**STATISTICS**

**Statistics is the grammer of science! (Karl Pearson)**

**14.04.2023**

Data Science bilimsel yontemini, dayanagini istatistikten alir.

Data Science istatistiksel yontemlerle analiz ederek, sayisallastirarak yani olculebilir hale getirerek anlamli insight’lar elde eder.

Yazilim teknolojilerini kullanarak analizini yapar

Data Science’in science olmasi istatistik’e dayalidir.

Istatistik data science’in kalbidir. Remember the beating heart! :)

#data preprocessing

#data visualization

#machine learning

#hypothesis tests

#prediction and insights

#data analysis

-istatistik modeller tahmin eder

-buyuk datalari analiz eder

**-datadaki patterni bulur. (oruntu tanimlama)**

Fundamentals Statistics

Probability

Distributions

Central Limit Theorem

Hypotesis Testing

YouTube channels:

Youtube StatQuest

YouTube CrashCourse

Data Manipulation as EDA

Data Visualizaton: 1. Tableu 2. Power BI

**AB test:**

After before. Mesela bir reklam yapilmis, satislar artmis, ama test ile bakip reklamdan mi kaynaklaniyor satis artisi onu ortaya koyuyoruz. Ona gore reklam harcamalarini revize eder firma. Iste bunu istatistiksel analiz ile yapiyoruz.

1. Collection
2. Characterization
3. Analysing
4. Visualisation
5. Inference-Insight
6. Presentation-Ilk 8 sn!

Impossible unlikely even chance likely certain

-tum bu surecler neticesinde tahminde bulunuruz.

-Bu tahminler belli bir guven araliginda yapilir.

-belli bir guven araliginda tahmin etmek istatistiksel bir yaklasimdir.

-unstructured datayi isliyoruz.

**Bias:**

Bias Variance Tradeoff

Yanlilik denebilir

**Randomization**

Bir seye bagli olarak degil de rastgele olmasi lazim bias olmamasi icin

Random olmazsa, belli bir sonuca dogru yonlendirme yapilmis demektir

**Istatistige gunluk yasamimizda en cok basvurulan alanlar:**

-arastirmalar

-hastalik teshisi

-Devlet islerinde

-Politika kampanyalarinda

Machine Learning’deki en onemli yaptigi sey, datadaki pattern’i yakalamak!

Statistics+sociology=sociometry deniyor mesela

Statistics+economy=econometry

Herbir bilim dalinin istatistikle birlesmesi ile yeni bilim dallari cikiyor

**Data science ve istatistik farki:**

**Istatistik:**

Istatistiksel analiz sistemi: sas programlama dili kullanilir

Significance testing yapilir

Istatistik daha akademik kalir

**Data science:**

Data science’da otomasyon sistemi vardir.

Machine learning’de mesela pipeline diye bir sistem kurulur.

Python ve SQL kullanilir

Software engineering yapilir.

Daha sektore yonelik

Istatistikte izlenen sira:

1. verilerin toplanmasi
2. Toplanan verilerin islenip duzenlenmesi
3. Duzenlenmis verilerin tablo veya grafik olarak sunulmasi

**Tipik olay:** birbirinin tam benzeri olan olaylar tipik olay denir.

Tek bir olay toplulugu temsil edebilir.

**Kolektif olay:**ortak ozellikleri olmakla beraber birbirine benzemeyen olaylar

Bir tanesini alinca toplulugu temsil ediyor diyemeyiz.

**15.04.2023**

**Statistics**

1. Descriptive
2. Inferential

Scipy kutuphanesinden istatistik methodlarini kullanacagiz. Python tabanli

**Biz daha cok inferential statistics kullanacagiz.**

**Descriptive statistics**

Pd.describe dedigimizdeki mean, max, min olanlari bize getiriyordu. Iste bu **descriptive** statistics

Big data icin ilk yapilmasi gereken islem summarize’dir. Biz bunu descriptive statistics ile yapariz. Biz bunu python ile yapacagiz. Dagilimini gormek, visualize etmek hep descriptive. Veriyi tanimak, fotografini cekmek

Descriptive statistics, Tumden gelim yontemiyle ifade edilir.

**-Position, central tendency, frequency, variation/dispersion→bunlari icerir descriptive statistics**

|  |  |
| --- | --- |
| **Descriptive Statistics** | **Inferential Statistics** |
| **Aciklayici/betimsel** | **Cikarimsal** |
| Kolektif olaylara iliskin verileri toplamak | Populasyon parametrelerini tahmin |
| Tumden gelim | Tume varim |
| Toplanan verileri isleyip duzenlemek | Orneklemden populasyon icin cikarim yapariz |
| Duzenlenmis verileri tablolar veya grafikler seklinde sunmak | Gelecege donuk tahminler yapiyoruz |
| Ve nihayetinde kolektif olaylarin egilimini ortaya cikarmak | **Hypothesis testing** yapiyoruz |
| Organize/summarize/data presentation | Generalize findings from samples to population/hytpothesis testing/asses relationships among variables/estimation |

**Inferential Statistics**

Istatistik randomly olan yerde vardir. **Randomization** yoksa istatistik yoktur.

Orneklemden ana kutleyi/populasyonu tahmin etmeye calisiyoruz

Ciro tahmini yapmak, reklam tahmini yapmak..

Kesinlik yoktur, yuzde 90 guvenilirlikle su aralikta cikmasini bekliyoruz seklinde ifade edilir

Burdaki cikarimlar bilimsel oldugu icin degerli

Bunun uzerine satis stratejisi, reklam stratejisi belirlenebilir.

Orneklem uzerinden populasyon uzerine cikarimlar yapmak icin onemlidir inferential statistics

Mesela Hindistan’in Banglore sehrinden data cektik, Hindistan’a genellenebilir mi ona bakacagiz, orneklem ulkeyi temsil edebilir mi, ulke populasyonuna genellenebilir mi, bunlara bakilir.

**Confidence interval** dedigimiz guven araliginda bir estimation yapilir.

**Unstructured-structured data**

**Structured data,** labellanmistir.

**Unstructured data**: duzenli hale getirilmemis data. Elimizde bir suru veri var, sutunlarin ayrimi yapilmamis olabilir. Duzensiz, etiketlenmemis

**Dependent-Independent Variable**

**Columns:** variables, degiskenler

**Rows:**data, observation, record

Surekli data ise regresyon, kategorik data ise siniflandirma yapariz

**Statistical data analysis**

Verileri sayisallastirmaya calisan ve istatistiksel analiz uygulayan bir sayisal arastirma

**EDA**-Exponential Data Analysis- kesifsel veri analizi

Nan degerler, droplar,..

DATA

Numerical Data (numbers)

Age, shoe size, number of childeren

Categorical Data

(words)

Nominal

No hierarchy

*Eye color*

Ordinal

Data has a hierarchy

*Rating, mood*

Discrete

Finite Options

*Number of childeren*

Continuous

Infinite Options

*Age, weight..*

**Qualitative Data/Categorical data:**

**Muhtemel kategorik siniflardan olusan datalar:**

marital status gibi

Political party

Eye color

Race

Bunlari sayisallastirabiliyoruz.

Kategorik datalar, birbirine gore ustunluk/hiyerarsi olup olmamasina gore **ordinal veya nominal** olarak ayrilir

**Ordinal** demek sirali demek. **Hiyerarsiktir**. Ornegin gomlek bedenleri. S,M,L,XL gibi

Ya da anket yapiyorsun, 1,2,3 diyorsun, 1 kotu, 2 orta, 3 iyi manasinda (encoding)

**Nominal: hiyerarsik olmayan data.**

**Quantitative/Numerical Data:**

**Sayisal datalardan olusur.** Continous olup olmamasina gore **continous/discrete** olarak 2’ye ayrilir.

Mesela **Height continous** bir datadir.

Machine Learning’de problemin analizini yaparken, classification problemi mi regresyon problemi mi diye bakiyor olacagiz.

**Encoding**

**Label Encoding**

Raw data (ham-orjinal data)

Country sutunu ML model icin encoding yapiliyor

Country’i encoding yapip, id sutununu droplayacagiz

**One Hot Encodering**

Machine laearning’teki modele uygun hale getirmek icin, yapiyoruz

Get\_dummies’de yaptigimiz seye benzeyen bu

Mesela ulkeleri alip tek tek sutun haline getiriyor, birine 1 digerlerine 0 vererek kodluyor.

**Parameters and Statistics**

Sample’dan populasyonu tahmin etmek mesele

Populasyonun tamamini incelemeden istatistiksel methodlar yardimiyla, rastgele/**randomly** secilen gozlemlerden olusan **sample** uzerinden tum **populasyon** adina **inference** yapiyoruz. Tum mesele populasyonu bilmek. O yuzden sample uzerinde calisiyorum.

Randomly olunca da normal dagilima yakinsiyor.

Sample size de basari icin onemli bir parametredir.

**Temsili etkileyen en onemli 2 husus: 1. sample size 2. randomly selected**

ML’de overfit’i engellemek icin en onemli yapilan seylerden biri de sample size’i buyutmek.

Standart sapma sample icin s

Populasyon icin bu sigma

Yani farkli gostergelerle ifade ediliyor sample ve populasyon icin

**Sample have statistics**

**Populations have parameters**

Mesela biz Ankara’daki universite ogrencilerini calisacaksak, populasyon Ankaradaki tum universitelerin tum ogrencilerini kapsar. Bundan sample olusturulur.

Min alman gereken bir sample vardir.

Anket sonuclari istatistik, secim sonuclari parametre mesela.

**Probability vs Statistics**

Normal dagilim kismiyla ilgilenecegiz probility konusunun.

Istatistikte, drawing conclusion onemlidir.

Probility deyince, randomization sureci varsayilir. Ihtimaller degerlendirilir.

Istatistik, bu ihtimaller uzerinden sonuc cikarir.

**Probility:**Kovada bulunanlara gore, elime ne cektigimin ihtimalinin hesaplanmasi probability. Data predict

**Statistics** ise, elime rastgele cektiklerime gore, kovada ne oldugunun cikariminin yapilmasi statistics

**Statistics with Pyhton**

Scipy kutuphanesinden biz genelde stats modulunu cekecegiz.

Pandas ve numpy kullanilabiliyor.

Matplotlib gorsellestirme icin kullanilir.

**Level of Measurement**

**Nominal-**siyah mi beyaz mi, **gender**, religion, goz rengi, telefon tipi gibi, siralama yok, kategorize etmek var, aidiyet var, yes or no, **=, !=**

**Ordinal** comparison level, **level of satisfaction**, nitelikler siralanmistir, school grade (A,B,C,D) hot, hotter, hottest gibi, begendim, begenmedim, likert scale kullanilir genelde, strongly agree, agree, neutral, disagree, strongly disagree gibi, school grade gibi, **<, >**

**Interval** aralik vardir, **temperature**, 10 derece daha sicak gibi, true zero yoktur, arbitrary zero vardir. Yani keyfi/gelisiguzel secilen 0 noktasi vardir. 0 noktasi mesela C icin farkli fahreneit icin farklidir, IQ score, salary interval gibi, **+, -**

**Ratio** oransal degerlendirme, yuzde 10 daha buyuk gibi, en genis bilgiyi ratio icerir.

Degiskenin ozelliklerini tanimlamak icin kullaniliyor. True zero vardir. Mutlak 0 noktasi. Baslangici 0 noktasi demektir. Mesela yas 0’dan baslar herkes icin, bu Truezero dur. Yani Interval ile ratio arasindaki fark Truezero mu arbitrary zero mu olmasina goredir. **\*, /**