Homework 2

Non-Parametric Statistics

Langdon Holmes

Question 1 Wilcoxon

```r
urban <- c(1,0,1,1,0,0,1,1,1,8,1,1,1,0,1,1,2)
rural <- c(3,2,1,1,2,1,3,2,2,2,2,5,1,4,1,1,1,1,6,2,2,2,1,1)

get_ranks <- function(a, b){
  m <- length(a)
  n <- length(b)
  combined_ranks <- rank(c(a, b), ties.method="average")
  a_ranks <- combined_ranks[1:m]
  b_ranks <- combined_ranks[(m+1):(m+n)]
  return(list(a_ranks, b_ranks))
}

R <- get_ranks(urban, rural)
R1 <- sapply(R[1], sum)
R2 <- sapply(R[2], sum)

cat("R1:", R1, "\nR2:", R2)
```

```
R1: 246.5
R2: 614.5
```

The critical values for a two-tailed Wilcoxon Rank Sum test with sample sizes 17 and 24 are $W \leq 282$ or $W \geq 432$.

Based on the above calculations, there is sufficient evidence to reject the null hypothesis that the means are equal. Since R_1 (the sum of urban sibling ranks) is below the critical value, rural folk are more likely to have more siblings than urbanites.

```
est_var <- function(x) {
  return(sum((x - mean(x))^2)/(length(x)-1))
}

pool_var <- function(a, b) {
  numerator <- (length(a)-1)*est_var(a) + (length(b)-1)*est_var(b)
  denominator <- length(a) + length(b)-2
  return(numerator/denominator)
}

student_t_test <- function(a, b) {
  pooled_var_est <- pool_var(a,b)
  t = (mean(a) - mean(b)) / sqrt(pooled_var_est*(1/length(a) + 1/length(b)))
  return(t)
}

cat(student_t_test(urban, rural), "not less than", qt(p=.05, df=39), "\n")
```

```
-1.638335 not less than -1.684875
```

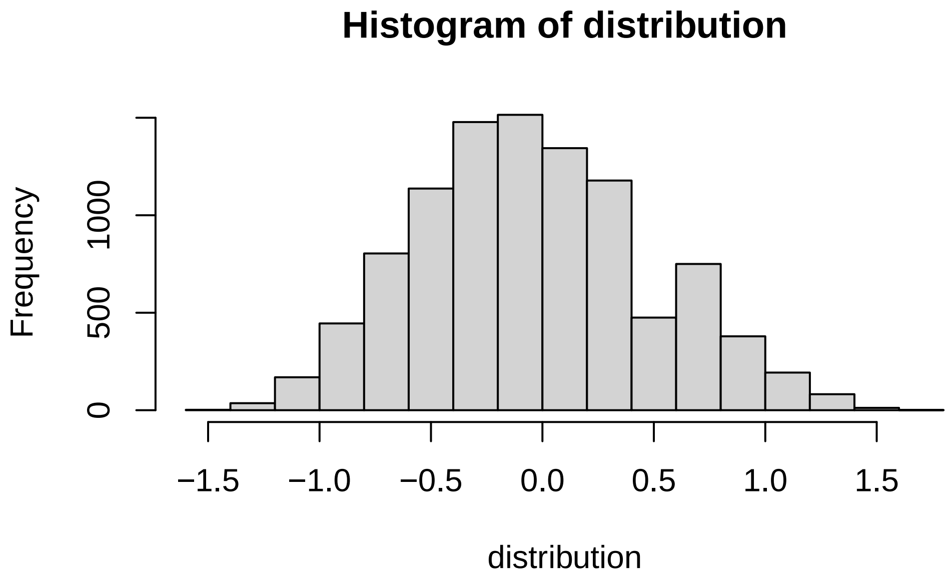
```
#equivalent to
t.test(urban, rural, var.equal=TRUE, "less")
```

```

    Two Sample t-test

data:  urban and rural
t = -1.6383, df = 39, p-value = 0.0547
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.02290674
sample estimates:
mean of x mean of y
 1.235294  2.041667
```

Assuming equal variance between the two samples, the critical value is $t_{.05,39} = -1.68$. The test statistic is -1.64, which is not extreme enough to reject the null hypothesis that the means are equal between the two groups. This seems to be a Type II error. There is one outlying urbanite with 8 siblings that may be causing undue influence on the T-test (while not having any undue influence on the rank-based Wilcoxon).

Question 2 Permutation and T-test

```r
mean.diff <- function(x, y){
  return(mean(x)-mean(y))
}

# we will reuse this for two-group data
permute.apply <- function(x, y, n_permutations, func, lower=TRUE){
  m <- length(x)
  combined <- c(x,y)
  N <- length(combined)

  distribution <- c()
  for(i in 1:n_permutations){
    sampled=sample(combined)
    distribution[i] <- func(sampled[1:m], sampled[(m+1):N])
  }

  observed <- func(x,y)
  result = quantile(distribution, c(.05,.95))
  hist(distribution)

  if(lower==TRUE){
  p_value = sum(distribution < observed)/n_permutations
  } else {
  p_value = sum(distribution > observed)/n_permutations
  }

  return(list(observed, result, p_value))
}

permute.apply(urban, rural, 10000, mean.diff)
```



```
[[1]]
[1] -0.8063725

[[2]]
        5%        95%
-0.8063725  0.8014706

[[3]]
[1] 0.0408
```

I used a permutation test to compare mean differences between the groups. There is a low probability ($p < 0.05$) of finding such a low sum of ranks for urbanites under H_0. While the permutation test was significant, its p-value is probably closer to the T-test than to the Wilcoxon test. I did not calculate an exact p-value for the Wilcoxon test, but it is probably much smaller than either the permutation test or the T-test. This result is not surprising, since the T-test and the permutation test are both considering mean differences, while the Wilcoxon test operates on rank-transformed data.

Question 3 Large Sample Approximation

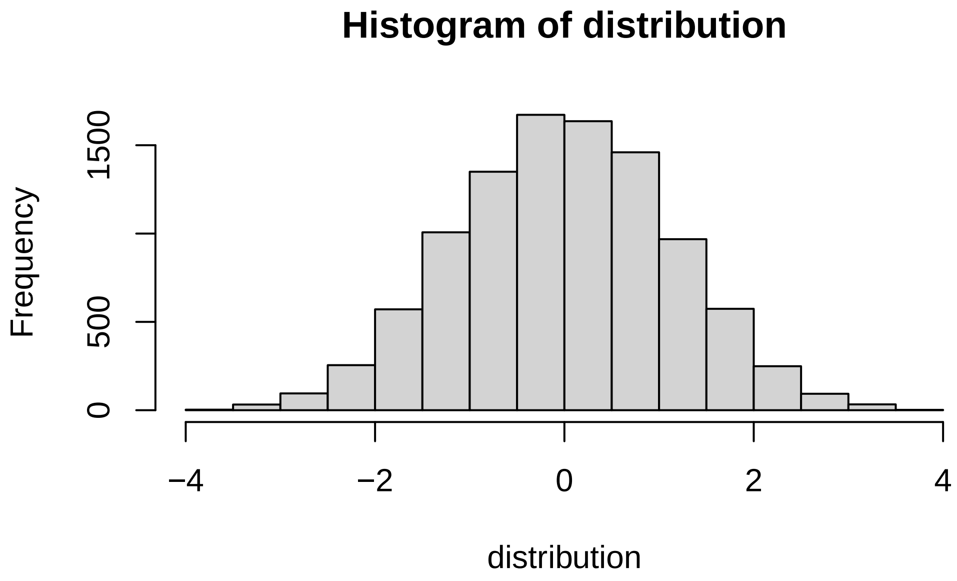
```
wilcox_approx <- function(a, b){
  m <- length(a)
  n <- length(b)
  combined_ranks <- rank(c(a, b), ties.method="average")
  T_1 <- sum(combined_ranks[1:m])
  N <- m+n
  mu <- sum(combined_ranks)/N
  t_expect <- m*mu
  variance <- (sum(combined_ranks^2) / N) - mu^2
  t_var <- m*n*variance / (N - 1)
  z <- (T_1 - t_expect) / sqrt(t_var)
  return(z)
}


wilcox_approx(urban, rural)
```

```
[1] -3.170701
```

The approximation returns a fairly extreme z-score that seems in line with the rank sum test in 1(a), indicating that the approximation is at least reasonably good.

Question 4 Permutation and T-test

```
C1 <- c(1.0,5.3,1.2,3.9,8.3,6.3,2.2,9.8,2.8,2.6)
C2 <- c(5.1,6.0,8.0,8.2,7.3,4.4,7.4,7.5,6.4,4.5,8.9)

permute.apply(C1, C2, 10000, mean.diff)
```



```
[[1]]
[1] -2.36

[[2]]
       5%       95% 
-1.863636  1.840000 

[[3]]
[1] 0.0167
```

```
t.test(C1, C2, var.equal = TRUE)
```

```

    Two Sample t-test

data:  C1 and C2
t = -2.2978, df = 19, p-value = 0.0331
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.5096724 -0.2103276
sample estimates:
```

```
mean of x mean of y
     4.34      6.70
```

The two sample t-test found a significant difference between the two groups. Since C2 has a higher mean, we can conclude that the population with a diet richer in corn exhibits more aggressive behavior. The permutation based approach also resulted in significant findings, and with a lower p-value (.0154). It seems that both approaches were fairly capable of modeling the central tendencies in this dataset, despite that the C1 (less corn) group has a few highly aggressive participants.

Question 5 Equality of Variance F-test, Siegel-Tukey, and Higgins

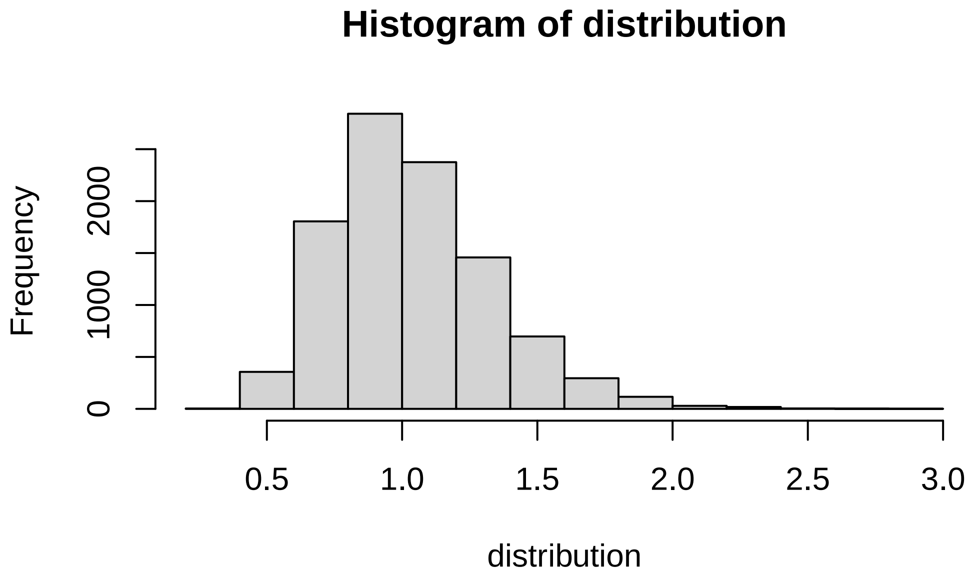
```r
# F-test for Equality of Variance
f_variance <- function(x, y){
  F_stat <- var(x)/var(y)

  F_crit <- qf(p=.05, df1=length(x-1), df2=length(y-1), lower.tail=FALSE)
  return(cat("F-test:", F_stat, "is above", F_crit, "\n"))
}
f_variance(C1,C2)
```

```
F-test: 3.844237 is above 2.853625
```

```r
# Siegel-Tukey
labels <- c(rep("A", length(C1)), rep("B", length(C2)))

# do not sort, just get indices
sorting_indices <- sort(c(C1, C2), index.return=TRUE)$ix

# hard code this
st_ranks <- c(1,21,20,2,3,19,18,4,5,17,16,6,7,15,14,8,9,13,12,10,11)

# indexing is quite tricky here...
w <- sum((1:21)[labels[sorting_indices[st_ranks]]=="A"])
cat("Siegel:", w, "is not below the critical value of 81.")
```

```
Siegel: 86 is not below the critical value of 81.
```

```r
# Higgins
mad <- function(x){
  return(sum(abs(x-median(x)))/length(x))
}

rmd <- function(x,y){
  return(mad(x)/mad(y))
}

higgins <- permute.apply(C1, C2, 10000, rmd, lower=FALSE)
```



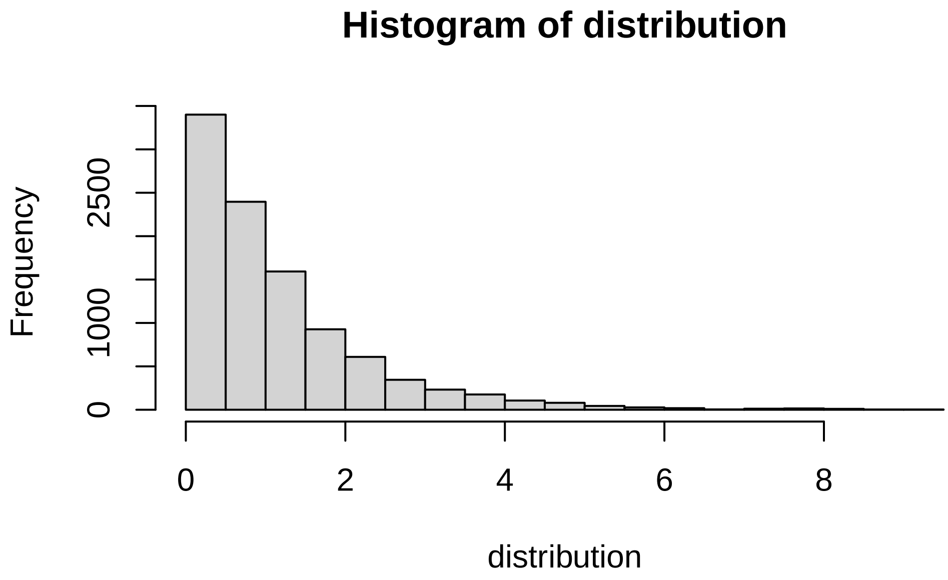
```r
cat("Higgins:", "observed:", higgins[[1]], "critical:", higgins[[2]][2], "p-value:", higgi
```

```
Higgins: observed: 1.925 critical: 1.58812 p-value: 0.008
```

$F_{.05,9,10} = 3.02$, so the F statistic of 3.84 is significant evidence that the variances are unequal. The Siegel-Tukey test, despite being the most challenging test to code, did not find a significant difference in the variances. The Higgins test, like the F-test, demonstrates evidence that the two groups do not share equal variance. So not all tests agree on whether there is a significant difference in the degree of variance between high and low corn populations.

Question 6 Permutation F-test and one-way ANOVA

```r
G1 <- c(2.9736, 0.9448, 1.6394, 0.0389, 1.2958)
G2 <- c(0.7681, 0.8027, 0.2156, 0.0740, 1.5076)
G3 <- c(4.8249, 2.2516, 1.5609, 2.0452, 1.0959)

df <- data.frame(
  group=factor(c(
    rep("1", times=length(G1)),
    rep("2", times=length(G2)),
    rep("3", times=length(G3))
  )),
  value=c(G1, G2, G3))

# this is a better approach, but I don't want to refactor previous code
permute.apply.df <- function(df, n_permutations, fun, lower=TRUE){

  distribution <- c()

  df$sampled <- df$value
  observed <- fun(df)

  for(i in 1:n_permutations){
    df$sampled <- sample(df$value)
    distribution[i] <- fun(df)
  }

  result = quantile(distribution, c(.05,.95))
  hist(distribution)

  if(lower==TRUE){
  p_value = sum(distribution < observed)/n_permutations
  } else {
  p_value = sum(distribution > observed)/n_permutations
  }

  return(list(result, observed, p_value))
}

get_f <- function(df){
  return(oneway.test(sampled ~ group, data=df, var.equal=TRUE)$statistic)
}

permute.apply.df(df, 10000, get_f, lower = FALSE)
```



```
[[1]]
        5%        95%
0.07359389 3.48816180

[[2]]
       F
2.990689

[[3]]
[1] 0.0735
```

```
oneway.test(value ~ group, df, var.equal = TRUE)
```

```

    One-way analysis of means
```

```
data:  value and group
F = 2.9907, num df = 2, denom df = 12, p-value = 0.08834
```

I ran a K-sample test one-way ANOVA 10,000 times on randomly sampled permutations of the original data. This approach did not find a significant difference across the three groups.

Using the same data, I tested the hypothesis that there is a difference between the group means using a simple one way ANOVA in R. This test also did not find a significant difference between the group means.

Question 7 Kruskal-Wallis

```
kruskal.test(list(G1, G2, G3))
```

```

	Kruskal-Wallis rank sum test

data:  list(G1, G2, G3)
Kruskal-Wallis chi-squared = 5.78, df = 2, p-value = 0.05558
```

The Kruskal-Wallis test in R did not quite reach the level of significance at our standard alpha level, but it is quite a bit closer than the methods of analysis in (6). None of the three tests provide sufficient evidence to reject the null hypothesis that the means are equal between the three groups.