Propulsion Academy

Reading notes

Day 3 - Inductive Statistics

Contents

1 Introduction to hypothesis testing

Before starting any inductive statistics, every research question should be transformed into reasonable hypotheses. Without hypotheses the whole concept of inductive statistics breaks down. The most important rule which needs to be known when it comes to hypothesis formulation is that you will never be able to accept any hypotheses. The only thing that you will be able to do in some successful situations is to reject the null hypothesis. Hence you need to formulate the null hypothesis in a way that it states what you would like to reject, e.g the treatment effect of the control group is equal to the treatment effect of the treated group. The hypothesis formulation is also closely connected to the whole experimental design.

2 Research Design

The goal of conducting studies is to answer research questions. The way research questions are asked is not different in data science then in the daily life. But how to transform these research questions into meaningful experiments and analysis is the art of Data Science. "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." R.A. Fisher The importance of a proper research design is very essential. The research design consists of choice of procedures and methods used in collecting and analysing measures of the variables specified in the research problem. Defining the study design means defining the study type, the research problem, the hypothesis, the independent and dependent variables, the experimental design, data collection methods and statistical analysis plan. We can distinguish different research designs. Here one possible categorization:

- Descriptive (e.g. case-study, survey)
- Correlational (e.g. case-control study, observational study)
- Semi-experimental (e.g. field experiment, quasi experiment)

- Experimental (e.g. randomized controlled trial)
- Review (e.g. literature review, systematic review)
- Meta-analytic (e.g. meta-analysis [10])

Every category has it's own specialties and methods. We will only cover most important and common methods and study types. But knowing these different categories you should always be aware in which category your current analysis is.

2.1 Experimental study design

The goal of an experiment is to test whether independent variables have an effect on the dependent variable. Experiments are conducted to learn about causality and not just correlation.
The most common cases of experimental design are

- A/B, two condition testing (split-run testing) - It is two-sample hypothesis testing scenario in which two versions (A and B) which are identical except for one variation are compared. A/B testing is very commonly used in web design (user experience design) to identify web pages that increase an outcome of interest (e.g. click-through rate for an advertisement).
- A/A, pre-post testing - A/A testing similar to A/B testing except the part that two identical versions with or without stimulus are compared against each other. Most commonly it is used to compare two identical versions of a page against each and acts as a method of double-checking the effectiveness and accuracy of A/B testing.
- Factorial design - Factorial design is to compare two independent variables with multiple number of levels in each independent variable. For example, if you want to study the effect of two independent variables, water and sunlight on plants there are two levels of sunlight (test with sunlight and control without sunlight) and two levels for water (test with water and control without water). Here both

independent variables have two levels (present/absent). This is the case of 2x2 design having 2 independent variables with 2 levels each. A 3x2 design will have 2 independent variables, one with 2 levels and one with 3.

A randomized controlled trial (RCT) is an experiment performed on human subjects (usually patients) in order to assess the efficacy of a treatment (or intervention) for some condition. An RCT has two key features:
1. The new treatment is given to a group of patients, the treated group, and another treatment (standard or placebo treatment) is given to another group of patients, the control group, at the same time.
2. Patients are allocated to one group or another by randomization.

2.2 Continuous and Categorical Variables

Variables are classified into multiple groups according to their characteristics. Broadly speaking there are two groups 1) continuous - measured on a continuous scale and have numeric values (e.g. temperature) 2) categorical - take discrete values (e.g. country of residence). The choice of the statistical method for hypothesis testing depends upon the type of variable you are investigating with your hypothesis e.g. for normally distributed continuous data a Student's t-test can be used whereas for ordinal (categories) data a Mann Whitney U test might be a better choice.

2.3 Independent and Dependent Variables

If the goal is to understand the causal relationship between two variables, then the variables having an effect are called independent variables whereas the variable being affected is called dependent variable e.g. vitamins make you more intelligent, vitamins - independent variable and intelligence - dependent variable.

2.4 Operationalization

The process of obtaining concrete measurements about conceptualized variables coming from research questions is termed as operationalization.

For some variables it is very straight forward e.g. weight is usually measured in kilograms. But for quite some variables this process is very essential, intelligence can be measured by an IQ test score but also by many other different indicators. Before starting an experiment a clear consensus needs to exists what should be measured and how. Otherwise the collected information is not comparable and cannot be used or needs to be transformed by troublesome work.

2.5 Levels of data

When thinking about an analysis, it is important to understand the existing level of your data: what does one observation in the data set represent and which observations you are comparing to each other. The levels of data are normally in a hierarchical order. This hierarchical order can be reflected in some methods, but unfortunately not in all methods. Ignoring and/or mixing different levels of analysis can lead to a bias in your results. For example when you do a study about school kids and collecting information about the child, some information about his family and some information about his school. The fact how big the school garden is, is not directly effecting the kids on personal level but it has its effect on school level and has the same effect for all kids in that school. The same is applicable for the family. A data scientist should always be aware of the existing levels in the underlying data.

2.6 Reproducibility

Reproducibility is one major goal of all research designs. In order to reach that reproducibility it is very important to document everything very clearly and accurate. The programs used for analysis should be straight forward and have clear instructions for a potential reuse in several years. The raw data should be stored as read-only copy and all changes to the data should be either documented on a audit history and/or performed by a comprehensible program.

3 Hypothesis testing

Hypothesis testing is a way to test some hypothesis on the basis of observed data. It is a method of statistical inference (deducing properties of an underline probability distribution). How you state your hypothesis depends on your experimental design.

3.1 Superiority or non-equality trial

In a superiority or non-equality trial your goal is to show that there is a difference between the treatment group and the control group. In a superiority setting the hypotheses should be stated like this:
Null Hypothesis : There is no difference between the groups. $C = T$
Alternative Hypothesis : There is a difference between the groups i.e. samples are drawn from different distributions. $C \neq T$

3.2 Non-inferiority trial

The goal of a non-inferiority trial is to show that the treatment group is not worse than the control group. This design leads to hypotheses like this:
Null Hypothesis : The treatment group differ from the control group by more than a certain pre-specified non-inferiority margin, often denoted as δ. $C - T \geq \delta$, if higher values are worse.
Alternative Hypothesis : The treatment group is not worse than the control group by means of the non-inferiority margin δ. ; $C - T < \delta$, if higher values are worse.

3.3 Equality trial

In a equality trial you would like to show that the treatment group and the control group are equal. This leads to the opposite hypotheses of the superiority trial.
Null Hypothesis : There is a difference between the groups by means of a pre-specified equality margin. $|C - T| \geq \delta$

Alternative Hypothesis : There is no difference between the groups by considering the equality margin. $|C - T| < \delta$

3.4 One-sided and two-sided hypotheses and tests

Depending on the formulation of the hypotheses you are applying a one-sided or a two sided test. The superiority/non-equality setting as well as the equality setting define a two-sided testing. Because the difference between the control and the treatment group can go in either direction, higher or lower values, and in both cases we would reject the null hypotheses if the difference is large enough. This is not true for the non-inferiority trial. We only reject the null hypothesis if the difference between the treatment and the control group is smaller than δ, if higher values are worse. Hence the non-inferiority trial requires a one-sided test setting. Typically the alpha value is 0.05 for a two-sided test and 0.025 for a one-sided test. There is no clear rule saying that, but it should always be true that the alpha value for a one-sided test is half as big as the alpha value of a two-sided test. The reason of this is that in a two-sided test you assume that the defined alpha value, e.g. 0.05, splits up into two parts, 0.025 on left end and 0.025 on the right end of the distribution, see figure 2. In a non-inferiority setting we only assume a type I error area on one end of the distribution. Assuming an alpha of 0.05 would lead than to a quite big type I area one this one end.

3.5 One-sample and two-sample hypotheses and tests

A two-sample test setting describes the situation when you have a treatment and a control group or group A and group B in your sample population. But in order to save money and time and sometimes also due to missing possibilities you cannot collect data of a control group and instead of it you just use a reference value. This reference value should normally be derived based on a systematic literature review and should represent the current estimate and knowledge of the effect you would like to investigate. Having a reference value instead of a control population can make your

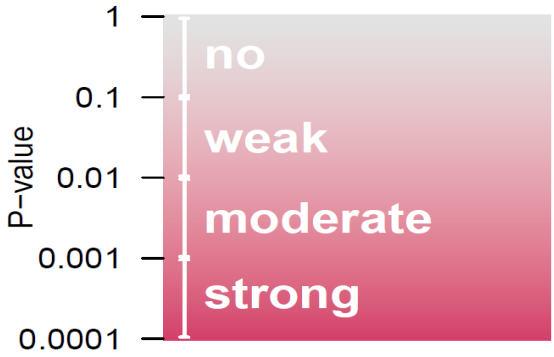

Figure 1: Evidence against the null hypothesis

hypothesis testing a bit easier because the reference value does not come with a certain standard error and has hence no underlying distribution.

3.6 p-Value

A statistical tests gives us a p-value that tells us if we can reject the null hypothesis or not. You never accept or reject alternative hypothesis. Only the null hypothesis can be rejected if the p-value is small enough. Normally the threshold of 0.05 is used to state if the p-value is small enough or not. But please keep in mind the p-value is a continuous measurement and should also be interpreted like this, see figure 1. Doing the categorization in either saying the result is significant or not, reduces information from your result. The P-value is the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed. The p-value is not the probability of the null hypothesis and the p-value is not the probability that the observed data occurred by chance. The p-value is calculated from the value that results out of the test statistic.

3.7 Test statistic

Every test has it's one test statistic. But a general formulation of the test statistic applies for all test statistics.

$t = \frac{Parameterestimate}{Standarderrorofthe estimate}$

The absolute value $|t|$ of the test statistic t quantifies, how many standard errors the estimate is away from the null. Most test statistics follow approximately a known reference distribution (often standard normal or t-distribution) under the null hypothesis. This can be used to compute the p-value, $p = Pr(|T| > |t|)$. In any case the larger —t—, the smaller the (two-sided) p-value.

3.8 Type I and type II errors

Using hypothesis testing can lead to two types of errors (image below, [7]). Type I error occurs when the null hypothesis is rejected although it is true. Whereas Type II error occurs when null hypothesis is accepted although alternative hypothesis is true. As mentioned above p-value is used for this testing however, even using a very low p-value does not guarantee 100% confidence in our results.

4 Parametric tests

Parameteric statistical models reply on making assumptions about the shape of data distribution. The hypothesis testing using parametric procedures involves asking which distribution F from a family of distributions generated a given dataset (data generating process). The distributions are definable with a set of parameters, θ e.g. normal distribution $F \in N(\mu, \sigma^2)$. The most famous representative of that class is the Student's t-test.

5 Non-parametric tests

Non-parametric procedures on the other hand do not rely on assumptions about data generating process. These procedures either do not assume presence of F or allow θ to be infinite. Some examples of parametric vs non-parametric tests are shown below [5].

The choice between parametric vs non-parametric test relies on the assumptions made about the data e.g. if the data is normally distributed you might wanna use t-test. If you do not see normality, stick to a non-parametric test.

6 Bootstrapping and Permutation

Bootstrapping and permutation are two resampling methods, where you either recalculated possible observation based on your underlying sample or where randomly pull a subsample out of your existing sample. Given that the independent variables are well balanced and that the sample is large enough, any permutation should result in similar results. If we don't see any treatment effect, e.g. no difference between the groups, permutation across the treatment and control group can be done and all these rearrangements should be equally likely. We don't need to worry about any assumption or what distribution our data follow.

7 Power Analysis

Power analysis is used to determine sample sizes, effect sizes, significance level, and power. Let's see what these things are.

Sample size - This is the number of observations in your data set. As more and more observations are added, the sample distribution approaches population distribution. P-value depends upon sample size.

Effect size - Effect size is defined as the magnitude of difference scaled by the standard deviation. For example when comparing group means (t-test) cohen's distance can be defined as

$$Cohen's\ \mathrm{d} = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}$$

Effect size measures strength of result and is solely magnitude based. P-value tells us that the results obtained are not due to chance (a high

sample size might lead to a good p-values anyway) however, it does not measure the practical significance of our results and for that we need to look at the effect size.

Significance level (α) - This represents the probability of finding significance when there is none (type I error, see image below) i.e. the result shows there is a difference between groups when in fact there is no true difference. α gives information about false positives. Most commonly a value of 0.05 is used for α.

Significance power ($1-\beta$) - β represents the probability of not finding significance though it is present (false negatives, type II error, see image below). $(1-\beta)$ represents probability of finding true significance (correctly reject null hypothesis). Usually 0.80 is used as a cut-off for $1-\beta$.

		Reality	
		H0 = true (no diff)	H0 = false (diff exists)
Our conclusion	H0 = true (no diff)	correct P | no diff = (1-α)	type II error (false negative) P | true diff = β
	H0 = false (diff exists)	type I error (false positive) P | no diff = α	correct P | true diff = (1-β)

Figure below shows all above for a randomly generated data ([2]). [2] also provides an interactive way to learn about power analysis.
Before running the study, information about effect size, significance level, and statistical power should be supplied to obtain the sample size that should be used in the study. After the study has run, significance level, sample size, and effect size are used to estimate the statistical power.

8 Threads to validity

In an experimental setting we change the level of the independent variable(s), and want to keep all other factors the same. Otherwise, any differences we observe between conditions may be due to these other factors. This can be very difficult and some problems and bias already

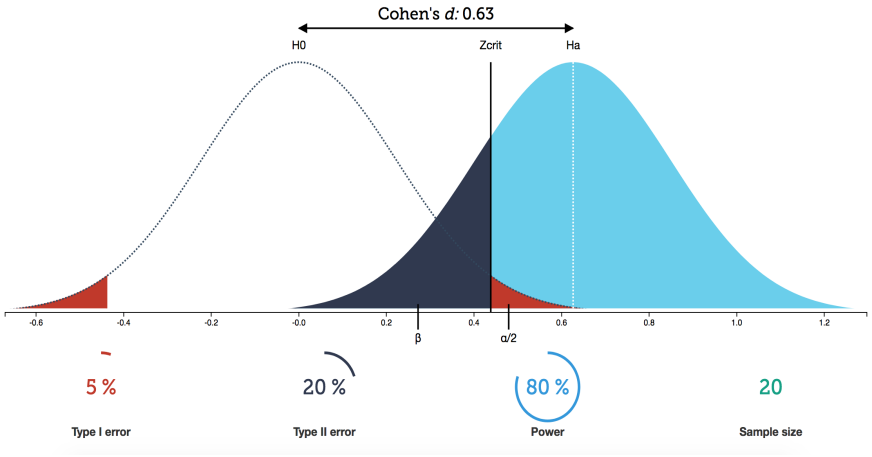

Figure 2: Visualization of Type I and Type II errors

appeared quite often.

8.1 Bias

- **Selection bias** will occur, if the decision to enter a patient to a trial is influenced by knowledge of which treatment the patient will receive.
- **Allocation Bias** will occur if the randomization process cannot ensure that the groups to be compared are balanced with respect to all relevant independent variables. This is a particular problem in trials with small sample size and many independent variables.
- **Assessment Bias** will occur if the observer knows the treatment being given to the patient, as this can influence the value of the assessment. This bias especially occurs in variables that have a subjective component of the assessor. This problem can be addressed by double blinding.
- **Stopping Bias** will occur if the outcomes of a trial are assessed repeatedly and the trial is stopped, once a promising result has been obtained. Interim analyses should be planed beforehand and considered in the experimental design. Naive repeated checking of "significance" will introduce bias.

- **Publication Bias** occurs if only successful results of trial are published. As most of the journals only publish experiments with a (significant) finding, this bias is very real. There are some attempts to address this problem, e.g. http://clinicaltrials.gov/, but there is no well-working solution until now.

- **Observation bias** occurs if subjects change their behaviour because they know that they are being observed or tested. This bias is also known as Hawthorne effect.

Other bias may occur if any events happen outside of the experimental setting between a first and second measurement. Depending on the time interval subjects may also change simply due to age/maturation. The way how a variable is measured, can also change over time and can result in different measures although everyone believes it is still the same measure. This is also the reason why all assessed variables should described very properly including the assessment method. For important variables which are difficult to assess a repeatability study may be useful. In a repeatability study you investigate the preciseness of a measurement and if and what external factors it is dependent on.

8.2 Multiple testing

Multiple comparisons or multiple testing problem occurs when one considers a set of statistical inferences simultaneously. The more inferences are made, the more likely erroneous inferences are to occur. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made. [9] The family-wise error rate (FWER) is one measure to quantify the effect of multiple testing. The FWER is defined as $1 - (1 - \alpha)^m$, if m independent comparisons/testings are done. Several method like Bonferroni or Bonferroni-Holm or Simes correction exist to adress the multiple testing problem. If some sequential testing needs to be

included into the trial per design Pocock and O'Brien-Fleming's design present a solution how to deal with multtiple testing and stopping rules.

9 Blocking

Blocking refers to the task of arranging experimental units in groups (blocks) that are similar to each other. Blocking is done when observations can be split into relatively homogenous groups. Blocking reduces variability as units in blocks are more nearly homogenous than units in different blocks.
For example if a new drug is tested on patients there are two levels of treatment, drug and placebo given to both male and female patients in double blind trial. In such a case the sex is used as a blocking factor to account for treatment variability between males and females (taken from [3]).
For those who are interested in learning some stats/math behind blocking please refer to [4].

References

[1] https://en.wikipedia.org/wiki/Research_design

[2] http://rpsychologist.com/d3/NHST/

[3] https://en.wikipedia.org/wiki/Blocking_(statistics)

[4] http://www.stat.ufl.edu/CourseINFO/STA6167/Analysis%20of%20Variance--Randomized%20Blocks%20Design.pdf

[5] http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test

[6] https://en.wikipedia.org/wiki/Student%27s_t-test

[7] http://www.vias.org/tmdatanaleng/cc_error_types.html

[8] http://www.epixanalytics.com/downloads/Zagmutt_PalisadeNOLA_2015_Final.pdf

[9] https://en.wikipedia.org/wiki/Multiple_comparisons_problem

[10] https://en.wikipedia.org/wiki/Meta-analysis