Lecture 2 (ch. 1)

As I have said, a great deal of stat 421 is about learning The language of experimental design. And like learning any language, That's a non-linear, complex process. As a result, it's hard to find a unique place where This page belongs. So, I'll put it here at The beginning; but be aware that you will have to go back and forth between The rest of The lecture(s) and This page.

Treatments:
Treatments are the different procedures we want to compare. Each treatment can be thought of a level of a single factor (called the treatment factor), or a combination of levels from multiple treatment factors. Other facfors will have other names, like nuisance factor.

Run:
1 run refers to performing one experiment at a given treatment combination level.

Experimental Unit (EU):
Experimental units are the things across which the treatments are randomly applied.

Randomization:
Randomization refers to the random assignment of treatments to EUs, (AND the random order in which the different runs are performed).

Experimental Error:
Also known as "measurement error," is the random variability in the observations of the response variable. It can result from variability across different EUs being subjected to a given treatment, or from variability across different treatments applied to a given EU.

Control:
That term arises in "control experiments," in which case it refers to the "controlled" (but still random) manner in which treatments are assigned to EUs. It also arises in "control group," in which case in refers to a special treatment with which other treatments are to be compared.

Replication:
Refers to the runs arising from a complete repeat of all possible treatment combination levels.

Repeated measurements:
skip (for now)

Blocking:
Refers to the situation where there is randomization within blocks of similar EUs.

Last time we talked about physical vs. empirical models. Most of what we do in 421 will deal with the latter (Ch. 3, 5). We also talked about observational vs. controlled studies. Although the models we will be developing in Ch. 3, 5 can be applied to data from both types of studies, most of what we do in 421 assumes we are dealing with controlled experiments. And, as I said, starting from Ch 6 we will learn about <u>how</u> to perform the controlled experiments.

We then started by looking at this example:

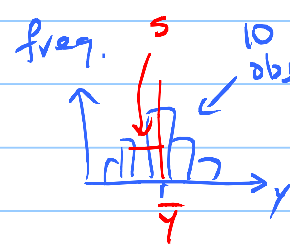
Example: How hard is a given chunk of metal?

We press a standard tip into a piece of the metal (called coupon), and then measure depth. $= y$

$\boxed{\text{Design 1}}$ Apply a tip to 1 coupon

Data: One y value. Answer: $y \pm ?$

$\boxed{\text{Design 2}}$ Apply a tip to 10 coupons:



Data:

coupon	1	2	...	10
	y_1		...	y_{10}

sample std. dev.

Answer: $\bar{y} \pm \sigma/\sqrt{10}$; estimate σ with s (Recall, $\underline{s}$ does <u>not</u> vary with sample size)

Assuming there is no difference across coupons.

Language: 10 runs,

A discrete random var. (r.v.) with 10 levels

1 factor (x) : coupon.

Chance are that in your past encounters with the data, this "coupon" factor is something you just called "case," and then ignored it. As you will find out, it's a factor that you do need to pay attention to, especially in controlled experiments.

So, start paying attention to it (even though we will find out, later, that it's a nuisance factor. And in Ch. 4, you will learn what to do with nuisance factors.)

$\boxed{\text{Design 3}}$: Apply a tip to 10 coupons, twice.

Data:

Coupon	1	2	---	10
	$Y_{1,1}$ $Y_{1,2}$	$Y_{2,1}$ $Y_{2,2}$	---	$Y_{10,1}$ $Y_{10,2}$

Language: 1 factor : Coupon (10 levels)
OR 2 factors : Coupon (10 levels) and Replication (2 levels)

Also: 10 runs, replicated 2 times. [or 20 runs!]

Answer : $\bar{Y} \pm \sigma/\sqrt{20}$ [ie. more precise that above].

Assuming there is no difference across coupons nor replications.

$\boxed{\text{Design 4}}$: Suppose we suspect the tip wears out. So, we test 2 tips, each on a different coupon.

Data:

		Coupon 1	Coupon 2
tip	1	Y_{11}	Y_{12}
	2	Y_{21}	Y_{22}

Language:

2 factors $\begin{cases} \text{coupon (2 levels)} \\ \text{tip (2 levels)} \end{cases}$

4 runs.

Design 5: Data:

		Coupon 1	Coupon 2
tip	1	$y_{111}\ y_{112}$	$y_{121}\ y_{122}$
	2	$y_{211}\ y_{212}$	$y_{221}\ y_{222}$

Language: 2 factors, 4 runs, replicated twice [or 8 runs!]

OR: 3 factors: coupon, tip, replication

First replicate: y_{ij1} $\quad i,j=1,2$

Second replicate: y_{ij2} $\quad i,j=1,2$

Note: each replicate has <u>all</u> the treatment level combinations.

This is an important part of the defn of replication

Note: y_{111} and y_{112} are actually on the same coupon.
You may be tempted to ask what if they are on diff coupons.
Then we're back to an unreplicated design:

		Coupon 1	Coupon 2	Coupon 3	Coupon 4
tip	1	•	•	•	•
	2	•	•	•	•

2 factors
8 runs.

In all of the above experiments, it's important to assign the tips to the coupons in a random manner. This is an example of what we mean by "randomization".

What's the answer to the question?!

Before we can answer, read on.

Data:

		Coupon 1	Coupon 2	
tip	1	y_{111} y_{112}	y_{121} y_{122}	→ mean
	2	y_{211} y_{212}	y_{221} y_{222}	→ mean
		↓ mean	↓ mean	

across tip   across coupon

We know that the y's fluctuate both row-wise and col-wise.
But in answering the question of hardness, we have to wonder

Does depth vary with tip?
Is there a tip effect?
Are the row-means different?

Does depth vary with coupon?
Is there a coupon effect?
Are the col-means different?

Is there a replication effect?

Suppose, we don't care about the replication effect at all. Then we must consider the replication factor as a nuisance factor. Then there only 2 treatment factors tip and coupon.

we do care about these.

If there is no tip effect
no coupon effect
no repl. effect. Then Answer: $\bar{y} \pm \frac{\sigma}{\sqrt{8}}$

Else, the answer will require estimating each of those effects. That's what 421 is about. Also, see below.

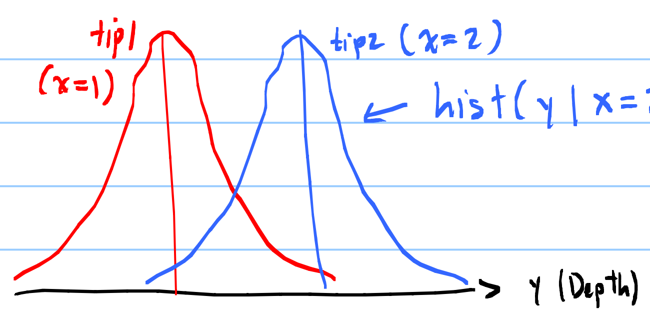
To simplify, suppose we have 1 treatment factor: tip (2 levels)
1 nuisance factor: coupon (say, 20 levels), and No repl.

To answer The question How hard is The metal? we also have to ask Is There a difference in The y values between tips?

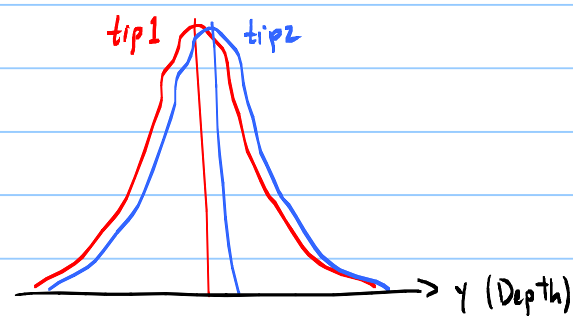
If There is a difference between tips, Then There are 2 answers to The original question; otherwise There is only 1 answer.

Q: But how do we find out if tip i.e. treatment has an effect on depth ?

A: Look at conditional histograms



There is tip effect



Cannot Tell!
[Wrong conclusion: There is no tip effect.]

So, if The hists are too wide, Then we cannot Tell if tip has an effect. That's not good! But we can fix That problem.

Let's ask a different question.

Q: Why are the hists wide at all?

A: Because there is variability across coupons.

Remember that the coupon factor is a nuisance factor.

This is what nuisance factors do: They introduce variablity into the data. And that variability can be so large that it prevents us from determining whether the treatment factor has an effect. (Hence the name "nuisance").

But we can eliminate that variability by a clever design: By applying 2 tips to each of the coupons, and comparing the depths pair-wise, we can essentially eliminate the variability due to coupons. In this design of the experiment, we say that we have "blocked the coupon factor," or we say "coupon is a block factor." ← $\boxed{\text{Language}}$. A generalization of a "paired design".

Blocking is an experimental design method that allows one to eliminate variability due to nuisance factors.

Such designs have higher power, i.e. they are more "likely" to find an effect (e.g. difference between 2 tips), if/when an effect truly exists (e.g. 2 tips really are different).

More on power, later.

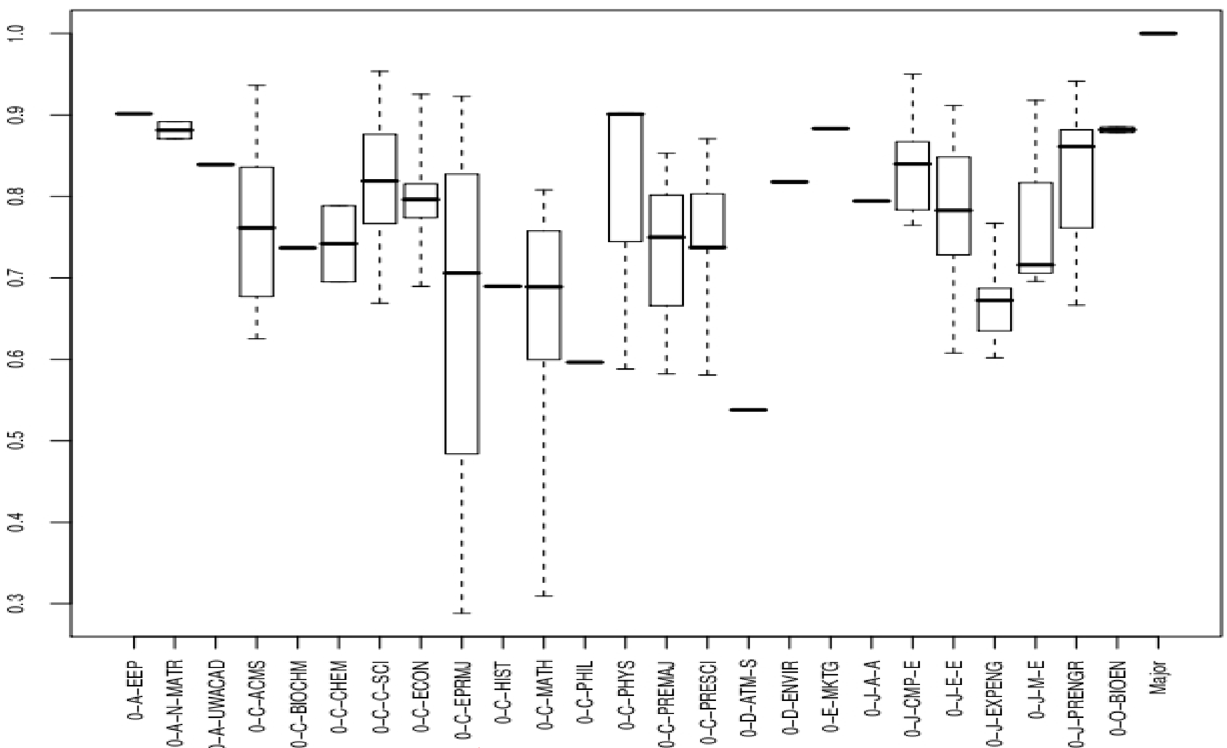
Much more on Blocking, in Ch. 4.

Q: What if There are more Than 2-levels in The treatment factor?

A: Comparative boxplots [trivial to make, but takes maturity to interpret.]

test Scores

Y



x = treatment fact = major.

For This example we would say x has an effect on y.

But, learn to discuss pictures like This, as we did in class.

hw-lect 2-1

Consider The design above leading to This data.

		Coupon 1	Coupon 2
trip	1	y_{111} y_{112}	y_{121} y_{122}
	2	y_{211} y_{212}	y_{221} y_{222}

a) Suppose This data are the result of a design in which The coupon factor was blocked. As discussed in class, such a design (with 1 nuisance factor with 2 levels) involves looking at differences between numbers. If The treatment factor is tip, write down The relevant differences.

b) Additionally, suppose There is no difference across replication. Then, the answer to the hardness question involves finding means. Write down The relevant mean (or means).

c) A notation That we will learn a lot more about (in Ch. 3) is to use a period to denote summed indices. E.g. $y_1 + y_2 = y_.$ Write The above average(s) in This notation.

This is an easy problem, so don't spend too much time on it. The main purpose is to give you more practice with The language.

hw_lect2_2

Consider the following data pertaining to a treatment factor, X, taking 8 levels denoted 1 through 8. There are 5 replications labeled 1 through 5. The response Y is a measure of accuracy (e.g. in weather prediction), and X denotes 8 different models used to make the predictions. Make comparative boxplots that allows one to compare the performance/goodness of the 8 models, and discuss the results as thoroughly as you can (within half a page!). By R.

	1	2	3	4	5
1	0.73	0.62	0.62	0.82	0.68
2	0.75	0.55	0.64	0.00	0.30
3	0.65	0.46	0.52	0.48	0.64
4	0.71	0.49	0.56	0.66	0.99
5	0.61	0.28	0.35	0.62	0.52
6	0.75	0.34	0.89	0.66	0.80
7	0.08	0.27	0.28	0.49	0.81
8	0.87	0.97	0.78	0.98	0.75

hw_lect2-3

In the lab I explained the idea of the sampling distribution very generally. But the one you saw in the lab is, technically, the *empirical* sampling distribution. The sampling distribution itself is a distribution - not a histogram. For some simple problems, it is possible to analytically derive the sampling distribution. That's something that you ought to know, but let's make sure you do.
Consider the followingprobability mass function:
x = 1, 2, 3
p(x) = 1/2, 1/4, 1/4
Consider samples of size 2.
a) Derive the sampling distribution of the sample geometric mean. Hint: geometric mean of x1 and x2 is sqrt(x1*x2).
b) Write R code to compute the empirical sampling distribution of the sample geometric mean.
c) Overlay the sampling distribution on the histogram in part b (to check that they are consistent).

Don't forget to finish reading ch. 1. !