Research Methods II, Module 2, Autumn 2002:
Two factor designs

Imagine you want to know whether ginseng speeds up reaction time more for males than for females. Design an experiment that answers this question. What would be your dependent variable (DV), independent variables (IV) and levels of the IVs?

The dependent variable would be reaction time in, say, seconds. There would be two independent variables: gender (with levels of male and female) and dose of ginseng (say, two levels: 0 mg and 200 mg).

You collect your data. How would you analyse your data to provide an answer to your question? What group means would you compare to which others?

You might suggest using one or more t-tests. For example, you could test males on 200mg ginseng vs females on 200 mg ginseng. Let say it is significant. That might just mean that males are faster on the task than females, even without ginseng. That t-test in itself does not answer the question of whether ginseng speeds up reaction time more for males than for females. So you do a t-test comparing males on no ginseng with females on no ginseng. But neither a significant result nor a non-significant result will definitively answer your question. If you got a significant result you can conclude that indeed males are faster on the task than females without ginseng. But both t-tests being significant are still consistent with ginseng speeding up reaction time for males more than for females. You still cannot answer the question with a yes or a no.

Let us say the t-test between the males and females when both are on ginseng is significant (and males are faster than females), but the test between males and females when neither are on ginseng is non-significant. Surely, you say, this shows ginseng affects the genders differently? Surely that shows that without ginseng there is no difference between the genders, with ginseng there is a difference, so ginseng must affect the genders differently? As tempting as that conclusion is, it is wrong. And it is a mistake that is frequently made, not only by students, but in published papers, so it is worth taking a digression to dwell on this point.

Imagine Smith et al run an experiment with two groups, conduct a t-test comparing the two groups, and get a significant result, <u>just</u> significant at the .05 level. Now let us say the difference between the group means that was found happened to be exactly the difference between the population means. And the sample variances within each group happened to be exactly the population variances. The population mean difference means there was a real effect there. And Smith et al detected it. You decide to replicate the experiment. You use exactly the same number of subjects as Smith et al, and follow their procedure exactly. What is the probability that you will get a significant result – that you will detect the population effect that really exists? Either put a number on the probability, or just decide approximately the sort of figure it should be.

Most people think the probability is very high, maybe 0.95. After all, there is an effect there, and an exact same previous experiment detected it. In fact, the probability that you will replicate the effect is 0.5. Only 50 out of 100 identical experiments would get a significant result at the .05 level. For each experiment, the difference between the sample means is unlikely to be exactly the population difference. It is likely to be a little bit more or a little bit less. If the mean difference is a little bit more than the population means then, other things being equal, you get a significant effect – the between group variability, the signal, is more than the noise (within group variability). But if the difference between sample means is a little bit less than the population difference, you get a non-significant result. Remember, the first experiment by Smith et al was only just significant; so if your mean difference is a little bit less than theirs, the result is non-significant. And there's even odds that the sample difference will be a little bit more than the population difference; and even odds it will be a little bit less. Similarly, for each experiment, the sample variance is likely to be a little bit more or a little bit less than the population variance. If it's a little bit less, than, other things being equal, the experiment will be significant. The noise (within group variability) through which you are trying to see the signal is lower, so the result is significant. But if the sample variance is a little bit more than the population variance, the result is non-significant. And there are even odds that your sample variance will be a little bit more; and even odds it will be a little bit less than the population variance. So you are poised on a knife edge – there are even odds that the sample mean difference and variance are each a little bit more or a little bit less than the population values, than the value needed to be just significant. So the probability of getting a significant effect is .50.

The probability of getting a significant result given there really is a population effect there is called power. The power in this example is .50. In fact, .50 is a very common power value for psychology experiments. So typically it is not very surprising if a null result is obtained in an experiment – even if an effect were there, maybe you only had even odds of detecting it anyway. Under those conditions, getting a null result is not informative. You shouldn't feel compelled to conclude there is no population effect. Maybe there is a population effect, maybe there isn't.

Now let's return to the ginseng/gender experiment. Imagine, in reality, men are faster than women on the task, and ginseng makes no difference whatsoever. Also imagine the power for your experiment is quite low, close to .50 – maybe .60. You run males and females without ginseng and you do not get a significant difference between the genders. You run males and females with ginseng and you do get a significant difference. In both cases there was a population difference between males and females, and of exactly the same size. You just detected it on one occasion and not the other. That's just what you would expect with the power of around 0.5. You get the effect half the time.

So a null result under one set of conditions and a significant effect under another set of conditions does NOT mean the effect has changed as conditions changed.

If you want to know whether the effect has changed then you have to test specifically whether the effect has changed. For every precise research question there is a specific test to answer that question. If we want to know whether ginseng speeds up reaction time for men more than women we need to test whether the effect of ginseng on men is greater than the effect on women.

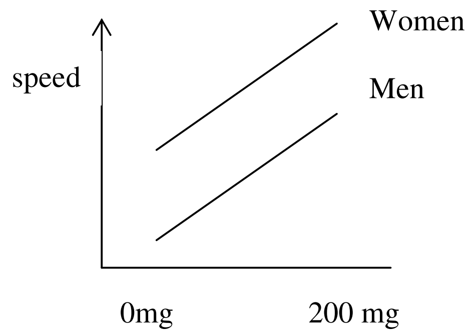
The effect of ginseng on men is a difference. An effect is a difference. We can not talk about the effect of the condition of men having ginseng. A condition by itself is not an effect. The effect of ginseng on men is the difference between the reaction times of men with ginseng and the reaction times of the men without ginseng, a difference between those two conditions.

The effect of ginseng on women is the difference between women with ginseng and the women without ginseng. If ginseng speeds up reaction times more for men than for women, then the effect of ginseng on men is greater than the effect of ginseng on women – there is a difference between the effect on men (which is a difference) and the effect on women (which is another difference). So we need to test whether there is a difference of differences. Only testing whether there is a significant difference of differences answers the question whether the effect of ginseng on men is greater than the effect of ginseng on women.

A difference of differences has a special name in statistics: It is called an interaction. In this case we are interested in whether gender and ginseng dose interact. If they interact, then the effect of gender at one level of ginseng is different from the effect of gender at another level of ginseng. There is a difference of differences. We can equivalently look at it the other way round. If there is an interaction, the effect of ginseng for males is different from the effect of ginseng for females.

Two independent variables can interact with each other, but it does not make sense to talk of the interaction of an independent variable and a dependent variable. If there is just one independent variable, there may be a difference in the mean amount of dependent variable in one level compared to another level of the independent variable. But that's just a difference. An interaction is a difference of differences. Also, if two variables correlate with each other, do not say they interact. A correlation is not a difference of differences. A correlation is one thing, an interaction is another.

If the results looked like this:



Someone might say "The women on ginseng have the highest speed. This is because gender affects speed and ginseng affects speed. They have a combined effect that makes women on ginseng the fastest condition. Because the effects combine, the two independent variables interact". That is wrong. A combination of effects is not an interaction. In the diagram the effect of gender at 0 mg of ginseng is exactly the same as the effect of gender at 200 mg. There is no difference of differences, no interaction. It is precisely because the effect of gender combines by simply adding onto the effect of ginseng in this example that there is no interaction.

In the next lecture we will explore further the kinds of effects one can analyse in a two factor experiment.