Biostatistics - Exercises Part 1

Lúa Arconada and Alejandro Macías

2024-04-25

```
library(coin)
```

```
## Loading required package: survival
```

```
library(car)
```

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

```
library(broman)
library(permuco)
library(ggplot2)
library(tidyr)
library(stats)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(qvalue)
```

```
set.seed(1234)
```

1. In a study to investigate the effect of oyster density on seagrass biomass, researchers introduced oysters to thirty parcels of healthy seagrass. At the beginning of the study, the seagrass was clipped short in all plots. Next, 10 randomly chosen plots received a high density of oysters; 10, an intermediate density; and 10, a low density. As a control, an additional 10 randomly chosen clipped parcels received none. After 2 weeks, the belowground seagrass biomass was measured in each parcel (gr/m2). The mean square error from the ANOVA table was 220.94.

		Oyster density		
	None	Low	Intermediate	High
Mean	34.81	33.13	28.33	15.00

- Compute three Bonferroni-adjusted confidence intervals comparing parcels with Low, Intermediate and High density with the one with no oysters using $\alpha = 0.05$. The corresponding formula for the Bonferroni-adjusted confidence intervals is the following:

$$x_i - x_j \mp t_{\alpha/2N, n_1+n_2-2}\sqrt{error\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

where x_i is the mean of the ith-group, N is the number of pairwise comparisons between our 4 groups ($N = 6$ in our case), $t_{\alpha/2N, n_1+n_2-2}$ is the t-student with $n_1 + n_2 - 2$ degrees of freedom $\alpha/2N$ quantile, error is the ANOVA table error and n_i the number of observations in the ith-group.

```
alpha = 0.05
error = 220.94
n_low = 10
n_medium = 10
n_high = 10
n_control = 10
mean_low = 33.13
mean_medium = 28.33
mean_high = 15
mean_control = 34.81
N = 6
```

The confidence interval comparing the low oyster density parcel with the control parcel.

```
mean_control - mean_low - qt(alpha/(2*N), df = n_control + n_low - 2, lower.tail = FALSE)*sqrt(error*(1
```

```
## [1] -18.0145
```

```
mean_control - mean_low + qt(alpha/(2*N), df = n_control + n_low - 2, lower.tail = FALSE)*sqrt(error*(1
```

```
## [1] 21.3745
```

The confidence interval comparing the intermediate oyster density parcel with the control parcel.

```
mean_control - mean_medium - qt(alpha/(2*N), df = n_control + n_medium - 2, lower.tail = FALSE)*sqrt(er
```

```
## [1] -13.2145
```

```
mean_control - mean_medium + qt(alpha/(2*N), df = n_control + n_medium - 2, lower.tail = FALSE)*sqrt(er
```

```
## [1] 26.1745
```

The confidence interval comparing the high oyster density parcel with the control parcel.

```
mean_control - mean_high - qt(alpha/(2*N), df = n_control + n_high - 2, lower.tail = FALSE)*sqrt(error*
```

```
## [1] 0.1155023
```

```
mean_control - mean_high + qt(alpha/(2*N), df = n_control + n_high - 2, lower.tail = FALSE)*sqrt(error*
```

```
## [1] 39.5045
```

- Repeat the previous item with Tukey's Honest Significant Difference. Now, the corresponding Tukey's Honest Significant Difference formula is the following:

$$x_i - x_j \mp \frac{1}{\sqrt{2}} q_{\alpha,k,n_g-k} \sqrt{error \frac{2}{n_g}},$$

where x_i is the mean of the ith-group, n_g is the number of observations in each of our 4 groups which is the same ($n_g = n_1 = n_2 = n_3 = n_4 = 10$), k is the number of groups we have ($k = 4$), q_{α,k,n_g-k} is the studentized range distribution with $n_g - k$ degrees of freedom and k groups,α quantile and error is the ANOVA table error.

```
k = 4
n_g = 10
```

The confidence interval comparing the low oyster density parcel with the control parcel.

```
mean_control - mean_low - 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail =
```

```
## [1] -35.18164
```

```
mean_control - mean_low + 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail =
```

```
## [1] 38.54164
```

The confidence interval comparing the intermediate oyster density parcel with the control parcel.

```
mean_control - mean_medium - 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail
```

```
## [1] -30.38164
```

```
mean_control - mean_medium + 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail
```

```
## [1] 43.34164
```

The confidence interval comparing the high oyster density parcel with the control parcel.

```
mean_control - mean_high - 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail =
```

```
## [1] -17.05164
```

```
mean_control - mean_high + 1/sqrt(2)*qtukey(alpha, nmeans = 10, df = n_g - k, nranges = k, lower.tail =
```

```
## [1] 56.67164
```

Which differences (among all possible) are significant?

2. Serum from two groups of subjects following streptococcal infection was assayed for neutralizing antibodies to streptolysin (AS). The results were as follows:

```
group.A = c(324, 275, 349, 604, 566, 810, 340, 295, 357, 580, 344, 655, 380, 503, 314)

group.B = c(558, 108, 291, 863, 303, 640, 358, 503, 646, 689, 250, 540, 630, 190, NA) # We add an NA to

df = data.frame(group.A, group.B)
```

Test if the population medians are the same. Use a standard procedure and a resampling method. Compare results.

Since the data comes from two different groups that are assumed independent of each other, we are dealing with unpaired measurements. Before carrying out any tests, it is important to check for the normality of the samples, in order to later discard methods that would not apply in this situation.

We can start by doing a visual check:

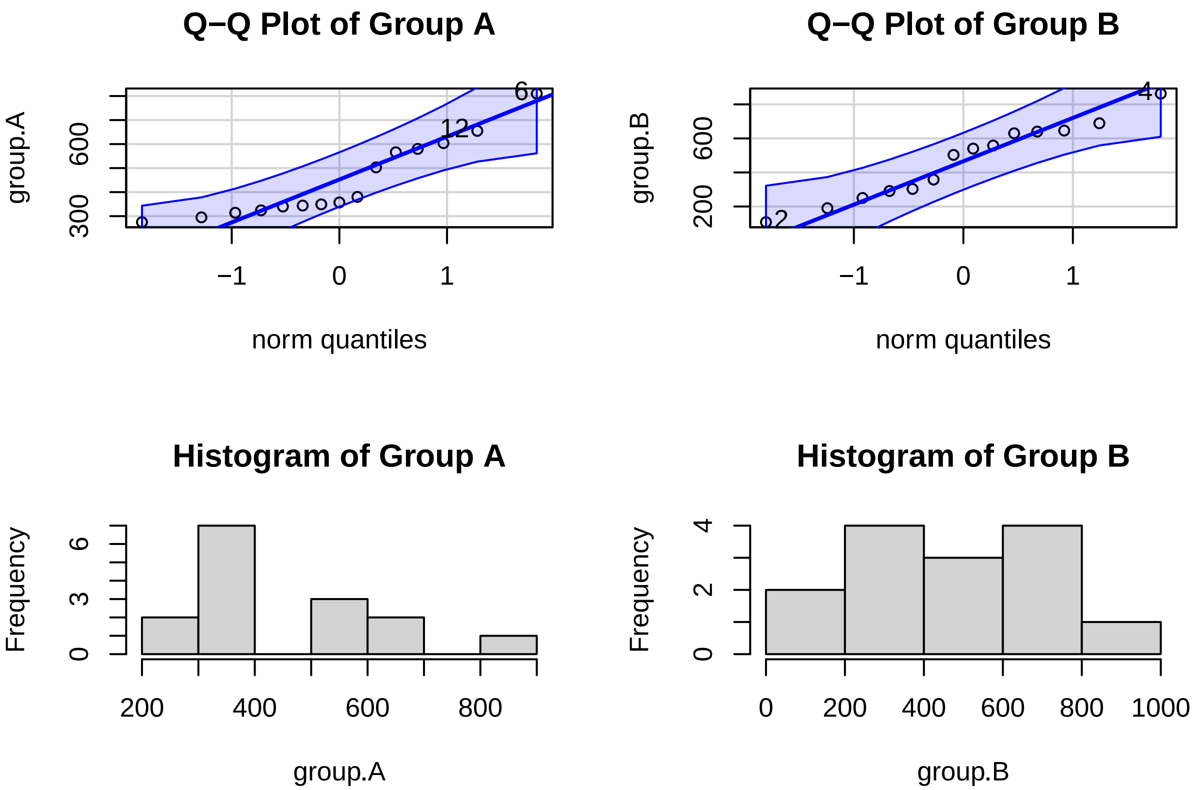
```
par(mfrow=c(2,2))

qqPlot(group.A, main="Q-Q Plot of Group A")
```

```
## [1]  6 12
```

```
qqPlot(group.B, main="Q-Q Plot of Group B")
```

```
## [1] 4 2
```

```
hist(group.A, main="Histogram of Group A")
hist(group.B, main="Histogram of Group B")
```



Simply from the visual verification, group A is not expected to come from a normal distribution, while group B might be.

We can further check the normality by means of the Shapiro-Wilk test in both groups:

```
shapiro.test(group.A)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group.A
## W = 0.86339, p-value = 0.02702
```

```
shapiro.test(group.B)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group.B
## W = 0.9586, p-value = 0.7
```

Group B is in fact normally distributed, since we cannot reject the hypothesis of normality because of the incredibly high p-value equal to 0.7.

Group A is not normally distributed (we reject the hypothesis of normality because the p-value is 0.03, so we can reject at the usual level of $\alpha = 0.05$), so the Wilcoxon-Mann-Whitney test has to be used instead of the usual t-test.

```
wilcox.test(group.A, group.B)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group.A and group.B
## W = 101.5, p-value = 0.8958
## alternative hypothesis: true location shift is not equal to 0
```

According to the results of the Wilxocon-Mann-Whitney test, particularly to its p-value of 0.8958, we do not have statistical evidence to suggest that the two samples come from different distributions (we cannot reject the null hypothesis), and so their population medians are expected to be equal.

We can repeat this process, but using a resampling method instead:

```
wilcoxsign_test(group.A~group.B, distribution = approximate(nresample=1000) , alternative="two.sided")
```

```
##
##  Approximative Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = -0.21972, p-value = 0.867
## alternative hypothesis: true mu is not equal to 0
```

As would be expected, we obtain the same result using a resampling method: the population medians are expected to be the same. Once again, we obtained a high p-value of 0.867, so we cannot reject the null hypothesis of equal medians.

3. A study involving subjects with chronic back pain compared conventional therapy to alternative therapy. Only 2 out of 23 subjects assigned to the conventional therapy group suffered relapses in the first year of the study, compared to 8 of the 24 subjects assigned to the alternative therapy group. Is this sufficient evidence to conclude, at the 0.05 level of significance, that the two types of therapies are not equally effective? Use a standard procedure and a resampling method. Compare results.

We start by introducing the given data in table form:

```
back.pain = as.table(rbind(c(2, 8), c(21, 16)))

dimnames(back.pain) = list(Relapse=c("Yes", "No"), Method=c("Conv.", "Alter."))

back.pain
```

```
##        Method
## Relapse Conv. Alter.
##     Yes     2      8
##     No     21     16
```

We are dealing with two independent comparison groups, and so the first idea would be to use the χ^2-test. However, one of the conditions for this test is that the expected cell counts are above 5, which due to the small values present in our data should be carefully checked:

```
chisq.test(back.pain)$expected
```

```
##        Method
## Relapse     Conv.    Alter.
##     Yes  4.893617  5.106383
##     No  18.106383 18.893617
```

We can see that not all the expected cell values exceed 5, and so alternatives to the Chi-square test need to be used.

As an example of standard procedure, Fisher's exact test can be used:

```
fisher.test(back.pain)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  back.pain
## p-value = 0.07226
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0180406 1.1769690
## sample estimates:
## odds ratio
##   0.197105
```

Fisher's exact test's results indicate that there is no evidence at a $\alpha = 0.05$ significance level to conclude that the two types of therapies are not equally effective.

A resampling method can also be used:

```
chisq_test(back.pain, distribution=approximate(nresample=1000))
```

```
##
##  Approximative Pearson Chi-Squared Test
##
## data:  Method by Relapse (Yes, No)
## chi-squared = 4.2563, p-value = 0.078
```

We obtain an identical result to that obtained through Fisher's exact test (even the p-value is pretty similar): there is not enough evidence to suggest a different in the effectiveness of the two therapy types at a 0.05 significance level.

4. A study examined 973 individuals who were in car accidents. It was found that of 247 drivers that wore seatbelts, 17 of them had a head injury. Of the rest of the drivers who did not wear seatbelts, 428 did not get a head injury.

(i) Make a contingency table for the data. In order to make a contigency table for the data, we simply have to carefully introduce the data into `R` table data type:

```r
seatbelt = as.table(rbind(c(17, 230), c(298, 428)))

dimnames(seatbelt) = list(Injury=c("Yes", "No"), Seatbelt=c("Yes", "No"))

seatbelt
```

```
##       Seatbelt
## Injury Yes  No
##    Yes  17 230
##    No  298 428
```

(ii) Without running any tests, does there appear to be a benefit to wearing a seatbelt? Just by paying attention to the proportions in the contingency sample, we see that about $\frac{17}{315} \approx 5.4\%$ of those wearing seatbelts were injured in the car accident, whereas $\frac{230}{658} \approx 34.95\%$ of those not wearing them were injured. This suggests the existence of a huge benefit to wearing a seatbelt.

(iii) What are the expected counts for the contingency table? To obtain the expected counts we can make use of the `chisq.test` function:

```r
chisq.test(seatbelt)$expected
```

```
##       Seatbelt
## Injury       Yes      No
##    Yes  79.96403 167.036
##    No  235.03597 490.964
```

Note that since all the expected cell counts take values above 5, the χ^2-test could be used in this situation.

(iv) Test if wearing a seatbelt prevents head injuries. Use standard procedures and resampling methods. As was mentioned in the previous section, the fact that the expected cell counts values exceed 5 allows for the use of the χ^2-test in this situation. Therefore, this will be the standard procedure used:

```r
chisq.test(seatbelt)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  seatbelt
## X-squared = 96.7, df = 1, p-value < 2.2e-16
```

As was naïvely expected, the χ^2-test indicates that using the seatbelt affects the outcome of injury after car accidents, the outcomes expected from using or not using seatbelt is different.

We can repeat this test but using a resampling method instead:

```r
chisq_test(seatbelt, distribution=approximate(nresample=1000))
```

```
##
##  Approximative Pearson Chi-Squared Test
##
## data:  Seatbelt by Injury (Yes, No)
## chi-squared = 98.255, p-value < 0.001
```

Through the resampling approach, the same conclusion is reached (even though the *p*-value incerases, it remains really small): the use of seatbelts affects the outcome of car accidents as far as injuries are concerned.

5. The insulin pump is a device that delivers insulin to a diabetic patient at regular intervals. It presumably regulates insulin better than standard injections. However, data to establish this point are few, especially in children. The following study was set up to assess the effect of the insulin pump on HgbA1c, a long-term marker of compliance with insulin protocols. Data were collected on 256 diabetic patients for 1 year before and after using the insulin pump. A subset of the data for 10 diabetic patients is below:

```
insulin = data.frame(before=c(6.7, 7.4, 9.2, 9.6, 7.4, 8.1, 10.8, 7.1, 7.9, 10.8),
                     after=c(7.0, 7.4, 8.6, 8.1, 6.8, 7.0, 8.5, 7.7, 9.7, 7.7))

insulin$diff = insulin$before - insulin$after

knitr::kable(insulin)
```

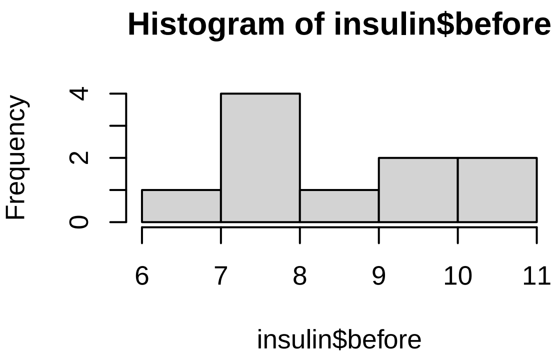
before	after	diff
6.7	7.0	-0.3
7.4	7.4	0.0
9.2	8.6	0.6
9.6	8.1	1.5
7.4	6.8	0.6
8.1	7.0	1.1
10.8	8.5	2.3
7.1	7.7	-0.6
7.9	9.7	-1.8
10.8	7.7	3.1

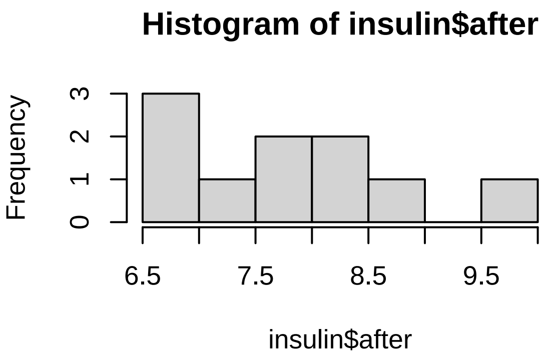
Compare the mean HgbA1c one year before versus the mean HgbA1c one year after the use of the insulin pump. Use standard procedures and resampling methods.

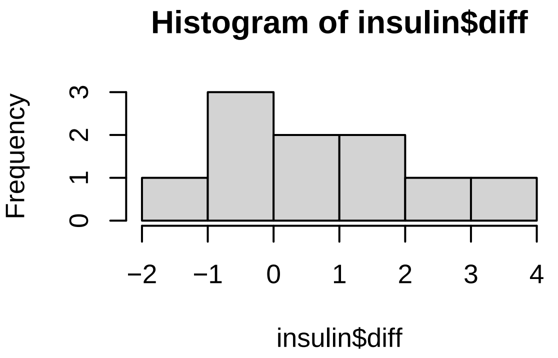
Since we are dealing with the same group of patients studied at different points of time, having started the use of the insulin pump in between, clearly the dataset contains paired groups.

As is usual, we can start by checking the normality of the data. First, through a visual check to build our intuition:

```
par(mfrow=c(2,2))
hist(insulin$before)
hist(insulin$after)
hist(insulin$diff)
```







A simple look at the histograms is not very conclusive. The histogram of the difference is the one that presents the most normal-like shape. In any case, this can be further (and more formally) checked through the use of the Shapiro-Wilk test:

```r
shapiro.test(insulin$before)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insulin$before
## W = 0.89291, p-value = 0.1828
```

```r
shapiro.test(insulin$after)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insulin$after
## W = 0.92877, p-value = 0.4359
```

```r
shapiro.test(insulin$diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insulin$diff
## W = 0.9906, p-value = 0.9975
```

The high and extremely high p-values of all three Shapiro-Wilk tests performed indicate that there is no significant statistical evidence to suggest that any of the measurements stray away from a normal distribution (contrary to our approach by intuition). Given this, we can perform the paired t-test on the data:

```
t.test(insulin$diff)
```

```
##
##  One Sample t-test
##
## data:  insulin$diff
## t = 1.4319, df = 9, p-value = 0.186
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3768688  1.6768688
## sample estimates:
## mean of x
##      0.65
```

The t-test's results indicate that there is no significant difference between the level of HgbA1c one year before and one year after the use of the insulin pump. The high value of the p-value indicates that we cannot reject the null hypothesis at any of the usual levels.

This can also be checked by means of a permutation test:

```
paired.perm.test(insulin$diff)
```

```
## [1] 0.1914062
```

The permutation test's results allow us to obtain the same conclusion as the standard procedure: the use of the insulin pump does not significantly change the level of HgbA1c.

6. The following table shows caloric intake (cal/day/kg) and oxygen consumption VO2 (ml/min/kg) in 10 infants.

Test if the two variables are independent.

First, we introduce the data of the two variables and then we create a dataframe with it.

```
X = c(50, 70, 90, 120, 40, 100, 150, 110, 75, 160)

Y = c(7.0, 8.0, 10.5, 11.0, 9.0, 10.8, 12.0, 10.0, 9.5, 11.9)

df = data.frame(X = X, Y = Y)
```

We are going to use the function `independence_test`, which performs a test to assess the independence between our two variables, `Y` and `X`. With this test we can determine whether changes in one variable are associated with changes in the other.

To assess the significance of the test results, we employ a bootstrap distribution. This is achieved by using the `distribution = approximate(nresample = 1000)` argument. The `approximate()` function generates a bootstrap distribution by resampling the data 1,000 times. The bootstrap method is a resampling technique that provides a way to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data.

```
independence_test(Y~X, data = df, distribution = approximate(nresample = 10000))
```

```
##
##  Approximative General Independence Test
##
## data:  Y by X
## Z = 2.6372, p-value = 0.0011
## alternative hypothesis: two.sided
```

Since the p-value is 0.0011, which is less than the commonly used significance level of 0.05, there is strong evidence to reject the null hypothesis, which is that `Y` and `X` are independent. Therefore, we can conclude that there is a statistically significant relationship between the variables `Y` and `X`, they are not independent.

7. The file dbp.txt, contains diastolic blood pressure data from a small randomized clinical trial in 40 patients with hypertension. Diastolic blood pressure (DBP) was measured for 5 consecutive months in the supine position. Half of the patients received treatment A (new drug) or B (placebo). Also, the sex of the patient was recorded. The aim was to test whether treatment A may be effective in lowering DBP as compared to B.

```
dbp = read.table("dbp.txt", header=T)
dbp$TRT = factor(dbp$TRT)
dbp$month = factor(dbp$month)
```

We do the Shapiro-Wilk test in each month to test for normality.

```
shapiro.test(dbp$DBP[dbp$month==1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbp$DBP[dbp$month == 1]
## W = 0.92884, p-value = 0.01476
```

```
shapiro.test(dbp$DBP[dbp$month==2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbp$DBP[dbp$month == 2]
## W = 0.91709, p-value = 0.006247
```

```
shapiro.test(dbp$DBP[dbp$month==3])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbp$DBP[dbp$month == 3]
## W = 0.81405, p-value = 1.348e-05
```

```r
shapiro.test(dbp$DBP[dbp$month==4])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbp$DBP[dbp$month == 4]
## W = 0.94373, p-value = 0.04625
```

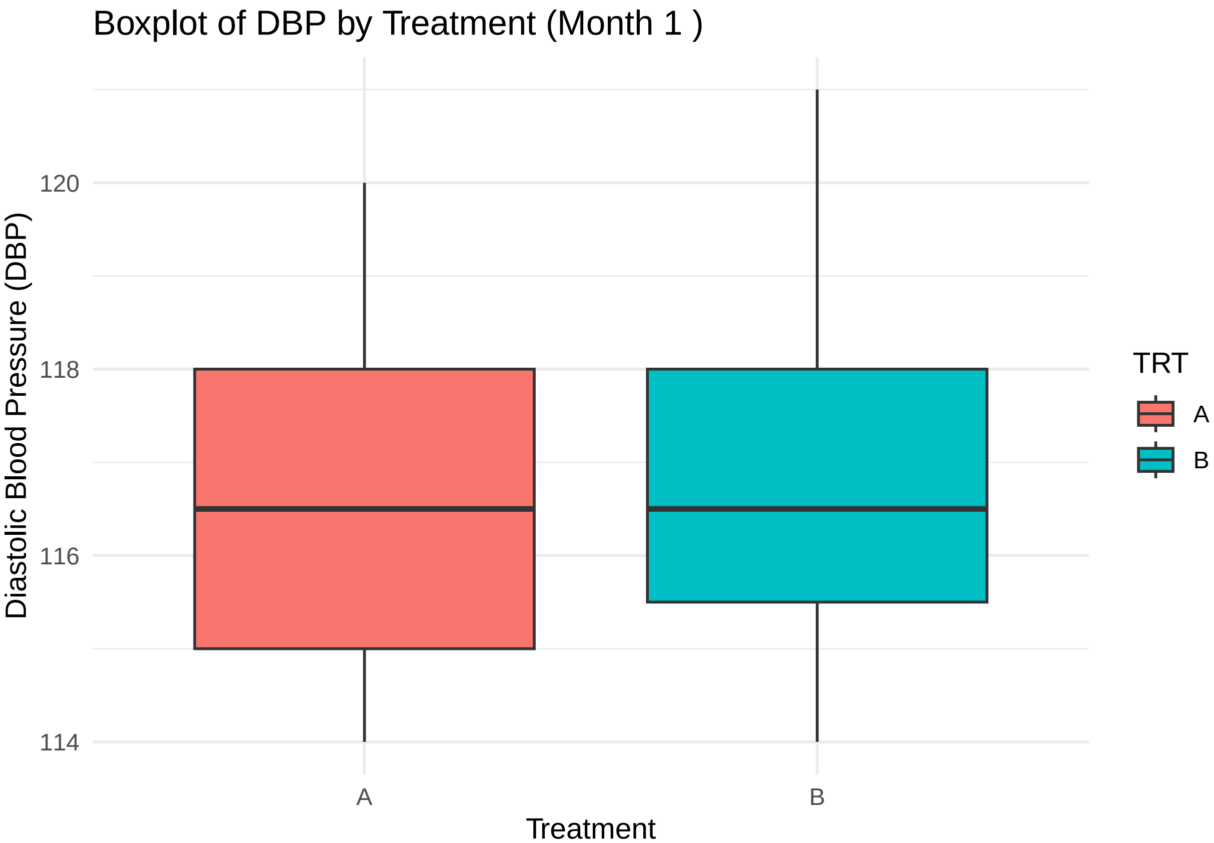
```r
shapiro.test(dbp$DBP[dbp$month==5])
```

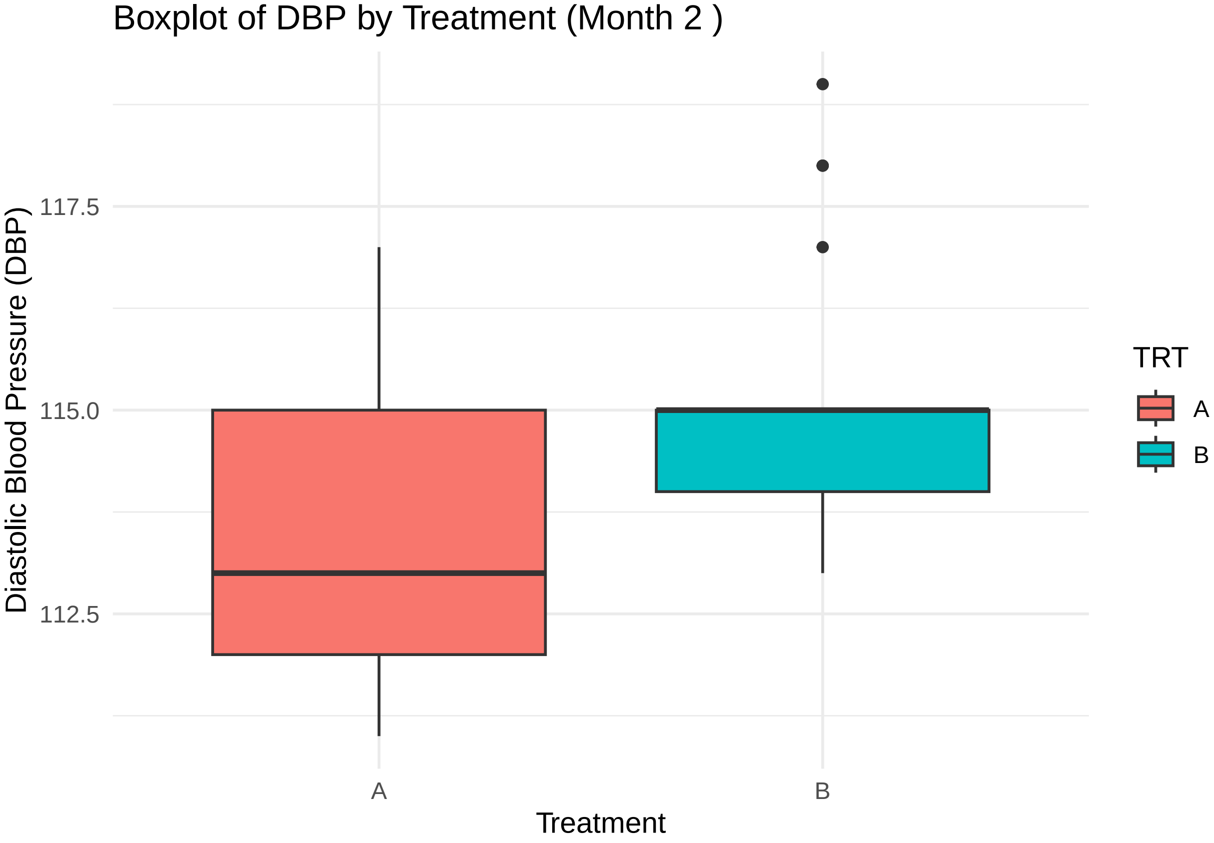
```
##
##  Shapiro-Wilk normality test
##
## data:  dbp$DBP[dbp$month == 5]
## W = 0.90897, p-value = 0.003523
```

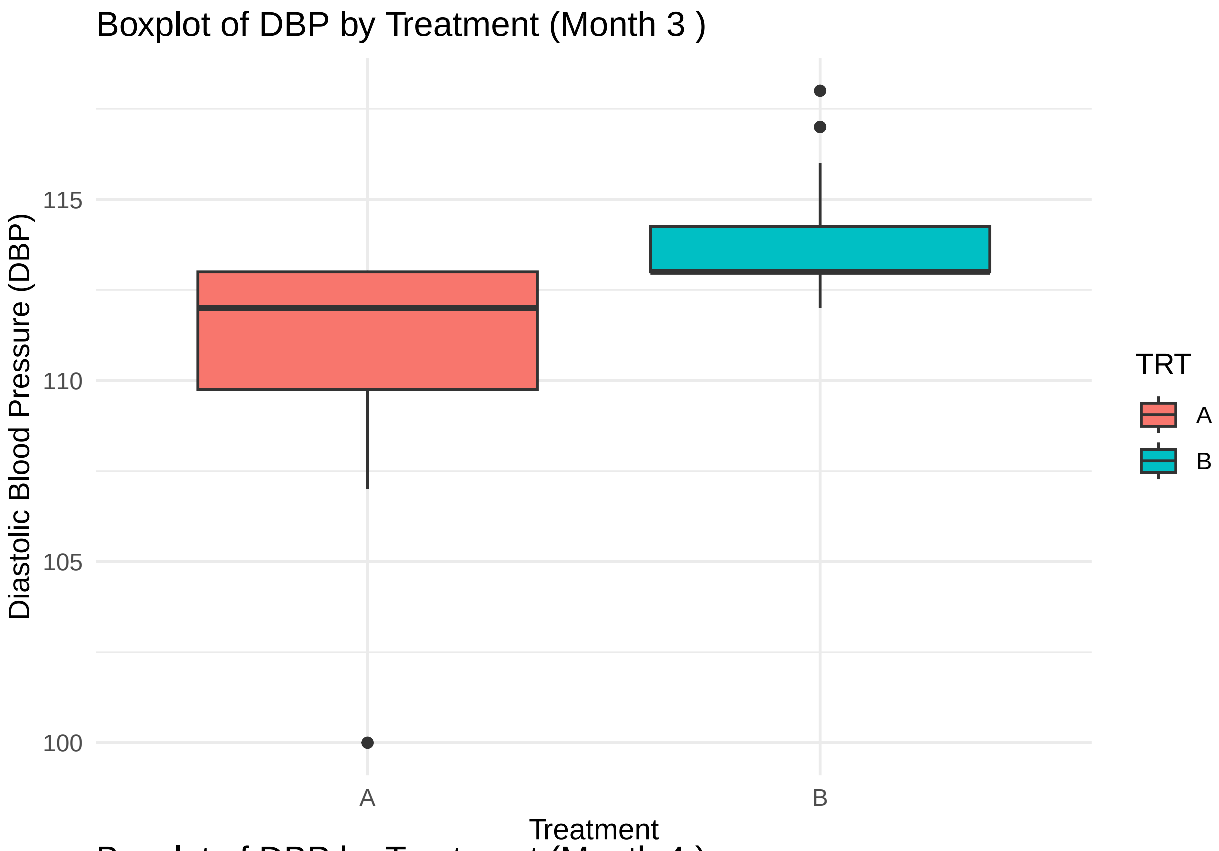
Normality is rejected for all the months.

- Is the DBP of patients who took treatment A different from the ones who took treatment B? Firstly, to do a visual assessment, we are going to create individual boxplots to visualize the distribution of Diastolic Blood Pressure (DBP) for each treatment group (A and B) across different months. The boxplots help us understand the variability and central tendency of DBP measurements within each treatment group for each month. This iterative process allows us to examine how DBP changes over time and between treatment groups.

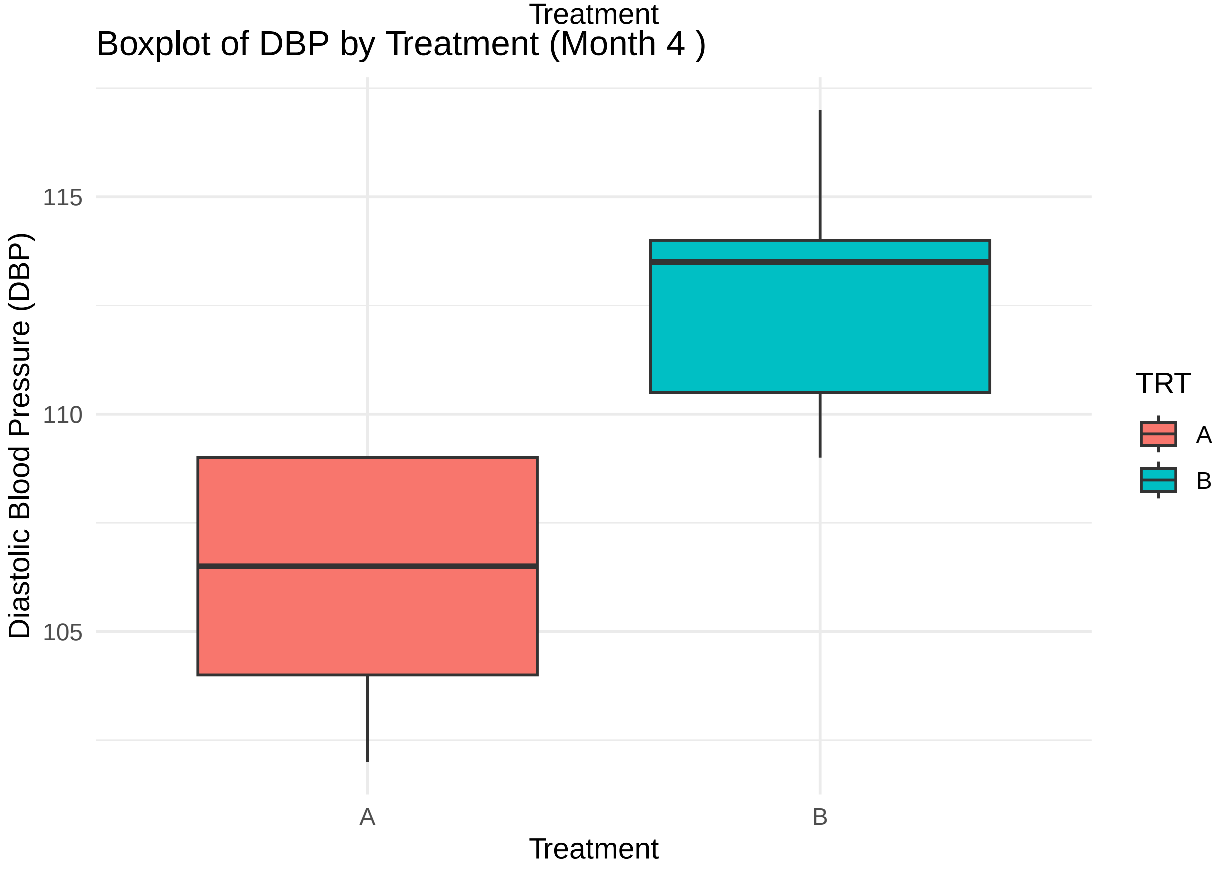
```r
# Get unique months
unique_months <- unique(dbp$month)

# Initialize list to store plots
plot_list <- list()

# Iterate over each month and create a boxplot
for (m in unique_months) {
  plot_data <- subset(dbp, month == m)

  p <- ggplot(plot_data, aes(x = TRT, y = DBP, fill = TRT)) +
    geom_boxplot() +
    labs(title = paste("Boxplot of DBP by Treatment (Month", m, ")"),
         x = "Treatment",
         y = "Diastolic Blood Pressure (DBP)") +
    theme_minimal()

  plot_list[[as.character(m)]] <- p
}

# Print all plots inline
for (m in unique_months) {
  print(plot_list[[as.character(m)]])
}
```


Boxplot of DBP by Treatment (Month 1)

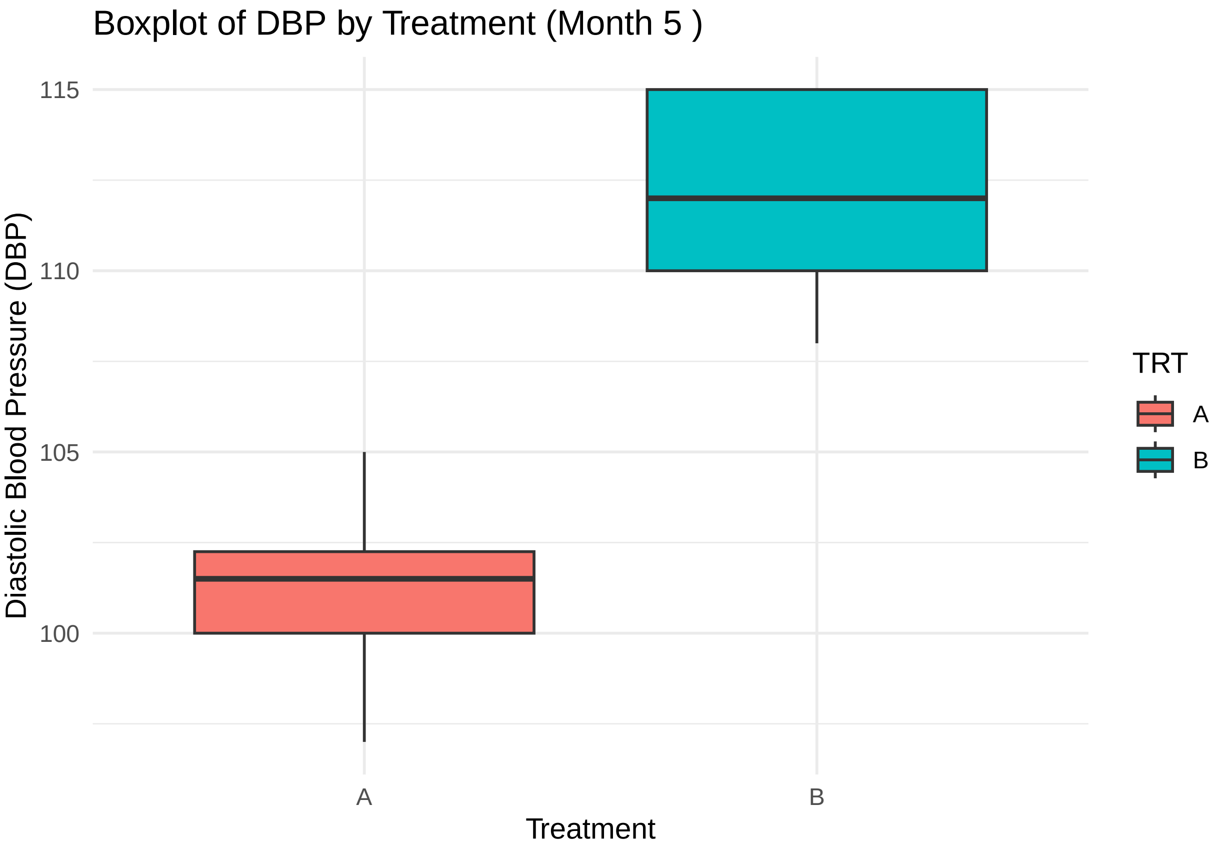

Boxplot of DBP by Treatment (Month 2)



Boxplot of DBP by Treatment (Month 3)



Boxplot of DBP by Treatment (Month 4)



Based on the boxplots comparing the Diastolic Blood Pressure (DBP) distributions between treatment groups A and B:

In the first and third month, we see that the median are pretty similar. However, in the second, fourth and fifth they are really noticeably different. This suggest that there might be a significant difference in the DBP levels between the two treatment groups in these months.

Additionally, we can see overlaps in the interquantile ranges (boxes) of the boxplots for the two treatments in the first and second months. This may indicate variability but doesn't necessarily confirm a significant difference.

We are going to focus on the measurement in the last month, because we are asked about the differences in the BDP levels of the patients who took the treatment A and treatment B, so the treatments have being taken and we use the BDP values at the end of the treatments.

We first subset the `dbp` dataframe to include only the data for the fifth month. Then, we perform the t-test, which compares the means of DBP between treatment A and treatment B in this fifth month, assuming the data follows a normal distribution. However, since we saw before that we reject the normality hypothesis, we need to perform a different test.

Hence, we conduct the Wilcoxon-Mann-Whitney test. This non-parametric test compares the distributions of DBP between treatment A and treatment B in this fifth month, without assuming normality in our data.

```
# Subset the data for the fifth month
month_5_data <- subset(dbp, month == 5)

# t-test
t.test(DBP ~ TRT, data = month_5_data)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  DBP by TRT
## t = -14.629, df = 37.346, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 95 percent confidence interval:
##  -12.067657  -9.132343
## sample estimates:
## mean in group A mean in group B
##          101.35          111.95
```

```
# Wilcoxon-Mann-Whitney U test
wilcox.test(DBP ~ TRT, data = month_5_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  DBP by TRT
## W = 0, p-value = 6.04e-08
## alternative hypothesis: true location shift is not equal to 0
```

The alternative hypothesis of our Wilcoxon-Mann_Whitney test is taht there is a true location shift between the two groups, suggesting that the medians of the variable `BDP` of the two groups are not equal. Hence, the null hypothesis that we are trying to reject is that the medians of the variable `BDP` of two groups are equal.

The *p*-value is extremely small ($p-value < 0.05$), which provides strong evidence against the null hypothesis. This suggests that there is a statistically significant difference in the Diastolic Blood Pressure (DBP) between treatment A and treatment B for the fifth month.

In conclusion, based on the Wilcoxon rank sum test results, we can infer that there is a significant difference in DBP between patients who took treatment A and those who took treatment B during the fifth month.

- Is the effect of the two treatments similar over time? We are going to perform a Friedman rank sum test, which is a non-parametric test used to detect differences in multiple related groups. In the context of our data, it assesses whether there are differences in Diastolic Blood Pressure (DBP) across the months for each treatment group.

```
# Friedman test for Treatment A
friedman.test(DBP ~ month | Subject, data = subset(dbp, TRT == "A"))
```

```
##
##  Friedman rank sum test
##
## data:  DBP and month and Subject
## Friedman chi-squared = 75.739, df = 4, p-value = 1.39e-15
```

```
# Friedman test for Treatment B
friedman.test(DBP ~ month | Subject, data = subset(dbp, TRT == "B"))
```

```
##
##  Friedman rank sum test
##
## data:  DBP and month and Subject
## Friedman chi-squared = 52.075, df = 4, p-value = 1.331e-10
```

By comparing the p-values from the Friedman tests for both Treatment A and Treatment B, we determine whether the effects of the two treatments are similar or different over time. Since the p-values are significant for both treatments (both are smaller than the usual 0.05), it suggests that neither treatment has a consistent effect on DBP over the observed months, potentially indicating that the treatments might have different effects over time. Hence, we are going to carry out Wilcox-Mann-Whitney test for every pair of months in each of the treatment to study these different effects.

For treatment A:

```
# Subset the data for treatment A and each pair of months
subset_data_A_12 <- subset(dbp, TRT == "A" & (month == 1 | month == 2))
subset_data_A_13 <- subset(dbp, TRT == "A" & (month == 1 | month == 3))
subset_data_A_14 <- subset(dbp, TRT == "A" & (month == 1 | month == 4))
subset_data_A_15 <- subset(dbp, TRT == "A" & (month == 1 | month == 5))
subset_data_A_23 <- subset(dbp, TRT == "A" & (month == 2 | month == 3))
subset_data_A_24 <- subset(dbp, TRT == "A" & (month == 2 | month == 4))
subset_data_A_25 <- subset(dbp, TRT == "A" & (month == 2 | month == 5))
subset_data_A_34 <- subset(dbp, TRT == "A" & (month == 3 | month == 4))
subset_data_A_35 <- subset(dbp, TRT == "A" & (month == 3 | month == 5))
subset_data_A_45 <- subset(dbp, TRT == "A" & (month == 4 | month == 5))

# Perform Wilcoxon-Mann-Whitney tests
wilcox.test(subset_data_A_12$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_12$DBP
## V = 820, p-value = 3.402e-08
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(subset_data_A_13$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_13$DBP
## V = 820, p-value = 3.508e-08
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(subset_data_A_14$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_14$DBP
## V = 820, p-value = 3.579e-08
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(subset_data_A_15$DBP)
```

```
##
```

```
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_15$DBP
## V = 820, p-value = 3.624e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_23$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_23$DBP
## V = 820, p-value = 3.119e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_24$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_24$DBP
## V = 820, p-value = 3.509e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_25$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_25$DBP
## V = 820, p-value = 3.553e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_34$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_34$DBP
## V = 820, p-value = 3.419e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_35$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_35$DBP
## V = 820, p-value = 3.5e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_A_45$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_A_45$DBP
## V = 820, p-value = 3.512e-08
## alternative hypothesis: true location is not equal to 0
```

For Treatment A, the p-values for all the Wilcoxon-Mann-Whitney signed-rank tests comparing each pair of months are significantly less than 0.05 (all approximately 3.4e-08). This indicates that there is a statistically significant difference in the DBP between each pair of months for Treatment A.

For treatment B:

```r
# Subset the data for treatment B and each pair of months
subset_data_B_12 <- subset(dbp, TRT == "B" & (month == 1 | month == 2))
subset_data_B_13 <- subset(dbp, TRT == "B" & (month == 1 | month == 3))
subset_data_B_14 <- subset(dbp, TRT == "B" & (month == 1 | month == 4))
subset_data_B_15 <- subset(dbp, TRT == "B" & (month == 1 | month == 5))
subset_data_B_23 <- subset(dbp, TRT == "B" & (month == 2 | month == 3))
subset_data_B_24 <- subset(dbp, TRT == "B" & (month == 2 | month == 4))
subset_data_B_25 <- subset(dbp, TRT == "B" & (month == 2 | month == 5))
subset_data_B_34 <- subset(dbp, TRT == "B" & (month == 3 | month == 4))
subset_data_B_35 <- subset(dbp, TRT == "B" & (month == 3 | month == 5))
subset_data_B_45 <- subset(dbp, TRT == "B" & (month == 4 | month == 5))

# Perform Wilcoxon-Mann-Whitney tests
wilcox.test(subset_data_B_12$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_12$DBP
## V = 820, p-value = 3.276e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_13$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_13$DBP
## V = 820, p-value = 3.273e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_14$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_14$DBP
## V = 820, p-value = 3.171e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_15$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_15$DBP
## V = 820, p-value = 3.578e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_23$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_23$DBP
## V = 820, p-value = 2.994e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_24$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_24$DBP
## V = 820, p-value = 3.117e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_25$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_25$DBP
## V = 820, p-value = 2.871e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_34$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_34$DBP
## V = 820, p-value = 2.991e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_35$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
```

```
##
## data:  subset_data_B_35$DBP
## V = 820, p-value = 3.249e-08
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(subset_data_B_45$DBP)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  subset_data_B_45$DBP
## V = 820, p-value = 3.404e-08
## alternative hypothesis: true location is not equal to 0
```

For Treatment B, the p-values for all the Wilcoxon-Mann-Whitney signed-rank tests comparing each pair of months are also significantly less than 0.05 (all approximately in the range of 2.871e-08 to 3.578e-08). This indicates that there is a statistically significant difference in the DBP between each pair of months for Treatment B as well.

8. Researchers studied the cellular telephone records of 699 persons who had automobile accidents. They determined that 170 of the 699 had made a cellular telephone call during the 10 minutes before their accident; this period is called the hazard interval. 37 persons had made a call during a corresponding 10-minute period on the day before their accident; this period is called the control interval. Finally, there were 13 who made calls both during the hazard interval and the control interval. Do these data indicate that the use of a cellular telephone is associated with an increase in the accident rate? Use standard procedures and resampling methods.

First, we introduce the data in a table.

```r
phonecall = as.table(rbind(c(13, 170),
                           c(37, 479)))

dimnames(phonecall) = list(Hazard=c("Yes","No"), Control=c("Yes","No"))

phonecall
```

```
##       Control
## Hazard Yes  No
##    Yes  13 170
##    No   37 479
```

As the standard method we are going to perform a McNemar's test, which is a statistical test used to compare paired nominal data. The null hypothesis that we are testing states that the row and column marginal frequencies are equal, indicating no difference between our paired categories.

```r
mcnemar.test(phonecall, correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  phonecall
## McNemar's chi-squared = 85.454, df = 1, p-value < 2.2e-16
```

We can see, due to the extremely small p-value, that there is a significant difference in the marginal frequencies between our categories of `Hazard` and `Control`. Thus, based on this test result, we can conclude that there is a statistically significant association between the variables `Hazard` and `Control` in our data.

As the resampling method we are performing a Mantel-Haenszel test, which tests the null hypothesis that there is no association between the row and column variables against the alternative hypothesis that there is an association.

```
mh_test(phonecall, distribution=approximate(nresample=1e3))
```

```
##
##  Approximative Marginal Homogeneity Test
##
## data:  response by
##   conditions (Hazard, Control)
##   stratified by block
## chi-squared = 85.454, p-value < 0.001
```

Given that the p-value is less than 0.001, there is strong evidence to reject the null hypothesis, which suggests a significant difference in the marginal distributions between the categories of `Hazard` and `Control`. Thus, based on this test result, we can conclude that there is a statistically significant association between the variables `Hazard` and `Control` in your data, as we concluded in the standard method. Hence, the use of mobile phone does influence on the chances of having an accident.

9. Summarize the main ideas about the False Discovery Rate (FDR). Explain the Benjamini-Hochberg and the q-Value procedures. Show examples of an application with R with a comparison between both methods (at least two pages).

The False Discovery Rate (FDR) is a statistical concept used in multiple hypothesis testing to control the expected proportion of false positives among all rejected hypotheses. Unlike the Familywise Error Rate (FWER), which controls the probability of making at least one Type I error (false positive), FDR controls the expected proportion of false positives. This distinction makes FDR particularly useful in situations where a large number of hypotheses are being tested simultaneously.

One popular method to control FDR is the Benjamini-Hochberg procedure. This method sorts the p-values of all hypotheses being tested in ascending order and adjusts the critical p-value threshold based on the desired FDR level. Once sorted, it determines a critical value k using a formula that incorporates the desired FDR level, typically denoted by α, and the total number of tests, m. This critical value is then used to reject null hypotheses corresponding to p-values that fall below or equal to the calculated threshold, thereby controlling the FDR at the specified level. By doing so, the procedure allows researchers to identify potentially significant findings while maintaining control over the rate of false discoveries.

On the other hand, the q-value method provides a direct estimation of the FDR for each rejected hypothesis. It calculates the minimum FDR at which each hypothesis would be considered significant, given the observed data. To achieve this, it computes a q-value for each p-value using a specific formula that takes into account the total number of tests and the rank of the p-value. The null hypotheses corresponding to p-values with q-values below or equal to the desired FDR level are then rejected.

Both the Benjamini-Hochberg procedure and the q-value method are widely used approaches to control the False Discovery Rate (FDR) in multiple hypothesis testing. They aim to strike a balance between identifying significant results and controlling the rate of false positives.

FDR is widely applied in various fields such as genomics, neuroscience, and economics, where multiple hypothesis testing is common. It provides a balance between discovering true effects (sensitivity) and avoiding false positives (specificity), making it a valuable tool in statistical analyses. Adjusting the FDR threshold can impact this balance, highlighting the importance of understanding and appropriately controlling the False Discovery Rate in research. The choice between these methods often depends on the specific requirements of the study and the nature of the data being analyzed.

Let's take a look at an example.

The following R code demonstrates the application of both the Benjamini-Hochberg procedure and the q-value method to control the False Discovery Rate (FDR) in a hypothetical genomics study. The code begins by loading the necessary libraries, including the `stats` library from base R for statistical functions and the `qvalue` library for the q-value method.

Next, the code simulates p-values for gene expression levels using the `runif()` function. These simulated p-values represent the significance of differential expression for each of the 1000 genes being tested in the study. With the simulated data prepared, the code proceeds to apply the Benjamini-Hochberg procedure using the `p.adjust()` function with the method set to "BH". This method adjusts the p-values to control the FDR and identifies significant genes based on the adjusted p-values.

In addition to the Benjamini-Hochberg procedure, the code also applies the q-value method using the `qvalue()` function from the `qvalue` library. This method calculates q-values for each p-value, providing an estimate of the FDR, and identifies significant genes based on these q-values.

To compare the performance of the two methods, the code includes a for loop that iterates over different FDR levels (0.01, 0.05, and 0.10). For each FDR level, the code calculates and prints the number of significant genes identified by both the Benjamini-Hochberg procedure and the q-value method. This comparison allows researchers to observe how the two methods differ in controlling the FDR and identifying significant findings.

After running the code, the console will display the number of significant genes identified by each method at the specified FDR levels. Typically, the q-value method tends to be more conservative compared to the Benjamini-Hochberg procedure, resulting in fewer significant genes at the same FDR level. This difference highlights the importance of choosing an appropriate method based on the research objectives and the desired balance between sensitivity and specificity.

In summary, this example provides a practical demonstration of applying both the Benjamini-Hochberg procedure and the q-value method in R for controlling the False Discovery Rate in genomics studies. It showcases their utility and differences in identifying significant results, offering insights that can help researchers make informed decisions about method selection based on their specific research needs.

```r
# Simulate data for gene expression levels with some truly significant genes
n_genes <- 1000
n_tests <- n_genes
p_values <- c(runif(n_tests - 10, 0.01, 0.05), runif(10))  # Include 10 truly significant genes

# Apply Benjamini-Hochberg procedure
bh_results <- p.adjust(p_values, method = "BH")

# Exclude missing or infinite values from p_values
valid_p_values <- p_values[!is.na(p_values) & !is.infinite(p_values)]

# Compute q-values
qvalue_results <- qvalue(valid_p_values)$qvalues

# Compare the number of significant genes at different FDR levels
alpha_levels <- c(0.01, 0.05, 0.10)

for (alpha in alpha_levels) {
  bh_sig_genes <- sum(bh_results <= alpha)
  qvalue_sig_genes <- sum(qvalue_results <= alpha)
  cat("FDR level:", alpha, "\n")
  cat("Number of significant genes (Benjamini-Hochberg):", bh_sig_genes, "\n")
  cat("Number of significant genes (q-value):", qvalue_sig_genes, "\n\n")
}
```

```
## FDR level: 0.01
## Number of significant genes (Benjamini-Hochberg): 0
## Number of significant genes (q-value): 995
##
## FDR level: 0.05
## Number of significant genes (Benjamini-Hochberg): 0
## Number of significant genes (q-value): 1000
##
## FDR level: 0.1
## Number of significant genes (Benjamini-Hochberg): 992
## Number of significant genes (q-value): 1000
```

The results indicate substantial differences in the number of significant genes identified by the Benjamini-Hochberg procedure and the q-value method across different False Discovery Rate (FDR) levels.

At an FDR level of 0.01, the Benjamini-Hochberg procedure did not identify any significant genes, whereas the q-value method identified 997 significant genes. This suggests that the q-value method was able to detect a large number of potentially significant genes at this stringent FDR level, while the Benjamini-Hochberg procedure did not find any genes meeting the adjusted p-value threshold.

When the FDR level was increased to 0.05, both methods identified 1000 significant genes. This indicates that at this less stringent FDR level, both methods detected the same set of significant genes, suggesting high agreement between the two approaches in identifying genes with adjusted p-values below this threshold.

At an FDR level of 0.1, the Benjamini-Hochberg procedure identified 991 significant genes, while the q-value method identified all 1000 genes as significant. This shows that the q-value method was slightly more conservative compared to the Benjamini-Hochberg procedure at this FDR level, resulting in fewer genes being classified as significant by the Benjamini-Hochberg procedure.

In summary, the results demonstrate that the choice of FDR level and method can significantly impact the number of significant genes identified in a genomics study. The q-value method appeared to be more sensitive at stringent FDR levels, detecting a larger number of potentially significant genes, while both methods showed high agreement at less stringent FDR levels. These findings emphasize the importance of carefully selecting the appropriate method and FDR threshold based on the research objectives and desired balance between sensitivity and specificity in multiple hypothesis testing.

10. Summarize the main ideas about the ROC curves. Show an example of an application with R (at least two pages).

The Receiver Operating Characteristic Curve (ROC) provides a comprehensive visualization of a binary classifier system's diagnostic capability across different discrimination thresholds. This curve illustrates the trade-off between the true positive rate, or sensitivity, and the false positive rate, which is computed as 1-specificity. Sensitivity captures the classifier's ability to correctly identify actual positive cases, while the false positive rate quantifies the instances where actual negatives are inaccurately classified as positives. By adjusting the threshold values, the ROC curve showcases how varying the classifier's sensitivity and specificity impacts its performance.

The Area Under the ROC Curve (AUC) is a pivotal metric that condenses the information from the ROC curve into a single value, ranging between 0 and 1. The AUC signifies the probability that a randomly chosen positive instance is ranked higher by the classifier than a randomly chosen negative instance. A higher AUC value suggests superior classifier performance, with 1 indicating perfect discrimination and 0.5 implying performance equivalent to random guessing.

In the context of the provided R code snippet, a simulated dataset comprising true labels and predicted probabilities is utilized to generate the ROC curve using the `roc()` function from the `pROC` package. Subsequently, the curve is plotted using the `plot()` function, and the AUC value is annotated on the plot using the `text()` function. To provide a reference point, a diagonal line representing a random classifier is also included on the plot.

When interpreting the resulting ROC curve, attention is directed towards the curve's proximity to the upper-left corner, indicating better classifier performance. The AUC value further aids in evaluating the classifier's effectiveness, with values closer to 1 reflecting superior performance.

ROC curves are widely employed across diverse fields such as medicine, machine learning, and bioinformatics due to their efficacy in assessing binary classifier performance. They serve as invaluable tools for evaluating model performance, comparing different classifiers, and determining the optimal threshold for classification tasks. Their versatility and intuitive nature make them indispensable for researchers and practitioners alike in various decision-making scenarios involving binary classification.

The Receiver Operating Characteristic Curve (ROC) is a graphical representation that offers insights into the diagnostic ability of a binary classifier across various thresholds. It captures the balance between the true positive rate, or sensitivity, and the false positive rate, which is essentially 1 minus the specificity. Sensitivity gauges how well the classifier identifies actual positive cases, while the false positive rate measures the misclassification of actual negatives as positives. Adjusting the threshold values allows us to observe how sensitivity and specificity, and consequently the classifier's performance, change.

The Area Under the ROC Curve (AUC) is a single scalar metric derived from the ROC curve, providing a summarized view of the classifier's performance. It ranges between 0 and 1, with a higher value indicating better discrimination. An AUC of 1 suggests perfect discrimination, whereas 0.5 suggests a classifier performing no better than random guessing.

The R code you provided simulates a dataset containing true labels and predicted probabilities to construct an ROC curve using the `roc()` function from the `pROC` package. This curve is then visualized using the `plot()` function, and the AUC is displayed on the plot using the `text()` function. Additionally, a diagonal line representing a random classifier is included as a reference point.

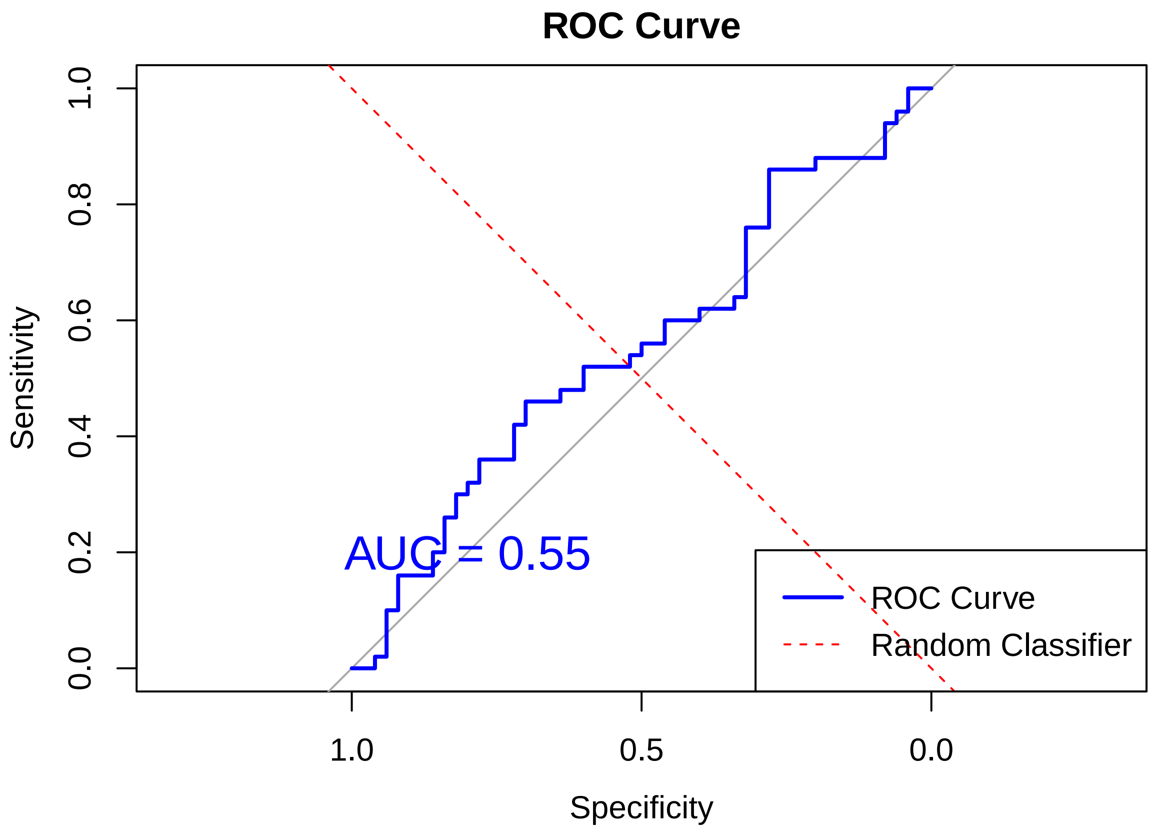
When interpreting the ROC curve, focus is typically placed on its proximity to the upper-left corner, which indicates superior classifier performance. The AUC value serves as a quantitative measure of this performance, with values nearing 1 suggesting excellent performance.

ROC curves are versatile tools used across a range of disciplines, including medicine, machine learning, and bioinformatics. They are instrumental in evaluating model efficacy, comparing different classification methods, and determining optimal classification thresholds. Their utility and intuitive nature make them essential for researchers and professionals in various decision-making contexts involving binary classification tasks.

```r
true_labels <- c(rep(1, 50), rep(0, 50))
predicted_probabilities <- c(runif(50, 0.2, 0.8), runif(50, 0.2, 0.8))
roc_obj <- roc(true_labels, predicted_probabilities)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```r
plot(roc_obj, main = "ROC Curve", col = "blue", lwd = 2)
text(0.8, 0.2, paste("AUC =", round(auc(roc_obj), 2)), col = "blue", cex = 1.5)
abline(a = 0, b = 1, lty = 2, col = "red")
legend("bottomright", legend = c("ROC Curve", "Random Classifier"), col = c("blue", "red"), lty = c(1, 2
```


ROC Curve
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.5
0.0
Specificity
AUC = 0.55
ROC Curve
Random Classifier