STAT 100 Chance ande Data Analysis Nov. 25, 2002

Today: Parameter Estimation and Hypothesis Testing (continued)
Follow up of Pulse Demo
Article about Human vs Ape (Tanur pp 68-76)

Follow up of Pulse Demo

Recall from last time

37.7 ± 2(.75) = 37.7 ± 1.5 pulses for females
35.8 ± 2(.83) = 35.8 ± 1.7 pulses for males

These [mean ± 2 SDs] are called **95% Confidence Intervals** for the population means. A CI is an interval estimate of a population mean.

So far it looks like 35.8 is not the mean for females and 37.7 is not the mean for males. (Why? Because 37.7-1.5 = 36.2 > 35.8 and 35.8+1.7 = 37.5 < 37.7.) One is tempted to conclude that the population mean pulse for females is greater than the population pulse mean for males. But we must remember that the means 37.7 and 35.8 are sample estimates and where they fall is not the question of scientific interest. We need to pose our question in terms of the population means. The relevant question is:
Are the two population means equal? Or, equivalently, does the difference in population means equal to 0?

Since the difference in population means is estimated by the difference in sample means, we need to use this difference – but to interpret this difference, we need the SD of the difference of sample means.

The SD of the difference is calculated as $\sqrt{.75^2+.86^2}=1.1$ and the sample difference is 37.7-35.8 = 1.9 so we could say the difference is
1.9 ± 2(1.1) or 1.9 ± 2.2.

This last interval does contain 0, so a 0 population difference is credible, and the data does not prove a difference in population means exists. We can conclude that the means of the two population distributions have not been shown to be different. (It is not true that we have shown the difference IS 0.)

The hypothesis testing approach ...

There is another way to report this result: What is the chance that a difference of means of 1.9 or more would result IF their were NO difference in the population means(*)? Because we are averaging here we can assume normality, and now we are asking how likely it is to get a value that is 1.9 pulses greater than 0? If we ask the question in terms of SDs above the mean - how likely it is to get a value that is 1.9/1.1=1.73 SDs greater than the mean? We only know that greater than 1 SD has a chance of 16% [16=(100-68)/2] and greater than 2SDs has a chance of 2.5% [2.5=(100-95)/2]. There are tables of the normal distribution that give the % for any number of SDs and in this case we can find from the table that the chance of a deviation greater than 1.73 SDs is about 4%.

What does this suggest? It suggests that a rare event has occurred, something that only happens 4 in 100 times. But there may be an more credible explanation. Actually the population means are NOT equal (contrary to our tentative assumption at (*)). If they are not equal (and for example the male pulse average were smaller than that for females) the sample results would be ordinary (not rare). Ergo, we conclude that a population difference exists!

This "contrapuntal" logic is what is used very often in statistical work, in the testing of hypotheses. The 4% that we came up with is called **"the p-value".** The tentative assumption of no population difference is called **"the null hypothesis".** The alternative to the null hypothesis is called **"the alternative hypothesis".**

The hypothesis testing approach goes like this:

State a null hypothesis.
Calculate the probability for the sample result, if the null hyp. true.
If this probability (the p-value) is small reject the null.
Otherwise conclude the null is credible.

Some people find the approach that uses the **Confidence Interval** of the difference of means to be simpler to understand than the **Hypothesis Testing** approach.

Another Hypothesis Test Example: Human vs Ape (Tanur pp 68-76)

2.5 million year old fossil elbow – human vs chimpanzee?

many measurements to attempt to quanitfy "shape" of the bone (Table 1 page 73 shows the 7 measurements and their averages for chimps and humans. Call them V_1, $V_{2,}$ ...., V_7. We want an index from these like

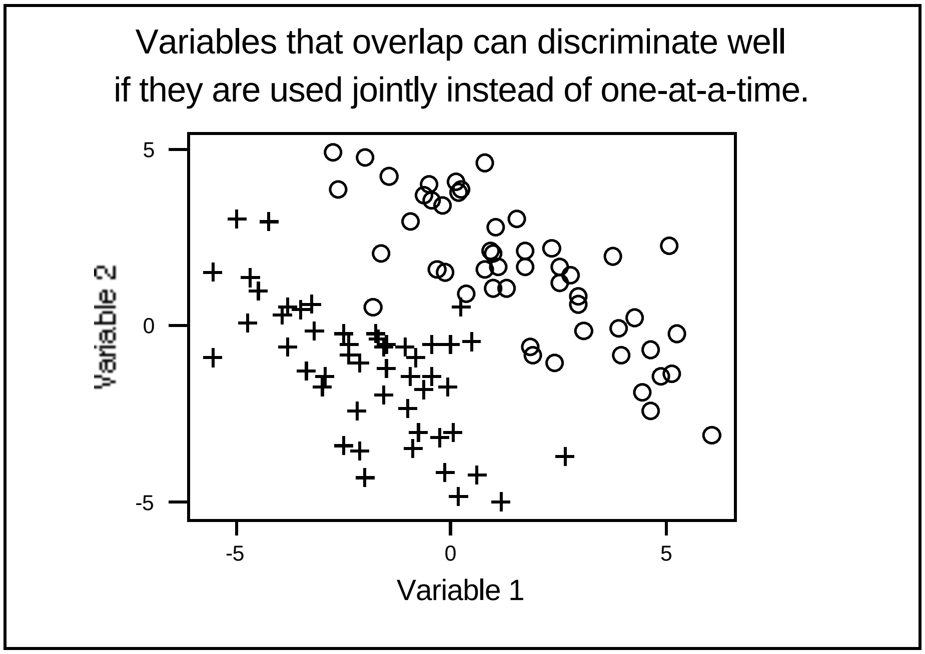
Index = -.09 V_1 + .40 V_2 + ...+ .56 V_7

so big values reflect chimpanzee elbows and small values reflect human elbows. (If the V_i are standardized, by subtracting the mean and dividing by the SD, then the coefficients -.09, +.40 etc will indicate the importance of the varous measurements in forming the index.)

The way the index is formed from the data is complex but easy to find using statistical software. The index is called **the discriminant function.** We calculate this index for any tooth we want to classify as human or ape.

Now we have our fossil elbow and we want to test whether or not it fits with the chimpanzee population (based on the 40 chimps data). It turns out that the chance of getting an index value as small as we did (it was 59.4 see top on p 75) is 1 in 500. IF the elbow is a chimp tooth, it is a very unusual one. A more reasonable explanation is that it is NOT a chimp elbow but more like a human elbow. This is the use of the hypothesis testing logic again.

Did we need all this machinery? Could we not have used one or other of the measurements, instead of all seven of them? Fig 1 in the article tells why. Here is another version of that figure:


Variables that overlap can discriminate well
if they are used jointly instead of one-at-a-time.

This figure shows that, when single variables do not discriminate well between two groups, using more than one such variable may allow very good discrimination. Note that, if one variable DID discriminate well, the index approach would be a waste of time.