MATH 2625: Biostatistical Methods
Homework 2, due Tuesday, February 11

Please submit a PDF or .doc version of your homework to Canvas by 3:30pm on the due date. Please type *all* responses. You are encouraged to use `R` for all calculations.

Theory

1. Suppose $m = \binom{G}{2}$ tests are conducted for $G \in \mathbb{Z}^+$ and > 2. Let X denote the RV that counts the number of rejected tests out of m. Under the null hypothesis that all m null hypotheses are true, characterize the distribution of X under each of the following rejection procedures:

   (a) Testing at the nominal α of 0.05.
   When a null hypothesis is true, this means that the probability of rejecting said test at the nominal level is $p = 0.05$. Suppose Y was a random variable describing this binary outcome of either rejecting or failing to reject the null hypothesis (for each individual test). Then we can observe that

   $$Y \sim Bern(0.05).$$

   Since we defined X as the random variable counting the number of rejected tests, this means that

   $$X \sim Bin(m, 0.05).$$

   (b) Using the Bonferroni correction with nominal α.
   When we apply a Bonferroni correction to the set of multiple hypothesis tests (with all null hypotheses still being true), then we observe that the random variable X which counts the number of rejected tests takes on a binomial distribution similar to that described in part (a):

   $$X \sim Bin(m, \frac{0.05}{m}).$$

   (c) Under the Šidák correction with nominal α.
   The Šidák correction results in X being binomial similar to the results demonstrated in (a) and (b):

   $$X \sim Bin(m, 1 - (0.95)^{1/m}).$$

   For each, justify your choice and state any necessary assumptions.

2. Show Pearson's correlation coefficient, r, is a function of the OLS estimate, $\hat{\beta}_1$, i.e. show result on slide 146

$$r = \frac{s_x}{s_y}\hat{\beta}_1,$$

where s_x and s_y are the sample standard deviations of x and y, respectively.
Pearson correlation coefficent is defined as

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

The simple linear regression formula is defined as

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\epsilon \sim N(0, \sigma_\epsilon)$. When we substitute in estimators, we get

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon,$$

If we take the variance of both sides of this equation, then it becomes

$$Var(Y) = Var(\hat{\beta}_0 + \hat{\beta}_1 X + \epsilon).$$

Since β_0 is a constant, we know that $Var(\beta_0) = 0$. We can rewrite this as

$$Var(Y) = \beta_1^2 Var(X) + \sigma_\epsilon^2.$$

When we take the square root of both sides, we get

$$\sigma_Y = \sqrt{\beta_1^2 Var(X) + \sigma_\epsilon^2}$$

When we plug this into the top equation, we get

$$r = \frac{Cov(X,Y)}{\sigma_X \sqrt{\beta_1^2 Var(X) + \sigma_\epsilon^2}}.$$

Case Study

For each of the following case studies, create a structured abstract no longer than 4 pages in length (including figures, tables, and references). The Background sections are provided for each and should be included in your write-up. You must write the Methods, Results, and Conclusion sections. Code should be included in an appendix as well.

1. In this case study, you will examine data on workers in the cadmium industry, focusing on a possible relationship between the level of exposure to cadmium and an indicator of lung health. The dataset is in the `ISwR` library and named `vitcap2`. It contains the variables `group` indicating cadmium exposure status (1 for exposed $>$ 10 years, 2 for exposure $<$ 10 years, and 3 for not exposed), each workers `age` in years, and each workers `vital.capacity`, measured in liters.

2. The second case study considers a study of the lung functioning of patients with cystic fibrosis and how it relates to body mass as well as age. The data is alongside this assignment in the file `cf.txt` and contains the primary outcome of interest `fev1` (the forced expiratory volume in one second). The primary predictors of interest are `bmp` (body mass as a percent of normal binned to `very low`, `low`, and `near normal`) and `weight` (in `kg`). A possible confounder of interest, age in years, is included as well.

Background

Vital capacity, a measure of lung volume defined as the maximum amount of air a person can expel from their lungs after taking a maximum inhalation, can be used to help diagnose underlying lung disease or dysfunction. If vital capacity is lower than expected, further testing for possible disease or other complications is warranted. Exposure to cadmium, even at low doses, may have impacts on kidney and bone health and at high levels may damage the lungs. Cadmium production increased considerably over the course of the 20th Century and is used across a number of different industries. To assess the potential impact exposure to cadmium might have, we examined the vital capacity in workers from industries that use cadmium in various production processes. Workers were randomly selected from nine different factories: eight PVC factories using cadmium stabilizers in the compounding of PVC and one cadmium-nickel battery factory. We classified the workers as having long term exposure (> 10 years exposure to cadmium), short to medium term exposure (< 10 years exposure), and no exposure. Our primary goal is to see if exposure to cadmium impacts the lung functioning of the workers in the industry. It is well known, however, that vital capacity also depends on a number of additional factors, including a person's age. Older workers may also be more likely to have longer periods of exposure, thus we also examine workers' age as a possible confounder.

Methods

We tested the vital capacity (VC) of workers who were identified as either having long-term exposure (exceeding 10 years) or short term exposure (less than 10 years). We also randomly measured vitals from workers who had no exposure to form a control group. 12 workers with long-term exposure, 28 with short-term exposure, and 44 with no exposure were selected for the study.

Analysis of the differences between vital levels was done nonparametrically using a Kruskal-Wallis Test at the nominal level. We also planned to conduct any necessary post-hoc tests using a Mann-Whitney U Test, with a Bonferroni correction to control for multiple tests.

As a secondary analysis, we checked the influence of age as a confounding variable using Spearman's rank coefficient. We checked for correlation between age and VC of all workers, and then separately in their respective groups (long-term exposure, short-term exposure, and the control group). A Bonferroni correction was applied to these four Spearman correlation tests, resulting in a corrected critical threshold of $\alpha^* = 0.0125$.

Results

There was a greater variation in the vital capacities of workers who experienced long-term cadmium exposure. Additionally, one worker who faced short-term exposure had an unusually low VC. This outlier is shown in figure 2, along with the distributions of the boxplots of all the study cohorts. We did not find evidence to suggest that exposure to cadmium made a difference in the vital capacity of workers ($p = 0.2477, H = 2.790, df = 2$).

Table 1: **Summary of Vital Capacity of Workers in Every Group**

	All Workers	**Long-term Exposure**	**Short-term Exposure**	**Control**
Number of Workers	84	12	28	44
Mean Age *(Years)*	40.55	49.75	37.79	39.80
Mean VC *(Liters)*	4.392 ± 0.164	3.949 ± 0.656	4.472 ± 0.682	4.462 ± 0.692
VC St. Dev. *(Liters)*	0.757	1.033	0.682	0.692
Median VC *(Liters)*	4.530	3.865	4.615	4.530
MAD *(Liters)*	0.756	1.460	0.667	0.689

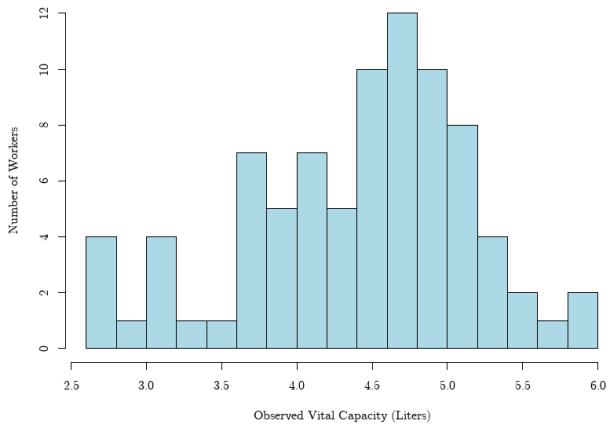


Figure 1: Histogram of Workers' Vital Capacity

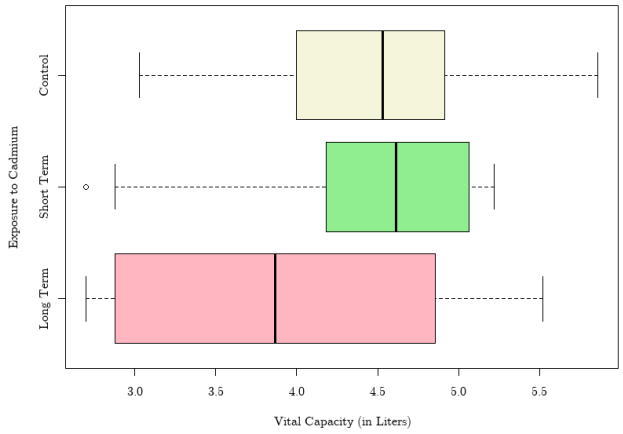


Figure 2: Boxplots of Vital Capacities Across Cohorts

We found evidence to suggest that age is a confounding variable for determining vital capacity in this study. Worker age and vital capacity appear to be negatively correlated ($p < 0.001, \rho = -0.544$). Figures 3 and 4 visualize this association, which also held true in all three cohorts of the study. Table 2 provides a summary of the negative correlation found between all exposure levels. Overall, age had a strong impact on the workers' vital capacity regardless of exposure to cadmium.

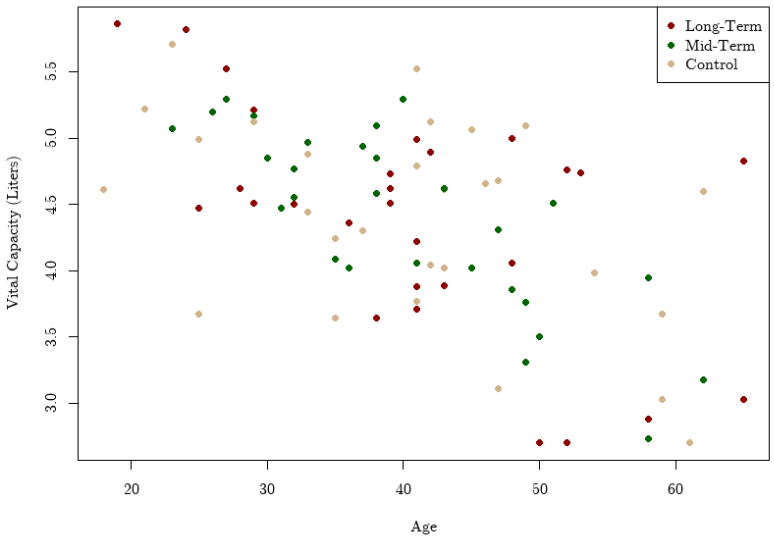
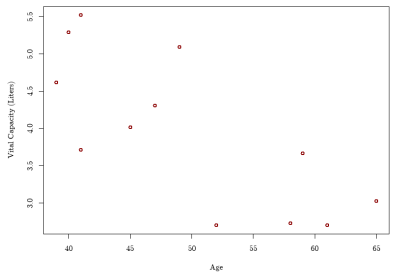


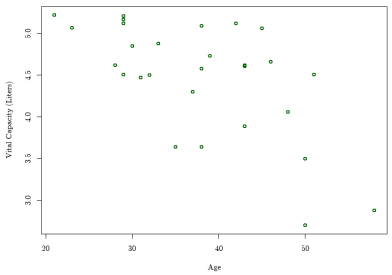
Figure 3: Worker Age and Vital Capacity Appear to be Negatively Correlated

Table 2: **Correlation Found Between Age and Vital Capacity Across all Exposure Levels**

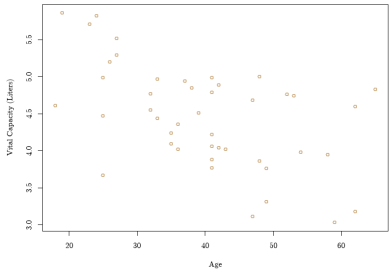
Group	**Spearman's ρ**	**p-value**	**Significant at $\alpha^* = 0.0125$**
All Workers	-0.544	< 0.001	Yes
Long-Term Exposure	-0.747	0.005	Yes
Short-Term Exposure	-0.557	0.002	Yes
Control	-0.494	< 0.001	Yes



(a) Long-Term Exposure



(b) Short-Term Exposure



(c) Control Group

Figure 4: Correlation between Age & Vital Capacity Found Among All Subgroups of Study

Discussion

Although no difference in vital capacity was observed as a result of the different exposure lengths to cadmium, we observed substantial impacts from a worker's age on their resulting vital capacity. This lack of statistically significant difference is corroborated by related findings in the literature. Though Yang and colleagues (2019) also did not find a significant difference in forced vital capacity (in short, a related but more effortful measure of vital capacity) per 1 μg/L blood serum cadmium in their multivariable regression models, they did find significant negative differences in other lung function metrics, patient-predicted FEV_1 and FEV_1/FVC, with each additional μg/L cadmium. Wang and colleagues (2024) support this result, finding in their meta-analysis that per 1 μg/L cadmium exposure, FEV_1/FVC level decreased by 47.54%, which was a statistically significant result. These findings may suggest that in order to capture the full picture of cadmium exposure and lung function, future research should explore including additional metrics like FEV_1 or FEV_1/FVC alongside vital capacity.

In terms of age and vital capacity, it is well documented that vital capacity declines with age over time (Cid-Juárez et. al., 2019). It is worth noting that the mean age of workers who faced long-term exposure was higher than both workers who only faced short-term exposure, workers with no exposure, and the overall group in general. This was one of the major limitations of the study and may have been due to the small number of workers who had long-term exposure. While it makes sense that workers who faced exposure for a longer period of time, a follow-up study should seek to address this by also finding older workers who faced short-term exposure and no exposure. It should also seek to sample a larger number of workers overall, as well as workers from a larger number of factories. This would help to address the issues related to internal and external validity that the study faces. Other factors found to influence vital capacity include gender, athleticism, and other pre-existing medical conditions that vary from person to person (Lemon & Moersch, 1924; Abboud et. al., 2024), so future research should monitor these variables as well for possible confounding.

Appendix

- Abboud, H. J., Hussein, H. K., & Fadhil, S. A. (2024). The effect of increasing the intensity of specialized endurance training on runners' ability in the advanced 1500-meter run in terms of vital capacity indicators (VC) and heart rate (SV). *TechHub Journal*, 7, 19–27.
- Cid-Juárez, S., Thirión-Romero, I., Torre-Bouscoulet, L., Gochicoa-Rangel, L., Martínez-Briseño, D., Hernández-Paniagua, I. Y., Delgadillo-Ruiz, O., Guerrero-Zúñiga, S., Del Río-Hidalgo, R., Cortés-Medina, D., Bapo-López, P. E., León-Gómez, P., Bautista-Bernal, A., & Pérez-Padilla, R. (2019). Inspiratory capacity and vital capacity of healthy subjects 9–81 years of age at moderate-high altitude. *Respiratory Care*, 64(2), 153–160. https://doi.org/10.4187/respcare.06284
- Lemon, W. S., & Moersch, H. J. (1924). Factors influencing vital capacity. *Archives of Internal Medicine*, 33(1), 136–144. https://doi.org/10.1001/archinte.1924.00110250139014
- Wang, Y., Wang, D., Hao, H., Cui, J., Huang, L., & Liang, Q. (2024). The association between cadmium exposure and the risk of chronic obstructive pulmonary disease: A systematic review and meta-analysis. *Journal of Hazardous Materials*, 469(133828). https://doi.org/10.1016/j.jhazmat.2024.133828
- Yang, G., Sun, T., Han, Y-Y., Rosser, F., Forno, E., Chen, W., & Celadon, J. C. (2019). Serum cadmium and lead, current wheeze, and lung function in a nationwide study of adults in the United States. *Journal of Allergy and Clinical Immunology: In Practice*, 7(8), 2653–2660. https://doi.org/10.1016/j.jaip.2019.05.029

R Code

```
# Install necessary packages
# install.packages('DescTools')
# install.packages('Rfit')
# Disclaimer: I used chatGPT to be able to display this in latex, hope that
    's okay
library(ISwR) # data source

# Formatting for LaTeX output
library(showtext)
library(DescTools)
library(Rfit)
font_add(family = 'ComputerModern', regular = 'cmunrm.ttf') # for
    consistent formatting in LaTeX
showtext_auto()
par(family = 'ComputerModern')

# Load dataset
data('vitcap2')

# Subset data by exposure level
noExposure = subset(vitcap2, group == '3')
midExposure = subset(vitcap2, group == '2')
longExposure = subset(vitcap2, group == '1')
exposures = list(vitcap2 = vitcap2,
                                longExposure = longExposure,
                                midExposure = midExposure,
                                noExposure = noExposure)

# Loop through each group and compute summary statistics
for (group_name in names(exposures)) {
  group_data = exposures[[group_name]]

  vital_sd <- sd(group_data$vital.capacity)
  n <- length(group_data$vital.capacity)

  margin = qt(0.975, df = n - 1) * (vital_sd / sqrt(n))  # margin for 95%
      CI

  # Print results
  print(paste('Group:', group_name, '( N =', n, ')'))
  print(paste('  Mean Age:', round(mean(group_data$age), 3)))
  print(paste('  Mean Vital Capacity:', round(mean(group_data$vital.
      capacity), 3),
```

```
                          '| SD:', round(vital_sd, 3)))
  print(paste('  95% CI MoE:', round(margin, 3)))
  print(paste('  Median VC:', median(group_data$vital.capacity)))
  print(paste('  MAD:', mad(group_data$vital.capacity)))
  print('')
}

# Histograms and boxplots
hist(vitcap2$vital.capacity, col = 'lightblue',
         breaks = 12, ylab = 'Number of Workers',
         xlab = 'Observed Vital Capacity (Liters)',
         main = '')

boxplot(vital.capacity ~ group, data = vitcap2,
                main = '',
                xlab = 'Vital Capacity (in Liters)',
                ylab = 'Exposure to Cadmium',
                col = c('lightpink', 'lightgreen', 'beige'), horizontal = T
                    ,
                names = c('Long Term', 'Short Term', 'Control'))

kruskal.test(vital.capacity ~ group, data = vitcap2)

# Scatter plots and correlation
plot(vital.capacity ~ age, data = vitcap2,
         col = c("darkred", "darkgreen", "tan"),
         pch = 16,
         xlab = 'Age', ylab = 'Vital Capacity (Liters)',
         main = '')

legend("topright", legend = c("Long-Term", "Mid-Term", "Control"),
           col = c("darkred", "darkgreen", "tan"), pch = 16)

# Individual scatter plots
plot(vital.capacity ~ age, data = longExposure, col = 'darkred',
         xlab = 'Age', ylab = 'Vital Capacity (Liters)', lwd = 2)

plot(vital.capacity ~ age, data = midExposure, col = 'darkgreen',
         xlab = 'Age', ylab = 'Vital Capacity (Liters)', lwd = 2)

plot(vital.capacity ~ age, data = noExposure, col = 'tan',
         xlab = 'Age', ylab = 'Vital Capacity (Liters)', lwd = 2)

# Spearman correlations
with(vitcap2, cor.test(age, vital.capacity, method = 'spearman' ))
with(longExposure, cor.test(age, vital.capacity, method = 'spearman' ))
with(midExposure, cor.test(age, vital.capacity, method = 'spearman' ))
with(noExposure, cor.test(age, vital.capacity, method = 'spearman' ))
```

Background

Cystic fibrosis is a genetic disorder that affects the lungs, among other major organs. A mutation in the cystic fibrosis transmembrane conductance regulator (CFTR) gene is the cause, resulting in the dysfunction of the CFTR protein which facilitates the transportation of chloride to the cell surface. Without the additional chloride, mucus lining the cells of vital organs becomes thick and viscous, clogging airways and resulting in difficulty breathing. Additional complications in increased risk of bacterial infection and respiratory failure. Poor growth or weight gain is a common complication as well, and often an early indicator of cystic fibrosis. To assess the relationship between growth and weight and lung functioning in patients with cystic fibrosis, we recruited 23 patients between the ages of 7 and 23 years old. Our primary outcome of interest is the forced expiratory volume in one second (FEV1), a measure of how much air the lungs can expel within the first second of exhaling. We wish to relate FEV1 to two predictors of interest related to growth and weight, specifically the body mass category of the patient and their weight (kg). Cystic fibrosis patients with poorer growth relative to normally developing children and lower weight may have lower lung functioning. Age may be potential confounder as lung functioning tends to improve with age and weight should increase as the patient gets older. Thus we will also examine potential age confounding in our analysis.

Methods

25 patients between the ages of 7 and 23 were recruited for this study. The participants' age, body mass as a percent of normal (aggregated into very low, low, & near normal), age, and forced expiratory volume (in one second) was recorded. 9 participants were under the age of 13, 12 participants were in their teens (13-19), and 4 were at or above the age of 20.

We used Spearman's rank coefficient to examine potential correlation between weight and forced expiratory volume. As a sensitivity analysis, we checked the results using Kendall's tau. To study the effects of body mass percentage (bmp), we utilized a Kruskal-Wallis test to determine whether there was a difference in FEV1 between patients with a very low bmp, low bmp, or near normal bmp. We planned to conduct any necessary post-hoc comparison tests using the Mann-Whitney U test, with a Bonferroni correction applied to these follow-up analyses ($\alpha^* = 0.0167$).

To check for the confounding impact of age, we also analyzed correlation between subjects' age and recorded FEV1 using Kendall's Tau. Aside from post-hoc analyses, all tests were conducted at the nominal level.

Results