INTRODUCTION TO STATISTICS

Measurements on a sample

An illustrative example. In a lab experiment on atmospheric particulate matter, a student used 1 cm^2 squares of scotch tape to collect particles from the air during a particular period of time at a particular place. She counted the particles on each of 5 squares twice, obtaining the following results:

square	count	
	1st	2nd
1	23	21
2	34	40
3	18	18
4	25	26
5	48	49

The purpose of her measurement was to get some indication of the concentration of particulate matter at that time and place. The counts of particles on scotch tape do not translate directly into concentrations in the atmosphere, but they do provide an indicator, which can be validly compared for different times or different places. The lower the counts, the cleaner the air.

Thus, she would like to use these numbers to estimate the "true value" of this particulate matter indicator, for the given time and place. An obvious estimate would be the average of her 10 counts, which is about 30.

> Use the Excel function MEAN to calculate this average (answer = 30.2); for practice, also calculate the averages for each of the two counts separately.

The (unknowable) true value. What is meant by the "true value" that she is estimating by calculating the average? It is actually quite an abstract concept. The best we can say is this: imagine doing this experiment with a very large number of squares, and counting each square very accurately. The average, for a very large number of accurate counts would constitute a close approximation to the "true value."

With this concept of true value (we can stop putting it in quotes now), it is natural to ask the following question: How good an approximation can one get from a more practical measurement, involving a relatively small number of squares (5, in this case), each counted fairly accurately (as judged by taking two counts per squares)?

The importance of knowing about variability. A little thought will tell you that the accuracy of any single measurement, or any small number of measurements, depends crucially on the homogeneity or inhomogeneity (variability) of what is measured. For example, if you wanted to measure the resistivity of copper wire of a certain type and thickness, you might cut off a piece say 10 cm long, measure the resistance with an ohmmeter, and divide by 10 to get resistivity in ohms/cm. You might not even take a second sample -- copper wire is copper wire, one sample is much the same as another. If you did take a second sample you would expect a measurement very close to the first one. On the other hand, if you wanted to measure the force that a person can exert with her hand, you might have her squeeze a dynamometer (balancing the force against a calibrated spring). But you would not take one measurement as definitive for people, even for people homogeneous in age, sex, height and weight. People differ in strength, and you would not expect a second person to give you a result identical to the first. To approximate the true population average force, say, for females, age 21, standing 170 cm tall, and weighing 55 kg, you would want to average the measurements for quite a few women with those characteristics. Even the force exerted by a particular woman might vary from one try to another, and so you might want to make several measurements (spaced to avoid fatigue).

Returning to particle counts on scotch tape: one might not have a very clear idea, a priori, how much they vary from one sample to another, nor how accurate any one count is, by a person exerting ordinary care. Taking several squares and counting each one twice give one some indication concerning both these issues.

Note that the different squares ranged from counts in the low 20s to counts in the high 40s: indicating considerable variability. One cannot trust any single square to give a very good approximation to the true value, and one might be skeptical about the average of just 5 squares. What if those 5 squares just happened to fall on the low side of the ensemble of all possible squares? Perhaps 70% of all possible squares are in the high 30s or 40s from this time and place, but it just happened that there were 3 (instead of 1 or 2 out of the 5) from the lower 30% of the ensemble. In short, the observed variability of the counts from the squares is important, because it tells us that this is a heterogeneous ensemble. One can get only limited accuracy from the results for a small sample.

How Statistics works. The discipline of Statistics was developed principally to give clear answers to questions of this sort, i.e., how well is a true average value approximated by the average calculated from a given set of measurements. It turns out to be feasible to do the following:

(a) get a rough estimate the variability of an ensemble of possible measurements from the observed variability of a few measurements,

and

(b) use this roughly estimated variability, in turn, to estimate the likely deviation of the observed average from the true value.

Statistics that express variability. The most obvious way to express variability is through the range, as was done informally above. For example, the 1st count (column 1 above) gives a maximum of 48 particles and a minimum of 18: so range = 30. For a small sample (5 number, in this case) the range is not a bad indicator of variability, but for larger samples it can be a very poor one, because it uses only the two extreme points (and thus does not tell you much about the bulk of the data) and for other reasons as well.

The most important way to express variability within a sample is through two closely related numbers, variance and standard deviation, corresponding to the functions VAR and STDEV in Excel. To get the variance of a set of N numbers, one first gets their mean, next one calculates the difference of each number from that mean, next one squares the differences. Finally, one very nearly averages those squared differences: one adds them all up, but instead of dividing by the total number, N, one divides by one less, N–1. (The reason for dividing by N–1 might become clear later.) The standard deviation is just the square root of the variance.

The following table shows this calculation for the 1st count.

sample	1st count	deviation from mean	deviation squared	mean	variance	standard deviation
1	23	-6.6	43.56			
2	34	4.4	19.36			
3	18	-11.6	134.56			
4	25	-4.6	21.16			
5	48	18.4	338.56			
Sum	148	0	557.2	148/5 = 29.6	557.2/4 = 139.3	11.8

Even though we will be using computer tools, such as the Excel functions VAR and STDEV to calculate the variance and the standard deviation most of the time, it is important to know how this calculation is done. The sum of the actual deviations from the mean must always be 0, by the laws of arithmetic. The standard deviation is a kind of average of the absolute values of the deviations, but one squares the deviations first, then takes the square root after averaging. Therefore, deviations that are small in absolute magnitude contribute little to the variance, thus little to the standard deviation. In the above calculation, most of the standard deviation is determined by the two extremes, which is why it is not that different in nature from the range. For larger samples, it is still the large deviations that matter most in calculating standard deviation.

To express all this in algebraic notation, let $x_1, x_2, \cdots, x_N$ be the observed values in a sample of size N. Let $\bar{x} = \Sigma x_i / N$ be the mean of the numbers. Then the variance is:

$$\frac{\Sigma(x_i - \bar{x})^2}{N - 1}$$

Do you expect the standard deviation for the 2nd count to be greater or less than that for the 1st count?

Calculate it by hand (calculator).

Confirm using Excel.

Using the standard deviation to compute error bars. We've been concentrating on step (a) of "how statistics works," i.e., estimating variability from the sample. Now it is time for step (b): using the rough estimate of variability to estimate how far the observed mean is likely to deviate from the true value. This estimate of maximum likely measurement error is what is often plotted on graphs as an error bar. There is a lot of quite deep mathematics behind the procedure used, but the procedure itself is unbelievable simple. One calculates a quantity called the standard error of the mean by dividing the standard deviation by $\sqrt{N}$, the square root of the sample size: in this case, $\sqrt{5} = 2.236$. The length of the error bar is then some appropriate or conventional multiplier of this quantity. One common and fairly simple convention is to take the width of the error bar (above and below the estimate) as twice the standard error. Using this convention, we would calculate the error bar for the 1st count as $2 \times 11.8 / 2.236$ or 10.6, leading to the following estimate for the mean:

$$29.6 \pm 10.6$$

There are a lot of different conventions about error bars. In physics, common practice uses a multiplier of 1, so that the estimate in this case would be 29.6 ± 5.3; in psychology, it is common to use a multiplier that is adjusted according to the amount of information (called degrees of freedom) that goes into the rough estimate of standard deviation. In this case, with the rough estimate based on only 5 measurements, the multiplier would be 2.78, leading to 29.6 ± 14.6 for what is known as a "95%-confidence" estimate. For large sample sizes (30 or more), the multiplier is about 2 for 95% confidence (that is where the convention of always using 2 comes from). For the time being, we will use a multiplier of 2, keeping in mind that very often, the approximation to the true value is well within this error estimate, but with bad luck (an atypical sample) the true value can be even farther away from the estimate.

Figure 1 shows this student's results for the data we have been discussing, as well as her results from 4 other places. At each of the other sites, she once again counted 5 squares of scotch tape; but each was counted only once. The data are displayed as means ± two standard errors.

You should remember how statistics works (steps a, b) and you should remember the fundamental rule for sampling error:

$$\text{standard error} = \text{standard deviation} / \sqrt{N.}$$

A large part of statistics just involves systematic application of these ideas.

The Pythagorean rule for independent samples. In Figure 1, it is easy to conclude that the true values of particle count are quite different at sites 2 and 3: the two intervals formed by the error bars don't come close to overlapping. However, the comparison of sites 3 and 4 is more problematic. The intervals overlap, but not so very much. Should we conclude that the air was cleaner at site 4 than at site 3?

Just as one can think about true values at each site, one can also think about a true difference in values. We are asking whether the true difference, site 3 minus site 4, is known to be positive. The standard error tells us how much the true value is likely to deviate from the observed mean, due to possible atypicality of the sample. Similarly, we can think about the standard error of the difference, which tells us how much the true value of the difference is likely to deviate from the observed difference in means.

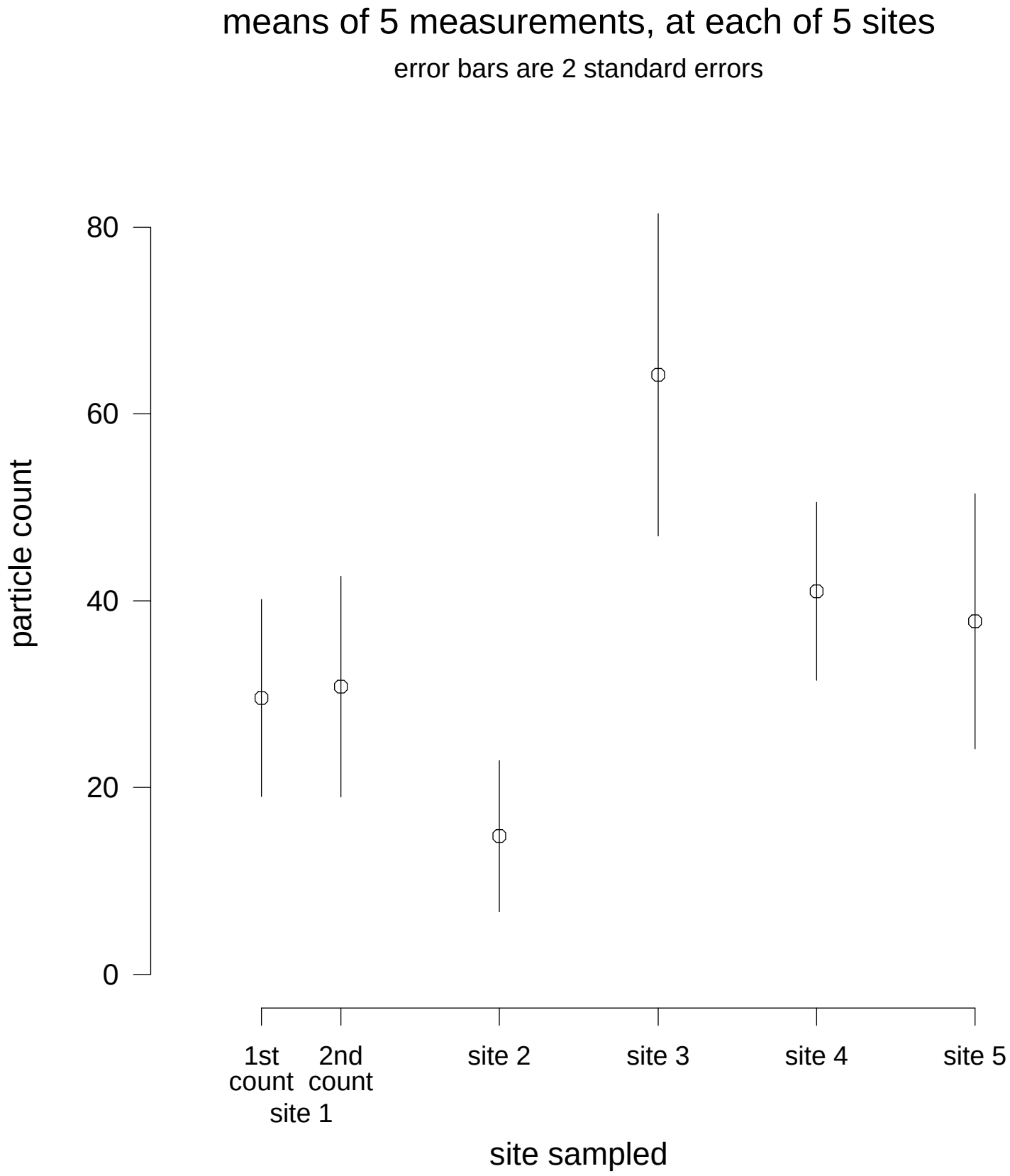
For independent samples, the standard error of a sum or difference is related to two separate standard errors, se_A and se_B, by the Pythagorean rule:

$$se_{diff}^2 = se_A^2 + se_B^2$$

For sites 3 and 4, it happened that the means are 64.2 and 41.0, respectively, and the standard errors are 8.6 and 4.8 respectively. Therefore, the difference is 23.2, and its standard error is calculated by taking the square root of 8.6^2 and 4.8^2, which comes to 9.9 (note that because the squares are added, the much smaller standard error has little influence). Thus, an interval estimate for the difference, using the convention of 2 standard errors, is:

$$23.2 \pm 19.8$$

We can be reasonably confident that the true difference is positive, but the interval estimate tells us that it could be much smaller (or much larger) than the estimate of 23.2 from the two samples.


means of 5 measurements, at each of 5 sites
error bars are 2 standard errors
particle count
80
60
40
20
0
1st count site 1
2nd count
site 2
site 3
site 4
site 5
site sampled