MAT 125 Elementary Statistics I
Lecture 11. Linear Regression

1 Distributions and Errors

1.1 Distributions

We have seen how to analyze and characterize data sets in one variable, or quantity. That is, we have a set of measured quantities for a population. Examples include the age distribution of people in the US, the rates of a particular illness in the population, the weights of tomatoes grown at a particular farm, etc. etc. etc. These can be analyzed to calculate means and dispersions: the average weight of the tomatoes and the spread in weights about that average, and so on. We've shown that the averages calculated for subsamples of the population have much smaller scatter than individual values (the central limit theorem), that their distribution approaches a normal distribution regardless of the shape of the population distribution (the central limit theorem), and approaches the population mean as the sample size increases (the law of large numbers).

In MAT126, the class following this one, we'll examine other distributions and use these and the normal distribution to see whether one distribution is significantly different from another. For example, do the tomatoes grown on your farm weigh more than those grown on my farm? This is a more complicated question than it first appears. We can calculate the average weight of your crop and compare that to the average weight of mine, and to be sure those numbers will be different; but are they *significantly* different? That is, is the difference large compared to the difference in the weights of the individual tomatoes? And how do you quantify the difference? This is an example of *analysis of variance*, which will be the topic of much of MAT126.

1.2 Random Errors

As well as data sets with an intrinsic distribution of a single variable (the weight of tomatoes, the height of children aged 4) there is a second kind of data set with a distribution of values: that of repeated measurements of the same quantity. For example, I might use a ruler to measure the length of a table. If I repeat the measurement *independently* a number of times (that is, each measurement is done without referring to any other measurement - I don't say, oh, this table must be 3 feet wide because that's what I got last time) I will get a slightly different answer each time because of limitations in the precision of the ruler, the precision with which I can read the ruler, and so on. The spread in measurements is smaller the more precise the measurement, but it will never go to zero. In this case, the **best estimate** of the true length of the table is the **mean** μ of the individual measurements and the dispersion, σ, of the distribution is called the **error**, or the **random error**, of my measurement. I will then report back to you that the length of the table is $\mu \pm \sigma$.

Remember that the dispersion in the sampling mean is $\sigma/\sqrt{n}$, where n is the number of measurements (the central limit theorem). This means that if I want to measure the length of the table to twice the previous precision, I will have to do four times more measurements.

1.3 Systematic Errors

A second type of measurement error is **systematic error** - what if the scale of my ruler is wrong? All the repeat measurements in the world won't fix this. To understand it, I have to use a different ruler and check the two measurement sets against each other. Identifying systematic errors is one of the most difficult things to do in science. We'll see more of this in MAT 126.

Measurements are thus characterized by *accuracy* and *precision*. These words are often used interchangeably in everyday speech, but have quite different meanings in statistics and science. *Accuracy* is how close the measurement is to the real value, and is affected by systematic errors. *Precision* describes reproducability and repeatability, and is affected by random errors.

2 Bivariate Data

The above discussion brings up the next topic for investigation: *correlation*. If your tomatoes have different weights than mine, what else is different? Do you give yours more water? More sun? More fertiliser? In other words one often knows several different things about the members of a population, and we want to know if these things are related to, or depend on, each other.

In the physical sciences, it is often the case that one measurable quantity is a *function of* or *depends on* the other. Examples include: the distance you travel depends on how long you are walking; the orbital speed of a planet around the Sun depends on its distance from the Sun; the amount of light radiated by a hot body depends on its temperature, and so on. These relationships are due to the underlying physics, and it is often the case that that scientific discovery proceeds by establishing and characterizing these relationships. Likewise, much investigation in the social sciences proceeds by establishing relationships between different quantities: you're likely to succeed in college if your parents went to college, etc. So often the production of new scientific knowledge involves the establishment of *correlations* between two quantities.

A data set with measurements of two quantities for the same object is then called a **bivariate data set**. The last two lectures in this course discuss how you use the data to measure the relationship between the two quantities.

Figure 1 shows some illustrative plots. The data that go into such plots are pairs of measured numbers. For example, you go for a run, Every few minutes, you read the time on your watch and read how far you've run on your meter, and write these numbers down. The list of pairs

of numbers tells you distance as a function of time. Such lists of pairs of numbers are called *bivariate data* (bi = two, variate means that the numbers are not fixed but are variables).

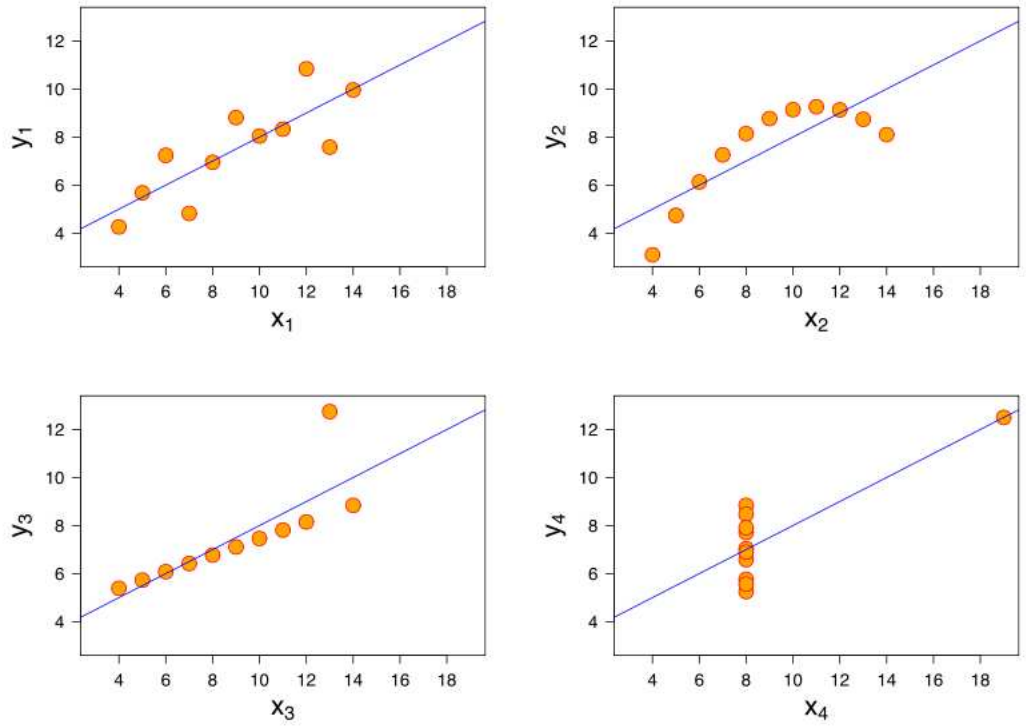

Figure 1. Plots of pairs of bivariate numbers. Such pairs of numbers are usually called (x,y).

If the two quantities are related to each other, so that one (y) *depends* on the other (x), the data will show a *correlation* between the two quantities. Figure 1a (top left panel) shows a *linear correlation* between the quantities x_1 and y_1. The line drawn on the figure is the *best fit* straight line, i.e. the line which best describes the relationship between y and x. But the points do not lie exactly on the line; they scatter about it. Why? The scatter is called *noise* or *error*, as discussed in the previous section, and is intrinsic to all measurements. If you measure a set of tabletops with a ruler, as described in the previous section, the values will sometimes be a bit high compared to the true value, sometimes a bit low. If you plot such measurements, as shown in Figure 1, the measurement uncertainties will show up as noise or scatter. A lot of the care you need to take when analyzing data is to be able to distinguish between deviations which are consistent with the expected error and those which are not.

Some examples are shown in the other panels of Figure 1. Figure 1b (top right panel) shows a correlation between x and y, but it is not linear, and the straight line is not a good fit. Figure 1c (bottom left panel) shows a tight linear correlation with one deviant point, which skews the best fit straight line. In data analysis, one needs an objective criterion for removing deviant data, but doing so involves the assumption that all the measurements are measuring the same thing. Sometimes, these deviant data points are telling you something important,

but most often they're due to mistakes. Removing deviant data is one of the thorniest topics of statistics; if you're not careful, you find yourself removing the points that don't "fit" your hypothesis, and thereby dishonestly biasing the data in favor of the conclusion you want to reach. And finally there's Figure 1d (lower right panel) with a bunch of y values measured for the same x and showing a lot of scatter, and one point at a different value of x. Can you really conclude any relationship from this plot?

As noted above, some phenomena are correlated and some not. Some examples of phenomena that aren't correlated:

– the speed at which you walk and the price of eggs in the supermarket

–the day of the week and the temperature

Are the following correlated?

–the time you spend studying and your grade for an exam

–the years you spend getting an education and your income

–the speed at which a planet goes around the Sun and its distance from the Sun

–the speed at which a planet goes around the Sun and the mass of the planet

–the speed at which a planet orbits a star and the mass of the star

–the time of year and the temperature outside

As you can see:

–some bivariate data are correlated and some are not

–sometimes the correlation is not perfect (actually it's never perfect in the real world) because other effects are affecting the values of the data, such as noise

–and a most important point: sometimes correlation means causation and sometimes it doesn't. We'll discuss this in the next lecture.

3 Straight Lines

We'll deal only with linear correlation in the remainder of the course. First, let's do a little brush-up on straight lines. A straight line is the shortest distance between two points. Its

mathematical equation is

$y = mx + c$

where m is the slope and c the y-intercept; x and y are variables, and m and c are constants (numbers).

For example

$y = 2x + 1$

is the equation of a straight line of slope m = 2 and y-intercept 1. The *y-intercept* is the point where the line crosses the y-axis and is found by setting x = 0. When x = 0, y = 1. The coordinate of the y-intercept is (0,1), i.e. x = 0, y = 1.

The above formula is, basically, a *function*, or recipe. You can choose any value of x, and this formula tells you how to calculate the corresponding value of y. Let's calculate a few values of y for input values of x, and plot the result on a graph:

x	y
0	1
1	3
-1	-1
2	5
-2	-3

The line is plotted in Figure 2.

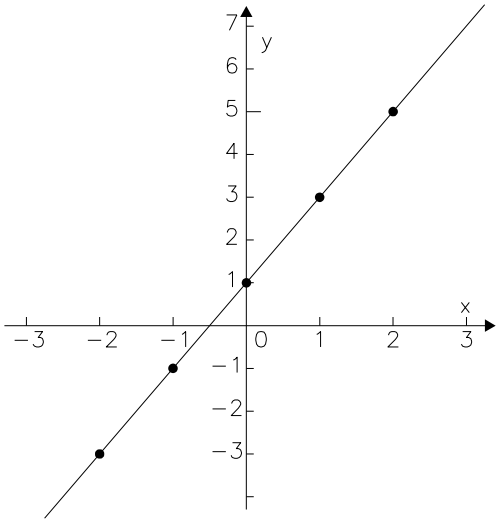

Figure 2. The line y = 2x + 1

A straight line is completely defined by two points on the line. When you're plotting a line like y = 2x + 1, it's a good idea to calculate 3 or 4 points just to check that they all lie on the line and you haven't made a mistake.

A line with a positive value of the slope m slopes upwards from left to right. A line with a negative value of m slopes downwards from left to right.

4 Linear Correlation

A *linear correlation* between two quantities means that as one quantity steadily increases, the other steadily increases or steadily decreases, and that if you plot one quantity versus the other, the points lie along a straight line - either exactly on the line, as in Figure 2, or scattered about it, as in Figure 3.

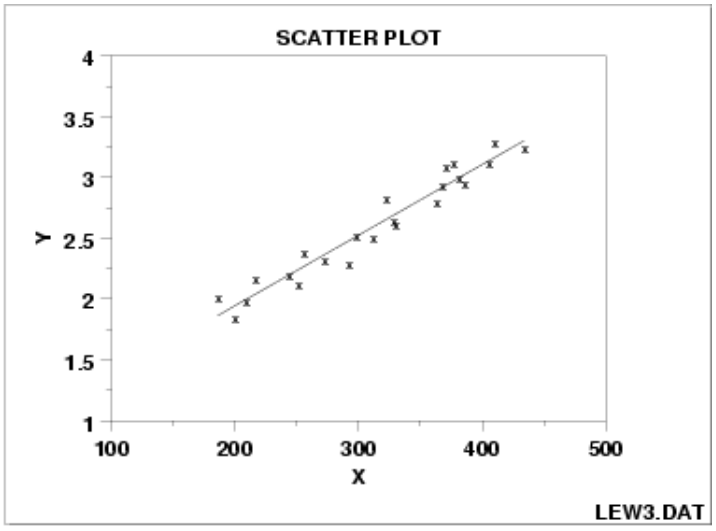

Figure 3. Scatter plot of measured values of x and y, i.e. bivariate data (x,y), where x is the independent variable and y the dependent variable. The line shows the *best fit* straight line to the data points.

Note that Figure 3 (correctly) does not extend the straight line beyond the domain (the span of values of x) of the data. We will discuss this later.

The *linear correlation coefficient* is called r. r is calculated from the data and measures the strength of the correlation. If r = 1 the two quantities are perfectly correlated. If r = -1 the two quantities are perfectly anticorrelated. If r = 0 the two quantities are uncorrelated. Thus -1 $\leq$ r $\leq$ +1. For data with noise, i.e. all real data sets, r will never be exactly $\pm$1. We will discuss how to calculate the linear correlation coefficient in the next lecture, but first we need to calculate the *best fit straight line*.

5 Linear Regression

To see if there is a linear relationship between two quantities (the data points for one are denoted by x and for the other by y), the first thing to do is make a plot of y against x. See, for example, the plots in Figures 1, 2 and 3. These, by the way, are called *scatter plots*.

Figure 2 shows a straight line with the data points lying on it perfectly (as well as one can tell from the size of the symbols, anyway). Figures 1a and 3 show a good straight line relationship with scatter. Figure 1b is not a straight line. Figures 1c and 1d may show a linear relationship.

All of the plots show the best-fit straight line. How do we determine it? First, we have to define *best fit*. The points do not lie exactly on the line because of *noise* or *uncertainties* in the measurements. We assume that the independent variable, x, is measured exactly with no noise while each measured value of the dependent variable y has an associated error. What

you measure is actually:

$y_i = mx_i + c + \epsilon_i$

as the i-th pair of data points (x_i, y_i), where y_i is the measured y-value for x_i, m and c are the slope and y-intercept of the straight line which best describes the set of n data points, and ϵ_i is the uncertainty or error for y_i.

As an example, consider the following data points, for how many hours you spend studying for a math test (call this x) and your resulting score (y). How do we find the best-fit straight line, and why would we do it?

–to *interpolate*, that is to use the line to calculate your expected score for a given number of hours of studying

–to *extrapolate*, that is to extend the line beyond the domain of the data and predict what your score would be if you spent longer times studying than any you have actually done. Extrapolation is dangerous; you don't know that the apparent straight line will persist outside the range and domain of the data you have. In this case it clearly won't; there's a perfect score for any exam, and no amount of studying will give you a score greater than 100%

–to *understand*. If you have a good estimate of the true numerical values of the slope and intercept, and a determination that the line fits the data to within the uncertainties, a hypothesis for explaining the data may suggest itself.

The following table shows the SAT math scores of a class of students and the amount of time they studied:

Hours	Score
4	390
9	580
10	650
14	730
4	410
12	600
22	790
1	350
3	400
8	590
11	640
5	450
6	520
10	690
11	690
16	770
13	700
13	730
10	640

How do you calculate the best fit straight line? You can do a not-bad job by simply plotting the points and laying a ruler on the points and drawing a line (this is sometimes colloquially called "χ by eye" where the Greek letter χ is pronounced "kigh" (see below).

Mathematically, what you do is minimize the sum of the distances between all the points and the straight line. In practice, since the values of x_i are defined to be exact, you're minimizing the y-distance from the line. Rather than sum the y-distances (which are of course both positive and negative) and finding the value of m and c which make this sum the smallest, you sum the *squares* of the distances and find the values of m and c which make this sum the smallest. Not only are these positive, but it gives extra weight to the points furthest from the line. So you find the values of m and c which make

$$\chi^2 = \sum (y_i - y)^2 = \sum (y_i - (mx_i + c))^2$$

the smallest possible. By definition, this is the best-fit straight line. This process is called χ^2 *minimization*. The sum, $\sum$, is carried out for all pairs of (x,y) points from i = 1 to N, the total number of pairs of points.

Figuring out how to minimize χ^2 is a snap using calculus and a chore using algebra, so let's just take our word for it. The values for m and c which give the best-fit straight line are:

$$m = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

and

$$c = \frac{\sum y_i - m \sum x_i}{N}$$

Notice that c is simply the mean offset from $y = 0$.

During the worksheet session you'll calculate the best-fit straight line to the data in the above table.

How do you know you have a good fit? You can clearly fit a straight line to any set of points just by grinding through the above arithmetic, but a straight line is clearly not always a good fit (see Figure 1). The best test of a good fit is whether

$$\sqrt{\frac{\chi^2}{N}} \sim \epsilon$$

i.e. that the mean deviation from the line is about the same as the noise.

In the next lecture, we'll use these results to calculate the linear correlation coefficient.