Displays for Statistics 5303

Lecture 2

September 6, 2002

Christopher Bingham, Instructor
612-625-7023
612-625-1024

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Components of an Experiment

Recall three components of an experiment of particular interest to statisticians.

- "**Treatments**"

  The statistician can help the experimenter choose which treatments will be included in the experiment.

  This is a particularly important issue when each possible treatment is determined by *several* factors (categorical variables) or numerical variables (amount of water, and length of time in the oven say, in comparing baking recipes).

  Choice of treatments means specifying exactly which combinations of factor levels and/or variable values will be used in the experiment. If you have 20 factors each at 2 levels (say presence or absence of each of 20 chemical additives), there are $2^{20} \tilde{=} 1.05 \times 10^6$ possible treatments to choose from.

- **Experiment units** (EU's), the entities to which a treatment will be applied
  1. Plots in a field
  2. Pens containing many pigs or other animal all getting the same diet (individual animal is *not* a EU)
  3. Classroom containing many students, all being taught the same way (student is not a EU)
  4. Individual patient in a clinic getting individual treatment would be EU.

Correct analysis depends on identifying the experimental units.

- **Assignment method** used to assign treatments to EU's (or EU's assigned to treatments)

In addition, of course, there are many details of experimental technique which may affect your choice of units or assignment method.

Features essential to a good design:

- A good design **avoids systematic error** by including randomness as a part of the assignment method
- A good design yields **precise results**, even in the presence of unavoidable variability.

  You can often reduce variability by careful experimental technique. Often more important is how your assign experimental treatments to experimental units to minimize the effect of variability among EU's.

- A good design must allow for the **estimation of error**. That is, it must allow you to estimate how accurate your comparisons actually are.

  Some designs that do a good job of minimizing error (comparisons are very accurate) don't allow you to estimate the error (quantify the accuracy).

  For example, excessive "balance" can make for very precise experiments whose actual error can't be estimated.

  Balance means, say, making sure each treatment appears equally often next to every other treatment, and in every corner of a field.

- A good experiment has **broad validity**. If units can be considered to be a sample from a population, then you can infer conclusions about the population, not just about the individuals actually studied.

More on Units

It's essential to distinguish between "measurement units" and "experimental units".

A *measurement unit* is the individual or piece of experiment material you actually measure.

- In comparing teaching methods, the experimental unit is usually a classroom, but tests are given to individual students, the measurement units.
- In animal nutrition experiments, the experimental unit is often a pen or a litter, but the measurement unit is an individual animal.
- In a study I worked with a few years ago, comparing different policies for providing reduced price meals in schools, the experiment unit was a school (I think there were 4), but the measurement unit was the individual student (there were hundreds).

- In experiments on particle board, treatments are applied to the boards (EU's), but compression strength measurements may be made on several samples (measurement units) taken from each board.

Caution: Don't treat measurement units as if they are experimental units.

If you do, you get unrealistically small estimates of errors, which can make small effects seem highly significant.

The effective sample size is the number of experimental units, not the number of measurement units.

The distinction between experimental units and measurement units is actually somewhat more complicated that this.

In **split plot experiments**, for example, there can be two or more types of experimental units with "Low level" EU's serving as measurement units for the high level EU's.

Example

Baking experiment to study the effect of high and low levels of three ingredients in a recipe ($2^3 = 8$ combinations) and the effect of baking temperature.

You have ovens that that can bake 8 batches at a time at *one* temperature.

- The **high level** EU's (whole plots) are the "runs" of the oven at a given temperature with 8 batches, one for each of the $8 = 2^3$ combinations of ingredients..
- The **low level** EU's (sub plots) are batches made up of the 8 combinations, randomly placed in the oven.
- The measurement units are the 8 batches in an oven, that is, the same as the subplots.

The analysis must reflect complications of this sort. Ignoring that there are two or more levels of EU's can lead to erroneous conclusions.

An important design problem is when there is a fixed amount resource available for experimenting:

- A fixed sized field
- A fixed amount of time
- A fixed amount of dollars to pay experts who will judge the results

Part of the design problem is how the resource should be divided up.

- Size and shape of plots in a field
- Number of samples per session for a judge and the number of sessions

A useful rule, not always applicable, is to *do more less well*. That is, have more experimental units (the real units of replication), instead of more measurement units.

It may be better to study one classroom in each of 20 schools, than 5 classrooms in 4 schools.

Another issue at the planning stage is to ensure that the response of one unit should not be influenced by the treatments or responses of other EU's, especially neighboring EU's. This may dictate substantial separation of units, or provision of "buffering" between units.

There can also be edge effects, both in time and space.

- Plants at the edge of a plot or a greenhouse flat, have other plants on only one side. This might make a difference in how they respond to treatment. (Space edge effect)
- It make take several runs in the morning for everything to be working properly. (Time edge effect)

I haven't said much about responses, the actual measurements made.

Often (usually?) there are several measurements made, each addressing a different question. That is, responses are potentially multivariate.

Possible responses from an experiment comparing ways to bake a cake.

- The amount of moisture in a cake.
- An assessment of the texture of a cake
- How good the cake tastes
- The shelf life of the product

A response that directly measures the quantity needed to answer a question is sometimes called a *primary response*. A direct measurement of how much H_2O is in a cake, would be a primary response.

In some cases, the you can't measure the actual primary response, perhaps because it would take too long. In such cases, you may measure *surrogate or proxy responses*.

- In comparing cancer treatments, the *primary response* might be the years of life after treatment. But that can take decades. A surrogate response with some of the same information is the fraction of patients still alive 5 years after treatment.
- You can't directly measure how good a cake tastes, but you can elicit numerical evaluations from trained judges as a surrogate response.
- You can get some information about shelf life by making measurements after 6 months, say, and comparing them with similar measurements of a fresh product.

Another class of responses are *predictive* responses. These are measurements that don't directly bear on the question of interest, but may help predict a primary response.

Such responses are often called *covariates* and are sometimes accounted for by using the **analysis of covariance**.

When a covariate helps predict or explain the *errors* (deviations from the mean), using it can *increase the accuracy* of an experiment.

Example

- Primary response might be this year's yield on plots treated differently.
- The predictive response might be last year's yield when all the plots were treated the same. This might give information about intrinsic differences among plots that were not related to treatments but that affected this year's responses.

Sometimes a covariate is itself affected by the treatments. In such a case, including it in the analysis can help in explaining how a treatment works.

Example comparing varieties of grasses

- Primary response is the total dry weight on a plot.
- Predictive response is the number of plants per plot, which also might differ among varieties.

If there is a difference of yield (effect of variety), much or all of the differences might be just a consequence of the number of plants. An analysis of covariance can show how much of the treatment differences in the primary response is a result of treatment differences in the predictive response. This might allow for full or partial explanation of observed differences.

Randomization

An experiment is *randomized* whn the method you use for assigning treatments to units involves explicit use of an understood probability mechanism.

Proper randomization is essential to ensure that uncontrolled or unobserved factors don't introduce systematic bias in your results.

Randomization is not "haphazard" assignment

You should not just assign treatments sequentially to the next experimental unit, even though it appears to you there is no pattern among the units.

In putting cakes in an oven, they should not be placed in the order you happen to pick batches up.

Examples from text of ways to assign four treatments A, B, C, D to 16 EU's

1. Use 16 slips of paper marked with letters, 4 A's, 4 B's, 4 C's, 4 D's. You mix them thoroughly in a box or basket. For each EU, you pick a slip and assign it the treatment indicated.
2. You assign A to the first 4 units you have at hand, B to the next 4, ...
3. If, when you get to an EU, the second hand is between 1 and 15, the unit gets A; if between 16 and 30, it gets B, ...

Only 1 is true randomization. The probabilities of any possible assignment can be determined (every possible assignment has P = 1/63,063,000).

The second is at best haphazard.

If there is a long time between runs, the 3rd may be almost random, but it's impossible to know exact probabilities.

Why randomize?

- Protection against "confounding", the effects of uncontrolled or unmeasured factors (lurking variables) that might affect a non-randomized assignment
- Randomization can form a basis for inference

Confounding

One source of bias is judgement of the possible outcome on the basis of what is known or seen before the treatment.

In assigning a medical treatment, a physician might guess a particular patient would be helped more by therapy A rather than B and therefore assign A. Now the condition of the patient is confounded with the treatment. There will be no way to know whether any difference in outcomes is due to the treatment or to the prior judgement, a lurking variable

However, if assignment is truly random, then any other factors (lurking variables) that might have affected a non-randomized choice, will tend to average out so they affect all treatments equally and have little or no net effect.

- If there are male and female patients, and assignment is random, there will be an approximately equal proportion of men getting each treatment.
- Similarly, average ages for each treatment will be close.
- Possibly more important, other factors that are *not* observed or measured will tend to affect all treatment groups equally.

There are often several possible randomization methods.

If there are two treatments you might:

1. Flip a coin for each patient, heads give treatment A, tails give treatment B.
2. Take patients in pairs, perhaps approximately matched by demographic or physical variables, flipping a coin to choose which gets A, then giving B to the other patient. This has the advantage you are sure to get an equal numbers of A's and B's.
3. If you have 30 patients, say, select a random sample (without replacement) of size 15 to get A and give B to the remaining 15 patients.

Important
Your analysis depends on how you randomize. For 2 you would probably use **paired t**. For 3 and probably for 1 you would use **two-sample t**.