## Objectives

Estimated Learning Time: 3h 1m

Inferential statistics and hypothesis testing are two types of data analysis often overlooked at early stages of analyzing your data. They can give you quick insights about the quality of your data. They also help you confirm business intuition and help you prescribe what to analyze next using Machine Learning. This module looks at useful definitions and simple examples that will help you get started creating hypothesis around your business problem and how to test them.

Learning Objectives:

* Understand statistical estimation and inference
* Recognize parametric and non-parametric approaches to modeling
* Become familiarized with the most common statistical distributions
* Fundamental understanding of frequentist vs. Bayesian statistics
* Hands-on experience communicating insights to peers and stakeholders
* Hands-on experience providing insights from data
* Hands-on experience assessing the quality of data
* Hands-on experience on Exploratory Data Analysis

Estimation is the application of an algorithm, for example taking an averave

Inference – involves putting an accuracy on the estimate (e.g. standard error of an average)

Machine learning and statistical inference are similar ( a case of cs borrowing from a long history in statistics.)

In both cases, were using data to learn/infer qualities of a distribution that generated the data (often termed the data-generating process).

We may care either about the whole distribution or just features (e.g. mean).

Ml applications that focus on understanding parameters and individual effects involve more tools from statistical inference (some apps are focused only on results)

Example: customer churn

Customer churn occurs when a customer leaves a company

Data related to churn may include a target variable for whether or not the customer left

Features could include:

* The length of time as a customer
* They type and amount purchased
* Other customer characteristics

Churn prediction is often approached by predicting a score for individuals that estimates the probability the customer will leave.

Estimation of factors driving customer churn involves measuring the impact of each factor in predicting churn.

Inference involves determining whether these measured impacts are statistically significant

Exmple dataset

Ibm cognos customer churn dataset:

* Data from fictional telelcommunications firm
* Includes acct. Type, cust. Charateristics, revenue per customer, satisfaction score , estimate of customer lifetime value.
* Includes info on whether customer churned (and some categories of churn type)

…

**Hypothesis Testing**

A hypothesis is a statement about a population parameter.

We create two hypotheses:

* The null hypothesis - h0
* The alternative hypothesis - h1

We decide which one to call the nulll depending on how the problem is set up.

The less specifc value is the alternative and the specific is the null hypothesis

A hypothesis testing procedure gives us a rule to decide:

* For which values of the test statistics do we accept h0
* For whikch values of the test statistics do we recect h0 and accept h1

You may hear some peiople say that you can reject h0 but that you never accept h1.

Here, this doesn’t matter very much , since we are hypothesis testing in order to decide which of two paths to take in the project.

In the bayesian interpretation (example to follow), we don’t get a decision boundary.

Instead, we get updated ((posterior) probabilites.

You have two coins:

* 1 coin has a 70% probability of heads
* Coin 2 has a 50% probability of heads

Pick one coin without looking

Toss the coin 10 times and record the number of heads

Given the number of heads you see, which of the two coins did you toss?

Give what we know about coins 1 and 2, we can make a table of the probability of seeing x heads out of 10 tosses

We can calculate ta likelihood ratio, based on the number of heads we saw when tossing the uniedentified coin.

Suppose we saw three heads:

Coin1 – 0.1117

Coin2 – 0.009

P1(3)/p2(3) = 0.117/0.009 = 13

Coin 1 was 13 rimes more likely to give us the output (3 heads) than coin 2

This is called the likelihood ratio

In the bayesian interpretation, we need priors for eACh hypothesis:

* In this case, we randomly chose the coin to flip
* P(H1 = we chose coin 1)= ½ and
* P(H1 = we chose coin 2) = ½

Since we have no way, before seeing the data, to determine the coin that was chose, we just assign ½ to each.

Priors: p(h1) = ½ = p(h2)=½

Updating priors after seeing the data 3 heads (bayes’rule);

Hypothesis testing: bayesian interpretation

The priors are multiplied by the likelihood ratio, which does not depend on the priors.

The likelihood ratio tells us how we cshould update the priors in reaction to seeing a given set of data.

Type 1 vs type 2 error

Neyman-pearson interpretation

The neyman-pearson paradigm(1933) is non-bayesian.

This gives an up or down vote on h0 vs h1

Terminology

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Decision | |
|  |  | Accept h0 | Reject h0 |
| truth | h0 | Correct | Type 1 error |
| h1 | Type II error | correct |

Type 1 error will inncorectly reject

Type 2 will incorrectly accept the null

Power of a test = 1 – P(type II error)

Example customer churn

Customer churn aoccurs when a customer leaves a company

Data related to churn may include a target variable for wheterh or not the customer left

Features could include:

The lenghtt of time as a customer

Thye type and amt. Purchased

Other customer characterisitcs

Churn prediction isw often approached by predictin g a socre for individuals that estimates the probability the cusomter will eave.

Suppose we use data on customer characteristics to predict who will churn over the next year.

In our data, customers who have been with the company for longer are less likely to churn.

This could be due to an underlying effect, or due to change

* A type 1 error occurs when this effect is due to chance, but we find it to be significant in the model
* A type II error occurs when we ascribe the effect to change, but the effect is non-coincidental.

Hypothesis testing terminology

The likilhood ratio is called a test statistic: w use it to decide whether to accep/reject h0

The rejection region: is the set of values of the test statitics that lead to rejection of J0

The acceptance region: is the set of values of the test statitics that lead to acception of H0

The null distribution: is test statitics’s distribution when the null is true

Testing marketing intervion effectiveness:

* For a new direct mail makerting campaign to exisiting customers, the null hypothesis suggest the campaign does not impact purchasing.

The alternative hypothesis h1, suggests it has an impact.

Testing a change in website layout:

* For a proposed change to a web layout, we may test a null hypothesis that the change has no impact on traffic.
* Here, we would look for evidence to reject the null in favor of an alternative hypothesis (h1, that there is an impact on traffic.)

Testing whether a product meets expected size threshold:

* Suppose a product is produced in various factories, with expected size S
* To confirm that the product size mieets the standard within margin of error, the company might:
  + Randomly sample from each product soruce
  + Establish h0, product size is not significantly different from S
  + And h1 there is a significant deviation in product size.
  + Test wheteher h0 can be rejected in favor of h1, based on the obvserved mean and standard deviation.

Significance level and p-value pt.1

We know the distribution of the null hypothesis

To get a rejection region, we calculate the test statistics.

We will chooose, before testing the data, the level at which we will reject the null hypothesis.

A significance level (alpha) is a probability threshold below which the null hypothesis will be rejected.

We must choose an alpha before computing the test statistic

If we don’t we might be accused of p-hacking

Choosing alpha is somewhat arbitrary but often .01 or .05

The alternative hypothesis h1, suggests it has an impact.

Testing achange in website layout:

* For a proposed change to a web layout, we may test a null hypothesis that the change has no impact on traffic
* Here we would look for evidence to reject the null in favor of an alternative hypothesis (h1, that there is an impact on traffic.)

Testing whether a product meets expected size threshold:

* Suppose a product in produced in various factors, with expected size S.
* To confirm that the product sizer meets the standard within a margin of error, the company might:
  + Randomly sample from each product source
  + Establish h0, product size is not significantly different from s
  + And h1 there is a significant deviation in product size.
  + Test whether h0 can be rejected in favor of h1, based on the objserved mean and standard deviation .

Significane level and p-values pt.1

Important terminology

* The p-value; smallest significance level at which the null hypothesis would be rejected.
* The confidence interval: the value of the statistic for which we accept the null.

Suppose we saw three heads in 10 rolls:

* P(3 or less heads)=p(0heads) + p(1 head) + p(2 heads) + p(3 heads) =17%
* Under the nuyll hypothesis, a value this extreme occurs 17 % of the time.

In the coin tossing example:

* H0: the coin is fair and p(H) = 0.5
* H1: the coin is unfair and p(H)<0.5

How can we test the null hypothesis if we observ 3 heads in 10 flips

Testing the null hypothesis:

* We know h0 is distributed binom(10, 0.5)
* Choose a p-value cutoff (more on p-values), say 5%
* Calculte the CDF of 3 heads from a binom(10, 0.5)
* CDF = 17.1% (above our cutoff
* This is > 5% so we don’t reject h0

Significance level and p-value pt.2

F-statistic

H0: the data can be modeled by setting all betas to zero.

Power and sample size

If ou do many 5% significance tests looking for a significant result, the chances of making at least one type 1 error increase

Probability of at least one type 1 error is approximately = 1 - (1-0.05) raised to the number of tests

This is roughly 0.05 x (# of tests). If you have 10 or fewer tests.

Bonferroni correction

The conferroni correction: says choose a p threshold so that the probability of making a type 1 error (assuming no effect) is 5%

Typically choose: p threshold = 0.05 / (# of tests)

Bonferroni correction allows the probability of a type 1 error to be controlled, but at the cost of power.

Effects either need to be larger or the tests need larger samples to be detected.

Best practice is to limit the number of comparisons done to a few well-motivated cases.

**Hypothesis testing demo – pt. 1**

Quiz:

You find through a grpah that there is a strong correlation between net promoter score and the visual tiem that customers spend on a website.

* There is an underlying factor that explains this correlation, but manipulating the time that customers spend on a website may not affect the net promoter score they will give to the company.

If you reject the null hypothesis, it means that the alternative hypothesis is true

* False

Type 1 error 1 is defined as:

* Saying the null hypothesis is false, when it is actually true

A p-value is

* The smallest significance level at which the null hypothesis would be rejected.