantities $Var(\bar{X}_1)$ and $Var(\bar{X}_2)$. 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:

Let σ^2 be the variance of the finishing time distributions for the two teams.

$$\begin{aligned}
\sum_{i=1}^{N_1}(X_{1i} - \bar{X}_1)^2 &= \sum_{i=1}^{N_1}[(X_{1i} - \mu_1) - (\bar{X}_1 - \mu_1)]^2 \\
&= \sum_{i=1}^{N_1}(X_{1i} - \mu_1)^2 - 2(\bar{X}_1 - \mu_1)\sum_{i=1}^{N_1}(X_{1i} - \mu_1) + N_1(\bar{X}_1 - \mu_1)^2 \\
&= \sum_{i=1}^{N_1}(X_{1i} - \mu_1)^2 - 2N_1(\bar{X}_1 - \mu_1)^2 + N_1(\bar{X}_1 - \mu_1)^2 \\
&= \sum_{i=1}^{N_1}(X_{1i} - \mu_1)^2 - N_1(\bar{X}_1 - \mu_1)^2
\end{aligned}$$

Since

$$\mu_1 = E(X_{1i}) = E(\bar{X}_1),$$

$$\begin{aligned}
E(\sum_{i=1}^{N_1}(X_{1i} - \bar{X}_1)^2) &= E(\sum_{i=1}^{N_1}(X_{1i} - \mu_1)^2 - N_1(\bar{X}_1 - \mu_1)^2) \\
&= \sum_{i=1}^{N_1} E(X_{1i} - \mu_1)^2 - N_1 E(\bar{X}_1 - \mu_1)^2 \\
&= \sum_{i=1}^{N_1} E(X_{1i} - E(X_{1i}))^2 - N_1 E(\bar{X}_1 - E(\bar{X}_1))^2 \\
&= \sum_{i=1}^{N_1} Var(X_{1i}) - N_1 Var(\bar{X}_1) \\
&= N_1\sigma^2 - N1\frac{\sigma^2}{N_1} \\
&= (N_1 - 1)\sigma^2
\end{aligned}$$

Similarly,

$$E(\sum_{i=1}^{N_1}(X_{1i} - \bar{X}_1)^2) = (N_2 - 1)\sigma^2$$

Let

$$s_p^2 = \frac{S_{hare}^2(N_1 - 1) + S_{tortoise}^2(N_2 - 1)}{N_1 + N_2 - 2}$$

Then,

$$\begin{aligned} E(s_p^2) &= E(\frac{\sum_{i=1}^{N_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{N_2}(X_{2i} - \bar{X}_2)^2}{N_1 + N_2 - 2}) \\ &= \frac{(N_1 + N_2 - 2)\sigma^2}{N_1 + N_2 - 2} \\ &= \sigma^2 \end{aligned}$$

So, s_p^2 is an unbiased estimation of the pooled variance.

And the variance of the sampling distribution, i.e. the squared standard error, is

$$\begin{aligned} E(Var(\bar{X}_1 - \bar{X}_2)) &= E[(\frac{S_{hare}^2(N_1 - 1) + S_{tortoise}^2(N_2 - 1)}{N_1 + N_2 - 2})(\frac{1}{N_1} + \frac{1}{N_2})] \\ &= (\frac{1}{N_1} + \frac{1}{N_2})E(\frac{S_{hare}^2(N_1 - 1) + S_{tortoise}^2(N_2 - 1)}{N_1 + N_2 - 2}) \\ &= (\frac{1}{N_1} + \frac{1}{N_2})\sigma^2 \\ &= Var(\bar{X}_1) + Var(\bar{X}_2) \end{aligned}$$

(d) Calculate the above quantity

```
n1 <- length(race$Hare) # N_1
n2 <- length(race$Tortoise) # N_2
N <- 1/n1 + 1/n2 # 1/N_1 + 1/N_2
df <- n1 + n2 - 2 # degree of freedom
# pooled variance
sp2 <- (sum((race$Hare - mean(race$Hare))^2) + sum((race$Tortoise - mean(race$Tortoise))^2))  / df
```

```
Var <- sp2 * N
paste('The variance of the difference between hares\' mean and tortoises\' mean is ', Var)
```

```
## [1] "The variance of the difference between hares' mean and tortoises' mean is  82.6225749518478"
```

(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

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

```
t_stat_observe <- mean_diff_observe / sqrt(Var)
paste('The test statistic for this dataset is ', t_stat_observe)
```

```
## [1] "The test statistic for this dataset is  -0.555094657013082"
```

```
pvalue <- 2* (pt(abs(t_stat_observe), df = df, lower.tail = F))
paste('The p-value for this dataset is ', pvalue)
```

```
## [1] "The p-value for this dataset is  0.585662843053947"
```

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

```
# Rejection region:
qt(c(0.025, 0.975), n1+n2-2)
```

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

Answer: The p-value is 0.586. The rejection region is below -2.10 or above 2.10. At the 5% significance level, there is no sufficient evidence to reject the null hypothesis that the true mean finishing time is the same between team tortoise and team hare.

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.

```
time <- c(race$Hare, race$Tortoise)
animal <- factor(c(rep(c(0,1), c(10,10))))
race_new <- data.frame(time, animal)
```

Levene's test of equal variance:

Null hypothesis H_0: The variances across the two teams are the same.

```
library(car)
```

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

```
leveneTest(race_new$time ~ race_new$animal, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  1  1.7466 0.2029
##       18
```

The p-value is 0.203. At the 5% confidence level, there is no sufficient evidence to reject the null hypothesis that the variances across the two teams are the same.

F-test for homogeneity of variance:

```
var.test(race$Hare, race$Tortoise)
```

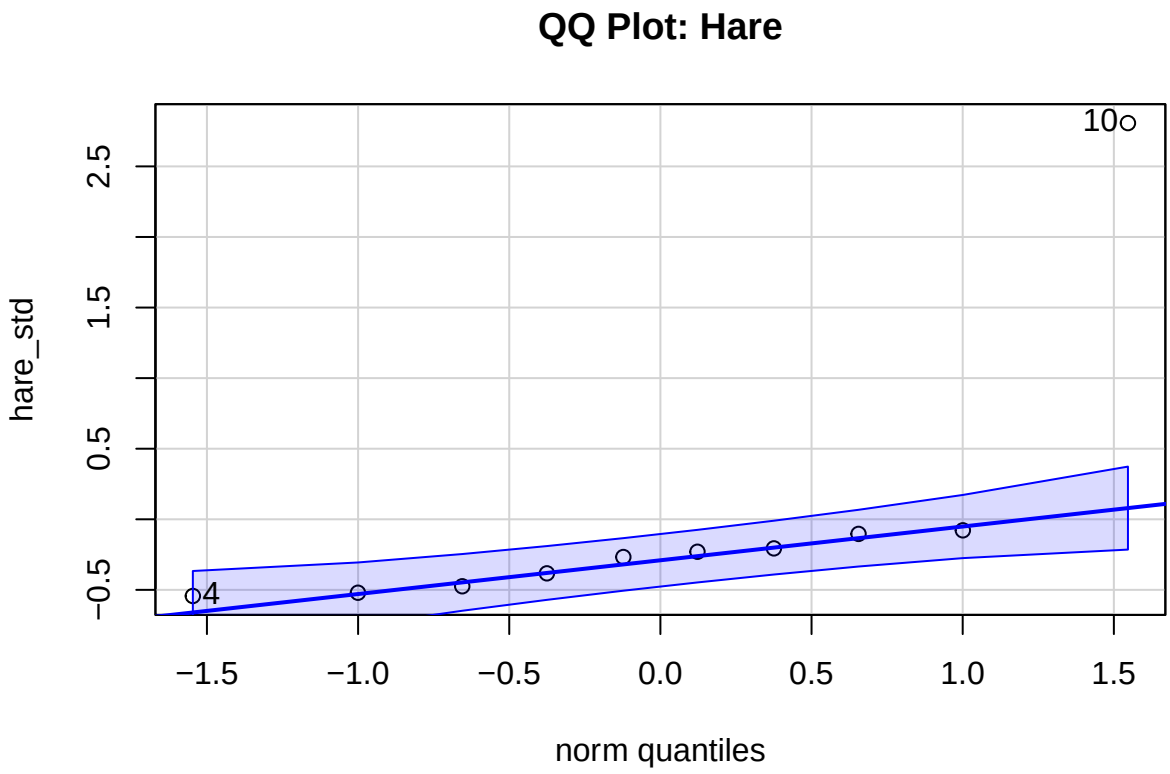
```
##
##  F test to compare two variances
##
## data:  race$Hare and race$Tortoise
## F = 9.039, num df = 9, denom df = 9, p-value = 0.00306
```

```
## 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
```

The result is different. According to the F-test, the variances across the two teams are not the same.

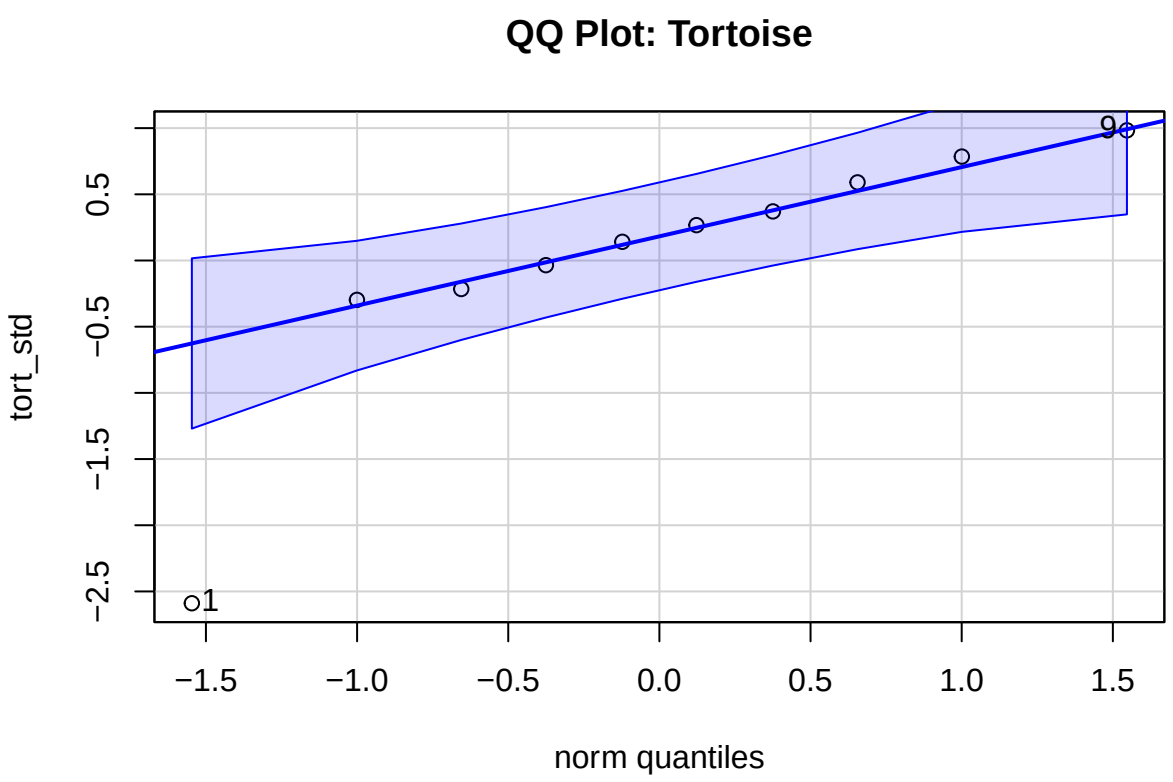
Test of normality:

```
# standardize the data
hare_std <- (race$Hare - mean(race$Hare)) / sd(race$Hare)
qqPlot(hare_std, main="QQ Plot: Hare")
```




```
## [1] 10  4
```

```
# standardize the data
tort_std <- (race$Tortoise - mean(race$Tortoise)) / sd(race$Tortoise)
qqPlot(tort_std, main="QQ Plot: Tortoise")
```




```
## [1] 1 9
```

From the Q-Q plots, I can tell that most of the data from hare and tortoise are normal. But, there is one outlier in hare and tortoise data respectively.

Meanwhile, the sample size is too small to resist the outliers. So, the t-test I used is not kind of appropriate, as the normality assumption is interfered by the outliers.

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(Xhare < Xtortoise) = P(Xtortoise < Xhare) should hold. Therefore, it is of interest to test the hypotheses:

(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 statistic for hare
U_hare_observe <- 0
for (i in 1:n1) {
  for (j in 1:n2){
    if(race$Hare[i] < race$Tortoise[j]){
      U_hare_observe = U_hare_observe + 1
```

```
    }
  }
}
paste('The U-statistic for hares: ', U_hare_observe)
```

```
## [1] "The U-statistic for hares:  81"
```

```
# U statistic for tortoise
U_tortoise_observe <- 0
for (i in 1:n1) {
  for (j in 1:n2){
    if(race$Tortoise[i] < race$Hare[j]){
      U_tortoise_observe = U_tortoise_observe + 1
    }
  }
}
paste('The U-statistic for tortoises: ', U_tortoise_observe)
```

```
## [1] "The U-statistic for tortoises:  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.

Under H_0, let

$$P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare}) = p$$

If there's no tie,

$$p + p = 1$$

i.e. $p = 0.5$

$$E_{U_{hare}} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(X_{hare,i} < X_{tortoise,j})P(X_{hare} < X_{tortoise}) = pN_1N_2$$

$$E_{U_{tortoise}} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(X_{tortoise,j} < X_{hare,i})P(X_{tortoise} < X_{hare}) = pN_1N_2$$

Since $N_1 = N_2 = 10$,

$$E_{U_{hare}} = E_{U_{tortoise}} = 50$$

(c) Under the null hypothesis, when the sample size is large enough, the U-statistic is approximately normally distributed. The mean for this distribution, mu(U0) , was calculated in the previous part (2.b). The standard deviation for this normal distribution is sigma(U0) = sqrt(n1n2(n1 + n2 + 1)/12). Therefore, we may use the following test statistic:

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

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

```
sigma_u0 = sqrt(n1*n2*(n1+n2+1) / 12)
z_stat_observe_hare = (U_hare_observe - 50) / sigma_u0
z_stat_observe_hare
```

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

```
sigma_u0 = sqrt(n1*n2*(n1+n2+1) / 12)
z_stat_observe_tortoise = (U_tortoise_observe - 50) / sigma_u0
z_stat_observe_tortoise
```

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

```
pvalue = pnorm(abs(z_stat_observe_tortoise), lower.tail = F)*2
paste('The p-value for the Z-test is ', pvalue)
```

```
## [1] "The p-value for the Z-test is  0.0191099222068444"
```

At the 0.05 significance level, the p-value of this test is lower than 0.05, so I can reject the null hypothesis that the probability of hare having shorter finishing time than tortoise is the same as the probability of tortoise having shorter finishing time than hare. The finishing time between hare and tortoise is significantly different.

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.

```
ranks <- rank(c(race$Hare, race$Tortoise))
w_hare_observe <- sum(ranks[1:n1])
w_tortoise_observe <- sum(ranks[n1+1:n2])
```

```
wilcox.test(race$Hare, race$Tortoise, exact = F, correct = F, alternative = 'two.sided')
```

```
##
##  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
```

The p-value of the Wilcox rank sum test is 0.01911, which is smaller than the 0.05 significant level. So, the null hypothesis can be rejected. Though Wilcox rank sum test is different from Wilcox signed rank test, the p-values from the two tests are very close.

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). The table below illustrates several permuted datasets:

(a)Generate 3000 permuted datasets as described above.

```
# mix the value of hare and tortoise into one column
all_value <- c(race$Hare, race$Tortoise)
group <- factor(c(rep('hare', 10), rep('tortoise', 10)))
```

```
permutation <- list()
# 3000 datasets
for (i in 1:3000){
  temp_vector <- sample(all_value, length(all_value))
  temp_df <- data.frame(Hare = temp_vector[1:10], Tortoise = temp_vector[11:20])
  permutation <- append(permutation, list(temp_df))
}
```

Through 3000 times sampling, I get a list containing 3000 data frames, in each of which there is one column for hare and one column for tortoise. We can have a look at the 1500th permuted dataset:

```
permutation[[1500]]
```

```
##         Hare  Tortoise
## 1   16.14952 10.497913
## 2   17.17253 25.792053
## 3   33.84349 31.853845
## 4    5.00000  8.624024
## 5   21.30704 26.534326
## 6  100.00000 37.408357
## 7   29.77001 30.912757
## 8   17.81809 35.610745
## 9   12.99890 28.169822
## 10  20.61959  9.251388
```

(b) For each permuted dataset, calculate:

- $\bar{X}_{hare} - \bar{X}_{tortoise}$

```
mean_diff <- c()
len_permut <- length(permutation)
for (i in 1:len_permut){
  temp <- mean(permutation[[i]]$Hare) -
    mean(permutation[[i]]$Tortoise)
  mean_diff <- append(mean_diff, temp)
}
head(mean_diff)
```

```
## [1]  12.738733 -12.428670   3.469984   8.139312   5.540638 -15.057805
```

- The t statistic as in equation (1)

```
t_stat <- c()
# traverse each permutated data set
for (i in 1:length(permutation)){
  df_current = permutation[[i]]
  # mean_diff <- mean(df_current$Hare) - mean(df_current$Tortoise)
  n1 <- length(df_current$Hare)
  n2 <- length(df_current$Tortoise)
  d_free <- n1 + n2 -2
  sp2 <- (sum((df_current$Hare - mean(df_current$Hare))^2) + sum((df_current$Tortoise - mean(df_current$
  Var <- sp2 * (1/n1 + 1/n2)
  t_stat_temp <- mean_diff[i] / sqrt(Var)
  t_stat <- append(t_stat, t_stat_temp)
}
t_stat[1:5]
```

```
## [1]  1.4707300 -1.4308041  0.3800387  0.9079821  0.6106282
```

- U_{hare} and $U_{tortoise}$

```
U_hare <- c()
U_tortoise <- c()
for (n in 1:length(permutation)){
  df_current <- permutation[[n]]
  U_hare_temp <- 0
  n1 <- length(df_current$Hare)
  n2 <- length(df_current$Tortoise)
  for (i in 1:n1){
    for (j in 1:n2){
      if(df_current$Hare[i] < df_current$Tortoise[j]){
        U_hare_temp = U_hare_temp + 1
      }
    }
  }
  U_hare <- append(U_hare, U_hare_temp)

  U_tortoise_temp <- 0
  for (i in 1:n1) {
    for (j in 1:n2){
      if(df_current$Tortoise[i] < df_current$Hare[j]){
        U_tortoise_temp = U_tortoise_temp + 1
      }
   }
  }
  U_tortoise <- append(U_tortoise, U_tortoise_temp)
}
# list(U_hare, U_tortoise)
```

- The Z statistics as in equation (2)

```
z_stat_hare <- c()
for (i in 1:n1){
  sigma_u0 = sqrt(n1*n2*(n1+n2+1) / 12)
  z_stat_temp = (U_hare - 50) / sigma_u0
  z_stat_hare <- append(z_stat_hare, z_stat_temp)
}
```

```
z_stat_tortoise <- c()
for (i in 1:n1){
  sigma_u0 = sqrt(n1*n2*(n1+n2+1) / 12)
  z_stat_temp = (U_tortoise - 50) / sigma_u0
  z_stat_tortoise <- append(z_stat_tortoise, z_stat_temp)
}
```

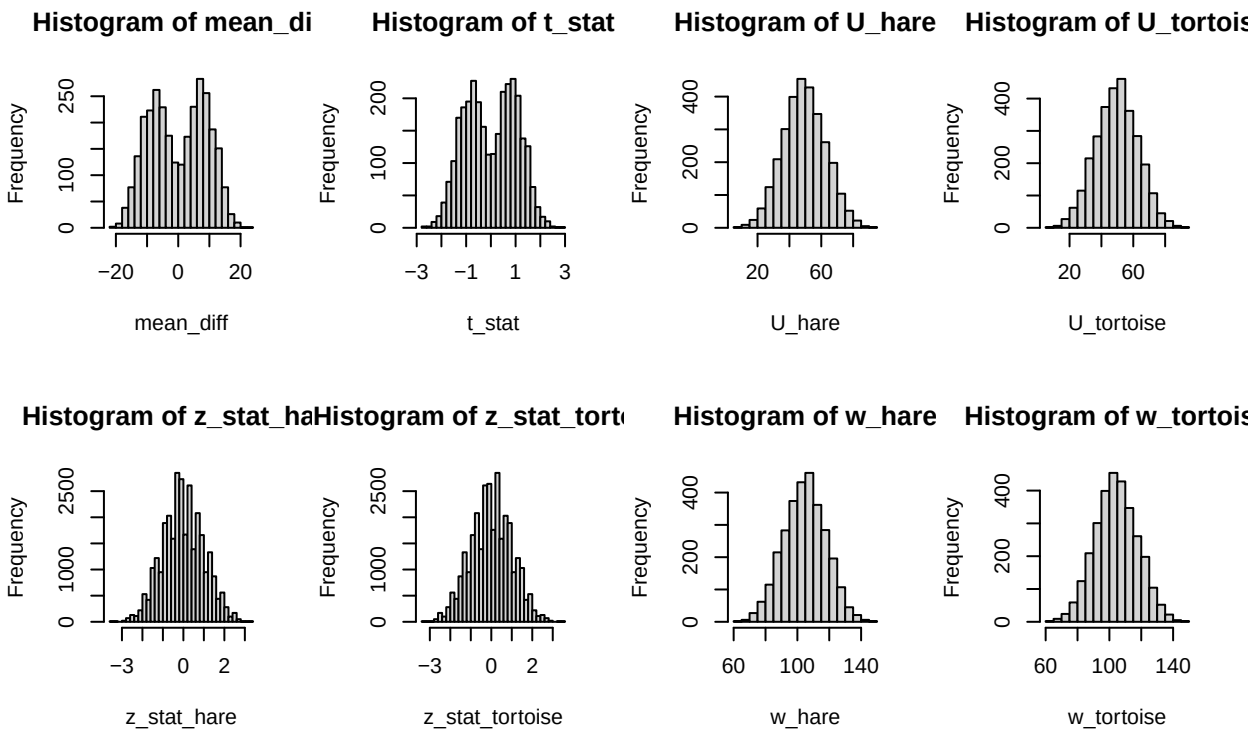
- Wilcox's rank sum statistics for team Hare and for team Tortoise

```
w_hare <- c()
w_tortoise <- c()
for (i in seq_along(permutation)){
  mix_values = c(permutation[[i]]$Hare, permutation[[i]]$Tortoise)
  ranks = rank(mix_values)
  w_hare_temp = sum(ranks[1:n1])
  w_tortoise_temp = sum(ranks[n1+1:n2])
  w_hare = append(w_hare, w_hare_temp)
  w_tortoise = append(w_tortoise, w_tortoise_temp)
}
```

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

```
par(mfrow=c(2,4), oma=c(0,0,3,0))
hist(mean_diff, breaks = 30)
hist(t_stat, breaks = 30)
hist(U_hare, breaks = 30)
hist(U_tortoise, breaks = 30)
hist(z_stat_hare, breaks = 30)
hist(z_stat_tortoise, breaks = 30)
hist(w_hare, breaks = 30)
hist(w_tortoise, breaks = 30)
# Title
mtext('Figure: Histograms of Different Statistics', side = 3, line = 0, outer = T)
```

Figure: Histograms of Different Statistics



- 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:

I used permutation to simulate a sampling distribution of different test statistics, under the null hypothesis that the true finishing time is the same between team tortoise and team hare. Given the null hypothesis, it should not cause much difference when assigning several values from one group to another group randomly.

The sampling distributions of these test statistics are almost symmetric. However, the distribution of mean difference and its t-statistic are bimodal, other test statistics' distributions are unimodal.

- What do you expect the mean value of each of the sampling distributions to be?

Answer:

Under the null hypothesis that the true finishing time is the same between team tortoise and team hare,

the expected mean value of $\bar{X}_{hare} - \bar{X}_{tortoise}$ will be 0,

the expected mean value of the t-statistic of $\bar{X}_{hare} - \bar{X}_{tortoise}$ will be 0,

the expected mean value of U_{hare} and $U_{tortoise}$ will both be 50,

the expected mean value of the z statistic of U_{hare} and $U_{tortoise}$ will be 0,

the expected mean value of Wilcox's rank sum statistics for team Hare and for team Tortoise will both be 105.

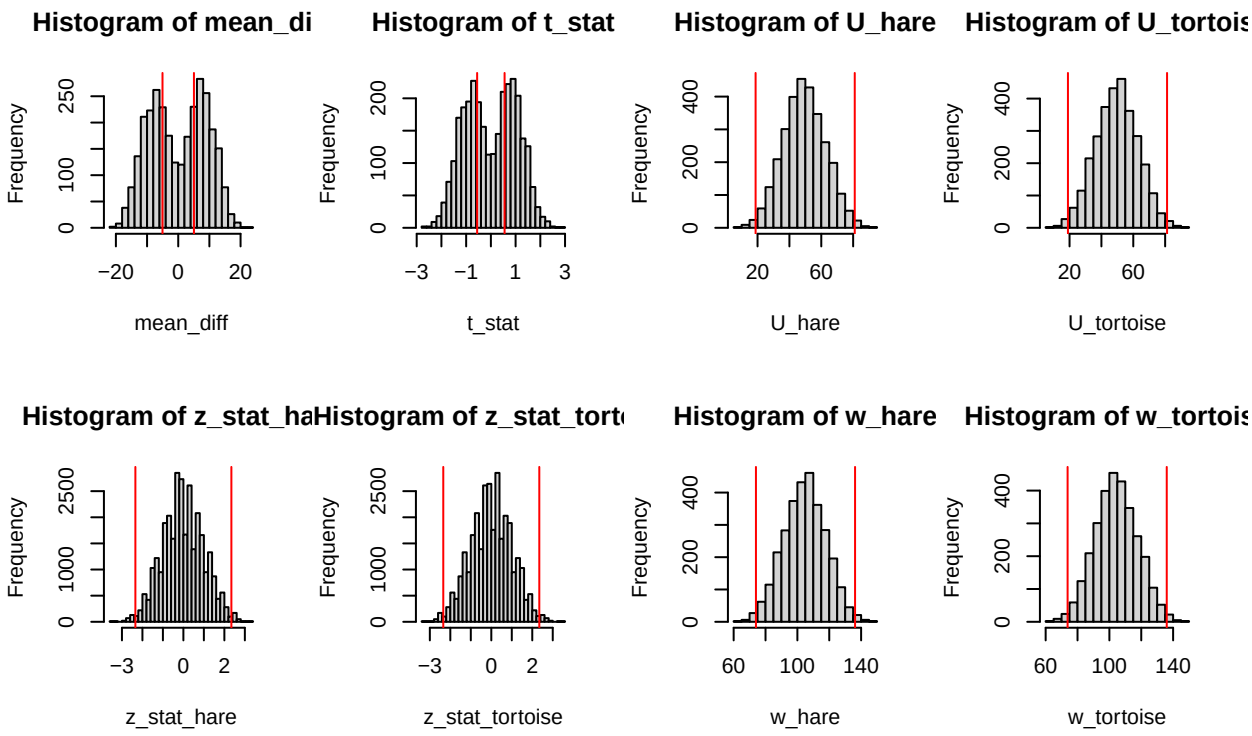
- How would you obtain a p-value from each of the distribution? Add vertical line(s) to the histogram to illustrate the p-value calculation.

Answer:

The p-value represents the probability of obtaining results at least as extreme as the observed result. In a permutation test, a p-value can be estimated as the proportion of the results that are as extreme as or more extreme than the observed result.

```
par(mfrow=c(2,4), oma=c(0,0,3,0))
# mean difference
hist(mean_diff, breaks = 30)
abline(v = mean_diff_observe, col = 'red')
abline(v = -mean_diff_observe, col = 'red')
# t-statistic of mean difference
hist(t_stat, breaks = 30)
abline(v = t_stat_observe, col = 'red')
abline(v = -t_stat_observe, col = 'red')
# U-statistic of hares
hist(U_hare, breaks = 30)
abline(v = U_hare_observe, col = 'red')
abline(v = 2*mean(U_hare) - U_hare_observe, col = 'red')
# U-statistic of tortoises
hist(U_tortoise, breaks = 30)
abline(v = U_tortoise_observe, col = 'red')
abline(v = 2*mean(U_tortoise) - U_tortoise_observe, col = 'red')
# z-statistic of U-statistic for team Hare
hist(z_stat_hare, breaks = 30)
abline(v = z_stat_observe_hare, col = 'red')
abline(v = -z_stat_observe_hare, col = 'red')
# z-statistic of U-statistic for team Tortoise
hist(z_stat_tortoise, breaks = 30)
abline(v = z_stat_observe_tortoise, col = 'red')
abline(v = -z_stat_observe_tortoise, col = 'red')
# Wilcox's rank sum statistics for team Hare
hist(w_hare, breaks = 30)
abline(v = w_hare_observe, col = 'red')
abline(v = 2*mean(w_hare) - w_hare_observe, col = 'red')
# Wilcox's rank sum statistics for team Tortoise
hist(w_tortoise, breaks = 30)
abline(v = w_tortoise_observe, col = 'red')
abline(v = 2*mean(w_tortoise) - w_tortoise_observe, col = 'red')
# Title
mtext('Figure: Histograms of Different Statistics with P-value Illustration', side = 3, line = 0, outer
```

Figure: Histograms of Different Statistics with P−value Illustration



- 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.

- mean difference

```
p_value_1 <- sum(mean_diff > abs(mean_diff_observe))*2 / length(mean_diff)
# if the sampling distribution is not symmetric:
# p_value_1 <- (sum(mean_diff < -abs(mean_diff_observe)) + sum(mean_diff > abs(mean_diff_observe))) / l
p_value_1
```

```
## [1] 0.7313333
```

At the 0.05 significance level, the p-value of $\bar{X}_{hare} - \bar{X}_{tortoise}$ is **too large** for us to reject the null hypothesis that the true mean finishing time of team tortoise and team hare have no difference.

- t-statistic

```
p_value_2 <- sum(t_stat > abs(t_stat_observe))*2 / length(t_stat)
# if the sampling distribution is not symmetric:
# p_value_2 <- (sum(t_stat < -abs(t_stat_observe)) + sum(t_stat > abs(t_stat_observe))) / length(t_stat
p_value_2
```

```
## [1] 0.7313333
```

At the 0.05 significance level, the p-value of the t-statistic for $\bar{X}_{hare} - \bar{X}_{tortoise}$ is **too large** for us to reject the null hypothesis that the true mean finishing time is the same between team tortoise and team hare.

- U-statistic

```
p_value_3 <- ifelse(U_hare_observe > mean(U_hare), sum(U_hare > U_hare_observe)*2 / length(U_hare), sum
p_value_3
```

```
## [1] 0.01666667
```

```
p_value_4 <- ifelse(U_tortoise_observe > mean(U_tortoise), sum(U_tortoise > U_tortoise_observe)*2 / leng
p_value_4
```

```
## [1] 0.01666667
```

At the 0.05 significance level, the p-value of the U-statistic of team Hare (and team tortoise) is **significant** for us to reject the null hypothesis that the number of hares passing the number of tortoises is the same as the number of tortoise passing the number of hares.

- Z-statistic

```
p_value_5 <- sum(z_stat_hare > abs(z_stat_observe_hare))*2 / length(z_stat_hare)
p_value_5
```

```
## [1] 0.01666667
```

At the 0.05 significance level, the p-value of the z-statistic for U-statistic of team Hare is **significant** for us to reject the null hypothesis that the number of hares passing the number of tortoises is the same as the number of tortoise passing the number of hares.

- Wilcox's rank sum statistics

```
p_value_6 <- ifelse(w_hare_observe > mean(w_hare), sum(w_hare > w_hare_observe)*2 / length(w_hare), sum
p_value_6
```

```
## [1] 0.01666667
```

```
p_value_7 <- ifelse(w_tortoise_observe > mean(w_tortoise), sum(w_tortoise > w_tortoise_observe)*2 / leng
p_value_7
```

```
## [1] 0.01666667
```

At the 0.05 significance level, the p-value of U-statistic of team Hare and the p-value of U-statistic of team Tortoises are **significant** for us to reject the null hypothesis that the the finishing time of hares and tortoise are the same.

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:

In the exploratory test of 1.(e).iii, it shows that the data is not rigorously normal, which was caused by one outlier in each column. Combined with the sample size being small, the outliers affect their columns' distribution seriously, in which the outlier in `hare` makes $\bar{X}_{hare}$ smaller, and the outlier in `tortoise` makes $\bar{X}_{tortoise}$ larger. Hence, the t-test in question 1 has no significant p-value at all.

In question 2, I used non-parametric tests. No matter the Wilcox's rank sum test or the Wilcox's signed rank test, the outliers would only affect the ranks that are related to themselves, while the rest majority of data points could get ranks as we expected. Even though the outliers are so far from the majority, as long as the proportion of outliers is not high, there will be less impacts on the statistics. So in question 2, both U-statistic W-statistic are significant at 0.05 significance level.

In question 3, I used permutation to reduce the effect due to the small sample. Through permutation, I get an approximation of the sampling distribution, under the hypothesis that the two groups of data has no difference. After calculated the statistics on the permutated data, the parametric statistics, i.e. t-statistic, become less significant, while the non-parametric statistics, i.e. U-statistic and W-statistic, become more significant.

Based on this dataset, which is small and not normal, permutation is a good choice to reduce the effect of the problems. But it is also subject to the outliers when used for parametric statistics. And, of course, it has a huge computational cost. For our data, 3000 random samples are even far from the full permutation.

Wilcox's rank sum test and Wilcox's signed sum test are good alternatives for t-test when the normal assumption is not meet. Especially when we are just want to dig out whether two groups are different or the treatment has effect compared with control. However, it does not reflect the magnitude of mean difference and the variance. For example, 1 vs 100 and 49 vs 50 can have same signed rank.

As long as the assumptions are satisfied, including normality, equal variance, homogeneity of variance, etc., t-test also has its benefits. It is sensitive to variance within a rational range, so it is more powerful to reject the null hypothesis when the null hypothesis is false. Meanwhile, it is simple to implement.