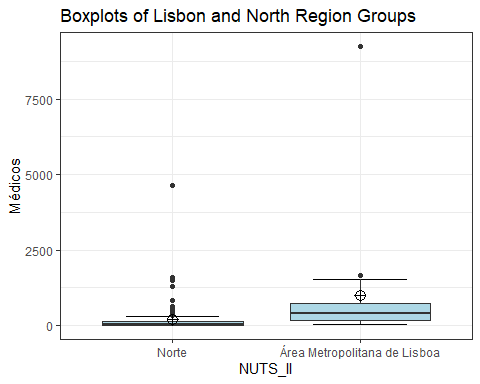
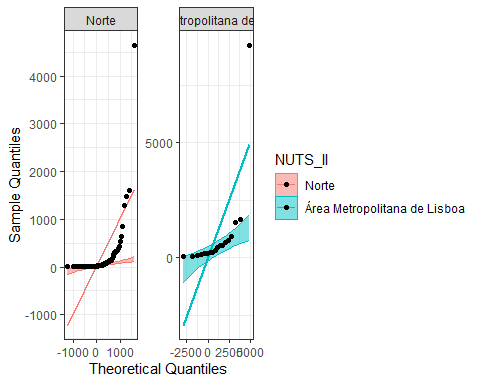
Homework 3 - Report

2023-02-10

### We intend to compare the North and Lisbon regions regarding the number of physicians. Please indicate, justifying, what inferential statistical methodologies would be appropriate for this purpose? Name the hypothesis tests used, comment on and interpret the results.



## # A tibble: 2 × 11  
## NUTS\_II n mean sd stderr LCL UCL median min max IQR  
## <fct> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Norte 86 198. 567. 61.1 76.7 320. 36 3 4650 123   
## 2 Área Metropolit… 18 988 2113. 498. -62.6 2039. 391 52 9240 552.

 First, we need to test for assumptions: Normal distribuition in each group, and homogeneity of variances

Starting with Shapiro-Wilk to test for normality:

## # A tibble: 2 × 3  
## NUTS\_II `W Statistic` `p-value`  
## <fct> <dbl> <dbl>  
## 1 Norte 0.341 8.15e-18  
## 2 Área Metropolitana de Lisboa 0.433 2.20e- 7

Both p-values are <0.05, meaning we reject the null hypothesis, so this data is not normally distributed We then need to do a non-parametric test: Mann-Whitney U

Before that, and for the sake of practice, lets test for homogeneity of variances

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 6.42 0.01281 \*  
## 102   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

we get a p-value of 0.01281, meaning we reject the null hypothesis, so between group variances are unequal

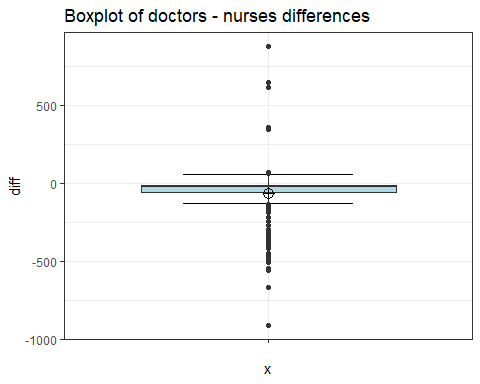
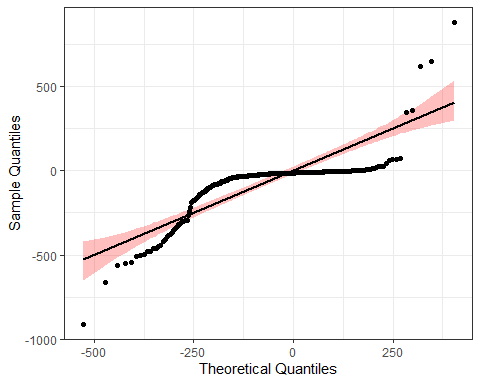
##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: Médicos by NUTS\_II  
## W = 233, p-value = 3.405e-06  
## alternative hypothesis: true location shift is not equal to 0  
## 95 percent confidence interval:  
## -507.0001 -131.9999  
## sample estimates:  
## difference in location   
## -229.0001

We get a p-value of 3.405e-06, meaning we reject the null hypothesis that distributions are equal and conclude that there is a significant difference in the number of doctors of both regions.

### A researcher argues that there are statistically significant differences between the average values per county for the number of physicians and the number of nurses in Portugal. Do you agree with the researcher’s statement? Identify the appropriate statistical methodologies to apply, comment and interpret the results.

## n mean sd stderr LCL UCL median min max IQR  
## 1 308 163.1136 667.677 38.04443 88.2528 237.9745 24.5 0 9240 79

## n mean sd stderr LCL UCL median min max IQR  
## 1 308 225.6039 694.2784 39.56018 147.7605 303.4473 48.5 0 9507 121



##   
## Shapiro-Wilk normality test  
##   
## data: BD$diff  
## W = 0.64009, p-value < 2.2e-16

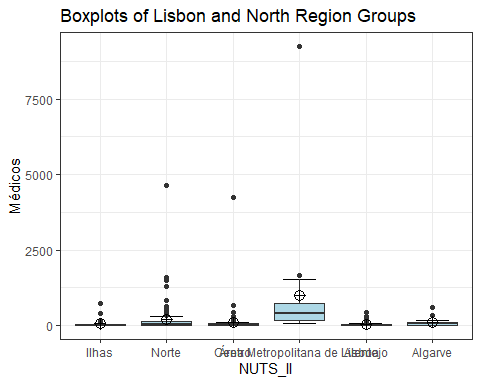
With a p-value of <2.2e-16, we reject the null hypothesis and conclude that the difference between groups is not normally distributed.

Since we rejected normality, we must perform a non parametric test: Wilcoxon test

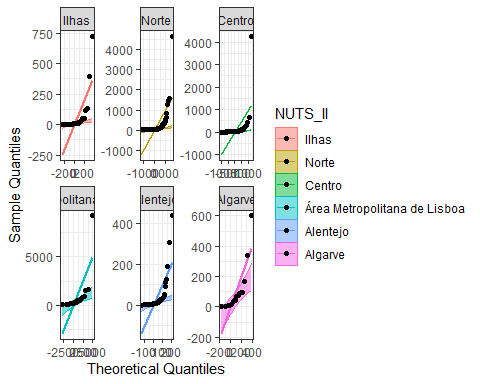
##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: BD$Médicos and BD$Enfermeiros  
## V = 3513.5, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0  
## 95 percent confidence interval:  
## -36.49999 -23.00001  
## sample estimates:  
## (pseudo)median   
## -28.0001

With a p-value of <2.2e-16, we reject the null hypothesis and conclude that there is a statistically significant difference between the number of physicians and the number of nurses.

### You want to compare all regions regarding the number of physicians. Please indicate, justifying, which inferential statistical methodologies would be appropriate for this purpose? Name the hypothesis tests used, comment and interpret the results.



## # A tibble: 6 × 11  
## NUTS\_II n mean sd stderr LCL UCL median min max IQR  
## <fct> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Ilhas 30 57.9 145. 26.5 3.65 112. 9 0 719 35.8  
## 2 Norte 86 198. 567. 61.1 76.7 320. 36 3 4650 123   
## 3 Centro 100 99.8 430. 43.0 14.4 185. 21.5 1 4239 46.8  
## 4 Área Metropo… 18 988 2113. 498. -62.6 2039. 391 52 9240 552.   
## 5 Alentejo 58 34.7 73.1 9.60 15.5 53.9 11 1 436 20   
## 6 Algarve 16 105. 156. 39.0 22.3 188. 56 4 600 82.5

 Since we want to compare multiple groups, we want to perform an ANOVA test or Kruskall-Wallis First, we need to test for assumptions: # Normal distribuition in each group; # Homogeneity of variances

Starting with Shapiro-Wilk to test for normality

## # A tibble: 6 × 3  
## NUTS\_II `W Statistic` `p-value`  
## <fct> <dbl> <dbl>  
## 1 Ilhas 0.419 8.24e-10  
## 2 Norte 0.341 8.15e-18  
## 3 Centro 0.181 4.42e-21  
## 4 Área Metropolitana de Lisboa 0.433 2.20e- 7  
## 5 Alentejo 0.443 1.70e-13  
## 6 Algarve 0.648 4.75e- 5

All of them present a p-value <0.05,we reject the null hypothesis, and conclude that this data is not normally distributed. We then need to do a non-parametric test: Kruskal-Wallis, but before that, and for the sake of practice, lets test for homogeneity of variances

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 5 5.1171 0.0001606 \*\*\*  
## 302   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We get a p-value of 0.0001606, meaning we reject the null hypothesis, so between group variances are unequal.

Let’s finally move onto the test we want:

##   
## Kruskal-Wallis rank sum test  
##   
## data: Médicos by NUTS\_II  
## Kruskal-Wallis chi-squared = 60.044, df = 5, p-value = 1.19e-11

With a p-value of 1.19e-11,we reject the null hypothesis, and conclude that there are differences in median values between groups

We must then move onto post-hoc analysis:

## Comparison Z P.unadj P.adj  
## 1 Alentejo - Algarve -2.7553672 5.862630e-03 4.690104e-02  
## 2 Alentejo - Área Metropolitana de Lisboa -6.7366767 1.620502e-11 2.430753e-10  
## 3 Algarve - Área Metropolitana de Lisboa -3.0255107 2.482136e-03 2.233923e-02  
## 4 Alentejo - Centro -2.5284374 1.145715e-02 6.874291e-02  
## 5 Algarve - Centro 1.3398184 1.803044e-01 5.409131e-01  
## 6 Área Metropolitana de Lisboa - Centro 5.4691014 4.523229e-08 5.880198e-07  
## 7 Alentejo - Ilhas 0.1618548 8.714202e-01 8.714202e-01  
## 8 Algarve - Ilhas 2.6309885 8.513692e-03 5.959584e-02  
## 9 Área Metropolitana de Lisboa - Ilhas 6.2185578 5.017452e-10 7.024433e-09  
## 10 Centro - Ilhas 2.1795854 2.928821e-02 1.464410e-01  
## 11 Alentejo - Norte -4.1954090 2.723796e-05 2.996176e-04  
## 12 Algarve - Norte 0.2395993 8.106409e-01 1.000000e+00  
## 13 Área Metropolitana de Lisboa - Norte 4.2622906 2.023421e-05 2.428105e-04  
## 14 Centro - Norte -2.0094802 4.448624e-02 1.779449e-01  
## 15 Ilhas - Norte -3.5334718 4.101399e-04 4.101399e-03

## Ilhas Norte Centro   
## Norte 0.00724 - -   
## Centro 0.18305 0.27392 -   
## Área Metropolitana de Lisboa 4.2e-06 4.9e-05 1.3e-07  
## Alentejo 0.99671 0.00032 0.05966  
## Algarve 0.12605 0.99993 0.68129  
## Área Metropolitana de Lisboa Alentejo  
## Norte - -   
## Centro - -   
## Área Metropolitana de Lisboa - -   
## Alentejo 3.1e-08 -   
## Algarve 0.00285 0.04508