**DATA SCIENCE**

**[TERMS]**

\* Analysis: past events.

\* Qualitative Analysis: explain how? and why?

\* Quantitative Analysis: data + quantity in the past.

\* Analytics: potential future events.

\* Qualitative Analytics: intuition + analysis.

\* Quantitative Analytics: formulas + algorithms.

\* Business Intelligence: analyzing and reporting historical business data. make decisions. extract insights. extract ideas.

\* Machine Learning: make predictions. analyse patterns. give recommendations.

\* Artifical Intelligence: simulating human knowledge. decision making.

\* Data: information stored in digital format. traditional/big data.

**[BUSINESS ANALYTICS]**

\* Business Experience-Driven

- Business Case Studies (Past)

- Qualitative Analytics (Future)

\* Business Data-Driven

- Preliminary Data Report (Past)

\*Business Intelligence

- Reporting with Visuals (Past)

- Creating Dashboards (Past)

- Sales Forecasting (Future)

\*Artifical Intelligence

- Creating Real-Time Dashboards (Past)

- Client Retention

- Fraud Prevention

- Machine Learning

- Symbolic Reasoning

\* Data Analytics

- Data Science

**[PRELIMINARY DATA REPORT]**

\* first step of any data analysis.

**[DATA]**

\* raw data >> processing >> information >> business intelligence

\* processing: pre-processing, case-specific

\* pre-processing: class-labeling, data-cleansing, missing-values

\* class-labeling: numerical, categorical, text, digital media

\* case-specific: data balancing, shuffling, mining, masking,

\* database diagram: entity relationship diagram, relational schema

**[BUSINESS INTELLIGENCE]**

\* preliminary step of predictive analytics.

\* analyze past data and extract useful insights.

\* create appropriate models.

\* observation >> quantification >> measure >> metric

\* metric = measure + business meaning

\* key performance indicator = metric + business objective

\* visualizations: reports, dashboards

**[TRADITIONAL METHODS]**

\* predictive analytics

\* regression: causal relationships among different variables.

\* logistic regression, clustering, factor analysis, time-series

**[MACHINE LEARNING]**

\* data, model, objective function, optimization algorithm, trial and error

\* supervised learning: labelling data, minimizing loss, support vector machines, neural networks, deep learning, random forests, bayesian networks

\* unsupervised learning: unlabeled data, k-means, deep learning

\* reinforcement learning: maximizing reward

**[SOFTWARE]**

\* Languages: Python, Infographic (R), C, C++, Java, Scala

\* Database: SQL

\* Calculation: MatLab

\* Traditional Data: Excel, SPSS

\* Libraries: Apache Hadoop, Base, Mongo-DB

\* Business Intelligence Visualizations: Power BI, SAS, Tableau, Qlik

\* Econometric Time-Series Model: Eviews

\* Academic Statistical and Econometric Search. Stata

**[BAYESIAN NOTATION]**

\* set: collection of elements.

\* empty set (ø): a set that contains no elements.

\* x ϵ A: element x is a part of set A.

\* A ϶ x: set A contains element x.

\* x∉A: element x is not a part of set A.

\* Ɐx: for all x such that

\* A⊆B: A is a subset of B. A⊆A and ø⊆A

\* AՈB: intersection of two sets. elements those are on both sets.

\* AՍB: union of two sets. A + B - AՈB. all elements in both sets.

\* mutually exclusive: AՈB = ø

**[PROBABILITY]**

\* likelihood of an event occuring. value between 0 and 1. probability 1; absoulte certainty of event occuring. probability 0; absoulte certainty of event not occuring.

\* **theoretical probability:** P(A) = (preferred outcomes) / (sample space)

\* **experimental probability:** P(A) = (successful trials) / (all trials)

\* **sample space:** all possible events. P(A) + P(B) + P(C) + ... = 1

\* **dependent event:** an event that is affected by other event.

\* **independent event**: an event that is not affected by other event.

\* **conditional probability**

- probability of dependent events. P(A|B) = P(AՈB)/P(B) ≠ P(A)

- probability of independent events. P(A|B) = P(A)

- probability of simultaneous independent events. P(A, B, C, ...) = P(A) \* P(B) \* P(C) ...

\* **total probability law:** P(A) = P(A|B1)\*P(B1) + P(A|B2)\*P(B2) + ... (A is union of mutually exclusive sets B1, B2, ...)

\* **additive law:** P(AՈB) = P(A) + P(B) - P(AՈB)

\* **multiplication rule:** P(AՈB) = P(A|B) \* P(B) (probability of simultaneous events)

\* **bayes rule:** P(A|B) = [P(B|A) \* P(A)] / P(B) (how two events affect each other)

\* **expected value:** specific outcome we expect to occur when we run an experiment.

- categorical: E(Y) = n\*p (n: number of trials, p: probability)

- numeric: E(Y) =

- σ2 = E((Y- μ2)) = E(Y2) - μ2 (mean and variance relationship)

\* **frequency:** number of times a given value or outcome appears in the sample space.

\* **frequency distribution:** probabilities for each possible outcome of an event.

\* **frequency distribution table:** table matching each distinct outcome in the sample space to its associated frequency.

\* **probability frequency distribution:** divide every frequency by the size of the sample space.

\* **complement:** everything an event is not.

- (A’)’ = A

- P(A) + P(A’) = 1

- (A + A’) = (sample space)

**[COMBINATORICS]**

\* **combinatoric:** combinations of objects from a specific finite set. permutation, variation, combination.

\* **factorial:** product of all integers from 1 to n. n!, 0! = 1, n < 0; n! doesn’t exist.

\* **permutation:** number of different ways we can arrange a number of elements. P(n) = n!

\* **variation:** number of different ways we can pick and arrange a number of elements. order is relevant.

- repetitive: (n, p) = np (n: number of elements, p: number of positions)

- non-repetitive: V(n, p) = n! / (n-p)!

\* **combination:** number of different ways we can pick a number of elements. order is irrelevant. no multiples.

- repetitive: (n, p) = (n+p-1)! / [p! \* (n-1)!] (n: number of elements, p: number of selections)

- non-repetitive: C(n, p) = V(n, p) / P(p) = n! / [(n-p)! \* p!]

- symmetry:

- seperate sample spaces: C = n1 \* n2 \* ... \* np (ni: size of ith sample space)

**[PROBABILITY DISTRIBUTIONS]**

\* **distribution:** possible values a random variable can take and how frequently they occur.

- notation: variable ~ type (characteristics)

- Y: actual outcome

- y: one of the possible outcomes

- p(y) = P(Y=y): probability function. assigns probability to each distinct outcome in sample space.

- mean: average value of distribution.

- variance: measure on how spread out the data is. Var(Y)

- standard deviation: positive square root of variance.

- disreete distribution: finite outcomes. bar graph. P(Y ≤ y) = P(Y < y+1). discreete uniform, bernoulli, binomial, poisson

- continuous distribution: infinite outcomes. smooth curve. P(Y = y) = 0. P(Y ≤ y) = P(Y < y). normal, students-t, chi-squared, exponential, logistic

\* **population:** includes all possible outcomes of an event.

\* **sample:** includes only a few possible outcomes of an event.

\* **probability distribution of population:** mean (μ), variance (σ2), standard deviation (σ).

\* **probability distribution of sample:** mean (), variance (s2), standard deviation (s).

\* **trial:** observing an event and record the outcome.

\* **experiment:** collection of one or multiple trials.

\* **experimental probability:** probability of an event based on the experiment.

**[DISCREET - UNIFORM DISTRIBUTION]**

\* **Y ~ U(a, b)**

\* all outcomes have equal probability. dice, cards.

\* expected value and variance have no predictive power.

**[DISCREET - BERNOULLI DISTRIBUTION]**

\* **Y ~ Bern(p)**

- E(Y) = p

- p > (1-p)

- σ2 = p\*(1-p)

\* single trial and two possible outcomes. true/false. coin.

**[DISCREET - BINOMIAL DISTRIBUTION]**

\* **Y ~ B(n, p)** (n:number of trials, p:probability of success)

- Bern(p) = B(1, p)

- p(y) = C(y, n) \* py \* (1-p)n-y

- E(Y) = p \* n (number of heads expected after flipping coin 10 times is 5.)

- σ2 = n \* p \* (1-p)

\* two possible outcomes per iteration and many iterations. sequence of identical bernoulli events.

\* how likely an event is to occur over a series of trials.

**[DISCREET - POISSON DISTRIBUTION]**

\* **Y ~ Po(λ)**

- λ: anticipated value.

- p(y) = (λy \* e- λ) / y!

- E(Y) = σ2 = λ

\* likelihood of a certain event occuring (frequency) over a given interval of time.

\* how likely a specific outcome is, knowing how often the event usually occurs.

\* out of ordinary or not.

\* only non-negative values.

**[CONTINUOUS DISTRIBUTIONS]**

\* **probability density function (PDF):** f(y)

\* **cumulative distribution function (CDF)**: F(y) = P(Y≤y)

\* **probability of individual event:** P(Y=y) = 0

\* **probability of interval:**

\* **PDF to CDF:**

\* **CDF to PDF:**

\* **expected value:**

\* **variance:**

**[CONTINUOUS - NORMAL DISTRIBUTION]**

\* **Y ~ N(μ, σ2)**

- E(Y) = μ

- outliers: {μ-σ, μ+σ}

\* natural events. size of animals in nature.

\* bell-shaped. symmetric. thin tails.

\* %68 of values should fall within (μ-σ, μ+σ). (68%, 95.99%, 99.7%)

\* shift left: decrement μ.

\* shift right: increment μ.

\* higher peak-thinner tails: decrement σ

\* lower peak-fatter tails: increment σ

**[STANDARD NORMAL DISTRIBUTION - Z-TRANSFORM]**

\* transformation: altering every element of a distribution to get a new distribution.

- add: shift-right

- subtract: shift-left

- multiply: shrink

- divide: expand

\* Z-Transform: standard normal distribution to z-score table.

- E(Y) = 0, Var(Y) = 1

- Y ~ N(μ, σ2) → Z ~ N(0, 1)

- z = (y-μ)/σ

- Z: standardized variable

- z: critical value (value read from table of known values)

**[CONTINUOUS - STUDENTS’ T DISTRIBUTION]**

\* **Y ~ t(k)**

- k: degrees of freedom

- k > 2 → E(Y) = μ, Var(Y) = (s2\*k)/(k-2)

\* small sample size approximation of a normal distribution.

\* bell-shaped. symmetric. fat tails.

**[CONTINUOUS - CHI-SQUARED DISTRIBUTION]**

\* **Y ~ χ2(k)**

- E(Y) = k

- Var(Y) = 2k

- Y ~ χ2(k) → ~ t(k)

- Y ~ t(k) → Y2 ~ χ2(k)

\* hyphotesis testing for goodness of fit.

\* asymmetric. skewed right. non-negative values.

\* contains a table of known values.

**[CONTINUOUS - EXPONENTIAL DISTRIBUTION]**

\* **Y ~ Exp(λ)**

- λ: rate parameter. how fast reaches to plateau.

- E(Y) = 1/λ

- Var(Y) = 1/λ2

- Y ~ Exp(λ), X = ln(Y) → X ~ N(μ, σ2) (transformation of exponential distribution to normal distribution)

\* rapidly changing events. online website traffic, radioactive decay.

\* starts off high. initially decreases. eventually plateauing.

\* not contains a table of known values. so transformation is necessary.

**[CONTINUOUS - LOGISTIC DISTRIBUTION]**

\* **Y ~ Logistic(μ, S)**

- S: scale (smaller scale closer to 1)

- E(Y) = μ

- Var(Y) = (s2\*π2)/3

\* how continuous variable intputs can affect probability of of a binary outcome. forecast analysis. team sports.

\* CDF picks up near mean. cut-off point.

**[STATISTICS]**

\* **data set:** population (N parameters, whole), sample (n statistics, partial)

\* **types of data:** categorical, numerical (discreet, continuous)

\* **levels of measurement:** qualitative (nominal:no-order, ordinal:order), quantitative (interval, ratio:true-zero)

\* **visualizations for categorical data:** frequency distribution table, bar chart, pie chart, pareto diagram, cross table

\* **visualizations for numerical data:** frequency distribution table, histogram, scatter plot

\* **measures of central tendency:** mean, median, mode

- **mean:** simple average of data set. population mean: μ, sample mean: . /n

- **median:** midpoint of the ordered data set. position of median in data set: (n+1)/2

- **mode:** value that occurs most often. value with highest frequency.

\* **skewness:** measure of asymmetry. dataset is concentrated on one side or not.

-

- **positive (right-skewed):** mean>median>mode

- **negative (left-skewed):** mean<median<mode

- **zero (symmetric):** mean=median=mode

\* **measures of variability:** dispersion of data around its mean value. variance, standard deviation, coefficient of variation

- coefficient of variation (CV): standard-deviation/mean

- population variance: σ2 = → population standard deviance: σ → CV = σ/μ

- sample variance: s2 = → sample standard deviance: s → CV = s/

\* **measures of relationship:** covariance, linear correlation coefficient

- **covariance:** measure of correlation between two variables.

- positive: two variables move together.

- negative: two variables move in opposite directions.

- zero: two variables are independent.

- population covariance:

- sample covariance:

- **correlation:** standardized measure of correlation between two variables. values between [-1, 1].

- perfect positive correlation (1): one variable can be explained by other entirely.

- perfect negative correlation (-1): one variable can be explained by other entirely.

- absolutely independent variables (0): one variable cannot be explained by other.

- correlation coefficient: cov(x, y)/[stdev(x)\*stdev(y)]

- population correlation:

- sample correlation:

- causality: correlation symmetrical with respect to both variables. cov(x,y) = cov(y,x). correlation doesn’t imply causation.

**[INFERENTIAL STATISTICS]**

\* **distribution:** function that shows the possible values for a variable and how often they occur.

\* **discreet uniform distribution:** all outcomes have an equal chance of occuring.

\* **central limit theorem:** sampling distribution of the means approximates a normal distribution.

\* **sampling distribution:** distribution formed by samples.

- ~ N(μ, σ2/n)

- distribution of

- k: number of samples.

- n: size of samples.

- more samples closer to normal. k → ꝏ

- bigger samples closer to normal. n → ꝏ

\* **standard error:** standard deviation of the distribution of sampling distribution.

\* **estimator:** function that approximates a population parameter depending only on sample information.

- mean(μ) estimator:

- variance(σ2) estimator: s2

- correlation(ρ) estimator: r

- smaller variance more efficient estimator.

- bias: expected value of an estimator is (parameter+bias).

- efficiency: most efficient estimator is the one with the smallest variance.

\* **estimate:** output of estimator.

- point estimate: single value. located in the middle of confidence interval.

- confidence interval: interval. more precise than point estimates.

**[CONFIDENCE INTERVALS]**

\* **confidence interval:** interval within which we are confident the population parameter will fall.

- level of confidence: (1-α). 0<α<1. common α: 0.01, 0.05, 0.1

- reliability factor (population, z-statistic):

- reliability factor (sample, ): , (, degrees of freedom)

- margin of error (ME):

- margin of error (population variance, z-table, narrow):

- margin of error (sample variance, t-table, wide):

- confidence interval: [ - ME, + ME]

\* **confidence interval of two samples**

- variance of the difference:

- dependent:

- population known/independent:

- population assumed equal/independent:

-

-

- population assumed different/independent:

**[HYPOTHESIS TESTING]**

\* **scientific method:** systematic observation, measurement, experiment, formulation, testing, modification

\* **hypothesis:** idea that can be tested. supposition made on limited evidence as a starting point for further investigation.

\* formulate hypothesis, find right test, execute test, make decision

\* **null hypothesis:** H0. hypothesis to be tested. status quo. (yerleşik inanç). decisions: accept or reject.

\* **alternative hypothesis:** all hypothesis except null. H1, HA. challenging. (aykırı söylem)

\* **rejection region:** region at the tails of standard normal distribution where we reject the null hypothesis.

\* **acceptance region:** middle region of the standard normal distribution excluding rejection region.

\* **significance level (α):** probability of rejecting the null hypothesis that is true. common: 0.01, 0.05, 0.1. if observed statistic is too far away from 0 depending on the significance level we reject the null, otherwise we accept it.

- two-sided test: null contains equality or inequality. =, ≠

- one-sided test: null doesn’t contain equality or inequality. <, >, ≤, ≥

\* **statistical errors:**

- type 1 error: false positive. reject a true null hypothesis. probability is α.

- type 2 error: false negative. accept a false null hypothesis. probability is β.

- reject false null hypothesis. probability 1-β is power of the test.

\* **p-value:** smallest level of significance at which we can still reject the null hypothesis given the observed sample statistic.

- notable p-values: 0.000 (reject all), 0.05 (cut-off line)

- closer to zero more significant result.

\* **hypothesis testing methodology**

- calculate a statistic .

- scale it .

- check if z is in the rejection region. Z closer to z, more acceptable hypothesis. accept if |Z|< z, reject if |Z| > z.

- same applies for t-score.

- z-statistic: big samples, known/unknown variances.

- t-statistic: small samples. unknown variances.

- critical value: number-from-table.

- reject hypothesis if;

- |test statistic| > |critical value|

- p-value < significance-level

\* **hypothesis testing formulas (one population)**

- variance known:

- variance unknown:

\* **hypothesis testing formulas (two population)**

- dependent:

- independent-variance known:

- independent-variance assumed equal:

-

-

**[REGRESSION ANALYSIS]**

\* regression analysis: prediction method for variables which have cause-relationship between them.

\* linear regression: linear approximation of a causal relationship between two or more variables.

\* simple linear regression model (population): (: intercept, : slope, : error)

\* simple linear regression equation (sample): (: estimated, x: observed, b0: intercept, b1: slope)

\* sum of squares total (SST): (Total Variability)

\* sum of squares regression (SSR): (Explained Variability)

\* sum of squares error (SSE): (Unexplained Variability)

\* SST = SSR + SSE

\* Ordinary Least Squares Model (OLS) Regression

- dependent variable: variable to be predicted. y.

- coefficient of the intercept: b0

- coefficient of the independent variable: b1

- p-value of t-statistic: P>|t|. independent variable significant or not. 0.000

- p-value of f-statistic: overall significance of the model.

- Method: least squares. minimum squares error. find minimum SSE by estimator.

- OLS Estimator of β for Simple Linear Regression:

- R-Squared: R2 = SSR/SST. variability of data. [0, 1]

- Adjacent R-Squared: considering number of independent variables also. negative values are interpreted as 0.

\* Other Regression Methods: generalized least squares, maximum likelihood estimation, bayesian regression, kernel regression, gaussian progress regression

\* dummy variable: used to include categorical data into a regression model.

\* standardization (feature-scaling): process of transforming data into a standard scale.

\* standardized variable: (x: original variable)

\* overfitting: training has focused on the particular training set so much , it has missed the point. captures all noise, high train accuracy, low test accuracy. prevent by splitting data into training (80%) and testing (20%) data.

\* underfitting: model has not captured the underlying logic of the data. low train accuracy.

\* variance inflation factor (vif): multicollinearity check. produces a measure which estimates how much larger the square root of the standard error of a estimate is when assumed that the variable is completely uncorrelated with other predictors.

\* residual: differences between the targets and the predictions. much lower than the mean: overestimation, much higher than the mean: underestimation. (y\_train - y\_hat)

\* machine learning jargon: intercepts = features, coefficients = weights. bigger weight bigger the impact. positive weight: feature increases in value. example: bigger engine higher price. negative weight: feature decreases in value. example: higher mileage lower price. dummy variable: assign a numeric value for a categorical data. positive weight: more expensive than the benchmark (benchmark is a categorical data that is not taken as dummy as a rule). negative weight: less expensive than the benchmark.

**[MULTIPLE LINEAR REGRESSION]**

\* multiple linear regression model (population):

\* multiple linear regression equation (sample):

\* Adjusted R-Squared: (n: sample size, observations, p: estimators, predictors)

\* F-Statistic: testing overall significance of the model. lower f-statistic, closer to non-significant model.

**[REGRESSION ASSUMPTIONS]**

\* linearity:

\* non endogeneity:

- ommited variable bias: difference between observed value and predicted value is correlated with independent variable. caused by missing relevant variable.

\* normality/homoscedasticity: , errors are normally distributed and having equal variances.

\* no autocorrelation:, errors are uncorrelated

\* no multicollinearity: