Statistical Analysis of The Risk Factors of Chronic Respiratory Diseases

Business Statistics 2024 - Final Project

at the
Faculty of Economics and Business

at the
Université de Neuchâtel

with
Prof. Alina Matei

as part of the
Bachelor of Science in Data Science
&
Pre-Master Program

Authors and student numbers:

Alessia Airaldi / 18-740-183
Greta Sitta / / 21-984-562
Mariia Ponomarova / 21-507-710
Laila Ibrahim / 23-504-897
Salsabil Mtiraoui / 20-506-630
Zied Ellouzi / 23-503-188

Table of Content

List of Figures

Executive Summary

Our project investigates the relationships between environmental and behavioral factors—specifically air quality and smoking rates — on public health — here chronic respiratory diseases (CRDs)— to understand their impact on public health outcomes. Utilizing data from 85 countries, key variables include PM2.5 concentrations (an air pollution indicator), smoking prevalence, and CRD mortality rates. The study employs statistical methods, including hypothesis testing and the Central Limit Theorem (CLT), to analyze these factors' interplay.

The analysis reveals a significant relationship between air pollution and CRD mortality rates. Countries with PM2.5 concentrations above the median (19.1 $\mu g/m^3$) exhibit higher CRD mortality rates compared to those below the median. Statistical analysis confirms a p-value far below the 5% significance threshold, strongly supporting the hypothesis that poor air quality exacerbates CRDs. However, smoking prevalence shows no statistically significant relationship with CRD mortality rates. Contrary to expectations, hypothesis testing yields a p-value exceeding 0.05, indicating insufficient evidence to link smoking rates directly to variations in CRD mortality at the population level within this dataset. The project highlights the utility of the CLT for handling non-normal data distributions, enabling robust inferential statistics. Key analytical tools include z-scores and correlation tests to evaluate hypotheses and determine statistical significance.

The study underscores the critical role of air quality in influencing respiratory health outcomes, highlighting the urgency for policies aimed at pollution mitigation. The findings on smoking warrant further investigation, suggesting the need to incorporate additional variables, such as smoking intensity and healthcare access, for a more nuanced understanding.

While the analysis provides valuable insights, limitations such as potential data biases, outliers, and regional disparities must be acknowledged. Future research could benefit from a broader dataset and the inclusion of additional socio-economic and healthcare-related factors.
Overall, this project contributes to the growing evidence linking environmental factors to public health and emphasizes the importance of targeted interventions to improve air quality and address preventable health risks.

1 Introduction

Chronic respiratory diseases (CRDs) are a significant global health issue, contributing heavily to morbidity and mortality. Conditions like asthma and chronic obstructive pulmonary disease (COPD) arise from a complex mix of environmental, behavioral, and genetic factors (World Health Organization (2019)).

This discussion focuses on two major contributors identified in preliminary research: air pollution & smoking

Air pollution, especially particulate matter (PM2.5) and ozone, is linked to increased mortality and respiratory illnesses, with significant health impacts evident even at low exposure levels (Brunekreef and Holgate (2002)). PM2.5 originates from sources like vehicle exhaust and fossil fuel combustion or forms via chemical reactions. Due to its small size, PM2.5 penetrates deeply into the respiratory system and bloodstream, posing serious cardiovascular risks and impacting air quality regionally and globally (Labaki and Han (2020); United States Environmental Protection Agency (2023))

Smoking is a leading cause of respiratory diseases, including emphysema and COPD, as identified by the US Surgeon General in 1964. Despite awareness, its prevalence drives respiratory illness and preventable deaths, highlighting the need for global tobacco control strategies. (Sethi and Rochester (2000)).

This project examines the relationships between these three variables across countries, emphasizing the impact of environmental and behavioral factors on health. It employs statistical tools, including the Central Limit Theorem (CTL), to address non-normal data and guide the development of our methodology. We expect a correlation between poor air quality and higher CRD rates due to pollutant exposure, and an especially strong correlation between smoking and CRD mortality, given our understanding of the significant harm caused by smoke.

2 Data

For our research, we have decided to focus on the relationship between air quality, smoking rates, and CRD death rates across 85 countries. This analysis includes the PM2.5 concentration ($\mu g/m^3$), the smoking rate, and the CRD death rate. The PM2.5 concentration represents fine particulate matter in the air, a key indicator of air quality, while the smoking rate and CRD death rates reflect behavioural outcomes, respectively. The PM2.5 values range theoretically from 0 (ideal air quality) to much higher levels (indicative of severe pollution), with the dataset capturing values that span a wide spectrum of pollution levels globally. The smoking rate with 0% indicating no smoking and 100% theoretically representing universal smoking prevalence within a population. The CRD death rate, provides insight into the severity of respiratory health impacts. We have chosen to concentrate our analysis on these three variables because they represent crucial aspects of the environmental and behavioural determinants of health. Our dataset offers a significant sample size but not representing the entire global population. The data provides a snapshot for specific years, primarily 2019, making it a sample across both space and time.

2.1 Data gathering

The data gathering process began by identifying datasets that were comprehensive, credible, and aligned with our study's objectives. We specifically sought datasets that allowed for a global perspective, incorporating diverse geographic regions and economic conditions. The selected datasets provided raw values, which we downloaded and cleaned for analysis.

The primary data was obtained from the following sources:
Air Quality Data: PM2.5 concentrations (renamed pm25), were sourced from IQAir's "World's Most Polluted Countries" dataset. This variable serves as an indicator of air pollution, with higher values indicating worse air quality.(IQAir (2023))
CRD Death Rate (renamed crd): This variable represents death rates due to CDR, sourced from Our World in Data. (Our World in Data (2024))
Smoking Rate (renamed smoking): This dataset represents the percentage of adults who smoke, and was also obtained from Our World in Data. (Our World in Data (2015))

2.2 Descriptive overview

The table (Figure 1) presents a summary of the three variables measured across countries worldwide in 2019. The pm25 ranges from 3.30 to 83.30, with a median of 19.10, showing extreme outliers. The crd is even more variable, ranging from 6.40 to 180.60, with a median of 27.80, reflecting severe health challenges in some regions. In contrast, smoking prevalence is more balanced (3.60 to 45.10, median 23.40). The extreme values of pm25 and crd suggest environmental and healthcare inequalities, while smoking, despite less variability, remains a key risk factor. Combined, high pm25 and smoking likely drive the severe respiratory disease burdens seen in outlier countries.

As shown in Figure (2), the distributions and normality assessments of deaths from CRD, pm25 concentrations, and smoking rates in 2019 highlight these global disparities. The distribution of deaths from respiratory diseases is heavily right-skewed, with most countries reporting death rates below 50 per 100K people. However, a few outliers exceed 150, indicating severe health crises in certain regions. This skewness is confirmed by the Q-Q plot, which shows strong deviations from normality in the upper tail. Similarly, pm25 concentrations show a moderately right-skewed distribution, with most values concentrated below 30 $\mu g/m^3$. A few countries experience extreme air pollution levels above 60 $\mu g/m^3$. The Q-Q plot reflects this skewness and the presence of outliers, although less pronounced than in the respiratory disease data. In contrast, smoking rates are more symmetrically distributed, centered around 20–30%, with few extreme values above 40%. The Q-Q plot indicates a near-normal distribution, suggesting less variability and fewer disparities compared to the other variables.

Even though our data is not normally distributed, our sample size exceeds 30. According to the Central Limit Theorem (CLT), this allows us to assume that the sample means follow a normal distribution. The histograms (Figure 3) demonstrate the CLT "in action": as the sample size increases, the distribution of sample means becomes increasingly normal. With smaller sample sizes, the sample means reflect the skewness of the original data, but larger sample sizes reduce this skewness and concentrate the means more tightly around the center. This transformation ensures that parametric statistical methods can be confidently applied, even when the underlying data deviates from normality.

3 Methodology

The study's analysis comprises three stages, examining variable relationships across two subsamples using rigorous inferential statistics to validate hypotheses. The dataset is introduced with the CLT applied to assume normality, given the sample size exceeds 30. This sets the foundation for hypothesis testing and correlation evaluation to ensure statistical rigor.

The methodology begins with data standardization via z-scores and p-value calculation for hypothesis testing. The null H_0 and alternative hypotheses H_1 are defined, with a 5% significance level (α) selected to balance sensitivity and error risk. H_0 is rejected if the p-value is below α; otherwise, it is not.
Then, correlation tests are conducted under two conditions: (1) a linear relationship, verified through scatterplots, and (2) a bivariate normal distribution, confirmed visually through elliptical scatterplot patterns. These prerequisites ensure test validity.
The `cor.test` function calculates correlation coefficients and confidence intervals, evaluating relationship strength and reliability. Hypothesis testing employs z-scores to derive p-values, with one-tailed tests using 1-pnorm(z-score) or pnorm(z-score), based on the direction of H_1. Visual aids like scatterplots and confidence interval graphs clarify results. The application of CLT ensures methodological robustness, and the study concludes with an acknowledgment of limitations and a discussion contextualizing the findings.

4 Analysis

4.1 Air Quality Analysis

For our analysis, we split our sample into two groups: 43 values representing good air quality and 42 values representing poor air quality. To achieve this, we used the median as a threshold, ensuring the groups were roughly equal in size. The first group, consisting of 43 countries with PM2.5 concentrations of 19.1 or less, represented areas with good air quality. The second group included 42 countries with PM2.5 concentrations greater than 19.1, representing poor air quality. According to our hypothesis, the bigger the concentration of PM2.5 in the air of the country, the higher the rate of deaths from chronic respiratory diseases. Therefore, we test :

- $H_0 : E(\text{crd} \mid \text{pm}_{2.5} > 19.05) - E(\text{crd} \mid \text{pm}_{2.5} \leq 19.05) = 0$
- $H_1 : E(\text{crd} \mid \text{pm}_{2.5} > 19.05) - E(\text{crd} \mid \text{pm}_{2.5} \leq 19.05) > 0$

We define the following:

- n_{good}: The number of countries with PM2.5 concentrations ≤ 19.1.
- n_{bad}: The number of countries with PM2.5 concentrations > 19.1.
- n: The total number of countries in the sample, calculated as $n = n_{\text{good}} + n_{\text{bad}} = 85$.

Next, we estimate the conditional expectations for CRD based on air quality:

- The estimator $\hat{E}(\text{crd} \mid \text{pm}_{2.5} \leq 19.1)$, representing the expected value of CRD in countries with PM2.5 concentrations ≤ 19.1, is calculated as 22.70698.
- The estimator $\hat{E}(\text{crd} \mid \text{pm}_{2.5} > 19.1)$, representing the expected value of CRD in countries with PM2.5 concentrations > 19.1, is calculated as 46.05476.

These values provide the basis for further analysis of the relationship between air quality and chronic respiratory diseases. Therefore, the value of the estimator for the parameter of interest is:

$$\hat{E}(\text{crd} \mid \text{pm}_{2.5} > 19.05) - \hat{E}(\text{crd} \mid \text{pm}_{2.5} \leq 19.05) = 23.34779$$

To determine whether this difference is significant enough to reject H_0, we must consider the variability of the estimator. Specifically,

$$\hat{E}(\text{crd} \mid \text{pm}_{2.5} > 19.05) \sim N(E(\text{crd} \mid \text{pm}_{2.5} > 19.05), \text{Var}(\text{crd} \mid \text{pm}_{2.5} > 19.05)/n_{\text{bad}})$$

$$\hat{E}(\text{crd} \mid \text{pm}_{2.5} \leq 19.05) \sim N(E(\text{crd} \mid \text{pm}_{2.5} \leq 19.05), \text{Var}(\text{crd} \mid \text{pm}_{2.5} \leq 19.05)/n_{\text{good}})$$

Under H_0, the estimator of the parameter of interest should theoretically follow the distribution:

$$\hat{E}(\text{crd} \mid \text{pm}_{2.5} > 19.05) - \hat{E}(\text{crd} \mid \text{pm}_{2.5} \leq 19.05) \sim N\left(0, \frac{\text{Var}(\text{crd} \mid \text{pm}_{2.5} > 19.05)}{n_{\text{bad}}} + \frac{\text{Var}(\text{crd} \mid \text{pm}_{2.5} \leq 19.05)}{n_{\text{good}}}\right)$$

In `R`, we estimated the variability of our parameter of interest as:

$$\text{var}_{\text{estimator}} = \frac{s_1^2}{n_{\text{bad}}} + \frac{s_2^2}{n_{\text{good}}} = 34.31528$$

The value of the standard error estimator is equal to: $\sqrt{\text{var}_{\text{estimator}}} = 5.86$.
With $z = 3.985675$, the p-value is calculated as $p = 1 - \text{pnorm}(z)$, resulting in: $p = 3.364425 \times 10^{-5}$. Since $p < 0.05$, we reject H_0 at the 5% significance level. This conclusion is supported by the graph 6, where the observed z-value (maroon line) lies far beyond the critical z-value (aquamarine dashed line) for $\alpha = 0.05$. The significant distance between these values, coupled with the extremely small p-value, provides strong evidence against the null hypothesis, confirming the result is highly significant.

Additionally, we studied the correlation between CRD and air quality. The sample estimate of the correlation is 0.53. A preliminary analysis in `R` confirmed that the test's assumptions are reasonably satisfied. The scatterplot (Figure 4) indicates a linear relationship and suggests a bivariate normal distribution. To confirm the latter, we used the `car` library's `dataEllipse` function (Figure 7).

The results of the correlation test are as follows: Pearson's product-moment correlation for the variables `pm25` and `crd` yielded a t-value of 5.7475 with 83 degrees of freedom and a p-value of 1.457×10^{-7}. The alternative hypothesis suggests that the true correlation is not equal to zero. Since the p-value < 0.05 we reject the null hypothesis H_0 at the 5% level, indicating that there is a significant correlation between the variables `pm25` and `crd`.
The 95% confidence interval for the correlation coefficient is $(0.3615556, 0.6704488)$. This means we are 95% confident that the true correlation coefficient lies between 0.36 and 0.67. Since the interval does not include 0, this suggests a statistically significant positive correlation between the variables.

4.2 Smoking Analysis

We divided the sample into two groups based on the median smoking rate of 23.4%.
The first group consists of countries with a smoking rate of 23.4% or lower, characterized by the following properties for chronic respiratory diseases. The second group comprises countries with a smoking rate exceeding 23.4%

Our hypothesis is that higher smoking rates in a country lead to higher rates of CRD. Thus, we state:

- $H_0 : E(\text{crd} \mid \text{smoking} > 23.4) - E(\text{crd} \mid \text{smoking} \leq 23.4) = 0$
- $H_1 : E(\text{crd} \mid \text{smoking} > 23.4) - E(\text{crd} \mid \text{smoking} \leq 23.4) > 0$

Based on intuition, we anticipate rejecting H_0.

We define:

- $n_{\text{small}} = 44$: The number of countries with a smoking rate $\leq 23.4\%$.
- $n_{\text{big}} = 41$: The number of countries with a smoking rate $> 23.4\%$.
- $n = n_{\text{small}} + n_{\text{big}} = 85$: The total sample size.

Since estim_crd_bar_small = 31.30455 and estim_crd_bar_big = 37.39756, the value of the estimator for our parameter of interest is: *estim_crd_bar_big* - *estim_crd_bar_small* = 6.093016.

The value of the variability of the parameter of interest is:

$$\text{var\_estimator2} = \left(\frac{s_{\text{small}}}{n_{1,\text{sm}}}\right) + \left(\frac{s_{\text{big}}}{n_{2,\text{sm}}}\right) = 41.76879.$$

Hence, we estimate the variability of the estimator of our parameter of interest as:

$$\hat{E}(\text{crd} \,|\, \text{smoking} > 23.4) - \hat{E}(\text{crd} \,|\, \text{smoking} \leq 23.4) \sim N(0, 41.76879).$$

To find the p-value of our test and determine if we can reject H_0, we calculated z and got $z = 0.9427712$. p is then equal to 0.2860453. As $p > 0.05$, we cannot reject H_0 at a significance level of 5%. Since our z-score does not reach or exceed the critical z-value, and the associated p-value is $p = 0.2860453$ (greater than 0.05), we fail to reject the null hypothesis H_0. This indicates insufficient evidence to suggest that the observed result is statistically significant at the 5% level.
Visually, this can be seen in Figure 8 as our z-score falling well within the bulk of the standard normal curve, far from the critical region (the tail beyond the aquamarine dashed line).

The next step of our analysis was the correlation test between `crd` and `smoking`.
Figures 5 and 9 provide evidence that the conditions for the correlation test are satisfied. The scatterplot in Figure 5 suggests a linear relationship between the variables, while the ellipse-shaped distribution in Figure 9 indicates a bivariate normal distribution, supporting the applicability of the correlation test. The value of the sample estimate of the correlation is cor = 0.1656527. Pearson's product-moment correlation for the data on smoking and CRD gives $t = 1.5303$ with degrees of freedom ($df = 83$) and a p-value of 0.1297. The alternative hypothesis tested is that the true correlation is not equal to 0.

The 95% confidence interval is $(-0.04920864, 0.36586027)$.
Therefore we are 95% confident that the mean of our estimator lies within this confidence interval. Since 0 is included in this interval, we fail to reject H_0. This suggests there is no statistically significant difference between the two groups at the 5% significance level ($\alpha = 0.05$).

4.3 Limitations

While this analysis provides valuable insights, several limitations must be acknowledged. One key limitation is the potential for distortions within the dataset. For example, data collection methods and reporting standards may vary between countries, leading to inconsistencies or inaccuracies that could affect the reliability of the results. Additionally, the dataset may not be fully representative of global trends, as it might exclude certain regions or demographics.

Another limitation is the impact of outliers on the findings. Countries with exceptionally high or low rates of air quality or CRD can disproportionately influence the calculated averages and overall trends, potentially skewing the results and leading to misinterpretations.

Finally, the generalizability of these findings to other contexts is limited. The relationships observed between smoking rates and CRD may differ in regions not included in the dataset due to variations in healthcare systems, environmental conditions, and cultural practices. These factors can significantly affect the dynamics of health outcomes, making it challenging to apply these results universally.

Recognizing these limitations is essential for drawing cautious and contextually informed conclusions from the analysis.

5 Discussion

The relationship between air quality and CRD mortality underscores the significant impact of environmental pollutants, particularly PM2.5, in worsening chronic respiratory conditions. These findings align with global research linking air pollution to CRDs, emphasizing the need for stricter air quality standards, especially in polluted regions.

Unexpectedly, no significant link between smoking and CRD mortality was observed, despite smoking's established role in conditions like COPD. This discrepancy may be due to confounding factors (e.g., healthcare access or genetic predispositions), regional differences in smoking behaviors, or limitations in using smoking prevalence instead of intensity or duration.

While both smoking and air pollution affect respiratory health, the results suggest air quality has a stronger population-level impact, likely due to its pervasive nature.

In conclusion, the study confirms the strong link between poor air quality and CRD mortality, highlighting the urgent need for pollution mitigation strategies. However, the unexpected findings on smoking suggest the relationship may be more complex or context-dependent. Future research should consider variables such as healthcare quality, socioeconomic factors, and urbanization to further clarify these dynamics.

6 Bibliography

Brunekreef, B. and S. T. Holgate, "Air pollution and health," *The Lancet*, 2002, *360* (9341), 1233–1242.

IQAir, "World's Most Polluted Countries in 2023 - PM2.5 Ranking," https://www.iqair.com/world-most-polluted-countries?srsltid=AfmBOopLcWmQQmjRHN_d5HNfaRxPrz9lgCC33AK2QYNpyEQknvij3hy1 2023. Accessed 2023.

Labaki, W. W. and M. K. Han, "Chronic respiratory diseases: a global view," *The Lancet Respiratory Medicine*, 2020, *8* (6), 531–533.

Our World in Data, "Share of Adults Who Smoke," https://ourworldindata.org/grapher/share-of-adults-who-smoke?tab=table&time=earliest..2019&country=USA~BRA~RUS~IND~CHN~European+Union~IDN~OWID_WRL 2015. Accessed 2024.

_ , "Chronic respiratory diseases death rate," https://ourworldindata.org/grapher/respiratory-disease-death-rate?tab=table&time=earliest..2019 2024. Accessed 2024.

Sethi, J. M. and C. L. Rochester, "Smoking and Chronic Obstructive Pulmonary Disease," *Clinics in Chest Medicine*, 2000, *21* (1), 67–86.

United States Environmental Protection Agency, "Particulate Matter (PM) Basics," https://www.epa.gov/pm-pollution/particulate-matter-pm-basics 2023. Accessed 2023.

World Health Organization, "Chronic respiratory diseases," https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab_1 2019. Accessed 2019.

A Appendix - Supplements

Figure 1: Summary of the variables

```
> summary(pm25)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.30   10.90   19.10   21.58   24.80   83.30
> summary(crd)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.40   17.20   27.80   34.24   36.20  180.60
> summary(smoking)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.60   17.10   23.40   23.16   29.60   45.10
```

Figure 2: Data distribution

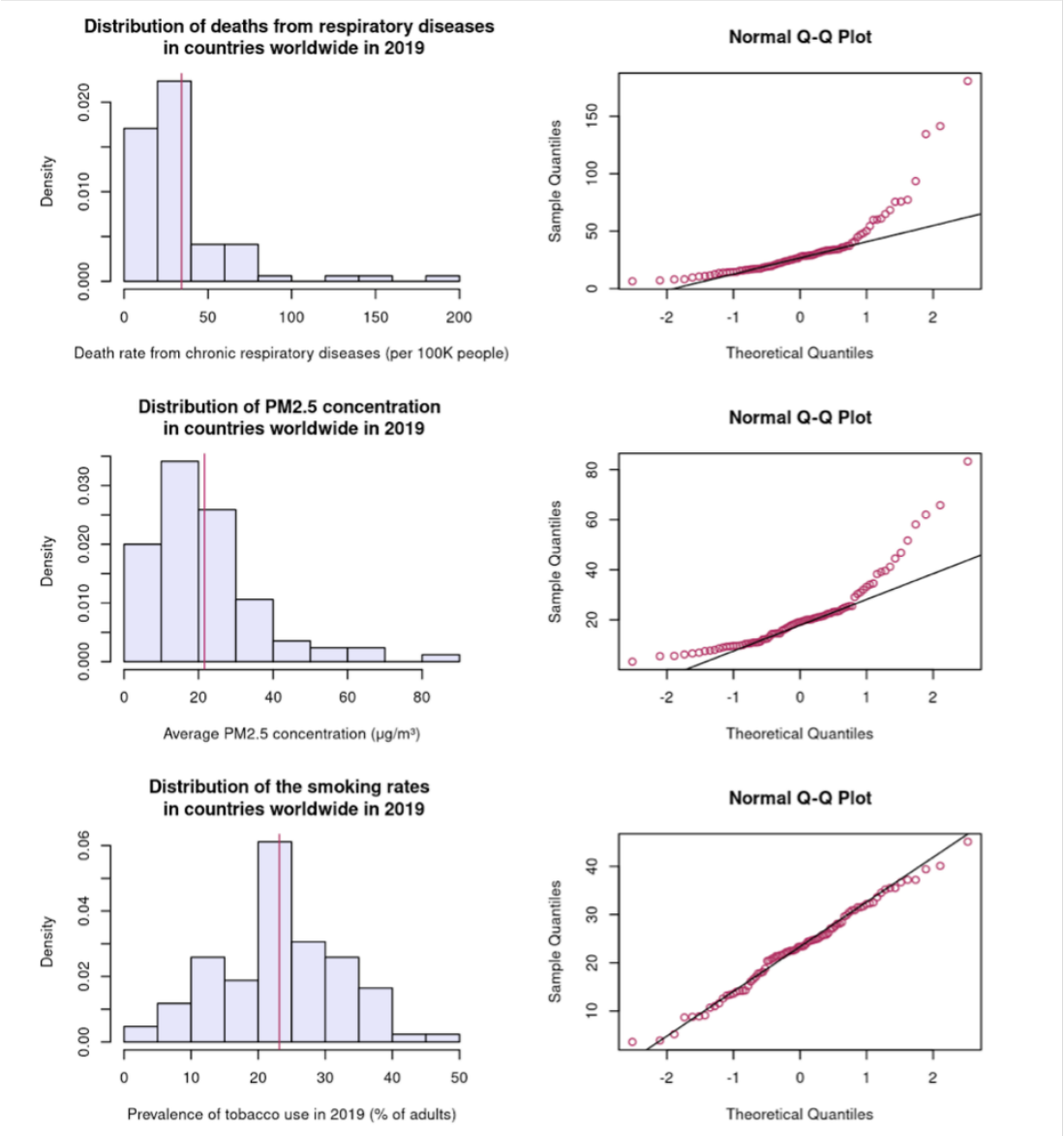


Figure 3: Histograms of sample means

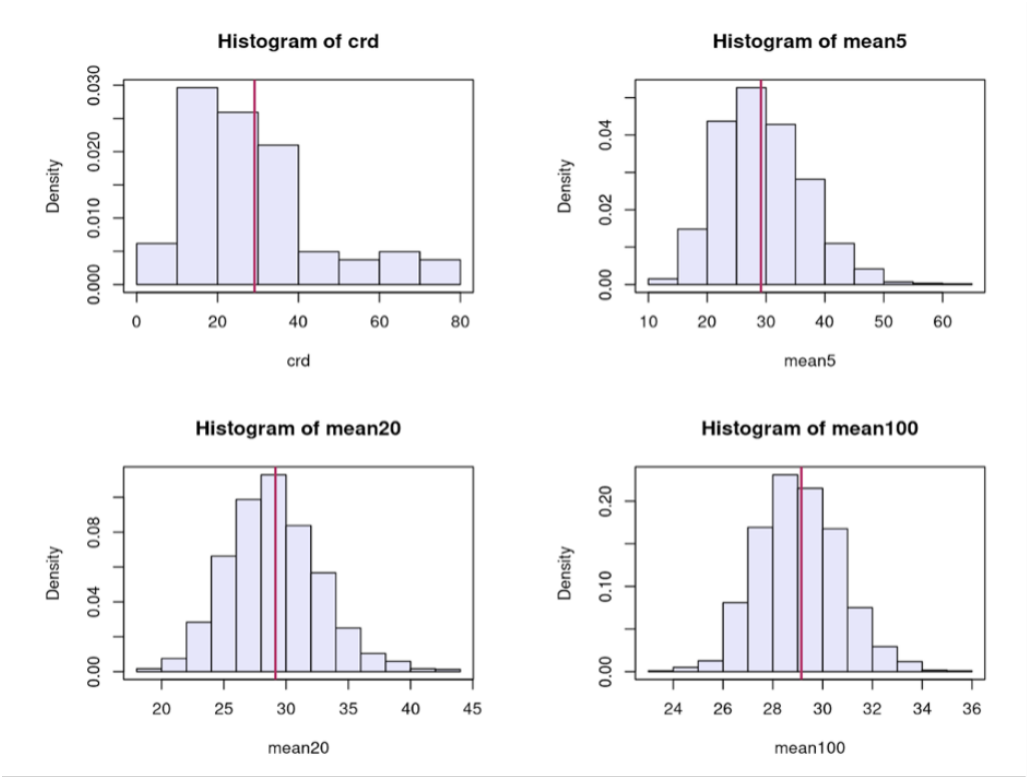


Figure 4: Scatterplot of pm25 vs crd

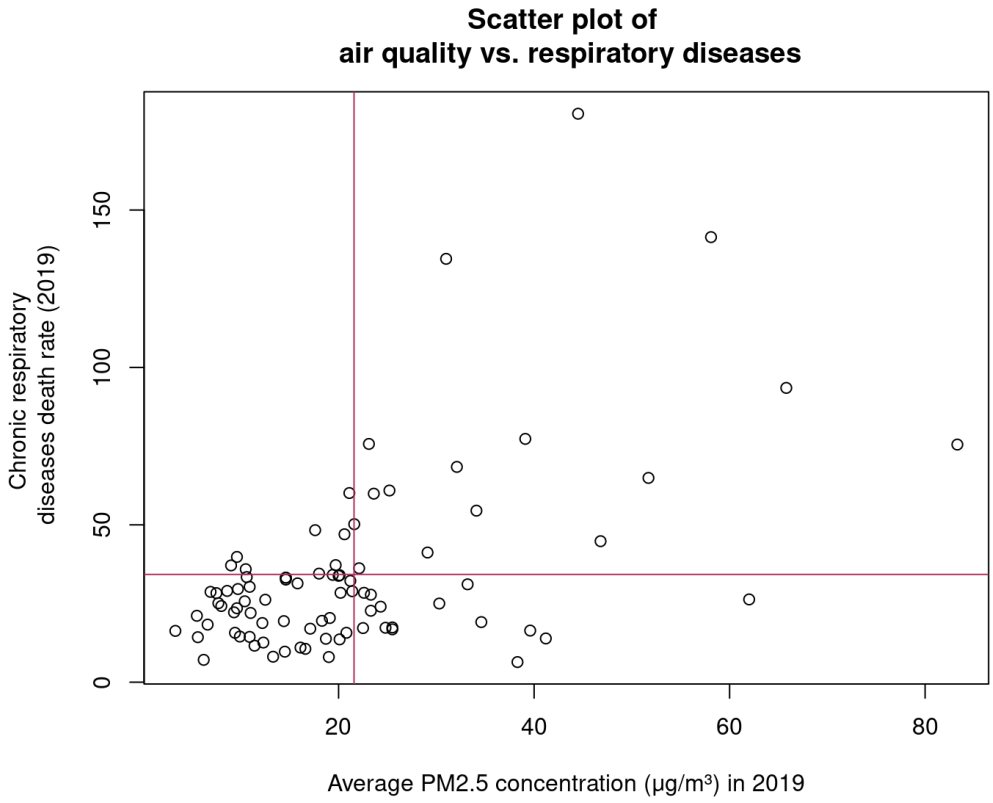


Figure 5: Scatterplot of smoking vs crd

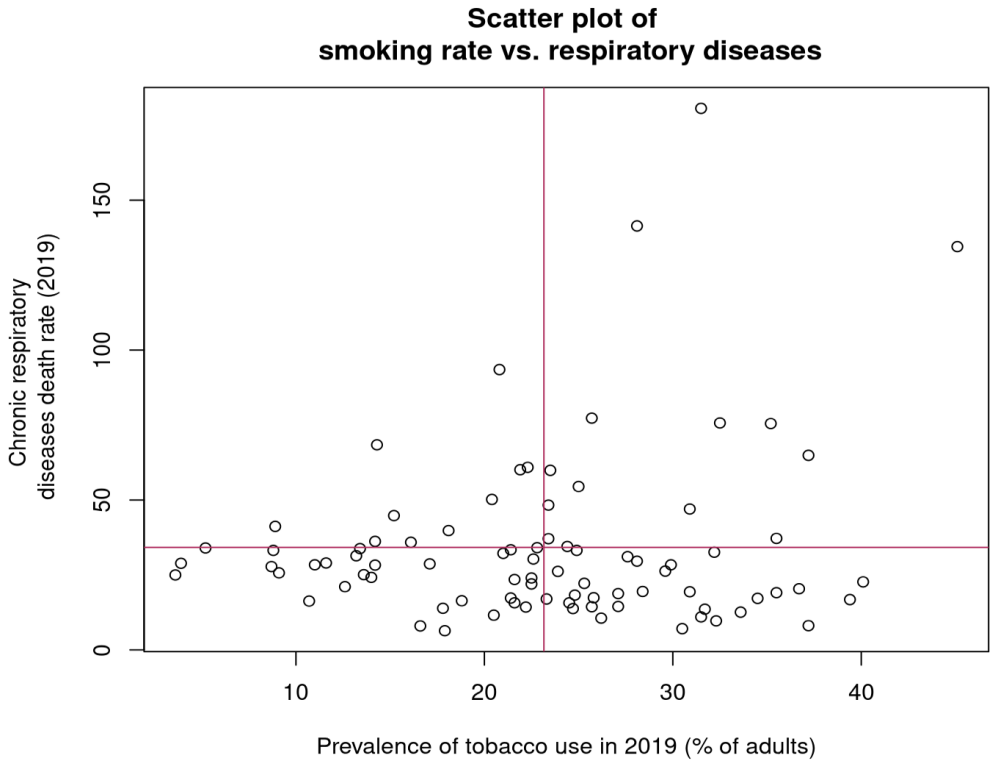


Figure 6: Normal estimator of air quality and CRD under H0

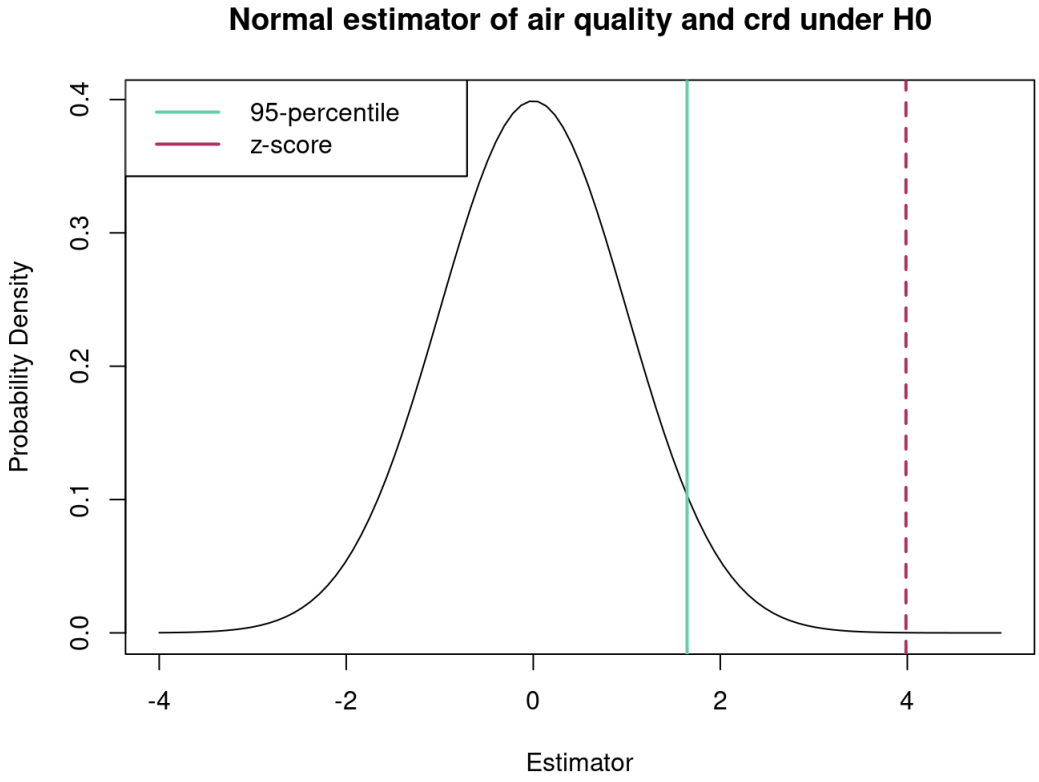


Figure 7: Correlation Between Air Pollution and CRD Death Rate

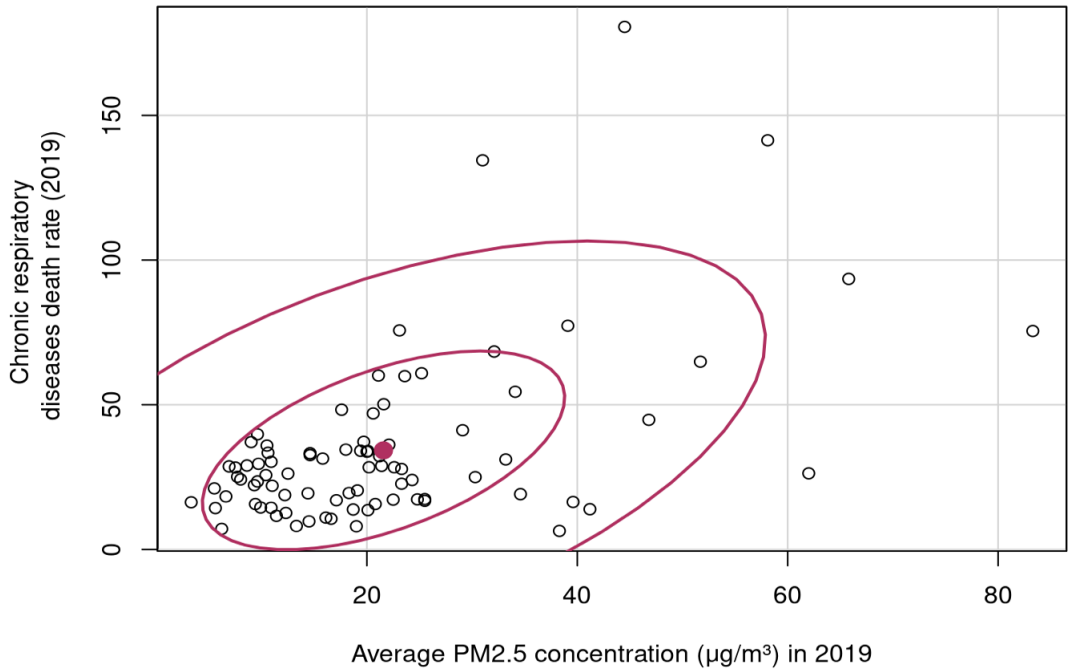


Figure 8: Normal estimator of smoking and CRD under H0

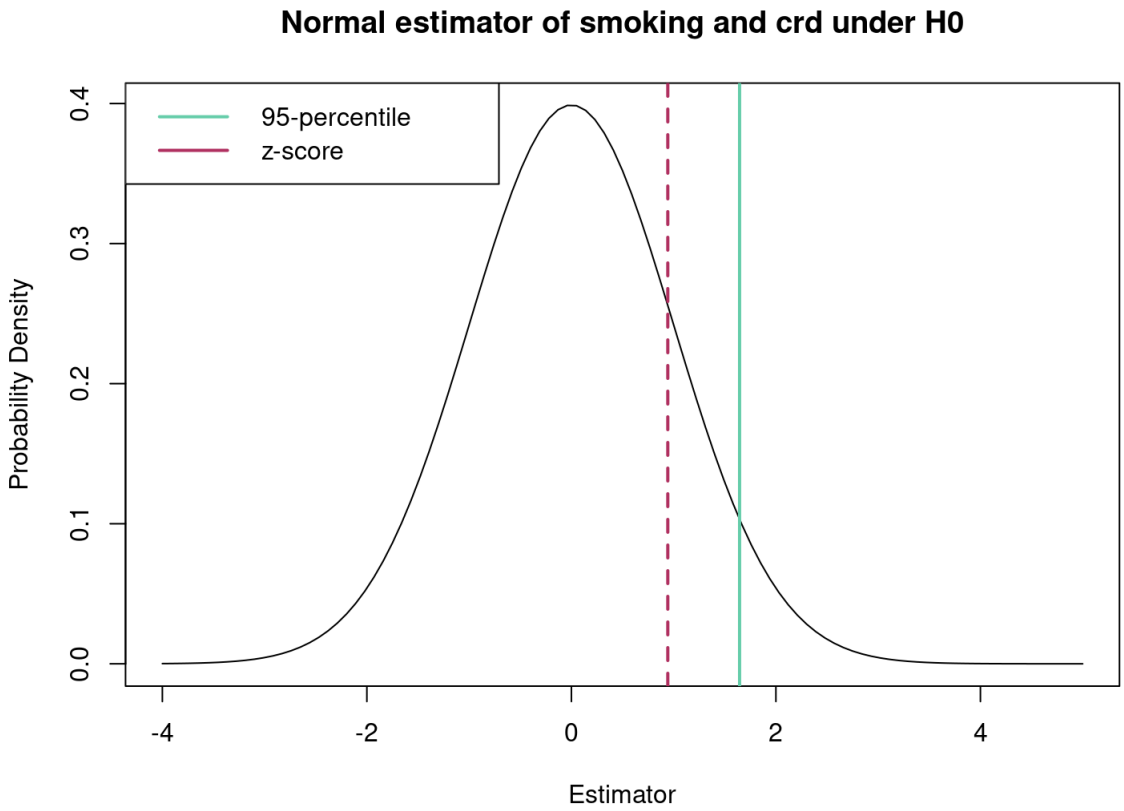
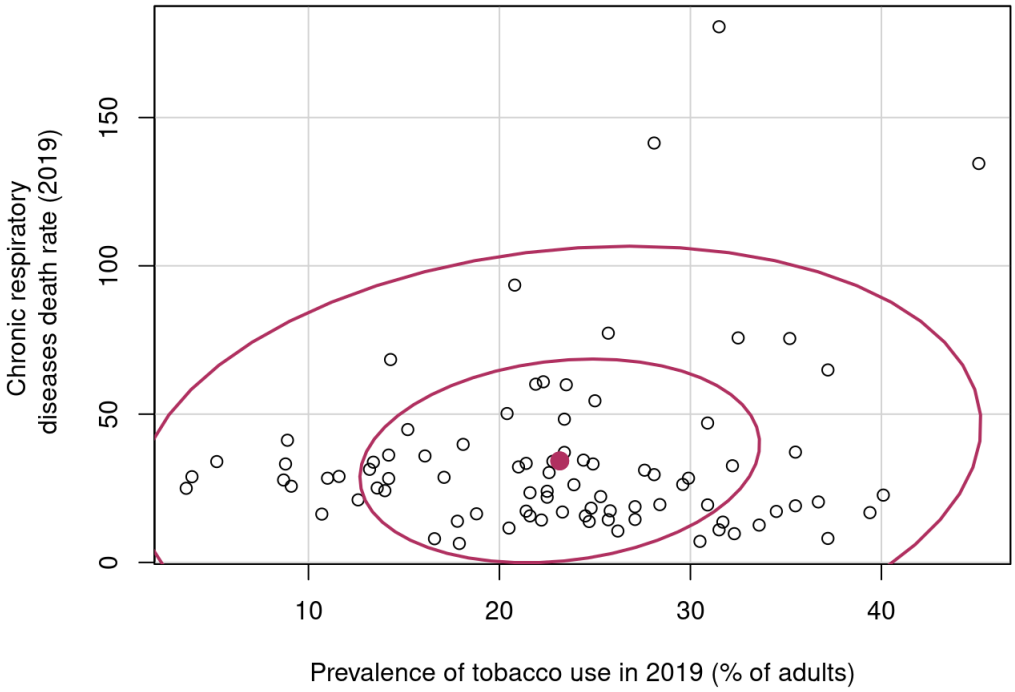


Figure 9: Correlation Between Tobacco Use and CRD Death Rate



B Appendix - R Code

```
# Import dataset
airquality = read.table("PM25.csv", header = TRUE, sep=",")
airquality
respdis = read.table("chron-resp.csv", header = TRUE, sep=",") # respiratory
    diseases
respdis
smoking=read.table("smoking.csv", header = TRUE, sep=",")
smoking

# Merging the air quality data and the respiratory diseases rates in one table
# using na.omit() to exclude NA values
data1 = na.omit(merge(airquality, respdis, by = "Countries"))
data

# Merging all of the data together
# We had to divide the merging in 2 steps, because the merge function couldn't
    handle combining
# all of the 3 columns at once
data = na.omit(merge(data1, smoking, by = "Countries"))
data

# Setting the variables for easier use
pm25 = data$pm25
crd = data$crd
smoking = data$smoking

# Excluding outliers to see if there was a significant difference in the
    relation
# between air quality vs. crd with outliers and without
subset_indices = which(data$pm25 < 80)
pm25v2 = data$pm25[subset_indices]
crdv2 = data$crd[subset_indices]
cor(pm25v2, crdv2) #0.5272442

# Checking differences between correlations with outliers (air quality)
cor(pm25, crd) #0.5335651
# It doesn't seem like there is a significant change in the correlation,
    therefore we decided to keep our outliers

### DESCRIPTIVE OVERVIEW ###
summary(pm25)
summary(crd)
summary(smoking)
par(mfrow=c(2,3))

# The Histograms of our 3 variables
hist(crd, col = "lavender", xlab = "Death rate from chronic \n respiratory
    diseases", main="Distribution of deaths \n from chronic \n respiratory
```

```
    diseases", probability = TRUE, cex.main=1)
abline(v=mean(crd), col="maroon", lwd=2)

hist(pm25, col = "lavender", xlab = "Average PM2.5 \n concentration
    ( g / m  )", main="Distribution \n of PM2.5 \n concentration", probability
    = TRUE, cex.main=1)
abline(v=mean(pm25), col="maroon", lwd=2)

hist(smoking, col = "lavender", xlab = "Prevalence of tobacco use \n in 2019
    (% of adults)", main="Distribution \n of the smoking \n rates",
    probability = TRUE, cex.main=1)
abline(v=mean(smoking), col="maroon", lwd=2)

# The qqnorms of our 3 variables
qqnorm(crd, col="maroon", main = "QQ-plot of deaths \n from chronic \n
    respiratory diseases", cex.main=1)
qqline(crd)

qqnorm(pm25, col="maroon", main = "QQ-plot of PM2.5 \n concentration",
    cex.main=1)
qqline(pm25)

qqnorm(smoking, col="maroon", main = "QQ-plot of the \n smoking rates",
    cex.main=1)
qqline(smoking)

# Histogram of the crd distribution
par(mfrow=c(1,1))
hist(crd, col = "lavender", xlab = "Death rate from chronic respiratory
    diseases \n (per 100K people)", main="Distribution of deaths \n from
    chronic respiratory diseases \n in countries worldwide in 2019",
    probability = TRUE, ylim = c(0, 0.030))
lines(density(crd), col="darkblue", lwd=2)

# Histogram of the pm25 distribution
hist(pm25, col = "lavender", xlab = "Average PM2.5 concentration ( g / m  )",
    main="Distribution of PM2.5 concentration \n in countries worldwide in
    2019", probability = TRUE,  ylim = c(0, 0.040))
lines(density(pm25), col="darkblue", lwd=2)

# Histogram of the smoking distribution
hist(smoking, col = "lavender", xlab = "Prevalence of tobacco use in 2019 (%
    of adults)", main="Distribution of the smoking rates \n in countries
    worldwide in 2019", probability = TRUE)
lines(density(smoking), col="darkblue", lwd=2)

# Scatterplot of pm25 vs. crd
par(mar = c(5, 6, 4, 2))
plot(pm25, crd, xlab="Average PM2.5 concentration ( g / m  ) in 2019",
    ylab="Chronic respiratory \n diseases death rate (2019)", main="Scatter
    plot of \n air quality vs. respiratory diseases")
```

```r
abline(v=mean(pm25), col="maroon")
abline(h=mean(crd), col="maroon")

# Scatterplot of smoking vs. crd
plot(smoking, crd, xlab="Prevalence of tobacco use in 2019 (% of adults)",
    ylab="Chronic respiratory \n diseases death rate (2019)", main="Scatter
    plot of \n smoking rate vs. respiratory diseases")
abline(v=mean(smoking), col="maroon")
abline(h=mean(crd), col="maroon")

#testing CLT
moyennes = function(donnees, taille) {
  moyennes = NULL
  for (i in 1:1200) {
    moyennes[i] = mean(sample(donnees, taille, replace=TRUE))
  }
  return(moyennes)
}
mean5=moyennes(crd, 5)
mean20=moyennes(crd,20)
mean100=moyennes(crd,100)

par(mfrow = c(2, 2))
hist(crd, probability = T, col="lavender") # crd
box()
abline(v = mean(crd), col = "maroon", lwd=2)

hist(mean5, probability = T, col="lavender") # mean5
box()
abline(v = mean(crd), col = "maroon", lwd=2)

hist(mean20, probability = T, col="lavender") #mean20
box()
abline(v = mean(crd), col = "maroon", lwd=2)

hist(mean100, probability = T, col="lavender") #mean100
box()
abline(v = mean(crd), col = "maroon", lwd=2)

### AIR QUALITY ANALYSIS ###
# We divided the sample in two sub-samples
# n_bad represents the number of countries with bad air quality (pm 2.5 <=
    19.1)
# while n_good is the number of countries in our sample with good air quality
# median(pm25) = 19.1

# H0: E(crd | pm25 > 19.1) - E(crd | pm25 <= 19.1) = 0
# H1: E(crd | pm25 > 19.1) - E(crd | pm25 <= 19.1) > 0
# In H1, we suppose that the expected value of of chronic respiratory diseases
# in countries with bad air quality is bigger than in the countries with good
    air qualities
```

```r
estim_crd_bar_good = mean(crd[pm25 <= 19.1])
estim_crd_bar_good
estim_crd_bar_bad = mean(crd[pm25 > 19.1])
estim_crd_bar_bad
estim_crd_bar_bad - estim_crd_bar_good

# The variability of the estimators of the expectations of each of the
    sub-populations
s_bad = var(crd[pm25 > 19.1])
n_bad = length(crd[pm25 > 19.1])
s_bad/n_bad
s_good = var(crd[pm25 <= 19.1])
n_good = length(crd[pm25 <= 19.1])
s_good/n_good
var_estimator = (s_bad/n_bad) + (s_good/n_good)
var_estimator
se_estimator = sqrt(var_estimator)
se_estimator

# Calculate the z-score
z = ((estim_crd_bar_bad - estim_crd_bar_good) - 0)/se_estimator
z

# Calculate the p-value
p = 1-pnorm(z)
p
# Here, p = 3.364425e-05, therefore p<0.05, therefore we can reject H0

par(mfrow=c(1,1))
curve(dnorm(x, mean = 0, sd = 1), from = -4, to = 5, xlab = "Estimator", ylab
    = "Probability Density", main = "Normal estimator of air quality and crd
    under H0")
abline(v = z, col = "maroon", lwd=2, lt=2)
abline(v = qnorm(0.95), col = "aquamarine3", lwd=2)  # Critical z
legend("topleft", legend=c("95-percentile","z-score"), col=c("aquamarine3",
    "maroon"), lty=c(1, 1), lwd=2)

# Correlation test
install.packages("car")
library(car)
dataEllipse(pm25,crd, xlab="Average PM2.5 concentration ( g / m  ) in 2019",
    ylab="Chronic respiratory \n diseases death rate (2019)", col=c("black",
    "maroon"))
cor.test(pm25, crd)

### SMOKING ###
# We divided the sample in two sub-samples
# n_sm corresponds to the number of countries in our sample with a small rate
    of smokers in the country (less than 23.4%)
# n_big stands for the number of countries where the share of smokers is
```

```
   bigger than 23.4
# median(smoking) = 23.4

# H0: E(crd | smoking > 23.4) - E(crd | smoking <= 23.4) = 0
# H1: E(crd | smoking > 23.4) - E(crd | smoking <= 23.4) > 0
# In H1, we suppose that the expected value of of chronic respiratory diseases
# In countries with a lot of smokers is bigger than in countries with lesser
   rate of smoking adults

estim_crd_bar_small = mean(crd[smoking <= 23.4])
estim_crd_bar_small
estim_crd_bar_big = mean(crd[smoking > 23.4])
estim_crd_bar_big
estim_crd_bar_big - estim_crd_bar_small

# The variability of the estimators of the expectations of each of the
   sub-populations
s_small = var(crd[smoking > 23.4])
n_sm = length(crd[smoking > 23.4])
s_small/n_sm
s_big = var(crd[smoking <= 23.4])
n_big = length(crd[smoking <= 23.4])
s_big/n_big
var_estimator2 = (s_small/n1_sm) + (s_big/n2_sm)
var_estimator2
se_estimator2 = sqrt(var_estimator2)
se_estimator2

# Calculate the z-score
z2 = ((estim_crd_bar_big - estim_crd_bar_small) - 0)/se_estimator2
z2

# Calculate the p-value
p2 = 1-pnorm(z2)
p2
# p2 = 0.172899, with p2 > 0.05, we cannot reject H0

par(mfrow=c(1,1))
curve(dnorm(x, mean = 0, sd = 1), from = -4, to = 5, xlab = "Estimator", ylab
   = "Probability Density", main = "Normal estimator of smoking and crd under
   H0")
abline(v = z2, col = "maroon", lwd=2, lt=2)
abline(v = qnorm(0.95), col = "aquamarine3", lwd=2)  # Critical z
legend("topleft", legend=c("95-percentile","z-score"), col=c("aquamarine3",
   "maroon"), lty=c(1, 1), lwd=2)

#correlation test
dataEllipse(smoking,crd, xlab="Prevalence of tobacco use in 2019 (% of
   adults)", ylab="Chronic respiratory \n diseases death rate (2019)",
   col=c("black", "maroon"))
cor.test(smoking, crd)
```

Déclaration d'honneur

Par la présente, j'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Neuchâtel et m'être renseigné-e correctement sur les techniques de citation. J'atteste par ailleurs que le travail rendu est le fruit de ma réflexion personnelle et a été rédigé de manière autonome.

Je certifie que toute formulation, idée, recherche, raisonnement, analyse ou autre création empruntée à un tiers est correctement et consciencieusement mentionnée comme telle, de manière claire et transparente, de sorte que la source en soit immédiatement reconnaissable, dans le respect des droits d'auteur et des techniques de citations.

Je suis conscient-e que le fait de ne pas citer une source ou de ne pas la citer clairement, correctement et complètement est constitutif de plagiat. Je prends note que le plagiat est considéré comme une faute grave au sein de l'Université. J'ai pris connaissance des risques de sanctions administratives et disciplinaires encourues en cas de plagiat (pouvant aller jusqu'au renvoi de l'université).

Je suis informé-e qu'en cas de plagiat, le dossier sera automatiquement transmis au rectorat. Au vu de ce qui précède, je déclare sur l'honneur ne pas avoir eu recours au plagiat ou à toute autre forme de fraude.

Neuchâtel, Décembre 2024 Les auteur-e-s.

Texte basé sur la ≪Déclaration sur l'honneur≫ du Rectorat de l'Université de Neuchâtel.