Hôpital général juif
Jewish General Hospital

McGill

INSTITUT LADY DAVIS DE RECHERCHES MÉDICALES / LADY DAVIS INSTITUTE FOR MEDICAL RESEARCH

LDI Seminar Series in Biostatistics: Lecture 5

Kevin McGregor

October 23, 2017

Last time

- Showed how to form the one-way ANOVA test statistic
- Covered some basic post-hoc tests
- Simple introduction to the factorial ANOVA

Today

Putting it all together:

- We have seen a few simple tools to do statistical analysis
- But you should be thinking about the statistical analysis right from the conception of the project! Not just after you've collected data.
- What sorts of things should you consider before collecting data?
- When you finally do get the data, what should you do before starting any formal testing procedures?

Sample size and power calculations

Need for sample size calculations

- A sample size calculation is a common initial step when undertaking a new research project.
- Allows researchers to get reasonable estimates of the amount of data that would need to be collected to have a high probability of detecting an association (if such an association does truly exist).
- Involves creating a set of "what if" scenarios... various parameters will determine the required sample size, but we don't know their true values.
    - Though there may be existing research to give an idea of reasonable values of these parameters to try.

Reminder: types of errors

		Reality	
		H_0 False	**H_0 True**
Test	**Reject H_0**	Correct rejection H_0 = Power = 1 - β	Type I error = α
	Do not reject ~~**Accept H_0**~~	Type II error	Correct acceptance of H_0

- Mainly interested in *power*, i.e. the probability of rejecting the null hypothesis given that it really is false.
- The type I error rate (probability of rejecting the null given that it's true) is actually equal to the significance level α.

Basic elements going into a sample size calculation

- Determining the type of test that will be performed is the first step to doing the sample size calculation.
  - To get a formula for sample size, we work backwards from the final test statistic (and its distribution under the two hypotheses) to see what values determine the required sample size.
- Basic important items: power we'd like to achieve, type I error rate (same as significance threshold), effect size, and standard deviation.
  - Will choose a range of reasonable values for each of these and see how the required sample size changes.
- Will illustrate through the independent-samples t-test.

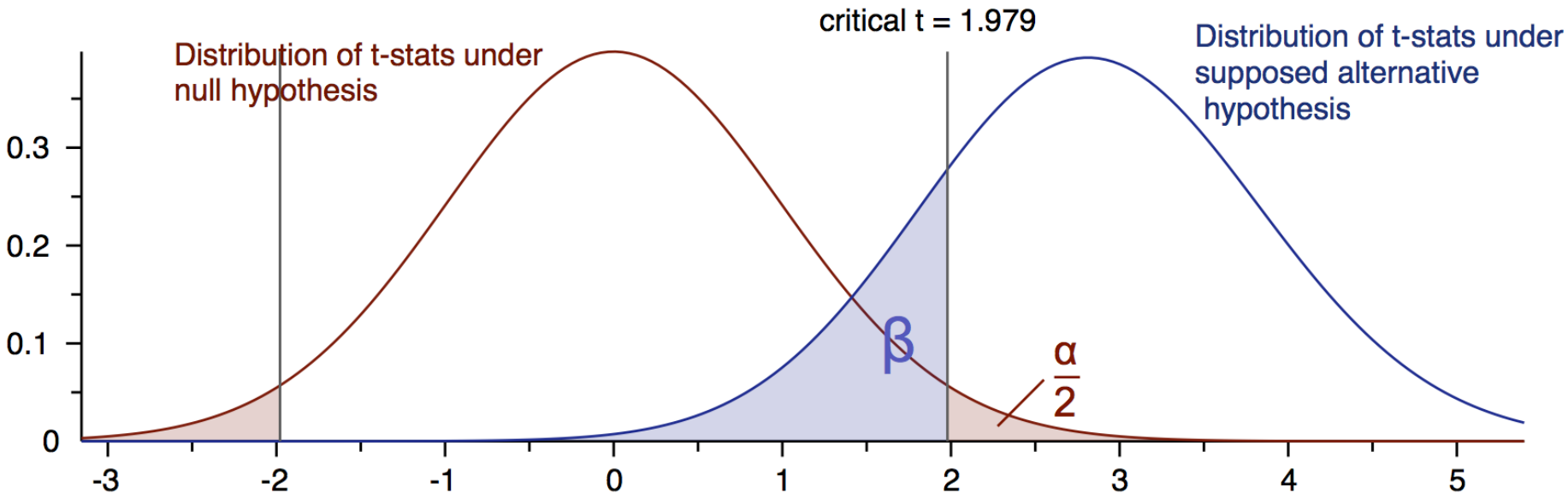
Independent-samples t-test sample size calculation

- Recall that in the independent-samples t-test we're looking at the difference between the means of two groups $\Delta = \mu_2 - \mu_1$.
- Standard deviation in each group is σ.
- Assume we want to have a power of $1 - \beta$ (common to choose $1 - \beta = 0.8$).
- Assume we will test at significance level α.
- The minimum sample size in each group is then calculated as:

$$n_{group} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

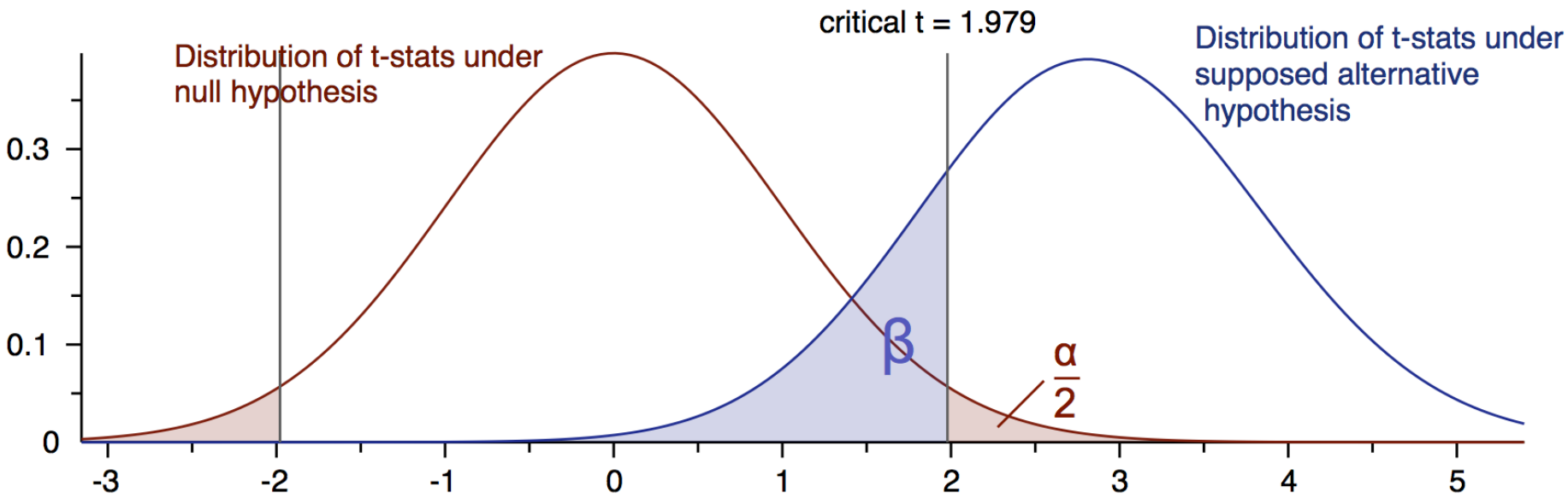
where $z_{1-\alpha/2}$ is the value on the standard normal distribution such that the upper tail probability is $1 - \alpha/2$.

Visualizing power



- The blue distribution is true distribution of t-statistics we would get under repeated samples from the population (fixed sample size).
- β is the probability of a type II error (do not declare significance when the null hypothesis is false).
  - Then $1 - \beta$ is the *power* (or *sensitivity*).

Visualizing Type I error

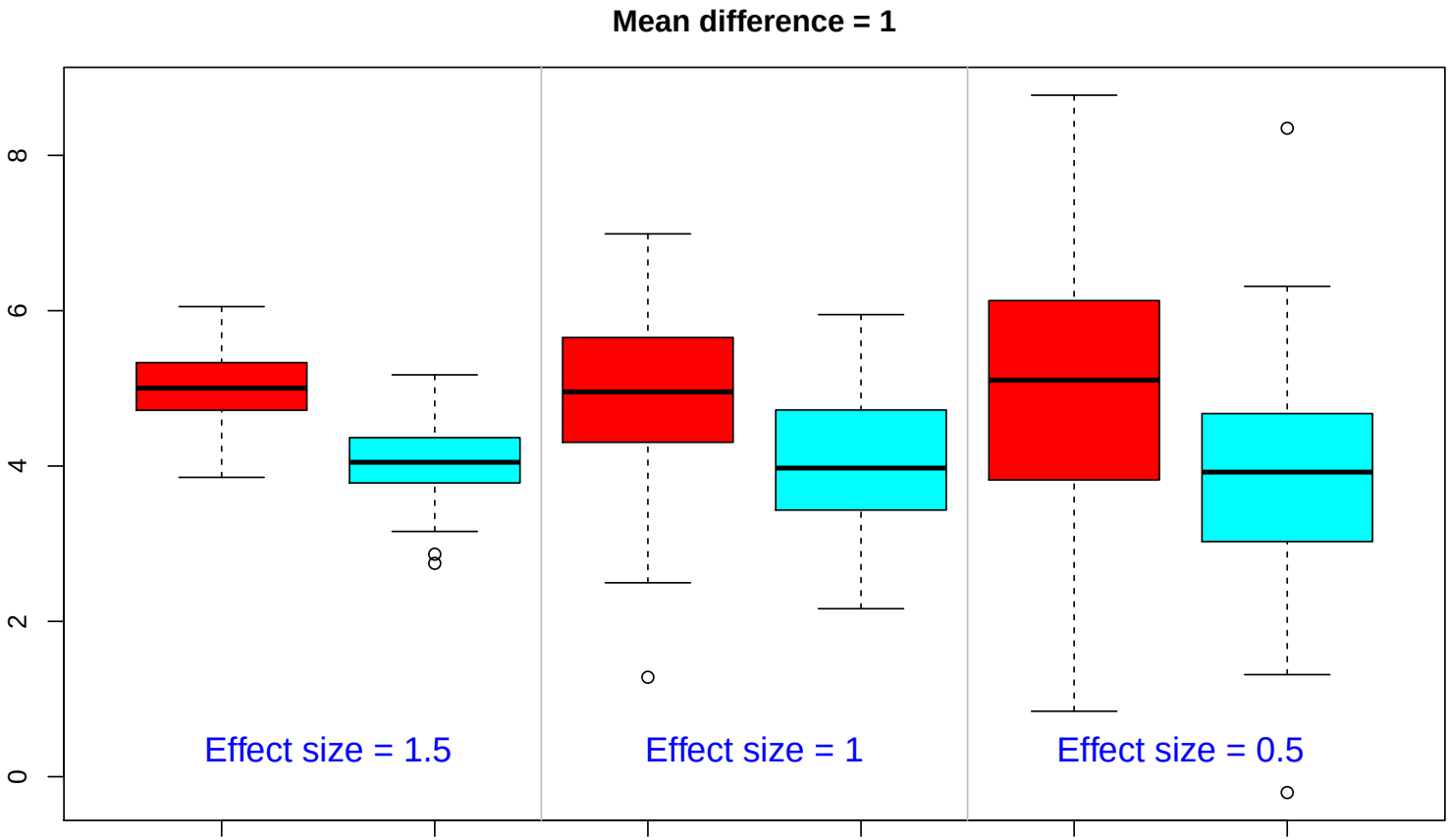


- The red distribution is the distribution of t-statistics if the null hypothesis were true.
- α is the probability of a type I error (declaring significance when the null hypothesis is true)
  - $1 - \alpha$ is the *specificity*.

Independent-samples t-test sample size calculation

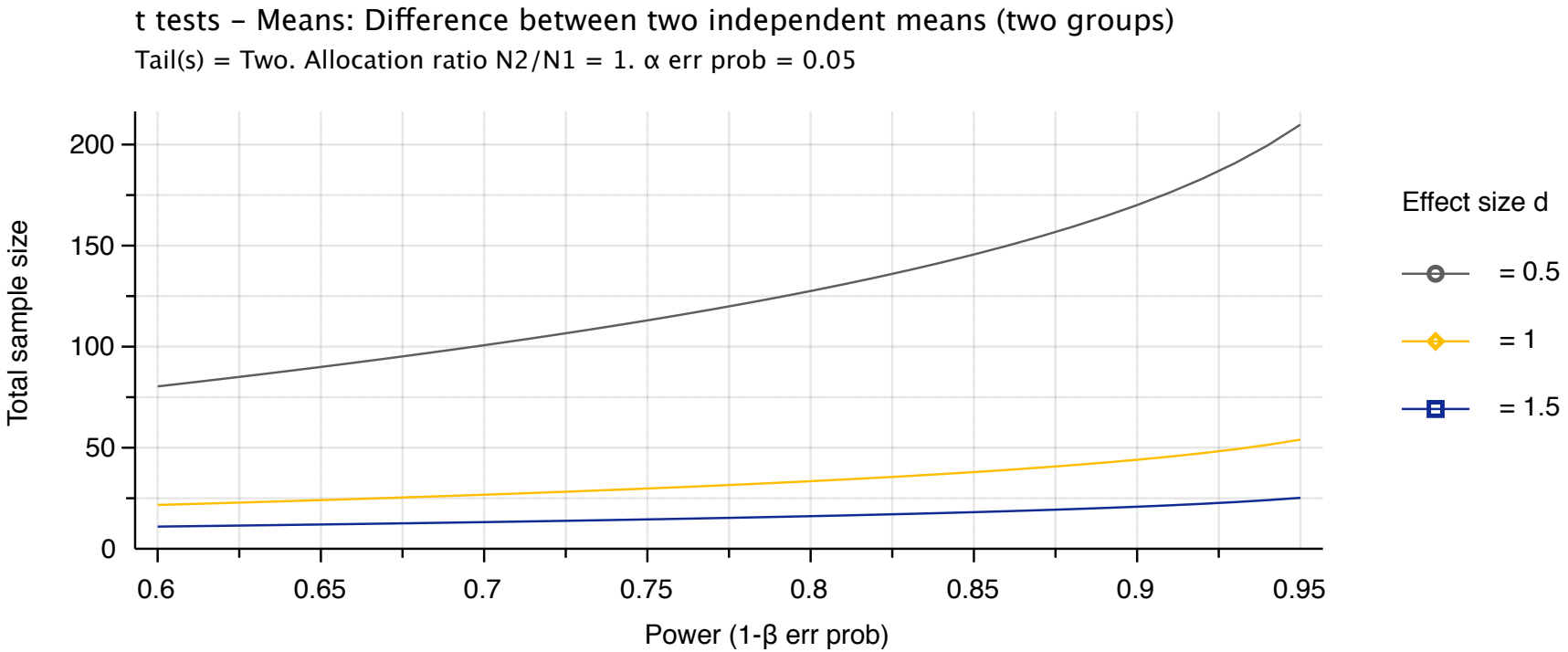
- To do the actual sample size study, we simply choose our power and significance level. Say, power $= 1 - \beta = 0.8$ and $\alpha = 0.05$.
- Choose a reasonable range of values for the mean difference Δ and the within-group standard deviation σ.
- Alternatively, you can specify values for the ratio $d = \frac{\Delta}{\sigma}$. This is the mean difference relative to the standard deviation.
  - Let's try three values for d: 0.5, 1, and 1.5.

Effect size



Same mean difference, but three different effect sizes.

Sample size vs. power



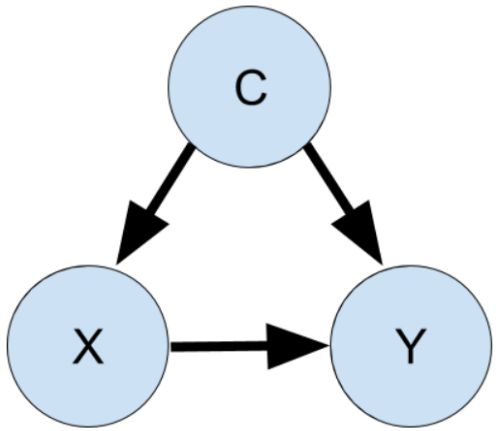
Sample size calculation in practice

- Sample size calculations can be performed for more complicated analyses. (Multiple regression, ANOVA).
- Can also take other items into account: dependent observations, correlation between variables, confounding factors.
- For the most standard statistical tests, a good (free) software choice is *G*Power*
  - Other online resources also available.
- For very complicated analyses, a simulation can usually be done to determine power.

Randomization

Confounding factor

- Suppose we're interested in the relationship between some predictor variable X and an outcome variable Y.
- In many observational studies there are factors which can induce a spurious association between X and Y.



- The variable C has an effect on both X and Y, and is therefore called a *confounder* or *confounding factor* in the relationship between X and Y.

Confounding factor

- A confounding variable can be dealt with by adjusting for it in the model.
- For example, in linear regression the confounder could be included as an additional predictor.
- There's an extension to ANOVA called ANCOVA (Analysis of Covariance) which would allow you to run an ANOVA while adjusting for the confounder.

Confounding factors in experiments

- In observational studies confounding factors are outside of the control of the investigator (e.g. age, sex, socioeconomic status, smoking, etc.)
- When designing an experiment, however, we have control over which individuals fall into different experimental groups.
- If we do a poor job of assigning individuals to the experimental groups we can end up with confounding variables through study design.
- Therefore, the assignment of experimental conditions (or treatments) must be *randomized*.
    - This way we know that no external variables could have affected the assignment of the experimental conditions. Theoretically don't have to adjust for anything.

Checking balance

- With a small sample size, randomization might not be enough...
- Could end up with imbalance of different kinds of individuals in the experimental groups just by chance.
  - e.g. Treatment group vs. control group in mice: could end up with more males in one group than the other.
- If imbalance is found before the experiment is started, you can rerandomize.
- Otherwise variables that are not balanced between experimental groups will have to be included in the analysis.

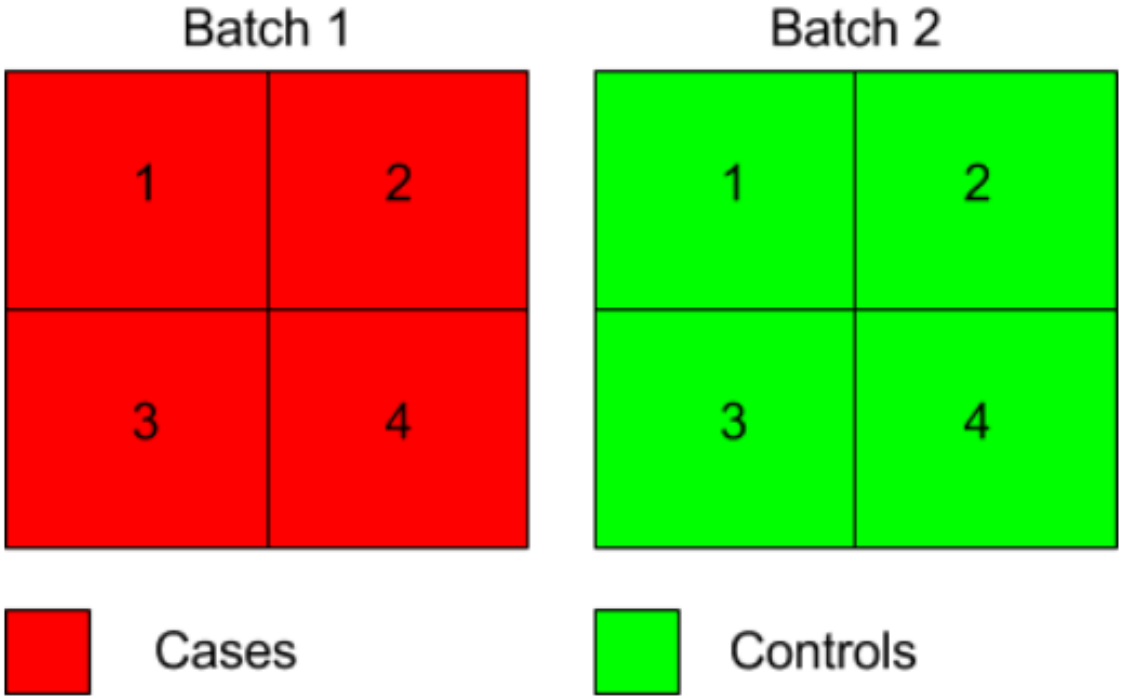
Technical factors

- Randomization is also important in mitigating the effect of technical factors.
- Factors in the lab can have an impact on measurements:
  - Processing samples in different batches (i.e. "batch effect")
  - The day that the samples are processed
  - The lab technician doing the experiment
- Randomize over everything! Don't process all samples from one experimental group in one go.

Batch effect example

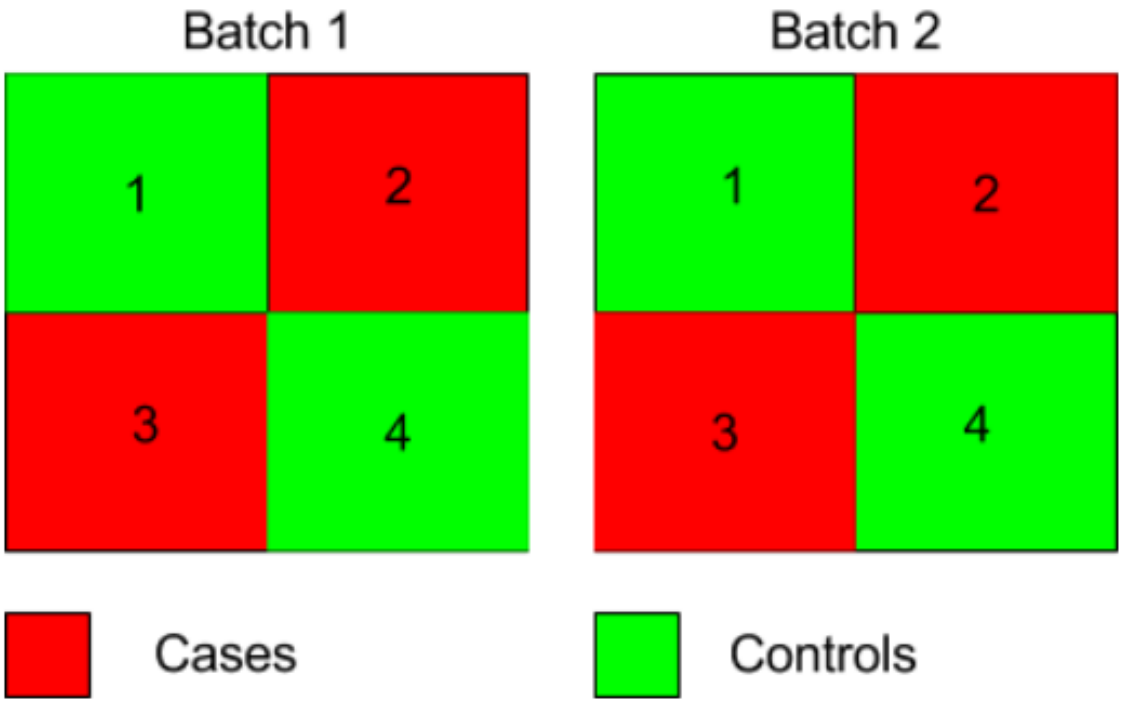
- Recent example: examining differences in DNA methylation on the Illumina 450K chip with respect to disease cases/controls.
- Wanted to see DNA more methylated in case samples than in control samples.
- 450K platform can handle batches of 96 samples at a time. However, the effect between batches can be strong.
  - Normalization methods exist to mitigate batch effect.
  - Study design is still very important!

Bad study design



- Confounding by design... can't separate out batch effect from case/control differences.
- Normalization does not solve the problem!

Better study design



- Better design... making sure there are an similar number of cases and controls on each plate.

After you have data

Creating a dataset

- Getting the data into a proper dataset can take a long time.
- Have to be very consistent with entering data
  - Make sure all factor variables are written exactly the same way... i.e. "male" would be considered a different group from "Male" in most software.
  - Dates should be in the same format throughout the file.
  - Use one kind of flag for missing data (I've seen empty cells, "NA", "N/A", "n/a", "na" all in the same dataset... the computer thinks they're different things!)
- Use informative variable names, and keep a separate file where you describe the different variables in more detail.

Think about confounders

- If you have observational data, you need to think about possible confounding variables.
  - Start by thinking about what variables could affect the outcome variable.
- If you have a randomized experiment, check for balance of your variables in the experimental groups. Might still have adjustment to do.
- Think about who or what the study applies to. Can results be generalized?

Running the test

- Think very carefully about what question you want to answer.
- Make sure you understand whatever test you're running... if not look it up or ask someone.
- Think about independence assumption... (remember that repeated measures from the same individual are not independent!)
- Check assumptions that can be checked.

Verify results

- Once you've done the test it's important to think about the vailidity of the results.
- Looking at effect sizes: does the estimated effect size make sense given the units of measurement in the data?
- Was a relationship between two variables the opposite of what was expected? Double check to make sure things are coded correctly (i.e. for a binary treatment variable maybe the 1's and 0's need to be reversed).
- Worried about an outlier? You can run the analysis with and without the outlier to see how sensitive the results are.
- **The final model you choose should be based on good statistical practice... not the one that gives** $p < 0.05$

Take-home message

- Statistical analysis is an incredibly important part of doing good research. You need to be thinking about it from the very beginning.
- It's very easy to run statistical analyses using modern software. However, you still need to have a good understanding of what the test is doing and how to properly interpret the results.
- Always a good idea to consult a statistician... and do so before you even begin an experiment!

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. -Ronald Fisher

Thank you! - Merci!

Questions?