Statistical Analysis Laboratory

BioRadio Laboratory Course System Version 7.0

Introduction

Throughout this laboratory course you will be collecting physiological data. Typically, the collection of physiological data is used for two purposes. It can be used to control another system such as in the case of recordings from voluntary muscles being used as an input to a myoelectric prosthetic for an amputee. More often, physiological data is collected in order to make a diagnosis. For example, an electrocardiogram (ECG) can be recorded to determine if someone has a particular cardiac arrhythmia. Often times, based on visual observations, you will be able to make judgments about theories and hypotheses. Statistics will allow you to make future guesses about a particular data set. For example, suppose that you find a new gene that exists in people with a particular type of cancer. You may want to be able to predict how many other people will get that kind of cancer based on if they have that gene. However, other times you will want to be able to state a conclusion with a particular certainty or state that a hypothesis is true within a particular tolerance. Statistical methods will allow you to quantitatively state the results of your data. These methods and tools are described in detail below.

Equipment required:

- Disposable Supply Kit
- Lab Course Software
- Microsoft® Excel, MATLAB® , or LabVIEW™

Background

There are several terms that you should become familiar with before starting the methods section of this laboratory course session. These statistical terms are described below:

Mean

The sample mean is the average of a set of *n* values of a given sample. It is given by the following formula:

$$m \equiv \frac{1}{n}\sum_{k=1}^{n} x_k,$$

In this example, *m* is the mean, *n* is the total number of samples, and *x* is the vector which represents all of the samples.

Variance

The variance of a set of data is given by the following formula, where *m* is the mean:

$$s_{N-1}^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

The term x(bar) refers to the sample mean.

Standard Deviation

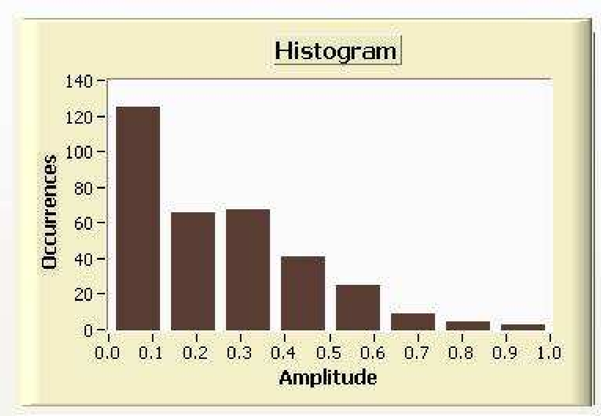
The standard deviation of a set of data is the square root of the variance, where x(bar) refers to the mean of the sample:

$$s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

The standard deviation is often used as the fluctuation for a sample when data is being collected in an experiment. The results taken from data are frequently written as the mean ± standard deviation.

Histogram

A histogram groups data into bins and then plots the numbers of elements or counts in each bin as a function of the maximum and minimum levels to qualify for that bin. A histogram can provide information about the probability density function of a data set. In other words, in what region of numbers will the data most likely be located?



Experimental Design

Experimental design is the process used in research to find answers to a question or problem. First a problem or question is identified. Next, a hypothesis is created based on that question. The hypothesis is then tested experimentally and conclusions are made based on the results.

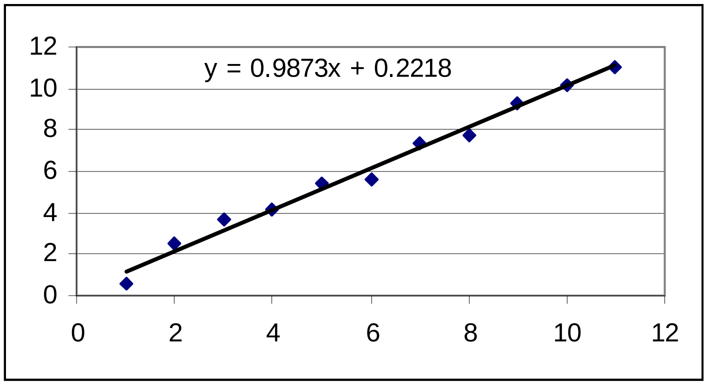
There can be several sets of factors in an experimental design. Factors can be thought of as variables in an experiment and each variable can have several different levels. For example, consider the problem of a pharmaceutical company completing a clinical study with a new drug that is supposed to prevent the flu. Therefore, the output of interest would be how many people get the flu. The factors in the experiment could be whether or not the person took the new drug, the dose of the drug, the age of the subjects, and so on. All of these factors must be taken into consideration when designing the experiment. You must consider any factors that you expect could have a significant effect on the outcome measure and include them as factors in the testing. Then with statistical methods you can determine whether each factor significantly contributed to the result or not.

Regression

Regression analysis is completed to try and fit a relationship between certain variables that are collected during an experiment. The developed relationship is explained in the form of an equation. The term regression refers to the method that one or multiple variables are regressed against another variable.

Linear Regression

Linear regression refers to when a straight line is fitted to a set of data points to determine the effect of one variable on another variable. The most common form of linear regression uses least squares fitting. An example of a linear regression of a data set is shown below:



Multiple Linear Regression

Multiple linear regression gives conditional values of one variable in terms of two or more other variables. It is generally of the form:

$$Y = \beta_o+\beta_1*x_{1+}\beta_2*x_2+\ldots+\beta_k*x_k$$

Least Squares Fitting

Least square fitting is a method for determining the best fit equation for a data set by minimizing the sum of squares of the residuals. A residual is the difference between the actual data point and the point calculated by a regression. In the linear regression above, the actual data for an input of 8 is 7.9. However, the equation calculated by the linear regression predicts that it will be 8.1. Thus, the residual here is equal to 0.2. By minimizing the squared sum of the residuals for all data points the best fit can be found.

Correlation

Correlation describes how two or more quantities are linearly associated. Types of correlation include cross-correlation and autocorrelation. To understand correlations, it is first helpful to understand convolution:

Convolution

Convolution expresses the amount of overlap of a function *g* as it is shifted over a function *f*. The equation for a convolution over an infinite range is shown below:

$$f * g \equiv \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau = \int_{-\infty}^{\infty} g(\tau)f(t-\tau)\,d\tau$$

In engineering, the output of a linear system is the convolution of the input signal and the system's impulse response.

Cross Correlation

The cross correlation of two functions is shown below:

$$f \star g \equiv \bar{f}(-t) * g(t),$$

The term f(bar) refers to the complex conjugate of f(t). Cross-correlation can be used to determine similarities between two signals or to remove artifact from a noisy signal.

Autocorrelation

Autocorrelation is defined by:

$$\rho_f(t) \equiv f \star f = \bar{f}(-t) * f(t) = \int_{-\infty}^{\infty} f(t+\tau)\bar{f}(\tau)\,d\tau,$$

The autocorrelation contains information similar to a power spectrum. It contains no information about phase and thus, is not reversible. Autocorrelation is most useful for finding the periodic components of data and computing power spectra.

Experimental Methods

Experimental Setup

You should make sure that your BioRadio receiver is connected to your computer before starting this laboratory session. If the receiver is not connected most of the functionality in the laboratory session will be disabled.

Procedure and Data Collection

1. Run the Lab Course software. Log in and select the "Statistical Analysis" laboratory session under the Engineering Basics subheading and click on the "Begin Lab" button.

2. Click on the tab labeled "Basics". This section utilizes previously collected force and EMG data to illustrate some basic statistical concepts.

3. First, click on "Run Stats". A plot of the EMG and force will be shown in the plots on the left hand side. Additionally, plotted in the lower right side will be another plot of the EMG signal.

4. The mean, standard deviation of each parameter can automatically be calculated by the software. The time frame over which each statistical parameter can be modified by adjusting the start time and the stop time of the statistical interval. Try adjusting these values to see the effect on the mean, standard deviation, and variance of each of the signals.

5. Next we will examine the effects of windowing on the EMG signal. A windowing function provides a smoothing of the data by averaging adjacent data points. The number of data points that are averaged has an effect on the amount of smoothing that occurs. This windowing function acts similar to a low pass filter. Set the "window size" control to several different values, click on Run Stats each time you reset the window size. Capture a few screen shots of different window sizes to your report.

6. Another useful tool for analyzing and presenting data is the histogram. Click on the sub tab labeled "Visualization" on the software menu. The histogram provides a convenient tool for visualizing where particular features of data lie. This section of the lab allows you to define different bin sizes for either the EMG or force data and plot it as a histogram. Select different bin sizes for both the EMG and force data and click on the "Visualize" button to update the plots. Notice any trends in either of the data that can be observed from the histogram.

7. Now click on the "Regression" sub tab. You will complete a regression analysis using the force and EMG data. Instead of plotting each recorded variable against time, the normalized EMG levels will now be plotted as a function of force. The regression will attempt to fit equations to the force versus EMG data. The regressions will fit a linear relationship, an exponential fit, and a polynomial fit to the data. The order of the polynomial fit can be adjusted with the poly order input. Make sure that the switch is set to "model only" and click on the green "Regress" button. Capture a screen shot of this to your report. Note the different shapes of the regression curves based on the fit type. The mean squared error (MSE) for each type of fit is shown next to the plot legend.

8. The results of a regression analysis are often used to make predictions about future data points. Existing data can be used to fit an equation to the data points and that equation describes a relationship between the data. An assumption can then be made that future data points will have a similar relationship. Therefore, if you plug the value of one variable of a newly collected data point into the equation you should be able to predict the other variable. For example, once you record a new force, you may be able to enter that number into the equation and predict what the EMG level was that produced the force. Therefore, when a person creates a model using collected data, they typically use a certain percentage of the data to create the model, but leave a certain percent aside to test the generalization of the model. Change the switch to "Model and Generalization". You will now be able to set aside a certain percentage of the data to create the model and the rest to test the generalization of the model. Set the percentage to 80 and click on "Regress".

9. The generalization of the model is applied only using the linear fit. A plot of the residuals for the model data is also now illustrated. Try using a different percentage of the model data as the model and generalization sets and notice the effects on the error between the actual and predicted data points.

10. Now click on the "Correlation" sub tab. There are several tools available to analyze correlations that exist in collected data sets. For this example you will be able to analyze two different sets of data, physiological waveforms or simulated waveforms by using the switch in the top left hand corner. In addition, you will also be able to add noise to your data files to examine its effect on the correlation analysis.

11. The physiological data set contains one minute of data from a recorded overnight sleep study. The available data channels include electrocardiography (ECG), blood pressure, electroencephalography (EEG), nasal airflow, abdominal breathing effort, electro-oculography (EOG), and electromyography (EMG). Let's start by examining the physiological data set, so make sure the switch is set to the "physiological waveforms".

12. Make sure that both channel 1 and channel 2 are set to ECG and that the noise amplitude is set to zero for both channels. Now click on the "Run Analysis" button.

13. Several plots are created by the software. These include the raw data signals plotted versus time (top left), a convolution between the two signals, a cross correlation between the two signals, a signal power spectrum for each channel (top right), a spectrum cross correlation, and an autocorrelation of channel 1. Because you set both channels to the same ECG data you should only notice one channel plotted in the top right plot. Additionally, you should notice several important features of the other plots since the same data channel is being used. Capture a screen shot of this for your report.

14. Now set channel 2 to be "Blood Pressure" and click on the "Run Analysis" button again. Notice what changes in the plots. Capture a screen shot of this for your report.

15. You should now compare the following list of signal 1 and 2 combinations and capture a screen shot of each for you report. You will need this data to complete the discussion questions that follow:

    a. ECG vs. blood pressure
    b. ECG vs. EEG
    c. ECG vs. airflow
    d. EEG vs. blood pressure
    e. Airflow vs. abdominal respiratory effort
    f. Airflow vs. EOG
    g. EOG vs. EMG
    h. EMG vs. EEG

16. Now change the switch at the top left hand corner to be simulated waveforms. Set both waveforms to be sine waves with a frequency of 2 Hz, an amplitude of 1, with zero noise and zero phase. Then click on run analysis and capture a screen shot of the result to your report.

17. Change the noise level on channel 2 to 5. Rerun the analysis, notice any changes in the plots, and capture a screen shot of this to your report.

18. Change the input signal type for channel 2 to a square wave and change the noise amplitude back to zero. Rerun the analysis and capture a screen shot of this to your report.

19. Now change the frequency of channel 2 to 10 Hz, rerun the analysis and capture a screen shot of this to your report.

20. Now click on the "Experimental Design" sub tab. Illustrated in this section is data concerning a pharmaceutical intervention for a cardiac condition known as premature

ventricular contractions. You can read the description at the top of the page. The hypothesis is that the dose of beta blocker that the subject receives has a significant effect on the amount of PVC's that the subject experiences per hour. You have control over several features in this example. For example, you can modify the number of PVC's per hour or the dose for each subject.

21. You also have controls over setting the dose level thresholds that will be used for the hypothesis testing. Medication doses will be considered a low level, mid level, or high level input to the hypothesis testing based on the thresholds that you define with the "Range Level Control" slider tool. Any dose above the blue slider bar will be considered a high level, and any dose below the blue slider bar and above the grey slider bar will be considered a mid level, and a dose below the grey slider bar will be considered a low level. The high, mid, and low level doses correspond to dose levels 2, 1, and 0, respectively, on the plot to the right.

22. You may also modify the significance level of the test using the "Select" significance digital input control.

23. Using all the default settings, click on the button labeled "Run Hypothesis Test". The result of the hypothesis test will be shown in the bottom right hand corner as either accepter or rejected. Additionally, the calculated significance will be shown. Capture a screen shot of this for your report.

24. You should have rejected the null hypothesis. In other words, when all the default settings are used, dose level has a significant effect on the number of PVC's that occur in a subject.

25. Now adjust each of the parameters to determine their effect on the results of the hypothesis test result. Start by slowly increasing the number of PVC's/hour for subjects taking the lower dose of medications. Notice at what levels the result of the hypothesis test state that you should accept the null hypothesis that the dose level does not affect PVC's.

26. Adjust the selected significance level and the level of the range control to better understand the inputs to a hypothesis test.

Discussion Questions

1. Describe what is meant by mean, standard deviation, and variance.

2. In general, what effect does windowing have on collected data? Specifically, what effect does the windowing size have on the collected data?

3. What useful information can a histogram provide about your data?

4. Describe the purpose of regression analysis and what purposes it could be used for.

5. When developing a model to predict future outcomes with a regression analysis, why is it important to use a generalization set to test the model?

6. When completing the correlation analysis, you completed an analysis between several different channels of physiological waveforms. When examining the ECG waveform, which other channel showed the highest spectrum cross correlation? Why do you think this is?

7. When examining the airflow channel, which other channel showed the highest spectrum cross correlation? Why do you think this is?

8. When completing a correlation analysis with the simulated waveforms, what effect did channel frequency have on the resultant plots? What effect did adding noise to a signal have on each of the plots?

9. Describe the factors and inputs that are important to consider during an experimental design and hypothesis testing?

References

1. Devore, J. L. Probability and Statistics for Engineering and the Sciences. Third Edition. Brooks, Cole Publishing, 1991.

2. www.mathworld.com