SDA Group Submission Assignment Assign5

Group Gr18

MengliFeng (2720589) and PepijnVanOostveen (2801582)

Exercise 1

```r
# Load the data
grades <- readRDS("grades.RDS")
on_time <- grades$on_time
late <- grades$late
```

a.

```r
# Shapiro-Wilk Test for normality
shapiro_on_time <- shapiro.test(on_time)
shapiro_late <- shapiro.test(late)

# Print results
cat("Shapiro-Wilk Test for on_time:\n")
```

```
## Shapiro-Wilk Test for on_time:
```

```r
print(shapiro_on_time)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  on_time
## W = 0.86829, p-value = 7.731e-14
```

```r
cat("\nShapiro-Wilk Test for late:\n")
```

```
##
## Shapiro-Wilk Test for late:
```

```r
print(shapiro_late)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  late
## W = 0.87859, p-value = 0.006463
```

for the on-time sample, the null hypothesis is rejected for the late sample, the null hypothesis is rejected Since the null hypothesis for Shapiro-Wilk test is that the data is normally distributed and it is rejected for both of the samples, we know that t-test (requires normal distribution of the samples) is not an approriate location test for the data

b.

```r
# Remove values equal to 7, as they don't contribute to the sign test
filtered_on_time <- on_time[on_time != 7]

# Count how many are greater than 7
n <- length(filtered_on_time)
num_positive <- sum(filtered_on_time > 7)

# Exact binomial test (sign test)
sign_test <- binom.test(num_positive, n, p = 0.5, alternative = "greater")
cat("\nSign Test Result (on_time > 7):\n")
```

```
##
## Sign Test Result (on_time > 7):
```

```r
print(sign_test)
```

```
##
##  Exact binomial test
##
## data:  num_positive and n
## number of successes = 170, number of trials = 250, p-value = 6.38e-09
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.6280437 1.0000000
## sample estimates:
## probability of success
##                   0.68
```

```r
# Normal approximation
# Mean and standard deviation under H0
mu <- n * 0.5
sigma <- sqrt(n * 0.5 * 0.5)

# Normal approximation (continuity correction)
z <- (num_positive - mu ) / sigma
p_norm <- 1 - pnorm(z)
cat("\nNormal approximation p-value:\n")
```

```
##
## Normal approximation p-value:
```

```r
print(p_norm)
```

```
## [1] 6.274324e-09
```

the null hypothesis under the sign test is rejected the p-value for the normal approximation is slightly larger than that under sign test, with magnitude of e-9. the null hypothesis for the normal approximation is also rejected.

c.

```r
# Remove values equal to 7
filtered_late <- late[late != 7]
```

```
# Count how many are greater than 7
n_late <- length(filtered_late)
num_positive_late <- sum(filtered_late > 7)

# Exact binomial test (sign test)
sign_test_late <- binom.test(num_positive_late, n_late, p = 0.5, alternative = "greater")
cat("\nSign Test Result (late > 7):\n")
```

```
##
## Sign Test Result (late > 7):
```

```
print(sign_test_late)
```

```
##
##  Exact binomial test
##
## data:  num_positive_late and n_late
## number of successes = 17, number of trials = 25, p-value = 0.05388
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.4963584 1.0000000
## sample estimates:
## probability of success
##                   0.68
```

```
# Normal approximation
mu_late <- n_late * 0.5
sigma_late <- sqrt(n_late * 0.5 * 0.5)
z_late <- (num_positive_late - mu_late) / sigma_late
p_norm_late <- 1 - pnorm(z_late)
cat("\nNormal approximation p-value (late):\n")
```

```
##
## Normal approximation p-value (late):
```

```
print(p_norm_late)
```

```
## [1] 0.03593032
```

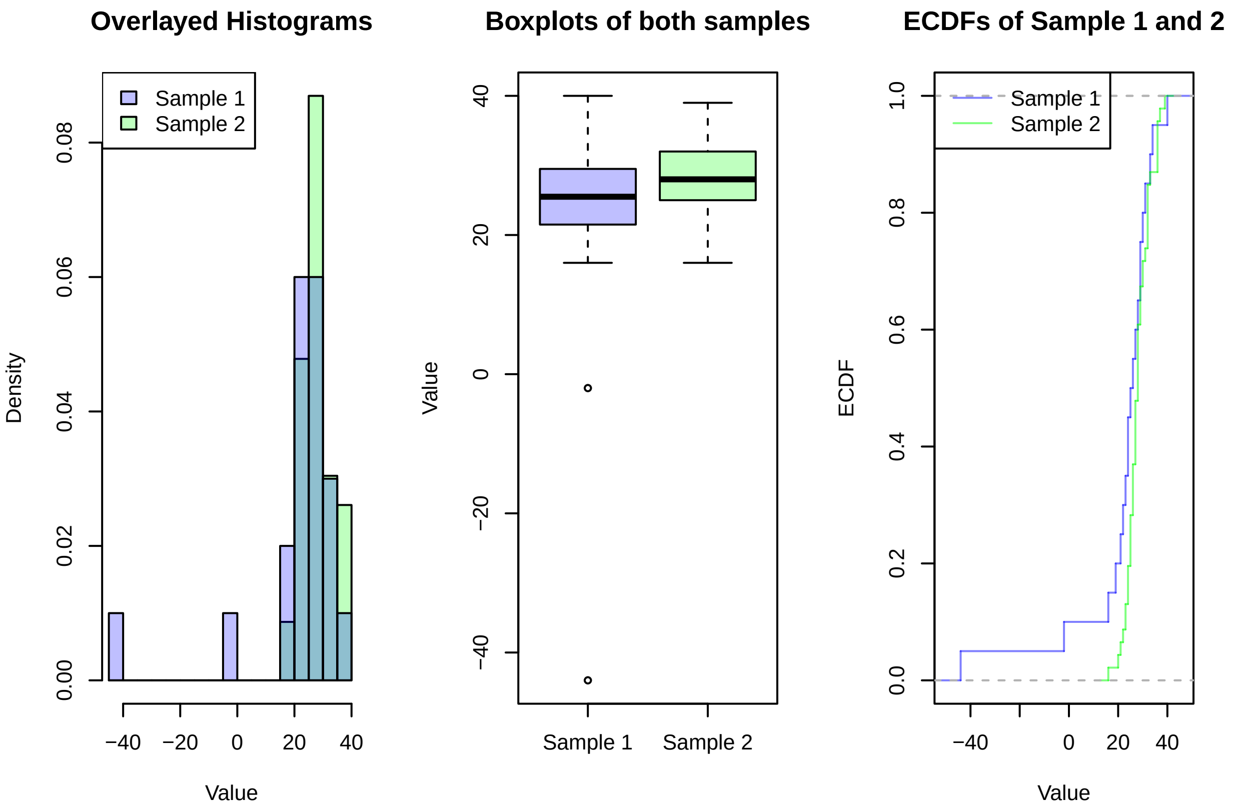
the normal approximation also lead to rejection of null hypothesis as the sign test does, but the p-value is off for around 0.0179. The reason is that binomial can be approximated by normal distribution with large sample size. As $n = 25$ (small sample size) and $p = 0.5$, $n \cdot p \cdot (1 - p) < 10$, the p-valule from normal distribution is less accurate.

Exercise 2

a.

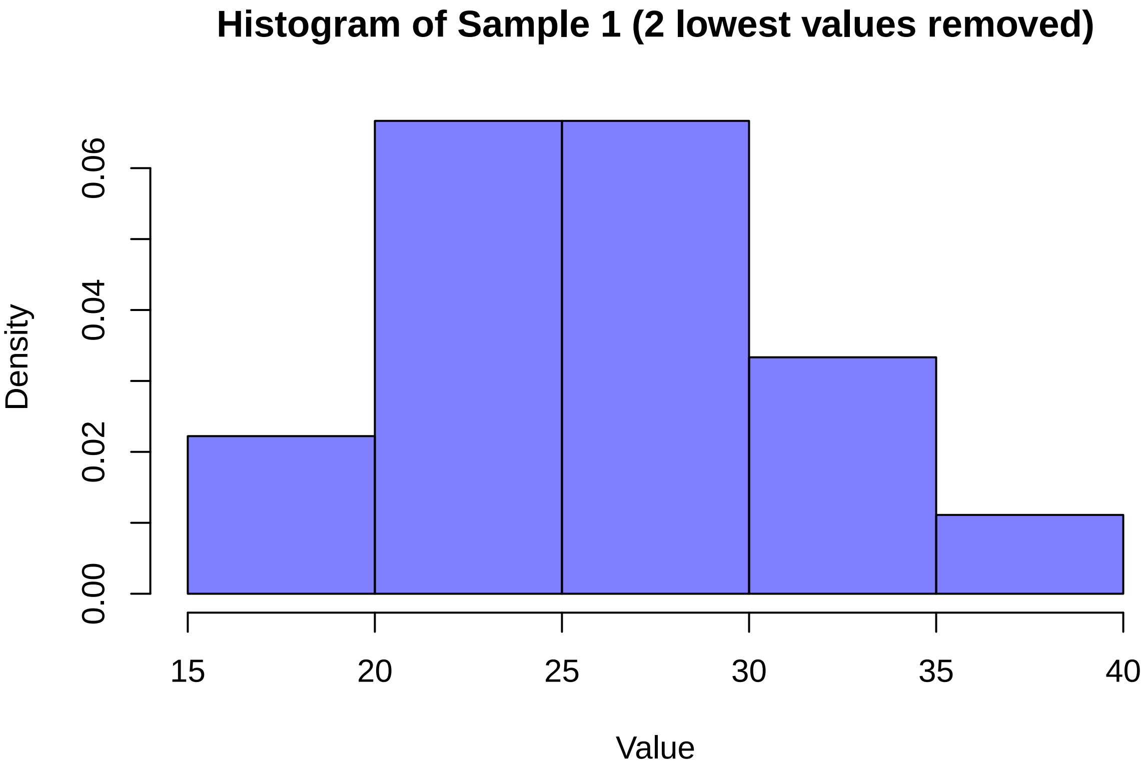
```
# load the samples and split it into the first 20 observations and the rest.
samples <- readRDS("newcomb.RDS")
sample1 <- samples[1:20]
sample2 <- samples[21:66]
```

```r
# set up plot layout. Two columns for the 2 sets of data and 3 rows for the 3 different plots
par(mfrow = c(1, 3))

# histogram
breaks <- pretty(range(samples), n = 20)

hist(sample2, freq = FALSE, breaks=breaks, col = rgb(0, 1, 0, 1/4), xlim = range(samples),
     main = "Overlayed Histograms", xlab = "Value")
hist(sample1, freq = FALSE, breaks=breaks, col = rgb(0, 0, 1, 1/4),xlim = range(samples), add = TRUE)
legend("topleft", legend = c("Sample 1", "Sample 2"),
       fill = c(rgb(0, 0, 1, 1/4), rgb(0, 1, 0, 1/4)))


# Boxplot
boxplot(sample1, sample2, names = c("Sample 1", "Sample 2"), col = c(rgb(0, 0, 1, 1/4), rgb(0, 1, 0,  1/4)),
        main = "Boxplots of both samples", ylab = "Value")


# empirical CDF
plot(ecdf(sample1), verticals = TRUE, do.points = FALSE, col = rgb(0, 0, 1, 1/2),
     main = "ECDFs of Sample 1 and 2", xlab = "Value", ylab = "ECDF")
lines(ecdf(sample2), verticals = TRUE, do.points = FALSE, col = rgb(0, 1, 0, 1/2))
legend("topleft", legend = c("Sample 1", "Sample 2"), col = c(rgb(0, 0, 1,  1/2), rgb(0, 1, 0,  1/2)), lty = 1)
```



Sample 2 looks like it could be the result of a normal distribution and the first 20 observations (sample 1) would also look like it comes from a normal distribution, but only if the 2 outliers were removed. These two outliers from sample 1 are the only outliers and make sample 1 asymmetric, while sample 2 isn't completely symmetric but is close to it. (I used rgb values as colors since I needed transparent colors for the histogram and wanted to use the same colors in the other graphs. I reduced the transparency in the ECDF for better visibility.)

```r
sample1_trimmed <- sort(sample1)[-c(1, 2)]
hist(sample1_trimmed, freq = FALSE, col = rgb(0, 0, 1, 1/2),
     main = "Histogram of Sample 1 (2 lowest values removed)", xlab = "Value")
```



This histogram shows that the first 20 observations do indeed resemble a normal distribution if the outliers are removed.

b.

The Wilcoxon two-sample test assumes that the distributions are the same shape and looks at the location of these shapes. This is probably a bad test for our sampels due to the outliers and we don't know if both samples come from a normal distribution.' So the Kolmogorov-Smirnov two-sample test is probably the best test since it tests if the samples come from the same distribution shape.

c.

```
ks_result <- ks.test(sample1, sample2)
ks_result
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  sample1 and sample2
## D = 0.25435, p-value = 0.1776
## alternative hypothesis: two-sided
```

The p-value is higher than our significance level of 0.05 and thus we don't reject our hypothesis that the two distributions come from the same distribution. With how low the p-value is I would say that we don't know if the two samples come from the same distribution without context.

With context I would say that a reasonable explanation for the difference is that the error of experiment is normally distributed and that the Newcomb became better at the experiment as he went along so the only 2 outliers happened in the first 20 observations. Let us remove this two possible mistakes and do the test again.

```r
ks_result <- ks.test(sample1_trimmed, sample2)
ks_result
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  sample1_trimmed and sample2
## D = 0.19324, p-value = 0.4917
## alternative hypothesis: two-sided
```

The p-value is 0.4917 and my conclusion is that we still can't be very confident one way or another.