Introduction to Data

QEA

Spring 2016

What is this about?

Another big important topic which relates to both linear algebra and to our study of facial recognition is the idea of data analysis. In this building block, we will discuss some of the basic concepts from data analysis and link them to linear algebra.

Resources to read and watch

There are lots of books about Data Analysis and lots of useful videos on the web. Here are some specific recommendations:

- Bevington's Data Reduction and Error Analysis for the Physical Sciences (available in the classroom)
- Your favorite statistics textbook.
- Wolfram Mathworld on Probability and Statistics

Many Measurements of the Same Thing

One of the simplest forms of data is a set of data which represents many measurements of nominally the same thing. Depending on what the goal is of our analysis, this might encompass measurements of the same quantity across many different situations, or many instances of the same situation.

Visualizing Measurements of the Same Thing

It's usually a good idea to *look* at data before you start calculating things associated with it.

You've surely encountered these ideas before, but for the sake of completeness, we'll highlight a couple of ideas here. If you have a large number of data points (say, for example, that you measured the heights of a bunch of different people), you might choose to simply plot the data versus the person number – the index. Note here that the data is plotted as individual points, since each point represents a measurement. Ideally we might also include error bars here to indicate our uncertainty in a given measurement, but for now, let's leave that out.

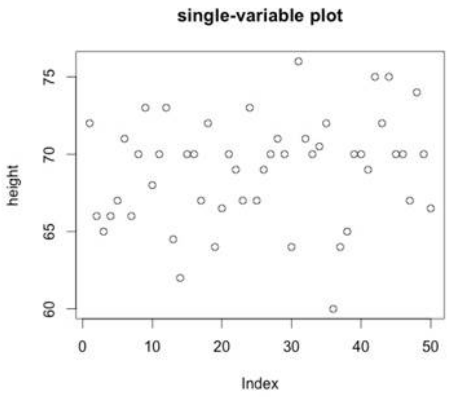


Figure 1: An example of a single variable plot

Alternatively, you could also visualize many measurements of the same thing by creating a *histogram*. This is a representation of how many measurements fall into different "bins": the height of a given bar is the number of samples that fall within the range associated with the bar. For example, in the figure, you can see that about 20 million people made between 0 and $5000 in 2008. You've likely seen this kind of thing before as well: it's not an uncommon way to represent test scores.

Note, of course, that how a histogram *looks* depends on what you choose for the bins - both how many there are, and where they are centered!

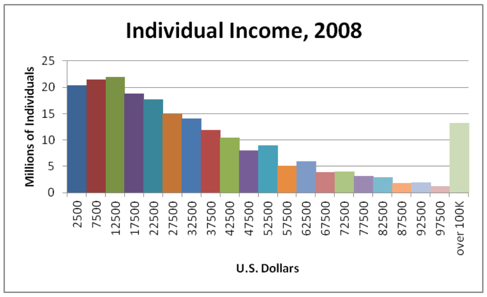


Figure 2: An example of a histogram.

Common Figures of Merit for the Same Thing

While looking at the data is certainly helpful, we can also extract or calculate a couple of important figures of merit of the data. The first is the average, or mean of the data, given by summing all the elements in the dataset $\{d_i\}$ and dividing by the number N of elements in the set:

$$\mu_d = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{1}$$

Note that if our data is a continuous function $f(x)$ over a range of the independent variable x as opposed to a set of discrete points, we can express the same thing as an integral:

$$\mu_d = \frac{\int_{range} f(x)dx}{\int_{range} dx} \tag{2}$$

The average captures the center or 'expected value' of the distribution of data. In addition to this, it is often helpful to capture the spread of the data around this average. There are a few different metrics which are used for this. A simple one is the variance from the mean R^2: the average of the squared difference between each data point and the mean.

$$R^2 = \frac{1}{N}\sum_{i=1}^{N}(d_i - \mu_d)^2 \tag{3}$$

Another commonly encountered measure is the standard deviation, which is simply the square root of the variance from the mean (often given the symbol σ:

$$R = \sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \mu_d)^2} \tag{4}$$

Exercises

1. Look at the single variable plot above. Estimate the value of the mean and the value of the standard deviation. What are the units of each?

2. Look at the histogram plot above. Estimate the value of the mean and the value of the standard deviation.

3. What is the mean and standard deviation of this data set (Do this in your head!)

$$\{1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3,1,3\}$$

4. Begin by considering the simple dataset of the high temperatures in Needham for ten days in March:

$$T = \{57,61,46,43,46,46,46,46,55\} \tag{5}$$

   (a) By hand, histogram this data. What size bin makes sense? What bin centering makes sense?

   (b) By hand, compute the mean temperature over these ten days. If you look at the data, does this mean make sense?

   (c) By hand, compute the variance and standard deviation of the temperature over these ten days. If you look at the data histogram does this make sense?

   (d) This dataset has a flaw, in that it has a small number of datapoints. What do you see as the possible effects of having such a small sample?

5. Now consider the larger dataset below of the approximated heights of the Olin faculty. In Matlab, create a vector which has this dataset as the entries.

$$\begin{aligned} H =& \{63,66,71,65,70,66,67,65,67,74,64,75,68,67,70,73,66,70,72,62,68, \\ & 70,62,69,66,70,70,68,69,70,71,65,64,71,64,78,69,70,65,66,72,64\} \end{aligned}$$

   (a) Computationally histogram this data. What size bin makes sense? What bin centering makes sense? Try a few different combinations. (The `hist` command will likely come in handy – and you'll definitely want to read the documentation!).

   (b) Computationally, find the mean of this dataset (do it both using the command `sum` and the command `mean`).

   (c) Computationally, find the standard deviation of this data (again, do it both "by hand" in the sense that you are actually calculating the expression using elemental operations, the command `sum`, etc., as well as and the command `std`).

(d) Do this mean and standard deviation make sense given the histogram of the data?

Correlation

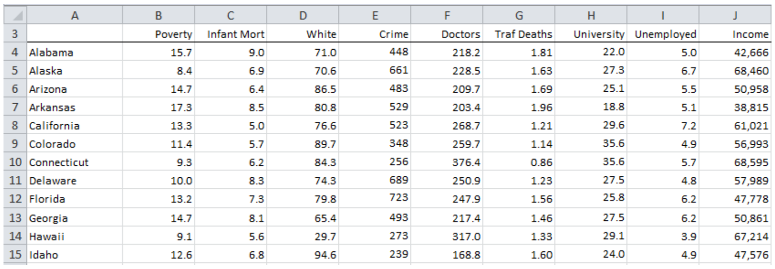
Now let's consider that we measure two different associated quantities and want to test whether these are linearly correlated (if one goes up, the other also goes up), anti-correlated (if one goes up the other goes down) or uncorrelated (the behavior of one cannot be predicted by watching the behavior of the other). (Please note that correlation has nothing to do with causality!). There are many different measures of correlation, but we will discuss here one of the most common, the Pearson Correlation Coefficient.

For a pair of associated datasets $X = \{x_i\}$ and $Y = \{y_i\}$, each with n elements, we define the Pearson Correlation Coefficient to be:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y} \tag{6}$$

where μ_x,μ_y,σ_x and σ_y are the means and standard deviations of the datasets. Essentially, for each pair of values, we take the product of the variations from the mean, then sum these products up over all pairs of values and normalize by the expected variation as characterized by the standard deviation. If the two values are consistently always on the same side of the mean, then each term in the sum will contribute positively, and the total value will be close to one, indicating positive correlation. If the two values are consistently on the opposite sides of the mean, then each term in the sum will contribute negatives, and the total value will be close to negative one, indicating anticorrelation. If, for every pair, it is just as likely that the two values will be on opposite sides of the mean as on the same side of the mean, then the sum will go to zero, and the two values are uncorrelated.

Consider the following data:



	A	B	C	D	E	F	G	H	I	J
3		Poverty	Infant Mort	White	Crime	Doctors	Traf Deaths	University	Unemployed	Income
4	Alabama	15.7	9.0	71.0	448	218.2	1.81	22.0	5.0	42,666
5	Alaska	8.4	6.9	70.6	661	228.5	1.63	27.3	6.7	68,460
6	Arizona	14.7	6.4	86.5	483	209.7	1.69	25.1	5.5	50,958
7	Arkansas	17.3	8.5	80.8	529	203.4	1.96	18.8	5.1	38,815
8	California	13.3	5.0	76.6	523	268.7	1.21	29.6	7.2	61,021
9	Colorado	11.4	5.7	89.7	348	259.7	1.14	35.6	4.9	56,993
10	Connecticut	9.3	6.2	84.3	256	376.4	0.86	35.6	5.7	68,595
11	Delaware	10.0	8.3	74.3	689	250.9	1.23	27.5	4.8	57,989
12	Florida	13.2	7.3	79.8	723	247.9	1.56	25.8	6.2	47,778
13	Georgia	14.7	8.1	65.4	493	217.4	1.46	27.5	6.2	50,861
14	Hawaii	9.1	5.6	29.7	273	317.0	1.33	29.1	3.9	67,214
15	Idaho	12.6	6.8	94.6	239	168.8	1.60	24.0	4.9	47,576

6. Look over the data. By eye, can you spot columns that you think look correlated? Anticorrelated? Uncorrelated?

7. By hand, compute the correlation coefficients for your three pairs. How did you do? Remember that values close to 1 indicate correlation, values close to -1 indicate anticorrelation, and values close to zero indicate no correlation.

8. Choose your three favorite columns of data from this dataset. Input these into vectors in Matlab. For each of these vectors, subtract off the mean, and then divide out the standard deviation.

   (a) With these vectors, how would you directly compute the correlation coefficient between any pair of them?

   (b) Now, consider one matrix which has these three vectors as columns, and another matrix which has these three vectors as rows (the transpose). From these two matrixes can you come up with an operation which computes correlation coefficients between every pair of vectors?

   A note of warning. Correlation does not imply causation.

9. To drive this point home, visit the Spurious Correlation Website. Follow the link at the bottom of the site to discover and plot a spurious correlation of your very own.

Linear Regression

If we have measurements of a quantity y at many different values of another quantity x for which the measurement takes different values, we can use this data to deduce a relationship between the two quantities. This relationship can take any functional form $y = f(x)$. Often times, we want to determine the functional relationship between x and y, where this function can include unknown constant parameters. If the function is linear in these unknown parameters (even if the relationship between x and y is not a linear relationship) then the process of finding the values of these constant parameters is known as Linear Regression.

TWO POINTS For starters, we will consider the case where the expected relationship between x and y is a line $y = ax + b$. Here a and b are our unknown 'fitting' parameters. As you probably learned in high school, it takes two points to define a line. If we have only two data points, and the expected relationship between the variables is linear, we have just enough information to define the line. If we have only two data points, we do not have enough information to fully determine any descriptive equation that has more than two unknown parameters.

10. Let's say that we have two measurements of the temperature in a room, which we expect to be changing linearly with time. At $x = 60$ seconds on our stopwatch, the temperature is $y = 22$ degrees, and at $x = 300$ seconds the temperature is $y = 24$ degrees. Find the parameters a and b to determine the equation of the line which best describes the temperature in the room.

Let's break down a little bit the process you went through to find a and b. Each pair of x and y values can be plugged into our trial equation $y = ax + b$, giving us two equations for our unknown parameters a and b, which you could then solve by the standard system of equations techniques you learned in high school. However, there is another approach.

11. Consider the following matrix equation:

$$\begin{pmatrix} 60 & 1 \\ 300 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 22 \\ 24 \end{pmatrix}$$

Do the matrix operation on the left hand side of this equation to find equations for the components of the vector on the right hand side. Do these equations look familiar?

12. We can write this system using a matrix $\mathbf{A}$, an unknown vector $\mathbf{x}$, and a known vector $\mathbf{b}$

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

To *undo* the matrix operation, we can simply find the inverse, and the solution we are looking for is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Go ahead and determine the inverse of the data matrix and solve for a and b.

MANY POINTS DEFINE A LINE The process above works well if we have measurements which are very very precise. Unfortunately, most of the time data is not so precise. Even if the temperature were constant, if we measured it multiple times we would likely get a spread of different numbers: from which we could extract the mean and standard deviation as discussed above. With measurements of changing values, there is likewise usually scatter around the expected values from any functional relationship.

When we have scatter in our data, the more data we have, the better we are able to find a function which accurately describes the relationship between the variables. However, if we have many data points and wish to describe this with a relationship linear in our

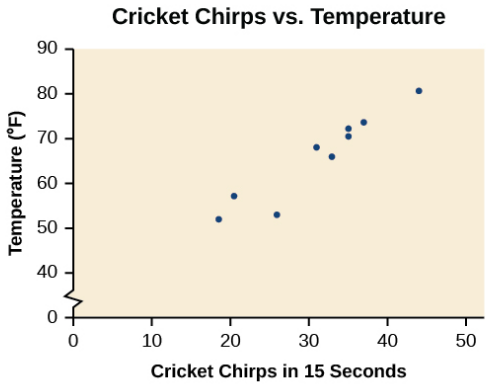


Figure 3: Example of Data with Scatter: an imprecise way of measuring temperature. From philschatz.com

unknown parameters, we need to use linear regression. The most commonly encountered method of linear regression, and the method we will discuss here, is the method of Least Squares. Again, for our example we will consider data where the relationship between our x and y data is described by a linear equation $y = ax + b$, though least-squares fitting will work for many functional forms.

In least squares fitting, we want to find the line which lies 'closest' to all our data points. To express this quantitatively, we want to find the line which minimizes the total variance of our data set from the line. Recalling our definition of variance from the mean above, we can rewrite that definition to find the variance from the equation for our line:

$$R^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - (ax_i + b))^2 \tag{7}$$

When R^2 is a minimum with respect to variable a, the first derivative of R^2 with respect to a will be zero. Likewise, when R^2 is at a minimum with respect to variable b, the first derivative of R^2 with respect to b will be zero.

$$\frac{\partial R^2}{\partial a} = -2\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - (ax_i + b)) = 0 \tag{8}$$

$$\frac{\partial R^2}{\partial b} = -2\frac{1}{N}\sum_{i=1}^{N}(y_i - (ax_i + b)) = 0 \tag{9}$$

This gives us two equations for our two unknown parameters a and b:

$$a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \tag{10}$$

$$Nb + a\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{11}$$

Note that these equations have coefficients which are easily calculable from our dataset! So let's give this a try on our data of cricket chirps vs. temperature. At temperatures (F) of:

$$T = \{53, 54, 58, 66, 69, 70, 71, 73, 81\} \tag{12}$$

our cricket scientists find corresponding rates of cricket chirping (chirps per 15 seconds) of:

$$C = \{19, 26, 21, 33, 31, 36, 36, 38, 45\} \tag{13}$$

13. Find the sums $\sum_{i=1}^{N} x_i$, $\sum_{i=1}^{N} y_i$,$\sum_{i=1}^{N} x_i^2$ and $\sum_{i=1}^{N} x_i y_i$.

14. Using these numbers, write out the system of two equations which will give you the slope and intercept a and b of the line which best describes this data.

15. By analogy with the simple two-point linear problem from above, write down the matrix equation which is equivalent to the set of two equations you just formulated. Determine inverse of the data matrix and find the best fit values for a and b. Plot the data in matlab and plot your best fit line on the same plot. How did you do?