## Introduction

The premise of each lesson is considering the ideal data science experience and then discussing what makes it ideal and more realistic settings.

1. The data pull is clean
2. The experiment is carefully designed, concepts
3. The experiment is carefully designed, things to do
4. Results of analyses are clear
5. The decisions are obvious
6. The analysis product is awesome

Let's drill down into each lesson to discuss the content.

1. The data pull is clean. In this lesson, we discuss the ideal setting where creating the analytic data set goes perfectly. Since this never happens, we give some strategies for managing the creation of analytic data sets. The first is creating summary tables with means, medians, other quantiles and standard deviations. The second is looking at diagnostics from regression output. The third is a strange little fact called Benford's law. The final suggestion is to perform sampling-based data quality queries.
2. The experiment is carefully designed, principles. In this lesson, we cover experimental design and what happens when we have observational data. We also discuss causality and randomization's role in uncovering it. Then we move on to discuss confounding, a central problem of observational studies and adjustment, the first line of defence for confounding.
3. The experiment is carefully designed, things to do. OK now that we know some of the concepts we can talk about things to do. First, we discuss A/B testing, the adjustment for confounding. We also discuss methods for combating sampling bias.
4. The results of analyses are clear. In this lesson, we cover issues when the results of analyses aren't easy to understand, especially focusing on hypothesis tests. First, we broadly discuss the need to consider more than hypothesis tests. Then we consider issues and strategies with testing and statistical results. Next, we discuss multiple comparisons and their role in hypothesis testing. Then we suggest comparing results of unknown phenomena with that of known. Finally, we discuss negative control analyses.
5. The decision is obvious. There are many reasons why decisions are not obvious. First, the data can be very equivocal about the hypotheses. We discuss power as it relates to this issues. Secondly, even in the presence of significant results, one may not have measured the right quantities.
6. The analysis product is awesome. In this module we discuss the final report. Here we make two recommendations. The first is enforcing reproducible of analysis documents. We suggest tools for doing this. Secondly, we suggest version controlling the software and give recommended solutions.

Throughout, we focus on practical recommendations that can be made to data scientists.

## The Data Pull is Clean

In almost every data science project, there is a large effort in organising an analytic data set. This often requires data munging, web scraping, pulling data from a larger more complex dataset, merging datasets and formatting changes. In the ideal, this process goes very smoothly and the analytic dataset is a clean representation of the desired process. In real life, this process is fraught with errors.

How do you keep on top of data quality without being in the trenches? In this lecture, we give three strategies.

1. **The construction of summary tables.** The summary tables are very useful for catching data errors. Especially useful is keeping track of units and recording several summaries (means, medians, maxima, minima and other quantiles and standard deviations). By contrasting reports over time, you can check to see if things are changing in the processing that shouldn't be.
2. **Regression diagnostics.** Regression is a universal first step in analysing data. Regression diagnostics are useful for catching data quality errors that manifest themselves in your analysis. Some useful regression diagnostics are:
   * **residuals** - the difference between the response and the fitted value
   * **hat diagonals -** these consider how variable a data row is among the space of predictors
   * **DF fits, DF betas, Cook’s distance** - these consider how much do fitted values and coefficients change when a point is not included in the fit?
   * **PRESS residuals, leave one out residuals** - how much do predictions change when a point is left out of an analysis?
3. The residuals consider:
   1. The difference between the response and the fitted value.
4. Summary tables should include:
   1. Quantiles
   2. Means
   3. Standard deviations
5. In leading digits of data that follow Benford’s, the digits (0-9) are all equally likely.
   1. False.
6. Using sampling in data quality queries, one can:
   1. Get a sampling-based estimate of the proportion of errant rows.
7. Merging (linking two datasets via a common index) errors can have a strong impact on subsequent analyses.
   1. True.

## The Experiment is Carefully Designed

In this module we discuss the ideal (the experiment is carefully designed) with what is often reality (the data is observational) and cover some solutions.

A key topic of the lecture is **confounding**. Confounding occurs when you want to compare two things and a third gets in the way. As an example, you want to look at ad performance and purchases. However, the ads ran on different sites, so were thus seen by a different audience. The different audiences may have different purchasing patterns, so any difference seen may not be due to the ad campaign but instead may be due to the audience.

An interesting component of randomisation in experiments is the ability to estimate **causal** effects. We define causal effects as the difference between the outcome for a subject observed at a particular treatment minus the outcome observed as a control. However, a subject can only receive one of the treatment levels, so only one of these two gets observed. The other, is called a counterfactual. We can estimate counterfactual effects because of this. However, we can, under assumptions estimated averaged counterfactual effects if we have randomization. The study of how to estimate causal effects using data is called causal inference. I find the most useful aspect of causal inference and counterfactuals is the way thinking about them helps me think about different experimental designs.

1. The definition of a causal effect (in our class) is the difference between what was actually observed and what would have been observed with the opposite treatment.
   1. True
2. If we see an association between two variables, it would be a good idea to:
   1. consider the possibility that the association is explained by a confounding third variable.
3. Some study designs allow us to:
   1. estimate the average causal effect under assumptions.
4. You see an effect of ice cream sales on the number of heat exhaustion cases. The effect is likely due to:
   1. The hot weather as a confounder.
5. Associations can imply causality:
   1. under a set of strict assumptions often as a result of design choices.

One way to combat confounding is at the stage of design. If we were to randomise the ad campaigns across sites, then (at least with high probability) the audiences would be similar. Of course, we might get unlucky, and imbalances of the audiences may still occur, but the chance of that gets smaller as we randomize across more sites. This is the premise of **A/B testing**. In A/B testing one formally designs an experiment with randomization to make the groups being compared (A versus B) as comparable as possible.

What can we do when we don't have randomisation? Also, if we know and collect a variable that clearly will be a confounder, shouldn't we incorporate that into our design rather than leave its balance across treatment groups up to chance? These questions are addressed by blocking and adjustment. In a **blocked experiment**, we randomise within levels of a potential confounding variable. **Adjustment** is a strategy that is used after the data has been collected. In adjustment, we look at the relationship between the predictor and outcome with levels of the confounding variable held fixed. So, if audience demographics confounds the relationship between our ad campaign and purchases, we look at the relationship within demographics. Regression models do this sort of adjustment for us automatically, with some assumptions. When putting a variable into a regression model, it can have all sorts of effects on the relationship of interest.

The final aspect of experimental design is **sampling**. Ideally one can use random sampling to obtain a sample that, with high probability, is a good representation of the population that you'd like to describe. Often, however, it's impossible to have control over the sampling process of an experiment. **Sampling bias** occurs when the sample is not indicative of the target population resulting in inferences that are off. We discuss three strategies to work around issues with the sample. First, is random sampling. The second is weighting, the process of allowing certain observations to carry more influence in models. The final is modelling. That is, trying to model the process that are biasing the sample.

1. In A/B testing randomisation of a treatment is used to:
   1. To make groups as comparable as possible
   2. Attempt to balance potential unobserved confounding variables
2. Three strategies to combat sampling bias are (check all that apply):
   1. Modelling
   2. Random sampling
   3. Weighting
3. It is generally a good idea to consider possible confounders when considering a significant effect:
   1. True
4. It's possible for a regression effect to reverse itself after the inclusion of another variable into the model:
   1. True
5. Blocking and adjustment are tools to:
   1. Account for variables potentially impacting the estimation of the effect of interest.

## Results of Analyses Are Clear

In this section we consider the results of our analysis. Ideally, the effects will be large and hypothesis test are significant.

Before we discuss further, we should make sure that you have a basic understanding of **hypothesis testing**. In hypothesis testing, we use a statistic to decide between two hypotheses. We set one as the default hypothesis (null hypothesis) and the other as the alternative. We make it difficult (require lots of evidence) to reject the null hypothesis and conclude the alternative. This is done by setting the chance that we reject the null and conclude the alternative by mistake is low (usually 5%). Often, the result of a hypothesis test is summarized with a **p-value**. A small p-value (close to 0) supports the alternative while a large one (close to 1) supports the null. We reject the null if our p-value is less than 0.05 if we want to control the probability of incorrectly rejecting the null at 5%.

The result of a hypothesis test is not to be confused with the **effect**. So, if we're testing whether the means between two groups are different the result of the hypothesis test would be the p-value or the conclusion of the test while the effect is the difference between the means. A **confidence interval** is an estimate of the effect that incorporates uncertainty.

In this lecture we discuss ways in which the results of our analyses are not clear and some ways to combat that.

First we discuss **multiple comparisons**. If one preforms lots of hypothesis tests, either from fitting a lot of models or because a lot of things are of interest, then the probability that we see apparently significant findings simply by chance even though they're not actually significant increases. In this lecture, I give one solution for combating multiple comparisons, the Bonferroni correction. Here, just multiply your p-values by the number of tests you performed and then consider the p-values. This is the most useful multiple comparisons rule for managers, since it's easy to do on the fly and can be done while discussing reports.

Next we consider the problem of not really knowing how to interpret the significance of an effect or its magnitude. If this is the case, a good strategy is to compare it to other variables with known effects. We consider an example of a hard effect, lead exposure on brain volume loss. Here, we compared the size of this effect to that of aging on brain volume loss, a better understood setting. Then we could make statements like "the lead effect is similar to 1/2 a month increase in age."

Finally, we consider settings where a rather complex processing and analysis stream has been executed. Often in such settings a concern is whether the effect observed is real, or an artifact of the process. Again, we used a brain imaging example for comparison. People often believe that effects in brain imaging are spurious. Therefore often people will run their analyses in the ventricles (areas of the brain where there is no brain matter, just cerebrospinal fluid). This is called running a **negative control**. A good negative control will be otherwise similar to the real study. For example, the ventricles work in our brain imaging example as they are right in the middle of the brain and subject to all of the processing of other brain regions. In contrast, running images of elbows through the analysis wouldn't be as compelling.

1. Potential problems with testing lots of hypotheses until a significant one is found include (check all that apply):
   1. Misrepresenting the strength of the findings
   2. Declaring effects that are not significant as significant by chance
2. Comparing your effects to familiar ones is useful for
   1. Mentally calibrating the size of an effect or its significance when a variable under study is not well understood
3. Negative control analyses are useful …
   1. as a validity check of an effect of interest by looking to see if similar effects occur with the same analysis on variables where an effect is known not to be present
   2. for evaluating processes to see if spurious effects are obtained
4. A good negative control analysis will
   1. have a negative control that is known not to have an effect but is otherwise similar to the variable under study

## The Decision Is Obvious

Ideally, when one is done with an analysis, the decision to be made from the data is obvious. Here we discuss two common instances where the decision is far from obvious.

First, we consider the instance where the results are equivocal. For example, p-values are around 0.05 where a 5% error rate is the standard. Secondly, we consider the instance where even if the decision is clear, the outcome can't be measured so that the analysis was performed on a surrogate variable.

These concerns over the reliability of data for making decisions are on top of all of the other concerns already addressed, such as unmeasured confounders, sampling bias and other forms bias that may have caused cloud a significant or non-significant finding. However, in previous lectures we gave some remedies for these problems so let's confine ourselves to the first two concerns from above.

It is interesting to note that there is no universal consensus on the role of sample size in the significance of a statistical finding. Does a marginally significant finding gain credibility from a large sample size or does it suffer because it should have been more clear with this much data to interrogate hypotheses? Richard Royall wrote a fascinating paper on this idea titled *The effect of sample size on the meaning of significance tests*. It's a somewhat challenging paper, but it covers this odd quirk of hypothesis testing well if you'd like to read it.

One statement about sample size is clear, if your results are null and your sample size was very low, the null results are not surprising. This is because the study wasn't set up for success. This is the idea of **power**. Power is the probability of rejecting the null hypothesis when it's false. You want more power. Low powered studies are likely not going to reject regardless of whether the null is true, just because of variability. Power is under control at the time of design of an experiment via the sample size. However, after the experiment is performed, there isn't much to be done about power. Often people get the idea of calculating power after the experiment has been done to try to differentiate between non-significant results due to lack of power or true non-significance. This is a bad idea. Potentially at your disposal, however, is conducting new studies, or getting more data for the existing study.

The second problem that we consider is one where the desired outcome is not measurable, but some surrogate is. This is extremely common. For example, BMI for body fat percentage, GDP for economic health, food frequency questionnaires instead of actual calorie consumption and so on. The use of a surrogate can occur for the outcome or the predictors or both. If the surrogate variable is used as a predictor, the field of **measurement error** has several tools available. A particularly useful one is called **SimEx**. As an outcome, the problem is called **surrogate outcomes**. Ideally, you'll know that the surrogate is unbiased around the true outcome and know the variance. This could occur if you could do a gold standard study. If you don't know it, the second-best case would be to have the data to estimate it. For example, in our BMI example, you could measure body fat percentage and BMI on a subset of your sample.

In the absence of any calibration data to evaluate your surrogate, you are left with either modelling via assumptions or **sensitivity analyses**.

Finally, if your surrogate variable is such an unreliable of an estimate of your actual outcome, one must come to the conclusion that it's better to not conduct the study at all.

1. Ideally, a surrogate variable for variable of interest will:
   1. be unbiased
   2. have a known or estimable variance around the desired measurement variable
2. If you get a null result it may be due to:
   1. low power
   2. that the null hypothesis is actually correct
3. A study with a very low sample size will likely have:
   1. low power
4. Calculating power after the study has been done and analysed is:
   1. problematic and should only be done by people well versed in the issues
5. When using surrogate variables, if possible, it's a good idea to:
   1. collect some gold standard data to evaluate the validity of the surrogate
   2. do a sensitivity analysis if a gold standard dataset can't be collected
   3. use a known variance of the surrogate around the gold standard in the analysis
   4. consider the role of the surrogate in interpreting the strength of conclusions

## The Analysis Product Is Awesome

In this lecture we consider the products of the analysis. This is usually some sort of report or presentation. If the analysis is far enough along, then it might be an app or web page.

If the product is a report, ideally, it would be clear and concisely written, with a nice narrative and interesting results. However, the needs of data science managers are variable enough that we focus on two components that make for good final products that are ubiquitous across all settings. These are making the report **reproducible** and making the report and code **version controlled**.

Analysis reproducible considers the fact that if we ask people to replicate their own analysis, let alone someone else's, they often get different results, sometimes, very different. The benefits of using the right tools for reproducible reports are many. They include dramatically helping achieve the goal of reproducibility, but also: helping organize ones thinking by blending the code and the narrative into a single document, they help document the code in a way that commenting doesn't achieve and they help automate the report writing process.

As for the content of the report, some other recommendations that come to mind.

Check the signs, magnitudes and units. Checking the signs means checking that your effects are in the direction that you expect. It's also helps enforce asking the people that you manage to do more than just reporting coefficients. Checking the magnitudes by comparison with other known effects (covered in an earlier lecture) is a really good thing to encourage. Finally, put units on everything (graph axes, coefficients, means, ...) And make sure to understand the units. I can't tell you how many times this small step has helped.

It's important to get the data scientists and analysts that you manage out of technical jargon speak and into interpretation and interpretability. For the former, I keep harping on this idea of comparison with other known effects. Secondly, I encourage the idea of **effect critiquing**. This is the idea of, instead of getting excited about an effect, become its biggest critic. Try to figure out every possible reason why it could be spurious. This almost always leads to new analysis that should be conducted. Finally, for interpretability, encourage parsimonious models with interpretable parameters. That is, place a premium on simplicity. This is, of course, if you're more interested in a data science experiment where you are trying to create new parsimonious knowledge. The situation changes if one is trying to only improve prediction (see the first lecture).

Finally, **version control** is just general good practice. Version control is the process of keeping versions of your software, code, data and reports. Modern version control software makes this process easy. Using good version control, the project can be reverted to any previous step, and all of the steps are commented. Tools like **git** make this possible in massively distributed settings where lots of people are working on the same project. Git is the version control software, and a git repository is the actual version control database. Collaboration on a git repository occurs on a server. Fortunately, there are tons of hosted git servers out there for general and business use.

Of course, git is one of many version control systems. The main point is recommend (or demand/force) a version control culture in your organization.

1. Reproducible report writing tools like knitr and ipython help:
   1. by documenting the analysis code with the project narrative
   2. by organizing ones thinking by blending the code and the narrative into a single document
   3. by advancing the goal of reproducibility
   4. by automating the report writing process
2. git is:
   1. version control software
3. Some easy things to double check reports include:
   1. verifying the signs of effect are in the obvious direction
   2. checking magnitude of effects by comparison with other known effects
   3. putting units on graphs and coefficients and generally keeping track of units
4. Some tools for helping reports have a more clear narrative include:
   1. eliminating jargon and focusing on interpretability
   2. critiquing significant effects by coming up with potential alternate explanations
   3. focusing on simpler models and parsimony
5. Version control software
   1. Keeps track of checked in versions of code, data and reports