**Statistics II**

AI may take over the technical part but won’t take the place of science and analysis.

Datalarin arasindaki iliskileri matematiksel isleme cevirir.

Statistics(yontem) – Programa Dilleri(Tools) – Tahmin, “insight” “decision” Knowledge(Goal) – Kapsam(Everything).

The base of algorithms is mathematics. That’s why we say our predictions are correct.

We analyze using statistical- why we make data mathematical; to make it measurable, bilimsel temele dayandirmak. Matematiksel kesinlik. We provide knowledge, to take decisions, using programming languages by statistical basis. DS statistical bilimi uzerine kurulmus yapidir. A picture containing text, screenshot, font

Description automatically generated

Features; factors affecting an event. Feature engineering 🡪 Machine Learning.

Frequency tables: Discrete and continuous data.

Bar chart: discrete data.

Histogram: continuous data, variation, skewness, and outliers

Boxplots: cont. data, variation skewness and outliers

Scatter plots: Two continuous variables. Or bivariate data.

Main goal is to extract insights and anlamli.

Scatter plot: ilişkinin yönünü ve büyüklüğünü gösterir. Shows the direction and strength of the relationship between two variables or features. What is the Pattern to find the relationship.

Bagimli ve bagimsiz degisken – Simple Linear regression – multiple regression.

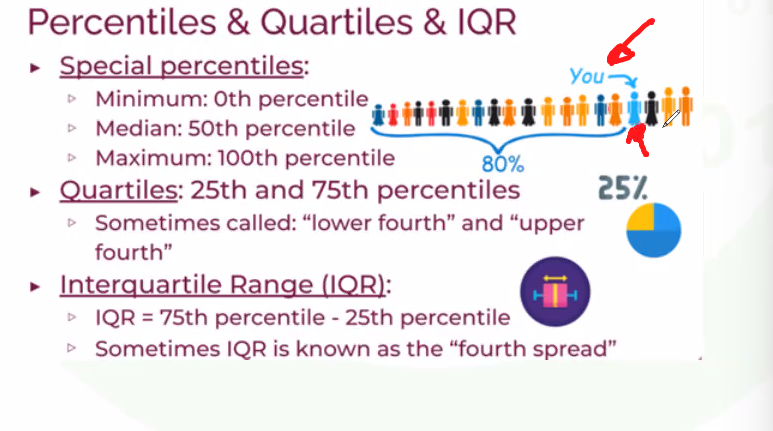
Linear, nonlinear, no relation. Fit – line of best fit; after studying error parameter. There’s no trash data, suyunu cikaracaksin.

* Linearity: checks if the data model is linear or not.
* Slop: either +ve or -ve. Undefined (x doesn’t change but y changes) and zero(x changes while y doesn’t).
* Strength: the degree of scatteration. İf the pattern is too visible or clear, then strong, otherwise weak. And differs the level from one area to other.
* Unusual Featres:

1. Clusters (Kumelenme): some here and some there.
2. Gaps (Bosluklar)
3. Outliers (Aykiri değerler): affects the best fit lines.

Visual bin(d)ing: if i can’t see a pattern then :  
a. Quantiles: quartiles, deciles.

b. percentiles.

 A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generatedA picture containing diagram, text, screenshot, line

Description automatically generated

IQR in some books called as fourth spread.

if the median is in the 50% or middle of IQR then normal distribution otherwise skewness.

Box and whisker plot is an EDA tool to show some important features of the data.

Covariance: the direction of change between two variables but not strength. 2 li degisen. Not as important as correlation. Ortalamadan ne kadar farkli. A picture containing text, font, screenshot, line

Description automatically generated

Can be between +infinity and -infinity. Covariance is not normalized.

Correlation: the level of relationship; both strength and slope. Values between +1 and -1. Correlation is the normalized (scaled between -1 and +1) covariance or covariance over standard deviation.

R(population) and r(sample)

Y 🡪 Dependent variable and X 🡪 independent variable. We check the variation of y using x. if there is no variance then no statistic.

Correlation doesn’t imply causation. So, if there is correlation it doesn’t mean cause and effect. Like the example of consumption of ice-cream and sweating in summer.

A picture containing text, screenshot, line, font

Description automatically generatedA picture containing text, screenshot, font

Description automatically generatedA picture containing font, text, handwriting, line

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

In Random patterns r = 0. Or correlation is 0, no correlation.

In different sectors the perfect correlation differs, for one it can be 0.85 while in other it can be 0.95.

R2 decides what model is the best. Supervised learning is the first and most important part of machine learning model. Correlation is the strength and direction of the relationship between two variables.

Covariance is not scaled hence not practical to use. We use heatmap for correlation, and can ignore the diagonal, as they show the correlation of same variables.

Multicollinearity, to check the relationship among the features. If check the correlation between Celsius and Fahrenheit we get a very high value of correlation, hence we can drop one of them, this can also be called feature engineering. Bias – variance trade off. Over fitting: if a lot of features added and machine wanting to revise all. If unnecessary or repetitive features are there, then it happens.

Simple Linear Regression: if the equation is dependent on only one variable, if it were multiple variables then it would be multiple linear regression.

The goal of SLR is to have a futuristic prediction based on the relation of two variables. The independent variable is the cause, and the dependent variable is effect or result.

Regression: A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model can show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

If and only if you can look at an incident from all perspectives, then only we can build a good machine learning model and a good prediction.

Y – **dependent** variable **is** the **target** variable.

While probing the relationship between two quantitative variables, the first way is correlation, second is to find the simple linear regression equation.

The least square method: (y = ax + b), to predict the value of the dependent variable in relation to the value of the independent variable. Can also be denoted as regression line. The equation of linear line.

The relationship between two variables written in mathematical form is called the least square method, this is the algorithm behind machine learning.

In the formula, b is called the intercept, it is the point where the line intercepts the y – axis, it is the value of the dependent variable when independent is zero, value of y where x = 0.

In the formula, a is the slope, the change in y in every change in x, or change of dependent variable or y over change of independent variable or x.

Error, or residual, is the perpendicular line from observed values to regression line, and our aim is to minimize them.

The regression line is the minimized sum of square of errors.

* Linear Regression requirements:  
  we should have at least a dependent and an independent variable.
* The relation between them must be linear.
* Should be scalable. Interval or ratio scale. Olculebilirlik.

Line of best fit, or best fit line 🡪 the regression line. The smallest or minimum sum of square of error.

In formula, b = SP/SSx and a = yorta – b\*xort.

Whenever we find a regression line, we must mention that that equation is applicable only on that solution space. Otherwise, it’d be called extrapolation. Stay away from extrapolation. And we must be able to analyze the equation. The slope in regression equation can take any value.

Error or Residual = Observed value – predicted value.

Coefficient of determination – R2: Explaining variance in the dependent variable. How good did it fit my model. R2 - What percent of independent variables can be explained by the dependent variable. Or the independent variable can explain only R2 - number of dependent variables. Can take values from 0 to 1.

If the value is closer to 0 it means our independent variable cannot explain our dependent variable, and if closer to 1 means it can explain as perfect as close.

Error Metrics in Machine Learning \*.

Adjusted R2 – makes you drop the columns with R2 values closer to zero. And now calculate the R2 again. This is called adjusted R2.

**Probability**

Bir olayin olma ihtimalini matematiksel olarak ifade etme.

Statistical Experiment: if:

* Bir den fazla sonuc iceriyorsa
* Bu sonuclar onceden tahmin edilebilir nitelikte ise
* Sonuclar şansa dayali ise burada istatistiksel deney sartlari saglanir.

If probable results are A, B and C then: P(A)+P(B)+P(C)= 1

P(S) = (Number of successes)/(Total number of outcomes) = r/n.

Hayatta kesin bir sey varsa o da kesin bir seyin olmayisi.

“Dunyada olum ve vergiler disinda hicbir sey kesin degildir” – Benjamin Franklin.

Law of Large Estimates:

In permutation siralama is important.

Conditional Probability – P(A|B): probability of A given B. Yani B verildiginde A’nin olmasi olasiligi.

P(A|B) = P(A and B) / P(B) = P(A intersect B)/P(B)

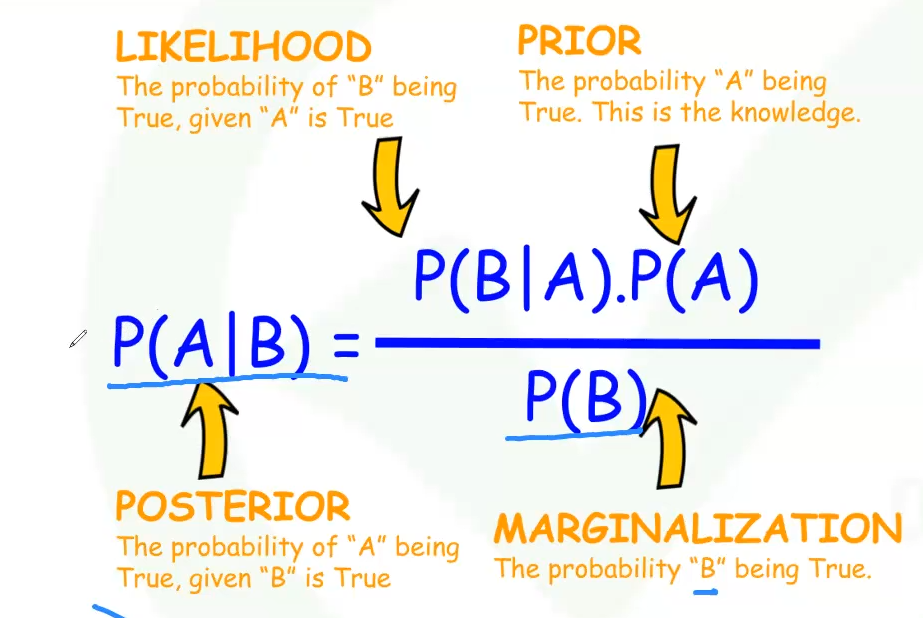
P(B|A) = P(B and A) / P(A) = P(B intersect A)/P(A)

Independence check: bir olayin sonucunun digger olayin sonucuna ya da olasiligina etkisinin olmamasi – independent. If any of the following rules applies, then it’s an independent event. If any of the following conditions hold, the other two will as well:  
- P(A|B) = P(A)

- P(B|A) = P(B)

- P(A and B) = P(A) \* P(B)

Bayes theorem: karar alicinin herhangi bir olayla ilgili subjective olasiliklarla dikkate alan ve sonucta belirsizlik kosullarini riske donusturmeye yardimci bir arac. Bu theorem bir sonucun sebebini bulurken, sonucun hangi olasilikla hangi sebepten kaynaklandigini bulmaya yardimci olur, bunu yaparken conditional probability kullanir. Verilen bilgilerin isiginda, onceden bilinen bilgileri nasil guncellenecegini belirliyor. Dolayisiyla belirsizlik varken, karar alma, machine learning ve statistics cikarim gibi alanlarda tip gibi kullanilabiliyor. This theorem is used to identify conditional probability.



Bu theorem olayla ilgili olabilecek kosullarin on bilgisine dayali olarak bir olayin olasiligini tanimlar.

Bir olayin meydana gelme ihtimalinin, olaya yonelik ek bilgi edinildigi durumda nasil degisecegini gosteriyor.

Bu theorem, diger belirli olasiliklari bildigimizde, bir olasilik bulmanin yoludur.

A picture containing text, screenshot, font

Description automatically generatedA picture containing text, screenshot, font, line

Description automatically generated

A picture containing text, font, electric blue, logo

Description automatically generated

**Random Variables (Rastgele Degisken)**

Olaylar ve rakamlar arasinda baglanti kurmaya yarayan arac, aldigi degerlerin dagilimi, olasilik dagilimlari, istatistiksel cikarimlarin ve hipotez testlerin temelini olusturur. Hipotez testleri yapabilmek icin kullaniriz istatistigi. Bir degiskenin degeri istatistiksel bir deneyin sonucuysa, bu degisken rastgele bir degiskendir.

Sansa dayali.

Bu degerleri onceden kesin olarak bilmemiz mumkun degildir.

Ordinary variable is known, like “the” article, but random variable is not precisely known, like a and an articles.

Random variables can take discrete and continuous values as it’s a quantitative variable. Knowing the type allows us to understand the analyzes of result, or distribution variables, and the way we comment. In discrete one the solution space is countable or whole number. If we want to model a continuous variable, we use regression analyzes but if it’s discrete then logistic regression.

Distributions: a. Discrete (Binomial, Bernoulli, Poisson) probability distribution and b. continuous (uniform, normal, standard, T) probability distribution.

**The probability distribution**

Bir random variable’in aldigi degerlerin olasiliklarinin nasil degistigini gosteren matematiksel bir fonksiyondur. Boylece hangisinin daha yuksek ihtimalinin oldugunu gorebiliriz. Data Science’cinin isi Belirsizlik altinda karar vermek, olasilik dagilimlari kullanrak belirsizlikleri azaltabiliriz.

Bir rastgele degiskenin aldigi degerlerin olasiliklarini listeleyen statistiksel bir fonksiyondur. Bunlari belirsizligi ortadan kaldirmak amaciyla kullaniriz. Istatistiksel bir deneyin her sonucunu, gerceklesme olasiligiyla iliskilerndiren bir tablo veya denklem. Rastgele bir degiskenin turu, olasilik dagilim turunu de belirler ya da, rastgele bir degiskenin degerlerinin dagilimi, olasilik dagilimlari tanimlari.

Discrete probability distribution is the probability distribution for discrete random variables.

Continuous probability distribution is the probability distribution for continuous random variables.

Why prob. Distr. Are important:

* Olasilik tahminlerinde bulunmaya yardimci olur.
* Istatistiksel analizler yapilabilmesini saglar.
* Bazi ML modelleri olasilik dagilimi varsayimlariyla calisir.

Yaygin kullanilan olasilik dagilimlari vesilesiyle, veriden cikarim yapmamiz kolaylasir, cunku dagilimini biliyoruz.

A diagram of a distribution

Description automatically generated

Discrete probability distribution:

The probability of every random variable must be between 0 and 1 and their sum should amount to 1. PMF probability mass function and CDF Cumulative distribution function.

P(X = xi) and P(X < x)

**Binomial:** en yaygin discrete, bi = two, two probable outputs as success p and failure q or (1-q), like head and tail.

Denemeler bagimsizdir. Onceki denemeler sonraki denemeleri etki etmemeli.

*Tekrarlanan* denemeler vardir.

Two parameters: p as success and n as number of experiments or test.

Formula: P(X=x) =A math equation on a green background

Description automatically generated

Bernoulli distribution: Special case of binomial.

Iki muhtemel sonuc vardir.

Basari olasiligi sabittir.

Sadece bir deney ya deneme.

Poisson distribution: most used in industry according to some.

Poisson Dağılımı Özellikleri

Poisson dağılımı, T zamanında meydana gelen X olay sayısının olasılığını verir.

Denek sayısı olan n büyük iken p de çok küçük ise binom dağılımı poisson dağılımına yaklaşır

Genel olarak np<0 olduğu zaman binom dağılımı yerine poisson dağılımı kullanılabilir.

Independent events.

Poisson dagilimi, belirli bir zaman diliminde sabit sayida kullanilir. Gerceklesen olayin, istenilen bir zaman araliginda ne kadar olasilikta gerceklesecegi ya da kac kere gerceklesecegini bulmak, ve iki olay ayni zamanda gerceklesmeyecek.

**Continuous probability distribution.**

if the random variables have continuous probability distribution. Sureklilik, bir aralikla ilgili konustugumuz zaman. 0 ile 1 arasindadir dagilimlari, egrinin altinda kalan alan, integral. In discrete we used PMF and CDF (in cumulative), here we use PDF or probability density function, bir araligin altinda kalan alanin hesabinni soyler, and in cumulative we use CDF here as well.

1. Uniform distribution: Rastgele bir degiskenin esit olasiliklarla meydana gelebilmesidir. Mean = (a+b)/2 and variance = 1/12\*(b-a)2. Butun modeller yanlistir, ama bazilari kullanisli ya da faydalidir. Bir deneyin esit olasilikla meydana gelmesidir.
2. Normal distribution: most important of continuous. Gauss, gauss – Laplace, or bell curve distribution. This is the basis for almost all statistical analysis and hypothesis tests. If normal, it means that it’s symmetrical, mean is 0 and standard deviation is one. Degerler Merkez etrafinda kumelenme egilimi gosterir. Dagilimin her iki ucu giderek yatay eksene yaklasir, ancak hicbir zaman bu eksene degmez (asimptomatik). Egri altinda kalan 1e esittir. 50 percent of the data is located to the right of the center, and the other 50 percent is to the left. Most of the incidents in nature have normal distribution.   
   Empirical rule: if a variable has normal distribution, 68 percent of the data is located within one standard deviation from mean, 95 percent within two standard deviations, and 99.7 percent within three standard deviations from the mean point.

**Z Tablolari ile alan hesaplama. Z-score, we standardize the normal distribution to make it easy to calculate.**