**BIOM 505-011: Biostatistical Methods I**
**Take-Home Project (Due Dec. 13, 2018 5 pm)**
**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://www.csun.edu/~msteele/classes/marine_ecology/handouts/writing%20scientific%20papers.pdf
3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3474301/pdf/ijspt-07-512.pdf

## Background on the data set:

The presented data are related to 27 adults of the ages 45 to 85 who were seen at the bronchoscopy suite at the UNM Hospital (UNMH), for endobronchial ultrasound bronchoscopy with fine needle aspiration biopsy (EBUS-FNA) of enlarged lymph glands located adjacent to the central airway. This convenience sample-data makes up a prospective single-center study for which investigators have collected dedicated EBUS-FNA biopsies from 27 patients with enlarged paratracheal and hilar lymph nodes. Malignant and benign histologic diagnoses were established in 15 and 12 patients, respectively (Saeed et al. 2016); of the malignant ones elven are lung cancer. For these 27 patients, 34 cytokines were assayed using Bio-Plex Pro human cancer biomarker panels, in a Bio-Rad 200 suspension array system. A mean cytokine value was taken from each subject with more than 1 lymph node station EBUS-FNA biopsies. From this pilot study, Saeed et al. have identified a novel panel of seven cytokines (sVEGFR-1, IL-6, VEGF-A, Angiopoeintin-2, uPA, sHER-2/neu and PLGF), with elevated levels in lymph nodes, as a biomarker for identifying lung cancer metastasis. These biomarkers will allow for an early diagnosis and development of an aggressive treatment plan for these patients with lung cancer metastasis, which will ultimately act to improve morbidity and mortality.

1. Saeed, A.I., Qeadan, F., Sood, A., VanderJagt, D.J., Mishra, S.I., Hill, D.A., Peikert, T. and Sopori, M.L., 2017. A novel cytokine profile associated with cancer metastasis to mediastinal and hilar lymph nodes identified using fine needle aspiration biopsy–A pilot study. *Cytokine*, 89, pp.98-104.

**Link to article: http://www.mathalpha.com/BIOM-505/Saeed_et_al-2017-Cytokine.pdf**

**Link to download the dataset: http://www.mathalpha.com/BIOM-505/cytokines.sas7bdat**

## The list of variables in the data set:

| # | Variable | Description |
|---|---|---|
| **1** | Serial_Number | Subject ID |
| **2** | Age | Age of subject in years |
| **3** | sex | Gender of subject (0=female, 1=male) |
| **4** | Smoking | pack year history of smoking |
| **5** | Lymph_Node_Station | 4R: right lower paratracheal lymph node station, 4L: left lower paratracheal lymph node station, 7: subcarinal lymph node station, 10L and 11L: left hilar lymph node station, 10R and 11R: right hilar lymph node station. |

| | | |
|---|---|---|
| **6** | Histology_Diagnosis | Benign, Small Cell, Squamous Cell Cancer (SCC), adenocarcinoma (AD) |
| **7** | cancer | 0=Benign, 1=Malignant |
| **8** | Angiopoietin_2 | List of 34 cytokines |
| **9** | EGF | |
| **10** | Endoglin | |
| **11** | HB_EGF | |
| **12** | FGF_basic | |
| **13** | Follistatin | |
| **14** | G_CSF | |
| **15** | HGF | |
| **16** | Leptin | |
| **17** | Osteopontin | |
| **18** | PDGF_AB_BB | |
| **19** | PECAM_1 | |
| **20** | Prolactin | |
| **21** | SCF | |
| **22** | sEGFR | |
| **23** | sHER_2neu | |
| **24** | sIL_6Ra | |
| **25** | sTIE_2 | |
| **26** | sVEGFR_1 | |
| **27** | sVEGFR_2 | |
| **28** | IGFBP_1 | |
| **29** | IL_18 | |
| **30** | IL_6 | |
| **31** | IL_8 | |
| **32** | PAI_1 | |
| **33** | PLGF | |
| **34** | TGF_a | |
| **35** | TNF_a | |
| **36** | VEGF_A | |
| **37** | VEGF_C | |

| | | |
|---|---|---|
| **38** | VEGF_D | |
| **39** | sCD40L | |
| **40** | sFASL | |
| **41** | uPA | |

**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 seven 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 SAS 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 focus and discuss at least one of the points in bold below:

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 <65, and 65+.
4. Create a smoking categorical variable with the categories ≤20 and 20+ pack year history of smoking.
5. Create a categorical variable indication 1. lung cancer, 2. Other types of cancer, and 3. No cancer.
6. Create a categorical variable as you see appropriate for the Lymph Node Station.

7. **Is there a health disparity in the cytokines profile for subjects by sex? Identify an expression of cytokines which differentiates between sex categories (either among all 27 subjects, or within only those with cancer or benign, or within only those with lung cancer, or any other group of your interest).**

8. **Is there a health disparity in the cytokines profile for subjects by age? Identify an expression of cytokines which differentiates between age categories (either among all 27 subjects, or within only those with cancer or benign, or within only those with lung cancer, or any other group of your interest).**

9. **Is there a health disparity in the cytokines profile for subjects by smoking status? Identify an expression of cytokines which differentiates between smoking categories (either among all 27 subjects, or within only those with cancer or benign, or within only those with lung cancer, or any other group of your interest).**

10. **Saeed et al. have identified optimal cut-offs for an expression/biomarker of seven cytokines (sVEGFR-1, IL-6, VEGF-A, Angiopoeintin-2, uPA, sHER-2/neu and PLGF) as shown in Table 3. These cut-offs were able to discriminate between malignant and benign cases of cancer with very high sensitivity and high discrimination power as shown in Figure 2 of their paper. (a)**

**Create 7 binary variables for sVEGFR-1, IL-6, VEGF-A, Angiopoeintin-2, uPA, sHER-2/neu and PLGF based on the cut-offs listed in Table 3, (b) test if any of these 7 binary variables are associated with either age, gender, and smoking status, (c) construct 3 graphs (i, ii, and iii) similar to than in Figure 2 where the x-axis has the frequency for age, gender, and smoking categories respectively for graphs i, ii, and iii.**

11. **Saeed et al. have identified an expression/biomarker of seven cytokines (sVEGFR-1, IL-6, VEGF-A, Angiopoeintin-2, uPA, sHER-2/neu and PLGF) that discriminate between cancer and benign cases. Their approach was based on comparing statistically the differences in medians of 34 cytokines between malignant and benign subjects. Instead, please identify an expression/biomarker of cytokines that discriminate between cancer and benign cases based on differences in the variability of cytokines rather than their concentration (i.e. compare variances/ standard deviations instead of means or medians).**

12. **Since some cytokines are highly correlated with each other, it's recommended to work with the ratio or difference of the correlated pairs of cytokines. Based on some correlation analysis, we have determined that the following pairs of cytokines are highly correlated and thus one need to repeat the analysis of Saeed et al. (Table 2) using either the ratio or difference of the following analytes:**

    1. il_6/IL_8
    2. VEGF_A/Angiopoietin_2
    3. uPA/IGFBP_1
    4. sHER_2neu/sVEGFR_1
    5. PLGF/Osteopontin
    6. TGF_a/sCD40L
    7. HGF/VEGF_C
    8. HB_EGF/Follistatin
    9. VEGF_D/sFASL
    10. Leptin/PDGF_AB_BB
    11. IL_18/sIL_6Ra
    12. sVEGFR_2/sEGFR
    13. EGF/Endoglin
    14. Prolactin/SCF
    15. PECAM_1/sTIE_2
    16. TNF_a/FGF_basic
    17. G_CSF/PAI_1

**NOTES:**

a. Due to sample size by groups, adjusting for multiple comparisons is not recommended for this project.
b. If you decide to compare means, you may need to log transform the data as most cytokines are not normally distributed. Test normality first.
c. If you don't transform the data, then use the data as is and compare medians instead unless the provided data are normally distributed to start with.
d. If you would like to graphically describe something like the differences presented in Table 2 of Saeed et al., then use Figures 1-3 in Fakhry et al.: **http://www.mathalpha.com/BIOM-505/Fakhry_et_al-2018-The_Laryngoscope.pdf**
e. As an extra credit, one could calculate the probabilities that a randomly selected person with similar characteristics to those in this study will exceed the cut-offs in Figure 2 of Saeed et al. which will act as a proxy for the chance of malignant cancer (recall shaded areas and assume normality).
f. This paper is expected to be written in a fashion similar to a manuscript to be submitted for publication.
g. Please don't use or distribute the provided data set for other purposes without a written permission from Dr. Qeadan and Dr. Saeed.