# Hypothesis Testing Project

Ramya Dhatri Vunikili

21 March 2017

## Hypothesis Testing

### Answer 1a

$H_0 : \mu_{Hare} = \mu_{Tortoise}$
$H_a : \mu_{Hare} \neq \mu_{Tortoise}$
Since we are interested in testing whether the true mean
finishing time is the same for team tortoise and team hare, I've chosen to perform a two-tail test. Unlike a one-tail test, when testing at a significance level of 0.05, a two-tailed test allots 0.025 of the alpha to testing the statistical significance in one direction and other 0.025 of the alpha to testing statistical significance in the other direction. Hence, by using a two-tailed test, we are impartial about testing the possibility of relationship in both directions (i.e we're not biased about true mean finishing times of team hare and team tortoise).

However, if we were to think in general terms, hares are faster than tortoises and hence the true mean finishing time of team hares is always less than or equal to that of the team tortoises'. In such a case, performing one-tail test would be beneficial. As I'm not being biased about the racing abilities of tortoise or hare I would choose to perform a two-tail test over a one-tail test.

### Answer 1b

$\bar{X}_{Hare}$ = smeanHare = 23.4439
$\bar{X}_{Tortoise}$ = smeanTortoise = 28.48954
$\bar{X}_{Hare} - \bar{X}_{Tortoise}$ = smeanDifference = -5.045642

```
                    ## Tortoise and Hare Racing Problem ##

## 1 b) Differnce in sample mean calculation
givenData <- read.csv("race.csv")
smeanHare <- mean(givenData$Hare)
smeanTortoise <- mean(givenData$Tortoise)
smeanDifference <- smeanHare - smeanTortoise
```

Listing 1: R code for calculating $\bar{X}_{Hare} - \bar{X}_{Tortoise}$

```
> smeanHare
[1] 23.4439
> smeanTortoise
[1] 28.48954
> smeanDifference
[1] -5.045642
```

Listing 2: Output

### Answer 1c

Derivation of the formula for pooled sample variance:

$Var(a\bar{X}_1 + b\bar{X}_2)$
$= E([a\bar{X}_1 + b\bar{X}_2 - E(a\bar{X}_1 + b\bar{X}_2)]^2)$
$= E([a\bar{X}_1 + b\bar{X}_2 - aE(\bar{X}_1) - bE(\bar{X}_2)]^2)$
$= E([a(\bar{X}_1 - \mu_{\bar{X}_1}) + b(\bar{X}_2 - \mu_{\bar{X}_2})]^2)$
$= E(a^2(\bar{X}_1 - \mu_{\bar{X}_1})^2 + b^2(\bar{X}_2 - \mu_{\bar{X}_2})^2)$
$= a^2E((\bar{X}_1 - \mu_{\bar{X}_1})^2) + b^2E((\bar{X}_2 - \mu_{\bar{X}_2})^2) + 2abE((\bar{X}_1 - \mu_{\bar{X}_1})(\bar{X}_2 - \mu_{\bar{X}_2})))$

Here a = 1 and b = -1 and since the finishing times of all racers are independent of each other, the co-variance term evaluates to 0, i.e., $2abE((\bar{X}_1 - \mu_{\bar{X}_1})(\bar{X}_2 - \mu_{\bar{X}_2}))) = 0$.

$Var(\bar{X}_1 - \bar{X}_2) = E((\bar{X}_1 - \mu_{\bar{X}_1})^2) + E((\bar{X}_2 - \mu_{\bar{X}_2})^2)$

[where $E((\bar{X} - \mu_{\bar{X}})^2) = Var(\bar{X})$]

Also, it is given that the variance of the finishing time distributions for the two teams are equal, i.e., $\sigma^2_{Hare} = \sigma^2_{Tortoise} = \sigma^2$
Now, since the sample sizes of the two teams are equal ($N_1 = N_2 = 10$), both sample variances act as equally good estimates of the respective population variance i.e.,
$S^2_{Hare} \approx \sigma^2$ and $S^2_{Tortoise} \approx \sigma^2$
Therefore the best estimate that we can obtain for $\sigma^2$ is an average of the two sample variances but weighted to take into account the possibility that they may be of different sizes and therefore provide different amounts of information. This weighted average is termed the

pooled sample variance, $Var(\bar{X}_1 - \bar{X}_2)$. The weighted variances are now obtained by multiplying each sample variance by its respective sample degrees of freedom. The pooled sample variance is then calculated by summing the two weighted sample variances and dividing the sum of the two sample degrees of freedom:

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

Thus, pooled sample variance is given by the formula,

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

### Answer 1d

$Var(\bar{X}_1)$ = varianceHare = 743.9244
$Var(\bar{X}_2)$ = varianceTortoise = 82.3013
$Var(\bar{X}_1 - \bar{X}_2)$ = pooledVariance = 82.62257

```r
## Tortoise and Hare Racing Problem ##

## 1 b) Differnce in sample mean calculation
givenData <- read.csv("race.csv")
smeanHare <- mean(givenData$Hare)
smeanTortoise <- mean(givenData$Tortoise)
smeanDifference <- smeanHare - smeanTortoise

## 1 d) Pooled Variance
varianceHare <- var(givenData$Hare)
varianceTortoise <- var(givenData$Tortoise)
n1 <- 10
n2 <- 10
pooledVariance <- ((varianceHare*(n1 - 1) + varianceTortoise*(n2 - 1))/(n1 + n2 - 2))*
    ((1/n1)+ (1/n2))
```

Listing 3: R code for calculating $Var(\bar{X}_1 - \bar{X}_2)$

```
> varianceHare
[1] 743.9244
> varianceTortoise
[1] 82.3013
> pooledVariance
[1] 82.62257
```

Listing 4: Output

### Answer 1e (i)

Test statistic is given by:
t = $\frac{\bar{X}_{Hare}-\bar{X}_{Tortoise}}{std.err(\bar{X}_{Hare}-\bar{X}_{Tortoise})}$
We know that

$$std.err(\bar{X}_{Hare}-\bar{X}_{Tortoise}) = \sqrt{Var(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{S^2_{Hare}(N_1-1)+S^2_{Tortoise}(N_2-1)}{(N_1-1)+(N_2-1)}}$$

t = $\frac{\bar{X}_{Hare}-\bar{X}_{Tortoise}}{\sqrt{\frac{S^2_{Hare}(N_1-1)+S^2_{Tortoise}(N_2-1)}{(N_1-1)+(N_2-1)}}}$

t = -0.5550947
p value = 0.5856628

```
 ## Tortoise and Hare Racing Problem ##

## 1 b) Differnce in sample mean calculation
givenData <- read.csv("race.csv")
smeanHare <- mean(givenData$Hare)
smeanTortoise <- mean(givenData$Tortoise)
smeanDifference <- smeanHare - smeanTortoise

## 1 d) Pooled Variance
varianceHare <- var(givenData$Hare)
varianceTortoise <- var(givenData$Tortoise)
n1 <- 10
n2 <- 10
pooledVariance <- ((varianceHare*(n1 - 1) + varianceTortoise*(n2 - 1))/(n1 + n2 - 2))*
    ((1/n1)+ (1/n2))

## 1 e) (i) Test Statistic and P Value
stdError <- sqrt(pooledVariance)
t <- (smeanHare - smeanTortoise)/(stdError)
pval <- 2*pt(t, (n1 + n2 - 2), lower.tail=TRUE)
```

Listing 5: R code for calculating test statistic and p value

```
> t
[1] -0.5550947
> pval
[1] 0.5856628
```

Listing 6: Output

### Answer 1e (ii)

At 18 degrees of freedom, the critical value at $t_{\frac{\alpha}{2}} = t_{0.025}$ is found to be 2.101. Thus the rejection region for two-tail test is given by:
t* < -2.101 and t* > 2.101

Conclusion
With a test statistic of -0.5550947, p value of 0.5856628 and critical value of Âś 2.101 at a 5% level of significance, we do not have enough statistical evidence to reject the null hypothesis. We conclude that there is not enough statistical evidence that indicates that the true mean finishing times of team hare and team tortoise differ.

### Answer 1e (iii)

Assumption 1: The difference in sample means is normally distributed
In order for this assumption to be true, we know that all populations from which we take samples should be normal. Considering the fact that the sample size n < 30, we cannot extend Central Limit Theorem to ascertain that these samples come from a normal distribution. Hence, it is unfair to assume that the difference in sample means is normally distributed.

Assumption 2: The population variances are equal
When the population variances are known, the difference of the means has a normal distribution. But we have not tested the equality of the two variances. Even if they were to be equal, the significance level at which the equality of variances is tested might be different from the significance level at which we have performed the t-test to test the difference in means (5%).

## MANN WHITNEY U / WILCOXON TEST

### ANSWER 2A

$U_{Hare} = 81$
$U_{Tortoise} = 19$

```
                    ## Mann Whitney U Test ##

givenData <- read.csv("race.csv")
totalData <- c(givenData$Hare,givenData$Tortoise)
n1 <- length(givenData$Hare)
n2 <- length(givenData$Tortoise)
U_Hare <- 0
U_Tortoise <- 0
for (i in 1:n1) {
  for (j in 1:n2) {
    U_Hare <- U_Hare + (givenData$Hare[i] < givenData$Tortoise[j])
    U_Tortoise <- U_Tortoise + (givenData$Tortoise[j] < givenData$Hare[i])
  }
}
```

Listing 7: R code for calculating U statistic

```
> U_Hare
[1] 81
> U_Tortoise
[1] 19
```

Listing 8: Output

### ANSWER 2B

Since $U_{Hare} + U_{Tortoise} = n_1 n_2$, the expected value of U statistic, under normal approximation, is considered as the mean of the U statistics of each team. Hence,
$U = \frac{n_1 n_2}{2} = \frac{100}{2} = 50$ (under normal approximation)
To explain it more intuitively, since the size of team hare and team tortoise is 10, there are 100 combinations of hare and tortoise that are possible. Under the null hypothesis, we know that $U_{Hare} = U_{Tortoise}$ which means that the expected value of U would be the mean of the two. Hence, the expected value of U statistic is 50.

### ANSWER 2C (I)

```r
givenData <- read.csv("race.csv")
totalData <- c(givenData$Hare,givenData$Tortoise)
n1 <- length(givenData$Hare)
n2 <- length(givenData$Tortoise)
U_Hare <- 0
U_Tortoise <- 0
for (i in 1:n1) {
  for (j in 1:n2) {
    U_Hare <- U_Hare + (givenData$Hare[i] < givenData$Tortoise[j])
    U_Tortoise <- U_Tortoise + (givenData$Tortoise[j] < givenData$Hare[i])
  }
}
sigma <- sqrt((n1*n2*(n1 + n2 +1))/12)
mu <- (n1*n2)/2
z <- (U_Hare - mu)/sigma
pval_U <- 2*(1-(pnorm(z, lower.tail = TRUE)))
```

Listing 9: R code for calculating z statistic and p value

```
z
[1] 2.34338
> pval_U
[1] 0.01910992
```

Listing 10: Output

### ANSWER 2C (II)

CONCLUSION

With a p value of 0.01910992 ($\approx 0.01911$) at a 5% level of significance, we have enough statistical evidence to reject the null hypothesis which states that the number of hares passing the number of tortoises to be roughly the same as the number of tortoise passing the number of hares.

## ANSWER 2C (III)

```
                                   ## Wilcoxon Rank Sum Test ##
givenData <- read.csv("race.csv")
wilcox.test(givenData$Hare, givenData$Tortoise, exact = F, correct = F)
```

Listing 11: R code for calculating z statistic and p value

```
Wilcoxon rank sum test

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

Listing 12: Output

exact=F
R describes the exact parameter as a logical indication whether an exact p-value should be computed. When there are ties in sample data the ranking cannot be unique. Thus, due to the lack of unique ranking R may not be able to compute exact p value even when it is given as true.

correct=F
It is a continuity correction factor used which is used to approximate a discrete distribution to a normal one. While the continuity corrections for large samples are negligible, they could be significant for small samples. This is because of the fact that we can get a better fit if we make the observed value of statistic closer to the expected value by half of the interval between adjacent discrete values.

## Permutation Based Test

### Answer 3a

```r
                    ## Generate 3000 Permuted Datasets ##

givenData <- read.csv("race.csv")
totalData <- c(givenData$Hare,givenData$Tortoise)
##Rows 1 to 10 contain the finishing times of Hares and
##Rows 11 to 20 contain the finishing times of Tortoises

perm.totalData <- data.frame(n = NULL, Hare = NULL, Tortoise = NULL)
for (n in 1:3000) {
  HareSample <- sample(c(1:20),10,replace=F)
  TortSample <- c(1:20)[!(c(1:20) %in% HareSample)]
  ##Includes only those data that are not included in HareSample

  perm.datasets <- data.frame(n = n, Hare = dat[HareSample], Tortoise = dat[TortSample])
  perm.totalData <- rbind(perm.totalData, perm.datasets)
  ##Combining the header with the dataset
}
rm(perm.datasets)
## As we have all the permuted datasets and header in perm.totalData delete the perm.
    datasets
```

Listing 13: R code for generating 3000 permuted datasets

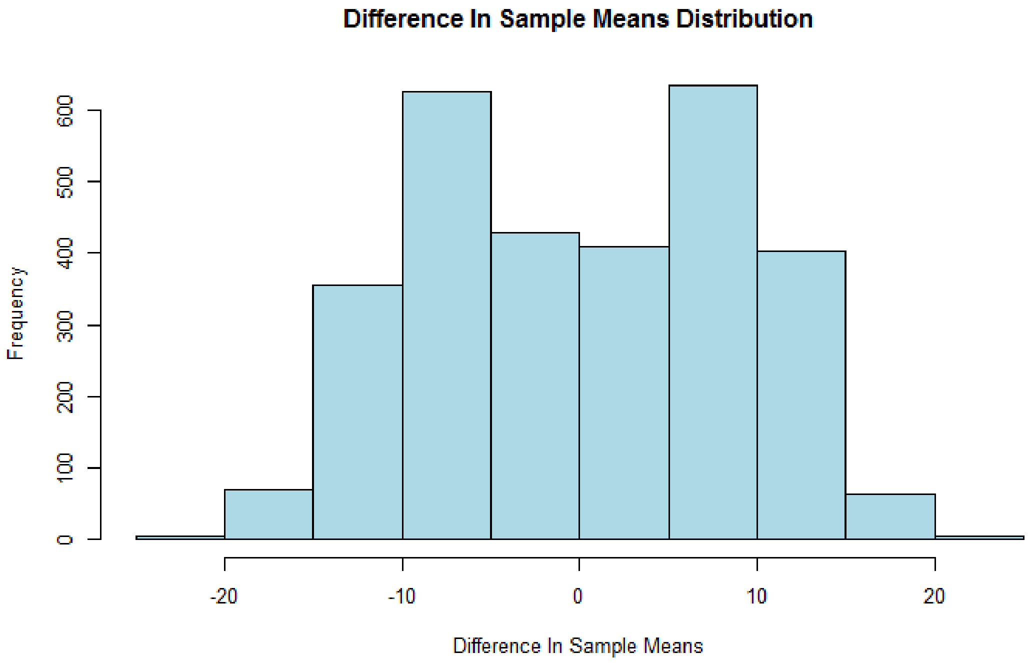
### Answer 3b

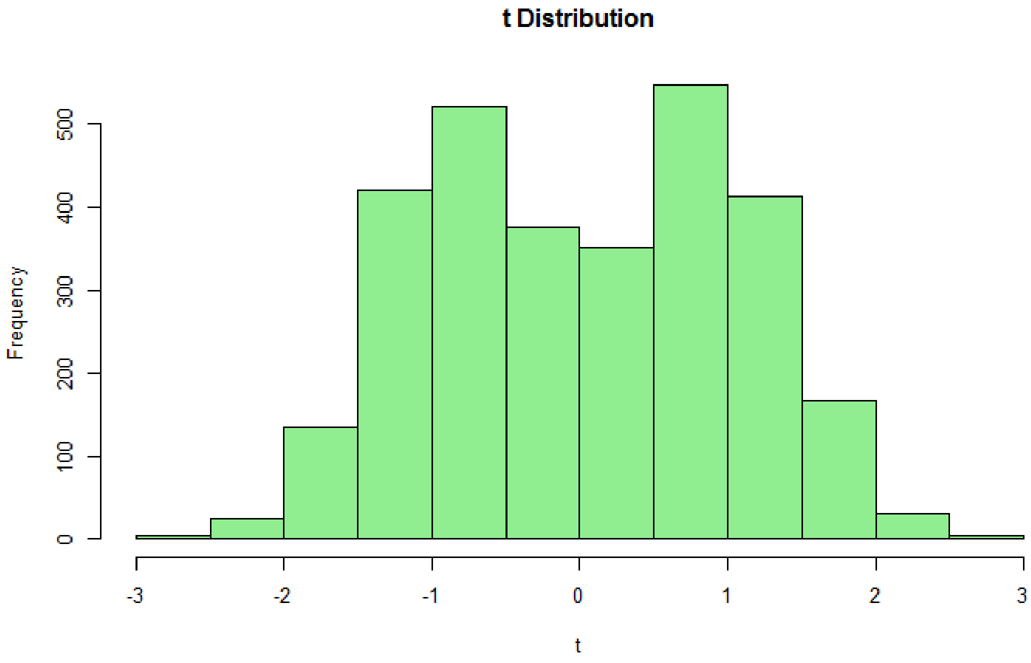
```r
            ## Evaluation Of Difference In Means And Test Statistics ##
perm.statistics <- data.frame(n = rep(NA, 3000), smeandiff = rep(NA, 3000),
                              t = rep(NA, 3000), U_Hare = rep(NA, 3000),
                              U_Tort = rep(NA, 3000), z = rep(NA, 3000),
                              W_Hare = rep(NA, 3000), W_Tort = rep(NA, 3000))
for (n in 1:3000) {
  rowindex = perm.totalData$n
  tempdata <- perm.totalData[which(rowindex == n),]
  ## Assigning the permuted dataset for particular value of n in tempdata

  ##Reading the value of nth iteration into the column n of dataset
  perm.statistics$n[n] <- n

  ## Calculation of the mean difference
  perm.statistics$smeandiff[n] <- mean(tempdata$Hare) - mean(tempdata$Tortoise)

  ## Calculation of pooled variance for particular n value
  tempvar_Hare <- var(tempdata$Hare)
  tempvar_Tort <- var(tempdata$Tortoise)
  n1 <- length(tempdata$Hare)
  n2 <- length(tempdata$Tortoise)
  pooledvar <- ((tempvar_Hare*(n1 - 1) + tempvar_Tort*(n2 - 1))/(n1 + n2 - 2))*((1/n1)+(1/n2))

  ## Calculation of t statistic
  perm.statistics$t[n] = perm.statistics$smeandiff[n]/sqrt(pooledvar)

  ## Calculation of U statistic
  perm.U_Hare <- 0
  perm.U_Tort <- 0
  for (i in 1:length(tempdata$Hare)){
    for (j in 1:length(tempdata$Tortoise)) {
      perm.U_Hare <- perm.U_Hare + (tempdata$Hare[i] < tempdata$Tortoise[j])
      perm.U_Tort <- perm.U_Tort + (tempdata$Tortoise[j] < tempdata$Hare[i])
    }
  }
  perm.statistics$U_Hare[n] <- perm.U_Hare
  perm.statistics$U_Tort[n] <- perm.U_Tort

  ## Calculation of z statistic
  perm.statistics$z[n] <- (perm.U_Hare - mu)/sigma

  ## Calculation of rank sum statistics
  perm.statistics$W_Hare[n] <- sum(rank(c(tempdata$Hare,tempdata$Tortoise))[1:10])
  perm.statistics$W_Tort[n] <- sum(rank(c(tempdata$Hare,tempdata$Tortoise))[11:20])
}
```

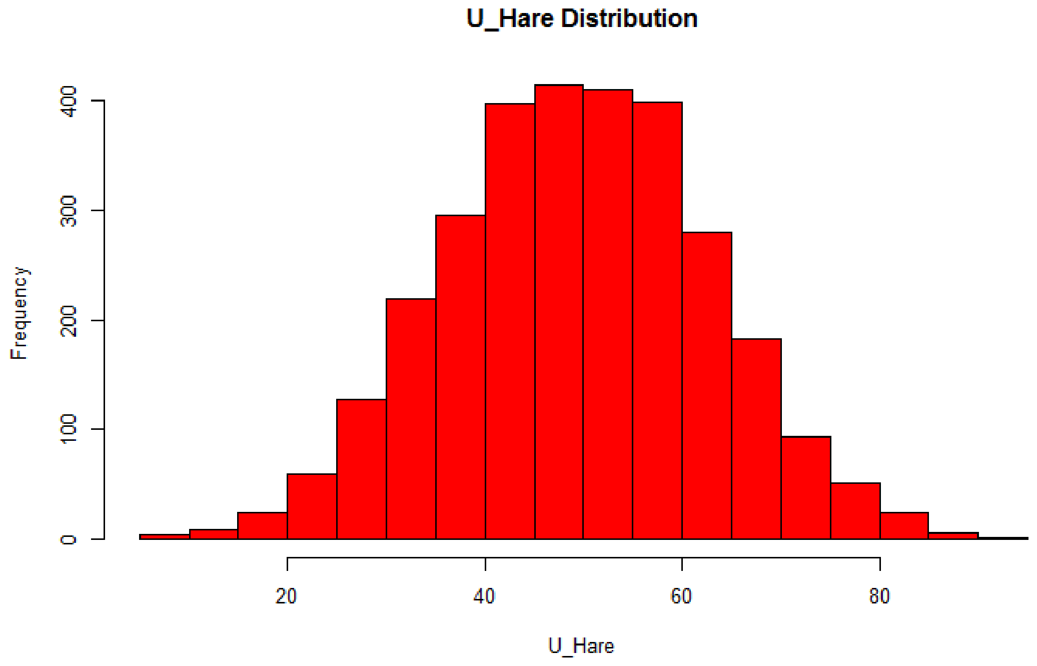
Listing 14: R code for evaluating statistics for each permuted dataset

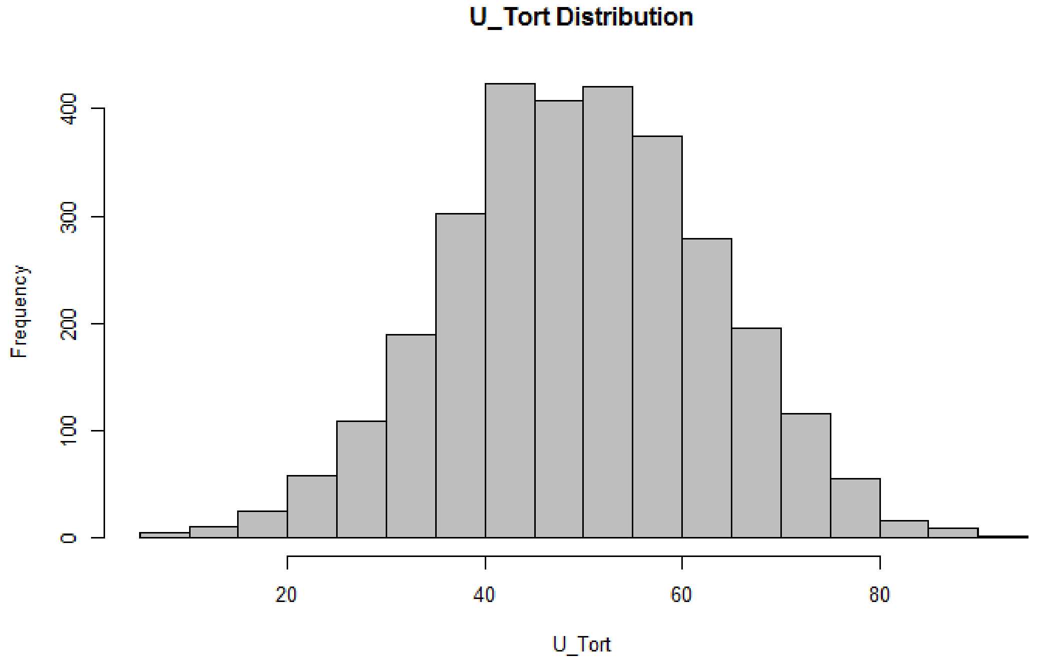
### ANSWER 3C

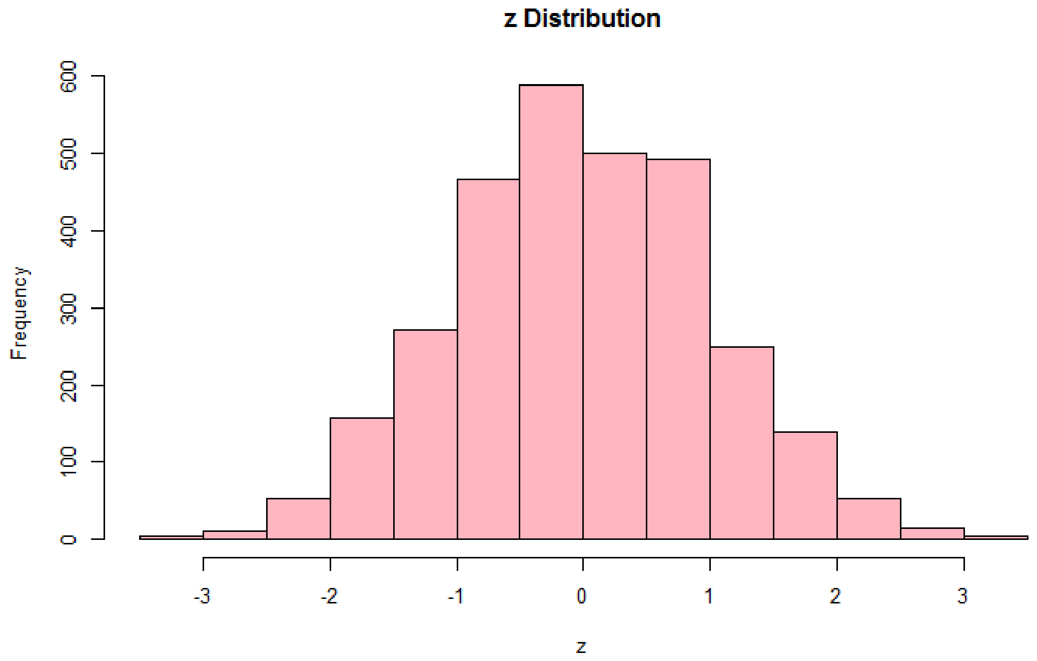
Below are the plots of each of the distributions:

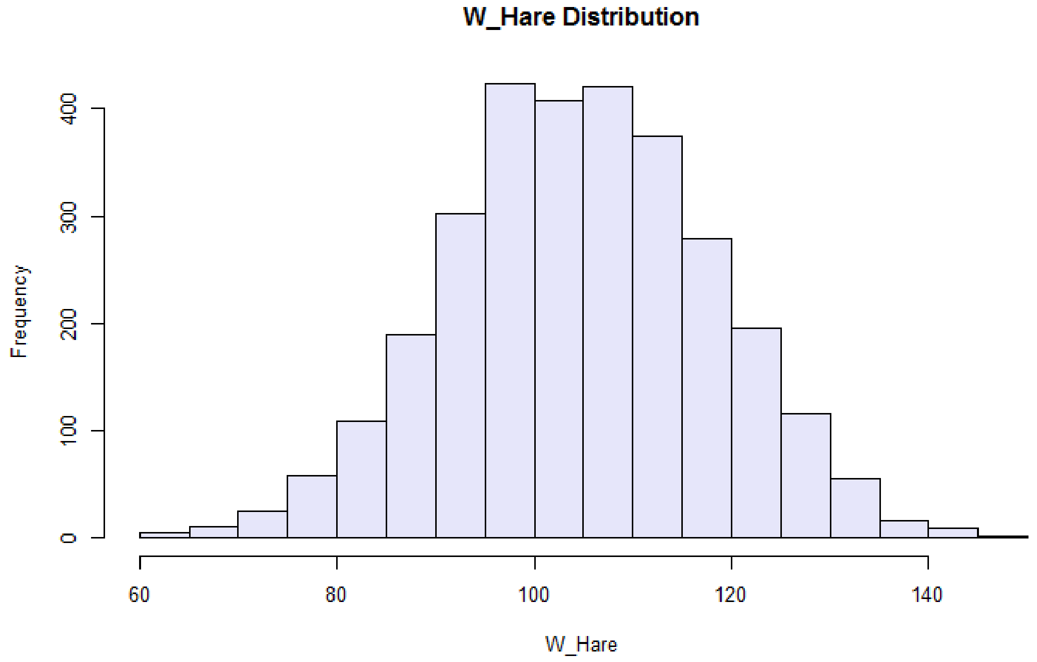


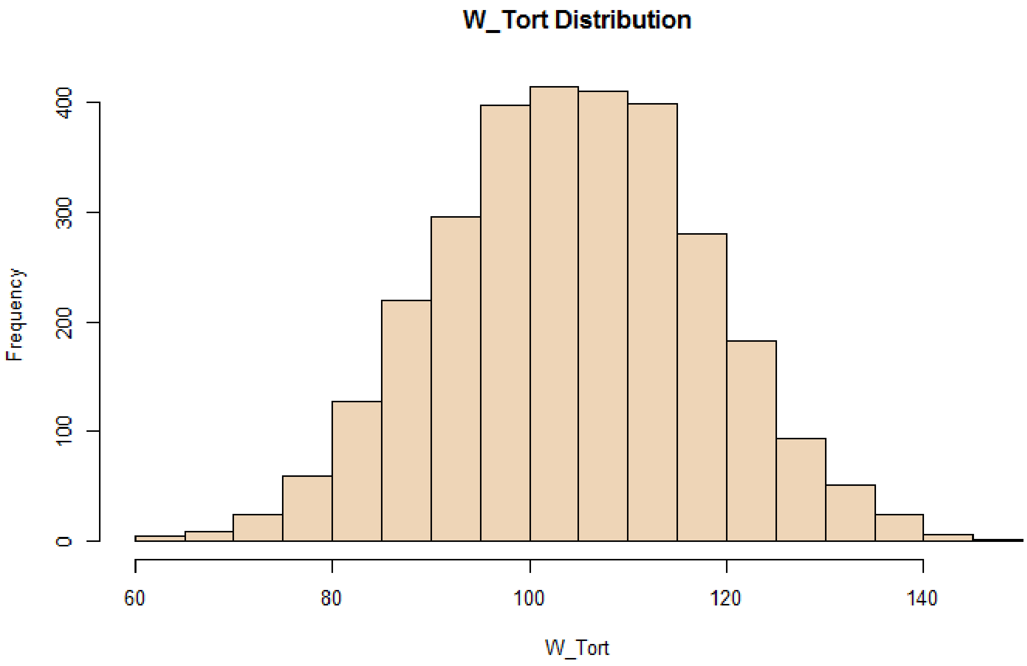



U_Hare Distribution
Frequency
0
100
200
300
400
20
40
60
80
U_Hare


U_Tort Distribution
Frequency
0
100
200
300
400
20
40
60
80
U_Tort


z Distribution
Frequency
0
100
200
300
400
500
600
-3
-2
-1
0
1
2
3
z


W_Hare Distribution
Frequency
0
100
200
300
400
60
80
100
120
140
W_Hare



Observations:
The distributions for $\bar{X}_{Hare} - \bar{X}_{Tortoise}$ and t-statistic resemble each other. This is due to the fact that they differ only by a constant factor, i.e., inverse of the standard error. Also, they do not approximate to a normal distribution. This can be proved by performing Shapiro-Wilk Test (The results have been attached below).

All of the other distributions approximate to a normal one. The mean of U statistic for both the teams is around 50 as expected. Since the Wilcoxon Rank Sum Test and the Mann Whitney U Test are similar kind of tests, the respective distributions resemble each other (with a change in the scale of X-axis).

## Results From Shapiro-Wilk Test (Normality Check)

```
                  ## Shapiro-Wilk Test For Each Of The Distributions ##

## (i) Differnece In Sample Means
shapiro.test(perm.statistics$smeandiff)
## data:  perm.statistics$smeandiff
## W = 0.96409, p-value < 2.2e-16
## As p < 0.05 this distribution does not approximate to a normal one

## (ii) T Distribution
shapiro.test(perm.statistics$t)
## data:  perm.statistics$t
## W = 0.97498, p-value < 2.2e-16
## As p < 0.05 this distribution does not approximate to a normal one



## (iii) U_Hare Distribution
shapiro.test(perm.statistics$U_Hare)
## data:  perm.statistics$U_Hare
## W = 0.99902, p-value = 0.09157
## As p > 0.05 This distribution can be approximated to a normal one


## (iv) Z Distribution
shapiro.test(perm.statistics$z)
## data:  perm.statistics$z
## 0.99902, p-value = 0.09157
## As p > 0.05 As expected, this distribution is a normal one


## (v) W Statistic Distribution
shapiro.test(perm.statistics$W_Hare)
## data:  perm.statistics$W_Hare
## 0.99902, p-value = 0.09157
## As p > 0.05 This distribution is a normal one.
```

Listing 15: Shapiro-Wilk Test

```
Shapiro-Wilk normality test

data:  perm.statistics$smeandiff
W = 0.96409, p-value < 2.2e-16


Shapiro-Wilk normality test

data:  perm.statistics$t
W = 0.97498, p-value < 2.2e-16


Shapiro-Wilk normality test

data:  perm.statistics$U_Hare
W = 0.99902, p-value = 0.09157


Shapiro-Wilk normality test

data:  perm.statistics$z
W = 0.99902, p-value = 0.09157


Shapiro-Wilk normality test

data:  perm.statistics$W_Hare
W = 0.99902, p-value = 0.09157
```

Listing 16: Output

### 3 C (II)

Expected Mean Values For The Distributions:

$$E(\mu_{\bar{X}_{Hare}-\bar{X}_{Tortoise}}) = 0$$

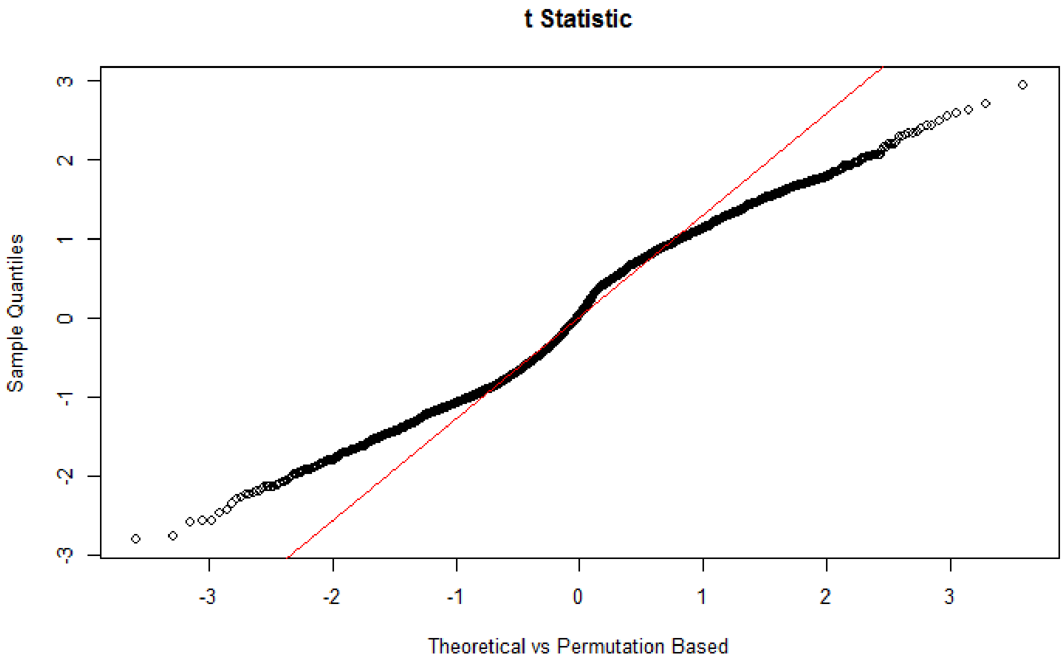
$$E(\mu_t) = 0$$

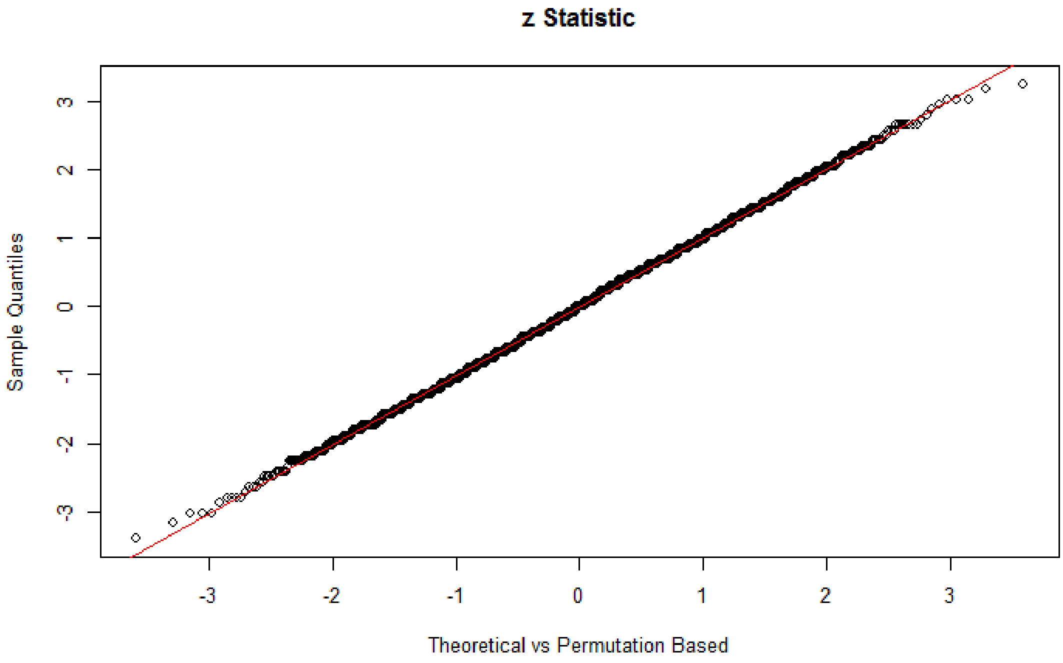
$$E(\mu_{U_{Hare}}) = E(\mu_{U_{Tortoise}}) = \frac{n_1 n_2}{2} = 50$$

$$E(\mu_z) = 0$$

$$E(\mu_{W_{Hare}}) = E(\mu_{W_{Tortoise}}) = \frac{n_1(n_1+n_2+1)}{2} = 105$$

3 C (IV)





In support of the results obtained from the Shapiro-Wilk Test, the QQ plot for t Statistic suggests that the distribution deviates from normality unlike that of the z Statistic.

## ANSWER 4

For the tortoise and hare racing problem, non parametric tests are proven to given better result than the t-test as the sample size n<30. However, there may be various other factors that we need to consider for choosing between these two kinds of test.

Reasons to Use Nonparametric Tests:
Reason 1: The area of study is better represented by the median
When the distribution is skewed enough, the mean is strongly affected by changes far out in the tail of the distribution whereas the median continues to more closely reflect the center of the distribution.
Reason 2: A very small sample size
When the sample size is really small (n<30), we might not even be able to ascertain the distribution of the data because the distribution tests will lack sufficient power to provide meaningful results. In such a case, it is better to choose a nonparametric test.
Reason 3: When we have ordinal data, ranked data, or outliers that cannot be removed
Typical parametric tests can only assess continuous data and the results can be significantly affected by outliers. Conversely, some nonparametric tests can handle ordinal data, ranked data, and not be seriously affected by outliers.

Reasons to Use Parametric Tests:
Reason 1: Parametric tests can perform well with skewed and non-normal distributions.
These tests can perform well with continuous data that are non-normal if the sample size requirement is satisfied.
Reason 2: Parametric tests can perform well when the spread of each group is different
While nonparametric tests do not assume that the data follows a normal distribution, they do have other assumptions that can be hard to meet. For nonparametric tests that compare groups, a common assumption is that the data for all groups must have the same spread (dispersion). If the groups have a different spread, the nonparametric tests might not provide valid result.
Reason 3: Statistical power
Parametric tests usually have more statistical power than nonparametric tests. Thus, it is more likely to detect a significant effect when one truly exists.