**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 has statistics**

**Populations have parameters**

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

Minimum 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. **\*, /**

**17.04.2023**

**Graphical Representation**

Datayi anlamli olmasi icin 2 boyutlu duzleme cekmemiz gerek

Patterni kategorik datalarda bile grafikle daha iyi gozlemliyoruz.

Grafik temsiller, datalardaki patterni goruntulemek icin verimli oluyor

**1.Center**

**2.Spread**

**3.Shape**

**4.Unusual Features**

**-gaps**

**-outliers**

Yonleriyle degerlendiriyoruz.

**1.Center**

Bir dagilimin merkezi, dagilimin median’inda bulunur. Median, siraladigimiz zaman o seriyi ortadan ikiye ayiran degerdir. Descriptive datada, onemli bir gostergedir. Bazen mean ile ustuste cakisir, bazen cakismaz. Merkezi deger acisindan onemli olan mediandir.

**2.Spread**

Verilerin varyasyonu anlasilir.

Atilan sut sayisina gore o maci kimin kazandigina dair bir denklem kurduk diyelim.

Burdaki sut sayisi golu etkiliyor.

**Narrow distribution** olunca, verilerin yogunlugu daha cok mean’e yakin dagilmistir. Daha dar bir aralikta, tek bir deger etrafinda toplanmis. Varyasyon daha dar.

**Broad distribution** olunca, ortalamadan daha genis araliga yayilmis demektir veriler. Gozlem kumesi genis bir araliga yayilmistir. Varyasyon daha genis.

Bir olay natural ise mesela, normal dagilima uygundur. Mesela insanin boylari, bir ortalama vardir ve degerler normal dagilmistir. Dogal gelisen seylerde normal dagilimdan bahsedilir.

**Kurtosis**

Basiklik- dar dagilimda daha dik, genis dagilimda daha alcakta olur.

Sadece mean degil, verinin merkezi de onemlidir, dagilim bicimi de onemlidir.

**Mean-based** incelersek ortalama ve standart sapma ile inceleyecegiz

**Median-based** incelersek ise IQR ile inceleyecegiz.

Dogal olaylarda, orta yol sahipleri daha coktur. Uclar daha azdir.

**3.Shape**

Asagidakiler kullanilarak dagilimin sekli tanimlanabilir.

**Symmetric**

**Number of Peaks**

**Skeweness**

**Uniform**

Datayi bir pattern’e benzetmeye calisiyoruz. Dagilimin bir yani diger yaninin ayna simetrisi seklinde ise normal dagilimdir.

Central Limit Theorem var. Ona uygun mu diye bakacagiz. Degilse datayi normalize etmeye calisacagiz ML’de. Hesaplamalarin kolaylasmasi icin

**Peak** sayisi, 1 veya 1’den fazla olabilir. Unimodal/bimodal

Unimodal veya bell shape(can egrisi) ise normal dagilim denir.

Birden fazla olay etkiliyorsa bimodal da olabilir

**Skeweness**

**Churn prediction**

Non-symmetric, bimodal bir dagilim dusunun. Isten ayrilmalarla ilgili olsun, 2. ve 7.yilda 2 kere peak yapmis olsun.

**Symmetric, unimodal, bell-shaped**’de mean degeri ortada, basari duzeyi sinifin normal dagilmis gibi.

Unimodal veya bell shape(can egrisi) ise normal dagilim denir.

**Skewed right(sagdan carpik)**

Grafigin bir tarafinda diger tarafina gore daha cok gozlem vardir.

Kuyruk ve tepe noktasi vardir. Kuyruk sagdaysa skewed right denir.

mesela ortalama not sinifta 40 diyelim, 40’in alti kaliyor diyelim ve cogunlugu 40’in altinda almis. Basarisiz sinif ornegini dusun

**Skewed left(soldan carpik)**

Kuyruk soldaysa skewed left denir.

bu da basarili bir sinif ornegi, cogu yuksek puan almis, gozlemler yuksek puanda daha fazla. Soldan carpik.

**Uniform**

Gozlem kumesinde tepe noktasi yok. Tekduze sekilde esit dagilmistir

Fabrikasyon uretimde mesela, hemen hemen aynisidir.

Zar attin, 6 gelme ihtimali 1/6’dir.

**4.Unusual Features**

Veri modellerinin ortak olagandisi ozellikleri, bosluklar, outlier (aykiri) degerlerdir

En cok mucadele edecegimiz sey outlier’dir.

**Gaps**

**Outliers-aykiri deger**

Mesela null bir hucre var, ama satir atamayacagimiz kadar onemli. Bazen mean ile ya da uygun baska bir seyle o degeri doldurararak, o satir elde tutulmaya calisilir. EDA calismasinin onemli bir boyutu. Kategorik datalarda mode, numeric datalarda mean kullanilarak genelde doldurulur.

**Word Cloud: kelime bulutu**

Kelime bulutlarinda olayi etkileyen ve en etkili olan kavramlar buyuk yazilir

Olayi etkileyen parametreleri belirler

Dunya kupasi ile ilgili bir world cloud hazirlasak en buyuk Messi yazilirdi mesela

Once outlier analizi mi? Yoksa missing value analizi mi?..

Outlier’lari eleyerek, analizimizi guclendirmeye calisiyoruz. Datasetini normalize ederek, daha anlamli sonuclar elde etmeye calisiyoruz. Outlier, modelin performansini 1. derecede etkileyen seydir. Atmaktansa logunu alarak da normalize ederek analize dahil etme imkani da olabiliyor outlier’i.

Rkare=0,75 iken Rkare=0,94’e cikti mesela outlier’lari eleyince. Yani %95 oraninda size anlamli sonuc verecek.populasyonla ilgili cikarimlar yapmak daha guvenli olacak

**Frequency**

**Frequency:**Her bir sinif araliginda ne kadar frequent oldugunu gosterir. Bir veri degerinin meydana gelme sayisi denir. Mesela 3 sayisindan 5 tane olmasi.

**Relative frequency** ise, toplam gozleme oranla ne kadar tekrar ettigini gosterir. Mesela toplam 10 sayi varsa, 3 sayisinin relative frequency’si 5/20=0.25

**Cumulative Frequency**: onceki relative frequency’lerin kumulatif sekilde toplanmasidir. Toplamda da 1 olur dogal olarak.

**Pie Chart**

Genelde nominal ve ordinal degiskenlerde kullanilir.

Data Science’da cok da kullanmayiz

**Bar Chart**

En cok bu kullanilir

Bar yuksekligi her niteligin **frekansini** gosterir.

Kategoriler vardir. **X ekseninde kategorileri** verirsiniz.

Discrete degiskenlerin karsilastirmasidir. Barlar arasi bosluk vardir.

Bar chart **kategorik degerleri** sunuyor.

**Histogram**

Interval/Ratio degiskenler icin kullanilir. Araligin sirali gitmesi gerektigi icin sutunlarin yerini degistiremezsin. Bosluk yoktur histogramda. Dagilimin **shape’ini** gostermek icin kullanilan cok onemli bir sekildir. **X ekseninde aralik**, **Y ekseninde frekanslar**i yazilir.

Birden fazla tepe noktasi varsa, datada heterojenlik var demektir. Ornek buyuklugu artirilmasi gerekir. Data sayisi arttikca normale gelmeye baslar, hatalar azalir, populasyonu daha cok yansitmaya baslar. Daha fazla data ile de ayni sonucu verirse ona gore yorum yaparim.

Histogram **sayisal degerleri** sunuyor

**Populations and Sample**

Istatistiki bir calisma cozum uzayinin incelenmesine dayanir. Neleri kapsiyor,

Ankaradaki universitelerdeki ogrenciler **cozum uzayi** mesela. Ankaradaki universite ogrencilerin ortalama gideri 1000 dolar diye bir sonuc cikardim diyelim. TR’deki universite ogrencileri bu olayin cozum uzayi degildir. Onunla ilgili yorum yapamam, benim cozum uzayim Ankara.

Populasyonun tamamina degil, bir sample’a ulasabiliyoruz. Bu sample uzerinden populasyon hakkinda cikarimlarda bulunuyoruz. Sample’i gozlemliyoruz.

population→parameters

Sample→statistics

**Central tendency(Measure of Tendency)**

**Merkezi egilim olculeri:**

1. Mean-ortalama
2. Median-ortanca
3. Mode-tepe noktasi

Tek bir degerle verileri en iyi sekilde tanimlamaya calisiriz. Tek bir metrik ile tanimlarsak, bunlar mean, median, mode olabilir.

**Dagilimi nasildir olculeri:**

1. Range
2. IQR
3. Standart sapma
4. Varyasyon

**18.04.2023**

-Verilerin merkezilestigi noktayi bilmek onemlidir.

-Dagilimin degiskenligi ne olcude bu da onemlidir.

**Mean**

Datalarin toplamini, toplam gozlem sayisina bolerek bulabiliriz.

Dagilimin yerinin belirlenmesinde kullanilir

Population mean: mu

Sample mean: x bar olarak gosterilir.

Median ve mode yerine ortalamayi kullanmanin dezavantaji var midir?

Bazen vardir: mean ile bulunca, outlier’a karsi hassas denir. Extreme degerlere karsi hassastir yani.

Orneklem 9 kisilik olsun diyelim. Maaslari gosteren bir data seti var. Ortalama maas da 37bin $olsun. 8 kisinin maasi aslinda 24-33 arasinda, ama 1 kisinin 102 oldugundan ortalama 37 cikiyor. Ortalama yani aslinda 9 kisilik grubun 8 kisinin maasindan yuksek

Boyle case’lerde mesela median’i kullanmak daha dogru bir yaklasim

**Median**

Mean ile median arasinda cok fark olabilir. O zaman outlier kanaatine varabilirim.

Mean ve median’a bakarak data’da outlier var mi, skewed right mi left mi diye anlayabiliriz.

Eger skorlarin kucuk bir kumesinde outlier varsa median daha iyidir.

Salary teklifinde “benzer konumlardaki maaslarin median degerine talibim”!

|  |
| --- |
| Mean→ aritmetic average  median→ positional average  Mean → central gravity  Median → mid-point |

**Mean median’in saginda kalirsa skewed right**

Box-plot’da da median solda kalir.

**Mean median’in solunda kalirsa skewed left**

Box-plot’da da median sagda kalir.

☑ Simetrik dagilimda median ve mean ayni yerdedir.

Mean ve median arasinda fark buyukse, outlier’lar var diyip, onlarla mucadele edecegiz. Outlier sorununu ortadan kaldirinca, genelde mean ile devam ederiz.

Mean ve median sadece numeric degerlerde kullanilir.

**Mode**

Data setinde en fazla tekrar eden, en populer degeri

Hem numeric hem kategorik degerlerde kullanilir.

Mode degeri olmasi icin fazla gozlem olmasi gerekir. Hepsinden birer gozlem var diyelim, bunun mode’u olmaz

Bazen de verisetinin birden fazla mode’u olabilir

Median gibi mode da outlier’lardan etkilenmez.

**Dispertion**

Merkezi egilim olculeri tek basina dagilimi karakterize etmez.

Iki data seti dusunun ikisinin de mean’i 100, ama biri broad distribution, biri narrow diyelim. Yani birinin standart sapmasi 40, birinin 10 diyelim. Yani biri mean’in daha genis araligina yayilmis gozlemler standart sapmasi yuksek olanlarda.

Dagilimi bu durumda, en iyi standart sapma ile aciklayabilmis oluruz.

1.Range

2. Standard deviation (sd)

3.IQR (interquartilerange)

**Range (aralik)**

Max-min

Bu bana bir dagilim bilgisi verir ama cok da degil.

Data nerde yogunlasmis bunun bilgisini range ile elde edemem

**IQR**

Outlier tespiti icin onemli

Bir sayi grubunu dorde bolen degerlerdir.

Q2 tum datasetinin mediani

Q1 medianin altinda kalan kisminin mediani

Q3 medianin ustunde kalan kismin mediani

IQR=Q3-Q1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 25% | 25% | | 25% | 25% | | 25% |
| Q1 | | Q2 | | | Q3 | |

Eger sub-set’te, median’in solunda ve saginda kalan yerde 4 tane data var diyelim. 2. ve 3. datanin ortalamasi buralarin median’idir.

Ama python, daha farkli yapiyor. Datasetinin medianini hem sagdaki hem soldaki, sub-setlere ilave ederek hesaplama yaptiriyor. IQR yapip, box plot cizdirerek yaptiracagiz.

Mesela Q1 50 cikti diyelim, ogrencilerin %75’i 50’den fazla not almis diye yorum yapabilirim.

**IQR’la Outlier bulmak:**

Outlier, 1.5 IQR uzerinden yapilabilir.

Outlier bu hesaplamaya gore, Q1’in altinda veya Q3’un uzerinde 1.5 IQR’dan fazla olan veri noktalaridir.

Bu outlier’lari ignore ediyoruz.

Bazi durumlarda 3 IQR alinabiliyor. Dataniz kiymetli ise mesela 3 IQR aliniyor.

Bazen 5IQR’a kadar genisletildigi de oluyor ama python’da default olarak 1.5 IQR aliniyor.

Mesela Q3+1.5IQR’in ustunde kalan noktalar ile, Q1-1.5IQR’in altinda kalan noktalar icin outlier diyecegiz.

**Varyans**

Varyans degisim demek. Degisim genisligi ile ilgili bir kavram.

**Herbir skorun ortalamadan ne kadar uzaklastigi**ni anlatan rakamsal bir deger.

X’de degisim oldugunda Y’de degisim varsa, arada varyans var demektir.

X’de degisim varken, Y’de degisim yok ise, varyans yok demektir.

**Variance, standart sapmanin karesidir.**

Standart sapma da varyansin karekoku dolayisiyla.

Varyans her bir skorun ortalamadan farklarinin karelerinin ortalamasi olarak tanimlanir.

Her bir skorun mean’den uzaklastigi miktardir.

Sample variance:sskare

Population variance:

Populasyonsa N’e boluyoruz. Gozlem sayisi

Sample ise n-1’e boluyoruz. O yuzden sample’in variance’i populasyonun varyansindan daha buyuktur denebilir.

Example

Data seti: 0, 1, 5, 6 degerlerinden olusuyor diyelim

Ortalama:3

Ss=(0-3)kare+(1-3)+(5-3)

**Standart Sapma(Standard Deviation)**

En yaygin kullanilan dagilim olcusudur.

**Varince’in karekokudur.**

**Standart sapma ortalamadan saptiginin numeric bir olcusudur.**

Sample icin **s** ile gosterilir.

Populasyon icin **sigma**

Standart sapmasi birinin 10, birinin 40 diyelim 2 sample icin

10 olanin dagilimi daha dar, gozlemler, ortalamaya daha yakin duruyor demektir.

40 icin ise daha genis araliga yayilmis demektir. Broad distribution.

**Standart sapma daha buyuk ise, daha genis yayilmis. Data setini olusturan degerler ortalamaya daha uzak aralikta yayilmis demektir. Standart sapma daha kucukse, veriler ortalama etrafinda daha dar bir dagilima sahiptir.**

Range(max-min) gibi standart sapma da outlier’dan etkilenen bir degerdir. **Standart sapma outlier’in onemli bir gostergesidir** yani. Ss buyuk ise outlier var demektir. Outlieri cikardigimizda ve yeniden hesapladigimizda, ss kuculur. Buyuklugunu ortalamaya gore ve min max’a gore olcmek mantikli.

**Emprical Rule**

**3 Sigma Kurali (Three Sigma Rule)(68-95-99.7 kurali)**

**Dagilimin normal dagilim olmasi sartiyla:**

1. **Kural**

**Gozlemlerin %68’i**

**(Ortalama-standart sapma) ve (ortalama+standard sapma) arasinda yer almaktadir.**

Gozlemlerimizin % 68’i x-s ile x+s araliktadir deriz.

Yani ortalamadan 1 ss saga 1 ss sola gidersen, bu aralikta verilerin %68’i bulunmaktadir

1. **Kural**

**Gozlemlerin %95’i**

**(Ortalama-2xstandart sapma) ve (ortalama+2standard sapma)arasinda yer almaktadir.**

Gozlemlerimizin % 95i x-2s ile x+2s araliktadir deriz.

Yani ortalamadan 2 ss saga 2 ss sola gidersen, bu aralikta verilerin %95’i bulunmaktadir

1. **Kural**

**Gozlemlerin neredeyse tamami, nearly all observations**

**(Ortalama-3xstandart sapma) ve (ortalama+3standard sapma)arasinda yer almaktadir.**

Gozlemlerimizin neredeyse tamami x-3s ile x+3s araliktadir deriz.

Yani ortalamadan 3 ss saga 3 ss sola gidersen, bu aralikta verilerin neredeyse tamami bulunmaktadir.

Skewed olan bir dagilimda bu yuzdeler gecerli olmaz. Bunlar sadece normal dagilim icin gecerlidir.

**Ex:** Mesela mu 50, standart sapma 10 ise,

Sinifin %68’i 40-60 arasinda not almis deriz

%95’i ise 30-70 arasindadir deriz.

%99.7’si de 20 ile 80 arasindadir deriz.

Elimizdeki data normal dagilim degilse, mumkunse normalize edip bunu kullanabiliriz

Regularization da deniyor bu surece. Normal dagilima cekmek.

Bazen logaritma alinabilir, karesi alinabilir, bazen veriyi atarak normallestirebiliriz.

Python arka planda bir takim komutlarla, datayi normalize edecek.

Normalize edilmis datayi modele uygulayacak.