IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Non-Parametric Tests: Practical

08 November, 2020

In this practical session, we will learn how to use R to calculate non-parametric tests for the differences in the median of a continuous, non-normally distributed variable between different groups. We will also see how to examine the association between two or more non-normally distributed continuous variable. More specifically,in this session we will:

1. Calculate the median of continuous variables that are not conforming to the normal distribution pattern. We will get insight about this lack of normal distribution through histograms.
2. Examine whether there are differences in the median of an outcome variable that is continuous not normally distributed between two groups of an independent variable (Mann-Whitney test or Wilcoxon rank sum test).
3. Extend this to look at differences in the median outcome between three or more groups of a categorical independent variable (Kruskal-Wallis test).
4. Assess the correlations between twoor morecontinuous variables that do not follow the normal distribution (rank correlation).

We’ll be working with the ELSA dataset again. Let’s also load the tidyverse and summarytools packages as we will using them in this session.

Remember to use a script to save your code and to change your working directory so you can load the ELSA dataset easily.

load("elsa.Rdata")  
library(summarytools)  
library(tidyverse)

**Q1. Calculate the median, the mean and the corresponding 95% confidence interval for the following variables: blood CRP level in mg/l (crp), total units of alcohol per week (alco). Please also look at the histogram for each of those variables to see whether they are normally distributed. What do you observe?**

To get the descriptive statistics, use the descr() function from summarytools. To get the 95% CI for the mean, use t.test().

descr(elsa$crp)

## Descriptive Statistics   
## elsa$crp   
## Label: Blood CRP level (mg/l)   
## N: 3129   
##   
## crp  
## ----------------- ---------  
## Mean 3.79  
## Std.Dev 6.12  
## Min 0.20  
## Q1 0.90  
## Median 2.00  
## Q3 4.20  
## Max 96.40  
## MAD 1.93  
## IQR 3.30  
## CV 1.61  
## Skewness 5.67  
## SE.Skewness 0.05  
## Kurtosis 50.25  
## N.Valid 2171.00  
## Pct.Valid 69.38

t.test(elsa$crp)

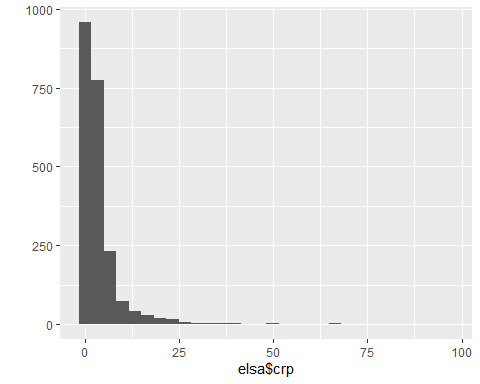
##   
## One Sample t-test  
##   
## data: elsa$crp  
## t = 28.874, df = 2170, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 3.534959 4.050117  
## sample estimates:  
## mean of x   
## 3.792538

It is clear to see that there is considerable difference between the mean and the median, with the former being much larger than the latter (mean: 3.8; median: 2). This implies that the variable does not follow the normal distribution, because a key property of the normal distribution is that the mean and the median (as well as the mode) have the same value. So, we would expect the frequency distribution of the blood CRP to be skewed. This is obvious by looking at the histogram. Use qplot() from ggplot2 which is part of the tidyverse.

qplot(elsa$crp)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 958 rows containing non-finite values (stat\_bin).

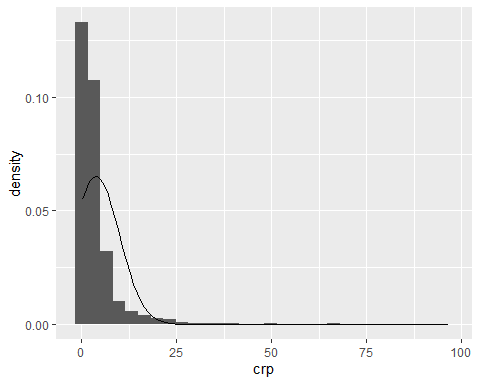


The distribution is clearly skewed. We can confirm this by overlaying a normal distribution on this graph that has the mean and standard deviation of the crp variable. The code for this is a little complex:

ggplot(elsa) +  
 aes(x = crp) +  
 geom\_histogram(aes(y = ..density..)) +  
 stat\_function(fun = dnorm,   
 args = list(mean = mean(elsa$crp, na.rm = TRUE),  
 sd = sd(elsa$crp, na.rm = TRUE)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 958 rows containing non-finite values (stat\_bin).

 ggplot() uses a *layered grammar of graphics* syntax. We take a dataset (elsa), map variables to visual properties of the graph (crp to a location on the x-axis, ..density.. which is derived within the function to the y-axis), and display the data as a geometry (in this case as a histogram). The last line overlays a normal distribution that has mean and standard deviation drawn from the crp variable. The graph confirms the variable is skewed - there are more observations with low levels of CRP.

Now let’s repeat this for the variable alco.

descr(elsa$alco)

## Descriptive Statistics   
## elsa$alco   
## Label: Total units of alcohol/week   
## N: 3129   
##   
## alco  
## ----------------- ---------  
## Mean 11.38  
## Std.Dev 16.19  
## Min 0.00  
## Q1 1.24  
## Median 6.37  
## Q3 15.35  
## Max 329.00  
## MAD 8.64  
## IQR 14.10  
## CV 1.42  
## Skewness 5.41  
## SE.Skewness 0.05  
## Kurtosis 69.21  
## N.Valid 2850.00  
## Pct.Valid 91.08

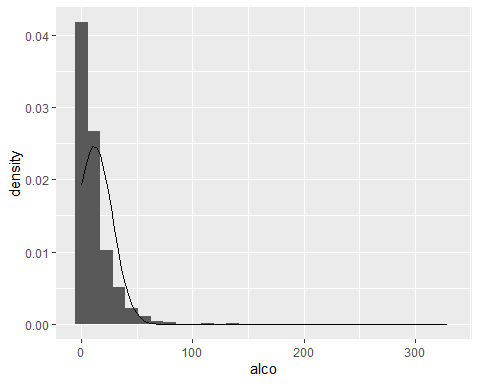
t.test(elsa$alco)

##   
## One Sample t-test  
##   
## data: elsa$alco  
## t = 37.533, df = 2849, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 10.78920 11.97863  
## sample estimates:  
## mean of x   
## 11.38392

ggplot(elsa) +  
 aes(x = alco) +  
 geom\_histogram(aes(y = ..density..)) +  
 stat\_function(fun = dnorm,   
 args = list(mean = mean(elsa$alco, na.rm = TRUE),  
 sd = sd(elsa$alco, na.rm = TRUE)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 279 rows containing non-finite values (stat\_bin).



We can again see that the mean is considerably higher than the median and the histogram shows a distribution that is highly skewed with a large majority of participants with scores close to 0 in terms of total units of alcohol consumed per week.

**Q2. Using appropriate tests, please assess whether there are differences in the CRP variable (crp) between: a) men and women (sex); b) manual and non-manual occupational social class (manual); c) between those that are current smokers or regular ex-smokers and those that are occasional ex-smokers or never smoked (smok\_bin). For this, it is advisable to also provide some summary figures in relation to CRP in the different groups of the independent variables**

Since the outcome variable is continuous but not normally distributed, we should employ non-parametric tests to look into the associations. In this exercise, all explanatory variables are binary, so the Wilcoxon rank sum (Mann-Whitney) test is used for all three associations (wilcox.test()). The syntax for the wilcox.test() function is similar to that for t.test() that we saw in previous weeks.

In order to have an understanding of the direction of any potential association, we will need to follow the test by summarising CRP for each of the groups of the independent variables. We can do that using the stby() function from summarytools (the stby() function works similarly to the by() function from Base R).

wilcox.test(crp ~ sex, elsa)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: crp by sex  
## W = 539572, p-value = 0.003537  
## alternative hypothesis: true location shift is not equal to 0

stby(elsa$crp, elsa$sex, descr)

## Descriptive Statistics   
## crp by sex   
## Data Frame: elsa   
## Label: Blood CRP level (mg/l)   
## N: 1368   
##   
## male female  
## ----------------- -------- ---------  
## Mean 3.46 4.06  
## Std.Dev 5.55 6.53  
## Min 0.20 0.20  
## Q1 0.90 0.90  
## Median 1.80 2.10  
## Q3 3.70 4.70  
## Max 64.50 96.40  
## MAD 1.63 2.08  
## IQR 2.80 3.78  
## CV 1.61 1.61  
## Skewness 5.17 5.83  
## SE.Skewness 0.08 0.07  
## Kurtosis 37.47 53.60  
## N.Valid 965.00 1206.00  
## Pct.Valid 70.54 68.48

The null hypothesis is that there is no difference in the rank of CRP by sex. The estimated p value (p=0.0035) indicates that there is hardly any support for the null hypothesis, thereby providing evidence about differences in CRP between men and women. We see from the stby() command that mean, median, and upper quartile (Q3) CRP levels are higher among females, so together the results suggest females have higher CRP levels than men in the population at large.

Now let’s look at the results for occupational class (manual).

wilcox.test(crp ~ manual, elsa)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: crp by manual  
## W = 459124, p-value = 1.313e-07  
## alternative hypothesis: true location shift is not equal to 0

stby(elsa$crp, elsa$manual, descr)

## Descriptive Statistics   
## crp by manual   
## Data Frame: elsa   
## Label: Blood CRP level (mg/l)   
## N: 1844   
##   
## non man manual  
## ----------------- --------- --------  
## Mean 3.53 4.15  
## Std.Dev 6.33 5.43  
## Min 0.20 0.20  
## Q1 0.80 1.10  
## Median 1.80 2.40  
## Q3 3.70 5.05  
## Max 96.40 54.10  
## MAD 1.78 2.37  
## IQR 2.90 3.93  
## CV 1.79 1.31  
## Skewness 6.32 3.73  
## SE.Skewness 0.07 0.09  
## Kurtosis 59.37 21.24  
## N.Valid 1329.00 800.00  
## Pct.Valid 72.07 65.79

Again, there is little evidence for the null hypothesis. Mean, lower quartile (Q1), median, and upper quartile CRP levels are higher among manual participants, so the results suggest CRP levels among manual workers in the population at large, compared with non-manual workers.

Now let’s look at the results for smoking status (smok\_bin).

wilcox.test(crp ~ smok\_bin, elsa)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: crp by smok\_bin  
## W = 500923, p-value = 1.93e-08  
## alternative hypothesis: true location shift is not equal to 0

stby(elsa$crp, elsa$smok\_bin, descr)

## Descriptive Statistics   
## crp by smok\_bin   
## Data Frame: elsa   
## Label: Blood CRP level (mg/l)   
## N: 1375   
##   
## never/ex occ ex/current reg  
## ----------------- -------------- ----------------  
## Mean 3.33 4.16  
## Std.Dev 6.25 5.99  
## Min 0.20 0.20  
## Q1 0.80 1.00  
## Median 1.70 2.20  
## Q3 3.50 4.80  
## Max 96.40 64.50  
## MAD 1.63 2.08  
## IQR 2.70 3.80  
## CV 1.87 1.44  
## Skewness 7.46 4.10  
## SE.Skewness 0.08 0.07  
## Kurtosis 79.59 24.03  
## N.Valid 970.00 1201.00  
## Pct.Valid 70.55 68.47

Again, there is little evidence for the null hypothesis. Mean, lower quartile (Q1), median, and upper quartile CRP levels are higher among former or current smokers, so the results suggest CRP levels among former and current smokers in the population at large, compared with non-smokers.

**Q3. Please find out whether there are differences inblood CRP levels between people with different levels of physical activity: low, medium, high (physact). Then, check whether there are differences in the same outcome (CRP) between people in the four different BMI groups as presented in variable bmi4. What can you conclude?**

In this case, the outcome variable is continuous but non-normally distributed (crp) and the explanatory variable has 3 groups (low-medium-high physical activity). The Kruskal-Wallis test is appropriate in these cases (kruskal.test()). We will again need to use stby() to understand descriptively how CRP levels differ across physical activity status.

kruskal.test(crp ~ physact, elsa)

##   
## Kruskal-Wallis rank sum test  
##   
## data: crp by physact  
## Kruskal-Wallis chi-squared = 15.758, df = 2, p-value = 0.0003787

stby(elsa$crp, elsa$physact, descr)

## Descriptive Statistics   
## crp by physact   
## Data Frame: elsa   
## Label: Blood CRP level (mg/l)   
## N: 1321   
##   
## Group 1 -low Group 2 - medium Group 3 - high  
## ----------------- -------------- ------------------ ----------------  
## Mean 4.25 3.42 3.55  
## Std.Dev 6.77 5.40 5.88  
## Min 0.20 0.20 0.20  
## Q1 1.00 0.80 0.90  
## Median 2.20 1.70 1.80  
## Q3 4.80 3.90 3.80  
## Max 96.40 65.60 65.30  
## MAD 2.08 1.63 1.78  
## IQR 3.80 3.10 2.90  
## CV 1.59 1.58 1.66  
## Skewness 5.63 5.32 5.71  
## SE.Skewness 0.08 0.09 0.11  
## Kurtosis 50.69 41.58 47.09  
## N.Valid 893.00 747.00 531.00  
## Pct.Valid 67.60 71.08 70.24

We see that there is little evidence that CRP is similar across the three physical activity groups (p=0.0004). There are higher CRP values for the low physical activity group, while the groups with medium and high physical activity have similar CRP profiles.

kruskal.test(crp ~ bmi4, elsa)

##   
## Kruskal-Wallis rank sum test  
##   
## data: crp by bmi4  
## Kruskal-Wallis chi-squared = 163.34, df = 3, p-value < 2.2e-16

stby(elsa$crp, elsa$bmi4, descr)

## Descriptive Statistics   
## crp by bmi4   
## Data Frame: elsa   
## Label: Blood CRP level (mg/l)   
## N: 63   
##   
## Under 20 Over 20-25 Over 25-30 Over 30  
## ----------------- ---------- ------------ ------------ ---------  
## Mean 2.73 2.92 3.72 5.07  
## Std.Dev 5.08 5.31 6.45 6.25  
## Min 0.20 0.20 0.20 0.20  
## Q1 0.40 0.60 1.00 1.70  
## Median 1.15 1.30 2.00 3.10  
## Q3 3.10 2.70 4.05 5.90  
## Max 27.40 65.30 96.40 50.30  
## MAD 1.11 1.19 1.93 2.82  
## IQR 2.60 2.10 3.03 4.20  
## CV 1.86 1.82 1.74 1.23  
## Skewness 3.56 5.33 7.07 3.71  
## SE.Skewness 0.35 0.10 0.08 0.11  
## Kurtosis 12.91 42.08 71.57 18.41  
## N.Valid 46.00 586.00 952.00 463.00  
## Pct.Valid 73.02 71.99 71.26 64.57

It is clear that there is no support for the null hypothesis of no difference (p < 0.0001). Therefore we conclude that there are differences in CRP values in the population according to BMI groups. Furthermore, by looking at the descriptive statistics, we can see that there are consecutively higher CRP levels across BMI groups.

**Q4. Finally, please calculate the correlation coefficients and respective p values for the associations between blood CRP levels (crp), total units of alcohol per week (alco) and deciles of household wealth (wealth10). For this, please use the sample with non-missing data on all these three variables.**

For this, we are looking at the correlation between these three variables. We already know that two of these variables (crp and alco) are continuous not normally distributed variables, while the wealth deciles variable (wealth10) refers to an ordinal variable with 10 categories. Therefore, we should opt for the rank correlation, which is a non-parametric test that makes no assumption about the distribution of the variables to be correlated. We can use the cor() function but we need to set the method as "spearman". To use the complete case say we also need to set the argument use to "complete.obs".

elsa %>%  
 select(crp, alco, wealth10) %>%  
 cor(use = "complete.obs", method = "spearman")

## crp alco wealth10  
## crp 1.0000000 -0.1288094 -0.1695439  
## alco -0.1288094 1.0000000 0.1641812  
## wealth10 -0.1695439 0.1641812 1.0000000

The negative sign of the respective value indicates that higher values of CRP are associated with lower weekly alcohol consumption and the same is also the case for the association between CRP and wealth (first column). The positive sign in the correlation between wealth deciles and alcohol consumption indicates higher levels of alcohol consumption of higher values in the wealth variable. The actual values of rho are rather low (rho= -013 for alcohol and CRP, rho=-0.17 for wealth deciles and CRP,and rho= 0.16 for wealth deciles and alcohol) highlighting that these associations are weak.