

PH 538: Biostatistical Methods I (Fall 2016)

Take-Home Project (Due date: Dec. 15, 2016)

Dr. Fares Qeadan: fqeadan@salud.unm.edu

Objectives:

In this project, students will write a scientific paper corresponding to the structure and contents below. More information about the structure of scientific papers and writing guidelines could be found in:

1. <http://www.nature.com/scitable/ebooks/english-communication-for-scientists-14053993/contents> [Unit 2]
2. http://classes.geology.illinois.edu/11SprgClass/geo440/Writing_Scientific_Papers.pdf
3. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3474301/pdf/ijspt-07-512.pdf>

Background on the data set:

This data set records are related to respiratory function and smoking for 654 children ages 3 to 19, who were seen in the Childhood Respiratory Disease Study in 1980 in East Boston, Massachusetts. "Forced Expiratory Volume" (FEV) is the amount of air a person can exhale in the first second of a forceful breath, and indicates the person's level of respiratory function. In particular, low FEV1 values are indicative of breathing problems and onset of asthma or possible respiratory dysfunction. The data is a subset of data collected by I. Tager, S. Weiss, A. Munoz, B. Rosner, and F. Speizer, and published in:

1. Tager, I., Weiss, S., Munoz, A., Rosner, B., and Speizer, F. (1983). Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function. *New England Journal of Medicine*, 309(12), 699-703.
2. Tager, I., Weiss, S., Rosner, B., and Speizer, F. (1979). Effect of Parental Cigarette Smoking on the Pulmonary Function of Children. *American Journal of Epidemiology*, 110(1), 15-26.

These studies were among the first to show clear evidence of the impact of smoking and exposure to second-hand smoke on respiratory health in children. One of the authors, Bernard Rosner, included this particular subset of the data in his book (1999), *Fundamentals of Biostatistics*, 5th ed., Pacific Grove, CA: Duxbury.

More information about this data set could be found at:

<ftp://linhrc3.hrc.es/pub/bioest/charlas/datasets.kahn.pdf>

The list of variables in the data set:

	Variable	Description
1	id	Subject ID
2	age	Age of subject in years
3	fev	FEV1, forced expiratory volume (liters)
4	height	Height of subject in inches
5	sex	Gender of subject (male and female)
6	smoke	Current smoking status (non-current smoke and current smoker)

Paper structure:

Your paper must be organized as follows:

1. Cover sheet with your names, class, date and title.
2. At least five typewritten (size 12 font and double-spaced) pages not including the appendix.
3. At least five references (Internet references are accepted with the understanding that the use of copy-and-paste method will be considered as plagiarism and may result in zero credit for the project).
4. Your report must have: Abstract, Introduction, Methods, Results, Discussion and Conclusion (plus an appendix for the used STATA code). [See the three links in the objectives section above for more help on how to write each of these components].
5. All tables and figures must be numbered and given captions.

Paper contents:

Your paper must answer and discuss the following points:

1. The population of interest
2. Numerical and Graphical descriptive statistics for the sample data. Make sure you make comments about the normality of the continuous variables in the data set. Also, make sure to provide the 95% C.I. around the true parameters we are trying to estimate.
3. Create an age categorical variable with the categories 3-9 (children), 10-14 (young adolescents), and 15-19 (teenagers)
4. Create an age categorical variable with the categories 3-9 (children) and 10-19 (adolescents)
5. Test if the mean FEV values is larger among adolescents as opposed to that among children?
6. Are gender and smoking independent? Also, is mean FEV higher among males than females (test the assumption of equality of variances first)?
7. Test whether the mean FEV values is larger among current smokers as opposed to non-current smokers among adolescents? Make comments on this question if $\alpha=0.10$ was used instead of $\alpha=0.05$. Before conducting the test, please examine the data for equal variances for FEV within the two groups (current smokers and non-current smokers) when the age is greater or equal to 10 (i.e. among adolescents).
8. Test whether age (larger or equal to 10 years old) and smoking are independent or dependent variables? Are older subjects in the study more likely to smoke? Conduct a chi-square test between smoking status and the categorical age variable if $\text{age} \geq 10$.
9. FEV of less than 2.5 L is regarded as showing some functional impairment (occasional breathlessness, inability to climb stairs etc.). Create a categorical variable for FEV such that FEV less than 2.5 is called Low-FEV and FEV of 2.5 or more is called High-FEV. Is this new variable independent from the smoking variable?
10. Create a categorical variable for height such that height values of 62 or more are called tall and height values less than 62 is called short. Is FEV dependent on height? Regardless whether it's or not, please test if the mean FEV is higher among tall subjects compared to short ones?
11. Is the proportion of current smokers higher among tall subjects if the age is 10 or more? You may need to create a numeric current smoking variable and a numeric categorical height variable before conducting this test to avoid errors in STATA.
12. Conduct similar tests as the ones above if it make sense among children (i.e. age is less than 10)
13. Make sure you examine the validity of the assumptions of all conducted tests above.