Business Analyst course

Introduction to the course

**Description**

**1** This video is dedicated to bureaucracies 

d

e

n

**2** The course has statistics as the base

c

a

n

e

s

r

c

i

e

r

f

t

n

e

**3** A business analyst needs to know 3 Analytics types

I

l

m

a

o

s

n

u

o

a

c

**4** The course is practice-focused

C

E

Business Analyst

n

o

i

t

a

t

n

e

m

g

e

S

e

v

it

ci

d

er

P

s

c

i

t

y

l

a

n

A

**5** The course materials are in the next lecture

Statistics and Descriptive Analytics

The Modern day Business Analyst

**Description**

**1** Being only good with numbers is a thing of past **2** Proficient with Statistics & Analytics methodologies **3** A bridge between technical and non-technical people

The impact of weather on sales

**Description**

**1** Weather influences seasonal industries 

**2** External factors are uncontrollable by nature **3** How to prove weather influences sales? If weather influences, then sales move when weather changes and are constant, all else being equal **4** 

**5** The Technique I used was Google Causal Impact

Predicting the future

**Description**

**1** Commercial teams are belief-driven in nature **2** Advanced Analytics give numbers to beliefs **3** But making the change is difficult 

**4** Simple and interpretable usually has high errors **5** One of the techniques used was Facebook Prophet

BASIC

STATISTICS

Game Plan

**Description**

**1** Backbone for the full course

**2** Master the principles to make it easier in the future **3** You will do an exercise for each statistic learned **4** Moneyball case study at the end

(Arithmetic) Mean

**Description**

Same thing as average

When we say mean, we refer to the arithmetic mean

Represents the expected value

**Methodological Representation** ��ҧ=σ ����

��

**Visualization**

Case Study

**Description**

**1** 

Briefing –

We have a dataset with baseball teams’ data from 1962 to 2012

Baseball

**2** There are 12 KPIs for each Team, League and Year **3** We will practice statistical concepts on the dataset

https://www.baseball-reference.com/

Mode and Median

**Mode Median**

The most frequent number in a set The central number of an ordered set

Fashion is a statistical term ☺

**Visualization**

| **X** | **2** | **3** | **3** | **5** |
| --- | --- | --- | --- | --- |

Mode is 3

If even numbers, then you average both middle points

**Used with skewed dataset**

**Visualization**

| **X** | **2** | **3** | **5** |
| --- | --- | --- | --- |

Median is 3

| **X** | **2** | **3** | **5** | **10** |
| --- | --- | --- | --- | --- |

Median is 4

(Pearson) Correlation

**Description Visualization**

Measures the relationship strength 

between 2 variables

Varies between -1 and 1

1 means strong positive relationship

-1 means strong negative relationship 0 indicates no relationship

Correlation does not imply causation

**Methodological Representation**

�� =σ ���� − ��ҧ ���� − ��ത σ ���� − ��ҧ2�� ���� − ��ത2

Standard Deviation

**Description**

**Methodological Representation**

Measures the variation or dispersion of

a set of values

High values mean higher variability �� =σ ���� − ��ҧ2 ��

**High variability Low variability**

**s e**

**c**

**n**

**e**

**r**

**u**

**c**

**c**

**o**

**#**

**se**

**c**

**n**

**er**

**uc**

**c**

**o**

**#**

**X X**

Moneyball Case Study

**Description**

**1** Money ball is set on the world of baseball **2** The A’s had success despite financial struggles **3** The team looked for undervalued players **4** Other teams did not look at statistics **5** The General Manager looked at specific statistics **6** With the right system, you can beat anyone

INTERMEDIARY STATISTICS

Game Plan

**Description**

**1** Level up our statistics game **2** Please use the Q&A **3** We have 2 datasets

Normal Distribution aka Gaussian Distribution

**Description**

Symmetric distribution with the mean in the middle

Data occuring near the mean is more frequent

Graph is similar to bell shaped curve

Statistical methods, i.e. regression assumes normalization of errors

In real-life, there will be some degree of similarity in most problems

**Visualization **

68-95-99 Rule in Normal Distributions

**Description**

Within +- 1 standard distributions, you can find 68% of observations

Within +-2 SD, you should encounter 95% of deviations

Within +-3 SD, it is 99.7%

**Key Idea**

A normal distributions is a pattern, and patterns enables us to categorize data

**Visualization**

99.7%

95%

68%

with more confidence **1SD 2SD 3SD**

Case Study – Wine Quality

**Description**

**You are a wanna be Wine Statistics Connoisseur**

**Challenge1** 

**1** Normal Distribution 

**2** Standard Errors

**3** Confidense Intervals

Paulo Cortez,

University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

@2009

P-value is all about likelihood **Description Examples**

The probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming the null hypothesis is correct.

It helps us understand what is the likelihood of “accepting” aka “fail to reject” the hypothesis

A small p-value (small probability) would mean we favor the alternate hypothesis

P-value threshold usually used: 0.05

H0: The average salary of business analysts is €60k H1: business analysts’ average salary **is not** €60k P-value = 0.2 -> We fail to reject the null hypothesis

H0: Blueberries prevent cancer

H1: Blueberries **do not** prevent cancer

P-value = 0.01 -> We reject the null hypothesis

Shapiro-Wilk test

**Description**

Quantifies how likely it is that the data was drawn from a Gaussian distribution

Created in 1965 and is one of many normality tests

**Interpretation**

H0: The distribution is gaussian

If p-value > 0.05

The distribution appears to have a normal distribution

If p-value < 0.05

The distribution does not look Gaussian -> reject the null hypothesis

Standard Error (of the sample mean)

**Description**

The standard error of the sample mean is an estimate of how far the sample

**Methodological Representation**

mean. ���� =����

mean is likely to be from the population

Standard deviation is the degree to

which individuals within the sample

differ from the sample mean.

Z-Score

**Description**

Gives you an idea of how far from the

**Methodological Representation**

mean is a data point.�� =�� − ��

Z-scores are a way to compare results to a “normal” population

It is a way to standardize values

��

**Example – Diogo College grades**

Uni: Z = (16 – 13) / (2) = 1.5

GMAT: Z = (680 – 560) / (120) = 1

Confidence Interval (when n > 30)

**Description Visualization – 95% CI Z-values table**

A range that gives a sense of how precisely a statistic estimates a

**Confidence Interval**

**Z-Value**

parameter.

The associated confidence level gives the probability with which an estimated interval

80% 1.28 85% 1.44 90% 1.65

**2.5% 97.5%**

will contain the true value of the parameterLower

Upper

**Methodological Representation** ���� = ��ҧ± �� ∗����

95% 1.96 99% 2.58 99.9% 3.29

bound

bound

T-Tests

**Visualization**

# of people

Non-Italians Italians

Hand usage while

talking

**T Test formally**

Test any statistical hypothesis in which the test statistic follows a Student's t-distribution under the null hypothesis.

**In practical terms**

Helps us understand whether one group is (statistically) different than from the other

**How do we know?**

If p-value less than 0.05, then the groups are statistically different

Challenge – Understanding Remote Work predictions

**Stack Overflow dataset**

**Challenge Worker‘s characteristics, and job related queries 1 1** T-tests 

**2** Chi-square tests

(Person) Chi-square test

**Visualization**

**Wears black**

Yes No

**Chi Chi--square squaretest test**

Determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies

**Lives in Berlin**

Yes No

|  |
| --- |

**Difference from t-test**

A t-test tests a null hypothesis about two means;

**Null hypothesis**

There is no relationship between variables

A chi-square test requires categorical variables, each having any number of levels.

Powerposing and p-hacking

**Description** 

**1** You put your body in a powerpose

**2** You would perform better in high-pressure moments **3** Powerposing is not backed up by science **4** Powerposing results were not replicated by others 

P-hacking is the removal of some individuals to achieve statistical significance **5**

LINEAR

REGRESSION

Game Plan

**Description**

**1** Building block in our learning capacity **2** I learned how to do it by hand. Yeah, really! **3** We will have a practice-focused approach

Case Study

Briefing – Pricing

**Description**

**1** We have a dataset of roughly 300 diamonds **2** We have the price, carats and other KPIs

Diamonds 

**3**

We want to understand how carats influence Diamond Prices

Package: Ecdat

(Linear) Regression crash course

**Visualization Definition**

Study of a relationship between a dependent

variable and at least one independent variable

e

c

i

r

P

**β**

Carat

**Intuition perspective**

Method for “What is the impact of X on Y?

**How is it different from a correlation?**

• Correlation studies the direction • Regression studies the impact

Linear Regression

**Visualization Methodological View** �� = �� + �� ∗ �� + �� 

**Interpretation**

If I increase X by 1, Y increases by b

If X happens, Y increases by b

How to read a Regression result

**Coefficients** 

One carat increases the price by 11.6k

**R-Squared and adj R-squared**

We can explain 89.3% of the variance

**Confidence interval (95%)**

The Carat coefficient is between 11.1k-12.1k

**Statistical Significance**

If P>|t| is less than 0.05, we have statistical

significance

Dummy variable trap

|  | Pepsi |
| --- | --- |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |
| 0 | 1 |
| 0 | 0 |
| 0 | 1 |

Observation Coca cola

a

b

c

d

e

f

g

h

j

**Multicollinearity**

Correlation between Coca-Cola and Pepsi is -1 Solution: remove one dummy variable

**Removing does not mean information is lost** A zero also represents information

The removed dummy variable becomes part of the intercept.

You can see it as being your baseline.

Dummy variable trap

|  | Pepsi | White  Label |
| --- | --- | --- |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

Observation Coca cola

a

b

c

d

e

f

g

h

j

**Multicollinearity**

Correlation between Coca-Cola and Pepsi is -1 Solution: remove one dummy variable

**Removing does not mean information is lost** A zero also represents information

The removed dummy variable becomes part of the intercept.

You can see it as being your baseline.

MULTILINEAR REGRESSION

Game Plan

**Description**

**1** Topics: outliers, assessment ad overfitting **2** Practice tutorial: Teacher’s salaries **3** Challenge: Retail Store drivers **4** Regression value adding is interpretability

Multilinear Regression

**Linear Regression Multilinear Regression Description**

Carat Diamond X Y

Carat

Color Diamond Shape

It is very rare that one input explains the output

We often need more predictors to improve the models

Beware of multicollinearity or overfitting

Case Study Briefing –

**Description**

The 2008-09 academic salary for Professors in a college in the U.S. **1** 

Professors‘ 

**2**

The data was collected to monitor salary differences between male and female faculty members.

salaries

**3** Use Multilinear Regression to study

Outliers

**Visualization Interpretation** Outliers can damage your analysis

s

e

i

r

a

l

a

S

Professional experience

You must distinguish between noise and valuable information

Consider using models good with outliers or non-linearity (e.g., Random Forest)

Modelling is finding the balance between under and overfitting

**Underfitting Overfitting**

**Insights**

Having a too simple model will get you nowhere

Too complex will not yield results in other testing scenarios

You should iterate based on results

Let‘s imagine this is our full data set

**Description**

****

Splitting between training and test enables an unbiased model assessment

**Training Set Test Set**

****

**Model Assessment**

Mean Absolute Error (MAE) vs Root Squared Mean Error (RSME)

**Visualization Key ideas**

Y

Model

MAE and RSME are performance indicators for Regression models with continuous outputs

������ =σ �� − ��ො ��

�������� =σ ��ො − ��2 ��

RSME is useful for models with extremes / outliers

MAE is more interpretable.

X

Challenge

**Description**

**Use Multilinear Regression to study a Store sales‘ drivers**

**1** Pick variables for your model

**2** Analyze the data i.e. summary statistics

**3** Correlation Matrix

**4** Create a Training a Test Set

**5** Use Multilinear Regression

**6** Assess Accuracy

Dataset: Ecdat package

LOGISTIC

REGRESSION

Game Plan

**Description**

**1** We now face a classification problem **2** The question influences the analytical technique **3** How do we measure accuracy?

**4** Case study: Which emails are spam? **5** Challenge: the sex of penguins

Case Study Briefing – Is it spam?

**Description**

**1** Dataset with ~5k emails

**2** What makes an email spammy? **3** Can we predict which emails are spam?

DAAG ´package

Logistic Regression crash course

**Visualization What is a Logistic Regression? Spam** Relationship study between a discrete dependent variable and at least one independent variable 

Y (1)

**From an Intuition Perspective?**

What is the impact of X on Y happening?

**How is it different from a Linear Regression**

N (0)

**# times the word**

**money occurs**

Linear is for continuous, logistic is discreteLinear we fit a straight line, logistic a curve

How to read a Logistic Regression coefficients1

**Linear Regression**

�� = �� + �� ∗ �� + ��

**Logistic Regression**

�� =ⅇ��+����

1 + ⅇ��+����

**Interpretation**

For each X unit increase, Y increases by ***b*** 

**B = 0.5**: For each X unit increase, Y increases by ***0.5***

For each X unit increase, the probability of Y happening increases by ***exp(b) – 1 \* 100 %*** 

**B = 0.5**: For each X unit increase, the probability of Y happening increases by ***64%***

The Confusion Matrix allows to access the results of a classifier **Accuracy**

**Confusion Matrix**

**Predicted**

Accuracy = (True positive + True negative ) / All

False True

Balanced dataset

False

| True Negative |
| --- |

False Positive

**F1-Score**

True

False Negative True Positive

F1-score = 2 \* TP / (2 \* TP + FP + FN)Unbalanced dataset

The Confusion Matrix allows to access the results of a classifier **Specificity or True Negative Rate**

**Confusion Matrix**

**Predicted**

True Negative / (True Negative + False Positive)

False True

When we focus in False values accuracy

False

| True Negative |
| --- |

False Positive

**Sensitivity, Recall or True Positive Rate**

True

False Negative True Positive

True Positive / (True Positive + False Negative)Focus is on True values

Challenge

**Description**

**Use Logistic Regression to predict the sex of penguins**

**1** Pick variables for your model

**2** Plot Histograms of the character variables **3** Transform the character variables into binary **4** Create a Training a Test Set

**5** Use Logistic Regression

**6** Assess Accuracy through the classification report

GOOGLE CAUSAL IMPACT

Why

Econometrics and Causal Inference

**Description**

**1** Decision-Making

**2** Understand and tacking biases

According to BNP Paribas, Sustainability-focused companies perform better



**Description**

**1** 

Are there other differences between sustaibanility focused companies and others?

BNP Paribas

**2** People define politics and decision

Not including all factors is falling into the omitted variable bias **3**

Does 

Smoking

prevent

Parkinson’s?

Smoking and Parkinson‘s

**Description**

The incidence of Parkinson’s in people between 55 and 75 is twice as significant in non-smokers **1** 

Is there a causal relationship between smoking and Parkinson’s? **2** 

People are more likely to get Parkinson’s the older they get **3** 

**4** Smokers’ life expectancy is lower than non-smokers’

**Non-smokers are more likely to have 5**

**Parkinson’s because they live longer, not because they don’t smoke**

Game Plan

**Description**

**1** Causal Impact was developed by Google **2** Practice Case Study: Paypal and Bitcoin **3** Challenge: Volkswagen CO2 scandal **4** Causal Impact is my most-used technique

Case study:

**Description**

**Use Google Causal Impact to estimate the impact of Paypal allowing crypto payments on Bitcoin price**

Paypal x Bitcoin

In October 21st, 2020, Paypal announced entered **1** 

the Crypto industry.

Given the bull market and other volatilities, we 

**2**

cannot compare the price before and after

**3** We need to find comparable control groups

What is Time Series Data?

**Visualization**

**Bitcoin Price**

22000

20000

18000

16000

14000

12000

10000

9/1/2020 10/1/2020 11/1/2020

**Key ideas** 

Sequence of data points in time order (oldest to newest)

Most commonly, it is data recorded in equally distanced time periods

Type of Panel Data

(multidimensional dataset)

Comparing before and after impact leads to omitted variable bias 

**Context Visualization**

How to measure the impact of Paypal on Bitcoin price?

This graph shows the sales in the market. The event started where the red line is

Comparing before and after would subject you to ommitted bias.

Bitcoin Price

22000

20000

18000

16000

14000

12000

10000

9/1/2020 10/1/2020 11/1/2020

Brodersen, Kay H.; Gallusser, Fabian; Koehler, Jim; Remy, Nicolas; Scott, Steven L. Inferring causal impact using Bayesian structural time-series models. Ann. Appl. Stat. 9 (2015), no. 1, 247--274. doi:10.1214/14-AOAS788. https://projecteuclid.org/euclid.aoas/1430226092

Difference-in-differences framework **Key ideas Visualization**

We use Google to create an artificial

control group

The delta between what actually happened and the what-if scenario is the **treatment impact**

e

e

c

c

i

i

r

r

P

P

**October 20th**

Bitcoin

Google

What actually happened with Bitcoin

**Paypal**

**impact**

Bitcoin, if same

evolution as Google

Google + Post

October 20th

Time

Causal Impact Step by Step

**Define pre and post period** 

****

**Retrieve the data we need**

****

**Check whether the variables are correlated in the pre period **

**Remove non-correlated data**

****

**Use Causal Impact**

Assumptions

**Parallel Trends Assumption Confounding Policy Change**

The Treatment and Control Groups are assumed to have the same evolution for the KPI

**Visualization**

Bitcoin **October**

There must be only one policy or initiative that differentiates the treatment from control groups.

You can only measure the impact of one treatment.

e

c

i

r

P

**20th**

Amazon

Google

Time

**How to Strengthen the assumptions** Use More control groups

Use a longer training period

Keep post-period to the bare minimum

Correlation in Time Series

**Description Visualization** 

Measures the relationship strength

between 2 variables

If the Time-Series grows over time, then

v

the correlation might be random

The data must be stationary

Stationarity

**Time dependent mean**

**Stationary Time Series Key idea**

Mean, variance and

covariance are not time

dependent

Stationary Time Series

have a defined pattern

**Time dependent variance Time dependent covariance**

Y

t

Y

t

**Statistical test:**

Dickey-Fuller test. If p value is less than 0.05, time series is

considered stationary

Making Data Stationary

**Time Series Differencing**

| 5 |
| --- |
| 9 |
| 1 |
| 7 |
| 3 |
| 7 |
| 4 |



| NA |
| --- |
| 4 |
| -8 |
| 6 |
| -4 |
| 4 |
| -3 |

Impact evolution

**Context Visualization**

Let‘s discuss what should be the impact of 

Paypal adopting Bitcoin:

Greater in the beggining

Impact gradually increases

You can also point out that the impact should

continue days after the announcement

**Causal Impact allows the impact variations**

**over time**

• Brodersen, Kay H.; Gallusser, Fabian; Koehler, Jim; Remy, Nicolas; Scott, Steven L. Inferring causal impact using Bayesian structural time-series models. Ann. Appl. Stat. 9 (2015), no. 1, 247--274. doi:10.1214/14-AOAS788. https://projecteuclid.org/euclid.aoas/1430226092

Challenge

**Description**

**Use Causal Impact to measure the impact of the CO2 scandal in Volkswagen stock Price**

**1** Pick Stocks for the control groups

**2** Perform a correlation matrix

**3** Measure the impact

MATCHING

Game Plan

**Description**

**1** There is no comparable control group **2** Helps us with (self)-selection bias

**3** How to measure referral programs?

**4** What is the incremental value of Mobile Shopping? **5** Practice case study: Catholic Schools and scores **6** Challenge: Remote work and career satisfaction

How do you figure out the value of Amazon Prime?

**Context**

Amazon Prime is a loyalty program that provides free shipping, discounts and other services

The goal of program is fourfold:

• Increase customer loyalty

• Increase revenue per customer

• Decrease marketing spendings in customer re-activation

• Decrease paid advertising in conversion

The subscription lasts 1 year

If you were to asked to provide the impact of Amazon Prime on its financials, how would you do it?

You cannot just simply compare the average Prime and non-prime subscriber

**Context**

Both groups may be inerently different from the start. Hence, they are not comparable. Beware of (self-)selection bias

A possible solution is Matching.

In a nutshell, you create a counterfactual group with similar characteristics to your treatment group

t

n

l

e

o

r

t

m

t

n

a

o

e

r

C

T

You cannot just simply compare the average Prime and non-prime subscriber

**Context**

Both groups may be inerently different from the start. Hence, they are not comparable.Beware of (self-)selection bias

A possible solution is Matching.

In a nutshell, you create a counterfactual group with similar characteristics to your treatment group

t

d

n

e

l

e

h

o

r

c

t

m

t

t

n

a

a

o

e

m

r

C

n

T

U

Case Study Briefing – Are catholic

**Description**

**Use Matching to understand whether catholic schoolsare better than others (from a standardized test score view)**

We have a dataset with kids‘ background, their parents upbringing among others **1** 

schools

The key metric of success is the standardized test 

**2**

scores

better?

**3** We need to re-create a comparable control group

Unconfoundedness

**Context Visualization**

The variables (confonders) used are enough

to describe the people or entities (W)

**W**

The characteristics affect the likelihood of

v

someone being part of the treatment (X)

The combination of the confounders and the

treatment leads to the outcome (Y) Meeting the Uncondoundedness assumption

**X Y**

is a tall order

v

Curse of Dimensionality

**Visualization Context**

Imagine you have a variable with 3 options

Then you had a second with 3 more

Finally, a third

The observations needed to fill each bucket

grows exponentially

The Matching outcome can be spurious, when

few elements belong to a “dimension”

**Key Idea** 

Make sure when you create a model as simple as possible

How to determine the Common Support Region

**Visualization**

Unconfoundness

List of confounders

**Examples**

y

t

i

s

n

e

D

**Key ideas**

We preditct whether someone is part of the treatment group

There will be people with high

Via Logistic Regression

Probability of being treated

Probability

likelihood of participating. You are not likely to find a

Treatment

Common Support

Region

Probability of the treated group being treated 



Probability of the non treated group bring treated

control group for them.

The greater the overlap, the higher the matching quality