Lecture 10 Cogsci 109

Fri. Oct. 19, 2007

Computing Basic Statistics II, variability

Outline for today

- Announcements
- A few matlab tips
  - **Ctrl-I**
  - **Close all**
  - **Clf**
  - **About 'loading data *into* a variable of your choosing'**
- The concept of probability density functions (PDF) reviewed
  - **The normal distribution is a PDF**

Announcements

- Reading
  - Monday, more reading will be assigned
  - Recordings - will post first two weeks for those who added late

Matlab tips

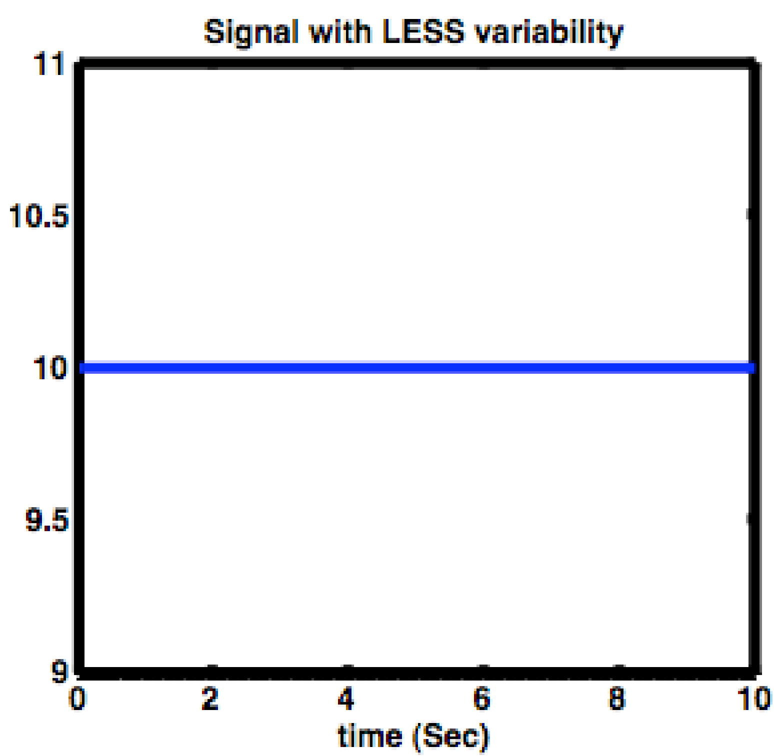
- Ctrl-I
- Close all, close
- Clf
- Loading data into a variable of your choosing issue

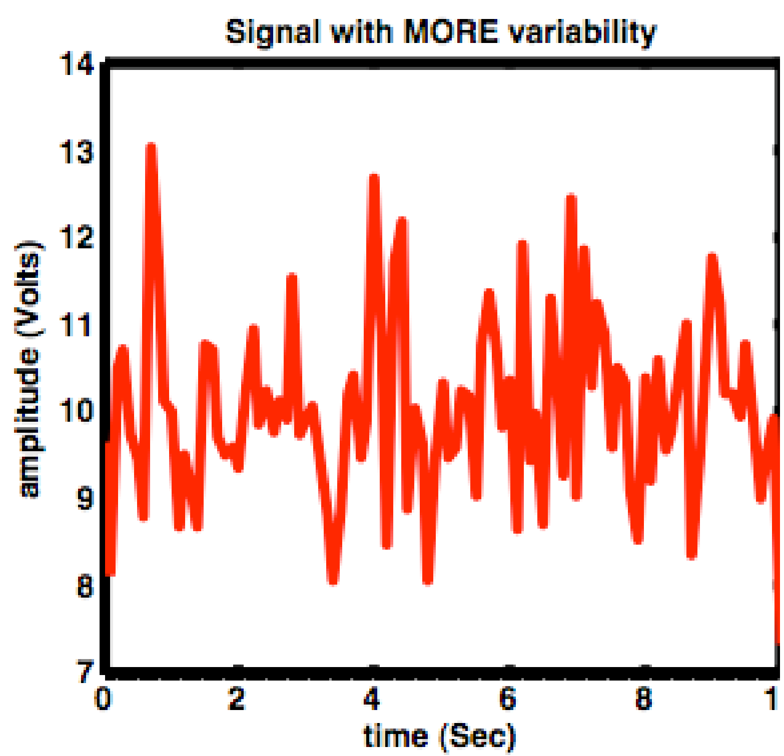
Outline for today II

- Measures of variability in terms of the normal distribution
  - **Variance**
    - Definition, properties, applications, how to compute in matlab
  - **Standard deviation**
    - Definition, properties, applications, comparison to variance, computing in matlab
  - **Covariance**
    - Definition, properties, applications, relationship to variance, computing in matlab
  - **Z scores and normalizing to unit variance**
    - How to perform this normalization
    - What are the applications and situations one might use this

Consider the following...

- Both signals have the same mean, but they are obviously different!
- One VARIES much more about the mean, can we create a quantitative measure of this?





We need a measure of Variability, here are a few…

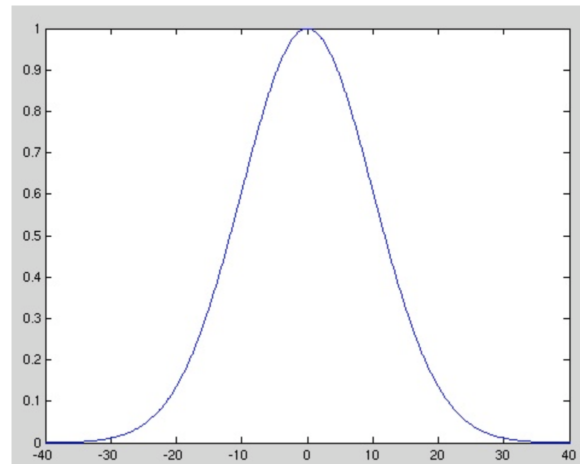
- Range
  - From math review, difference between max and min values of the data

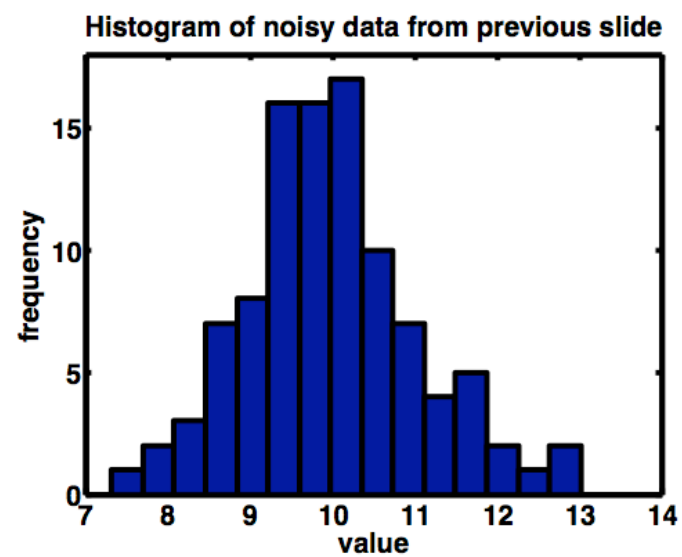
$$Range(x) = Max(x) - Min(x)$$

- Variance
  - Mean of squared deviations from the mean
  - In square units of the sample variable
- Standard deviation
  - Square root of variance
  - In units of the sample variable - sometimes easier to interpret

Returning to the normal distribution…and considering our data in terms of a histogram…

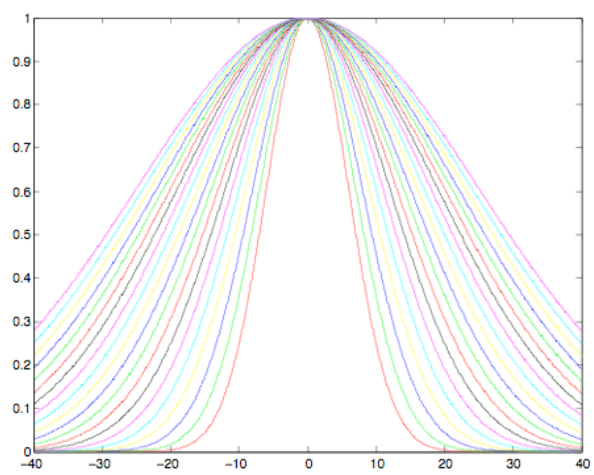
- The distribution of points about the mean can be considered in terms of probabilities
- How likely is a point to deviate from the mean?
- We call the normal distribution a *probability density function (PDF)* because it allows us to predict the likelihood that a sample will take on a particular value





Variance

- Whereas the mean defines a measure for the most likely point in state space (the center ‘location’ of a normal distribution)
- We can define the spread of the normal distribution about the mean by its *variance*



Variance (part II)

- Steps to compute the variance
  - **Compute the deviations from the mean for all the data**

    $$d_i = \left(x_i - \bar{x}\right)$$

  - **Compute the square of each of the deviations**

    $$sd_i = \left(d_i\right)^2$$

  - **Sum up all these squared deviations**

    $$ssqd = \sum_{i=1}^{N} \left(sd_i\right)$$

  - **Divide the mean squared deviations by N, the number of observations**

    $$Var = \frac{ssqd}{N}$$

How to compute the variance in matlab

- Function *var()*
- Example
- Matlab help: *help var*

Standard Deviation

- Typical 'deviation' from the mean
- Ie how far on average scores depart on either side from the mean
- Easy to compute after the variance - just take the square root of the variance

$$SD = \sqrt{Var} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

How to compute the standard deviation in matlab

- Function *std()*
- Example
- Matlab help: *help std*

Z scores

- A Z score is simply a measure of how many standard deviations away from the mean a score is
- Units are standard deviations

$$Z_i = \frac{X_i - \mu}{SD}$$

Covariance

- Covariance is very commonly used in statistical analysis as the basis for advanced statistics
- Gives a quantitative measure of the relationship between two variables

$$Cov(X,Y) = E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)^T\right]$$

$$E = \text{exp}ectation$$

$$\mu = mean$$

More Covariance

- If the two variables are independent, the covariance is 0
  - **(BUT IF COVARIANCE IS 0 THAT DOESN'T MEAN THE VARIABLES ARE INDEPENDENT!!!)**
- If they are totally dependent the covariance of data, can be arbitrarily large
  - **(AGAIN THE CONVERSE IS NOT NECESSARILY TRUE)**
- The diagonals are the variance of each variable
- If each row is an observation, and each column a variable…

$$\operatorname{cov}(X) = \left(\frac{1}{N-1}\right)\left(X - mean(X)\right)\left(X - mean(X)\right)^T$$

Matlab does it easily with

- Function: cov(X) where X is a matrix with rows being observations, columns being variables
- cov(X) where X is a vector yields the variance (a single scalar number)

As an aside: be careful about ‘sample’ vs. ‘population’ measures

- You can’t usually measure every possible subject or situation
  - **Can you measure the height of every SINGLE individual in the United States?**
    - Theoretically yes but it would take too long and too many resources
  - **Measure a representative group which is large enough to minimize the bias due to the fact that it is only a portion of the total possible measurements you could make**
  - **Can make some mathematical adjustments**
    - We won’t deal with this too much, since you learned about this in statistics, but you should know about the implications of each type of measure
- Matlab uses different equations to compute these statistics depending on you, but it has defaults of typically estimating populations

Trace

- Sum of the variances (the sum of the elements of the diagonal of the covariance matrix)