Statistics Knoledge Colledted.

for data sicence level

- source for data used

- # Explore the data analysis

  - data type structure / unstrucure.
  - ordinary is one type of ordered factor categorial data.
  - data: continous, discrete, categorical(binary, ordinal.)
  - database are more detaield in their classification of data types sql learning
  - rectangualar data -- data frame
  - python uses panda libratry `DataFrame()`
  - index created by default
  - R uses `data.frame`
  - R does not suupor use-sepecidied indexes , pyton does
  - non-rec data: times series data, spatial data for ammping and location analytics, knowledge graph(network optimization and reccomnederssystmens)
  - we care more about rec data - pred model.

- # Estimation of location

  - mean , weightedmeanm, median, weighted median, trimmed mean, robust robust not sensitive to extrene valuesm outliers.
  - data science and ba refer estimated for calues to draw actual and therotical as **METRIC**
  - Trimmed minus the extreme values's counts when doing mean
  - weighted x* w / w sigma
  - mean is more sensitive to data, but median is less likely to be affected.
  - median is not only robust estimate, trimmed mean, eg 10% top and buttom in real life. -- compromise for median and mean.
  - others more robust
  - good sources slides for calculation estimate
  - king of ds

Esitimate of variablity

- Deviations : differnece obs - estimate location. = errors = residuals. （残差）
- variance : sum of squared deviations from mean divied by n-1 = mean squaired errors.
- stanadard deviance: suqiree root of variances -- euclidean norm
- MAD: mean abs of dev
- range
- ranks : metriecs based on data values from low to high.
- percentile
- inrerquantile range. = IQR
- MAD= x - xi / n
- var = s^2 = (x- xi)^2 / n-1
- **degree of freedom** usually don't care, cuz n is large. (cue variable why n-1) _____ it is on premise that you want make estimate of pop based on samp.
- n --- biased var. (underestimate the var)
- n - 1 standard non biased vstimate
- In fact: consider constraints in computing estimates, n-1 === one constraint. [choosing K]
- **The above are not robost**
- STD > MAD >MEDIAN AD
- 后俩基于 constraint scale of factor， 前者基于norm dist

Exploring thedata dist.

- boxplot, freq.table, hist, density,

Exploring binary and cate. data.

- mode, expected, bar chars, pie
- bar chart similar to his. -- but x-axis not orderd.
- *expected value = a form of weighted mean in which the weights are prob.*
- cate data usually summarised in proportion
- distinct thigns, levels of fa, binned num.

Correlation

- modelling
- predict and target

    - coefficient correlation -- numerica 1 -1
    - correlation matrix
    - cc = 0 means no assoication
    - contingency tables, hexagonal binning, contour plots, volin
    - denssity
    - two categorial table

- ## data and smapling dist.

  -sample, population, n, rs, strata, sample bias.

    - data quality matters here
    - bias -- statistical bias to measurements of sampling erroes systemic and proces when coolec
    - observable or not
    - Ramdom selection -- avoid bis
    - in stratigied sampling
    - size matter
    - central limist therom
    - standard error = se = s/ root of n
    - standard deviation measures variability of individual data points
    - while stand error measure the variability of the sampling metric.
    - the lrger the sample size the normally it is
    - dist. of sample are normally sgagoed
    - error , standardzize by dividing sde. 根号下
    - z-scoe of standarlizing an indiviual pooint
    - errors are normally dist. usually while data might not.
    - possion dist. for per time period

- ## t dist. estimate of mean

    - degree of freedom
    - n sample size
    - t is similiar to z but thihcker tails for sampleing.
    - t used for sample mean, regression parameters and more.

- ## binomial dist.

    - tirl, sucess,

- n 变大的时候 可以被 normal approximate
- n * p(1-p)

Poisson dist.

- Many processes produce events randomly at a given overall rate.
- lamba- rate which occurs events
- position frequency distr. given time unit
- exponential dis.**the time or distance from one even to another**
- weibull dist. version of expo rate is allowed to shift over time.

statistical experiments and siinificace testing.

- pm uses alot
- interferce
- limitted sample to larger population
- formualte the hyphthesis ---- design experince ---- collet data ---- inference and conclusion
- typical hp : a treatment is better than control

ab testing

- treatment
- traetment group
- control group
- randamization
- susbjects
- test statistics
- eg: seed germination for product, profit, produces more clickm, web ads conversisions
- bliing study / double blind

hypo

- why hupothesis ? 1. mis understadnd a random evens as a pattern thing. 2. failure to anticipate extreme
- mean difference is more extreme or not.

resampling

  - resampling helps with ml algorithm.
  - **two types: bootstrap(with repalcement) / permutation(random test) **

- permutation : combine results from different groups; shufffle and randomnly draw with NO replacementof sae size as group A; repeat that with size equals to group B *with remianing data*; C,D,E, if any; calculate the stats and constitues one permutation iteration; repeart R times to form a permutation dist. if the overseved difference lies out side most of the permuation distribution, than we conclude that chance is not reponsible.

- ## statsi sig and p-value

  - P-value: the prob. of obtaiing reuslts as unsuall or extme
  - al[ha unsual
  - type 1 reject when h0 true
  - type 2 accpet when h0 falls
  - **pvalue** is the prob. that the result is due to chance.
  - It is the prob. given a chance model, results as extreme as the observed results could occur.
  - pvalue < alpha leads to reject null hypothessis.
  - t test
  - overfitting - fititng to the noise

- ## anova

  - pairwise
  - F statistics
  - SS sum of square
  - based on the ratio of variance by group means
  - ms(treatment) / me(error) gievs f statistics
  - For illustration, suppose that you wish to test the hypothesis that p p coefficients are zero, and thus these variables can be omitted from the model, and you also have k k coefficients in

CHi-squ !!!

- CHIsquare stat == measure of extend to whcih some observed data departs from expectation
- df
- pearson residual = obs - exp / root of exp
- (r-1)(c-1)
- shuttle resampling test
- chi-squre!!! more as a filter to determine effectr or feature is wortht of further consideration than as a formal test of signiicance.
- **feature selection** in machine larning
- chi which assumeption of independence.

power and sample

- efffect size
- power
- sigifican level

reg.

- fitted values -- estimate y head
- resi -- difference obs - fitted.=== errors
- yi = b0+b1x+ei
- yihead = b0head +b1heardx
- rss residual sum
- regr ---- predict / explain
- r-suqare == the proportion of variance explained by model
- t= b/ se error of coefficient
- higher t, lowewr pvalue means significant
- minimize AIC BIC from bayesian
- weighted reg 1 inverser-variance weight
- RMSE / Rsuqare 最重要俩指标
- stanadarlized error of coefficent can be used to measure the reliabitlity of the variable contributin to a model.
- factor used: dummy 0-1
- factors needed to be converted to numeric in use.
- confounding variable

  - cooked distance -- lever+ residual size

- ## classi

  - naive bayesian 查查
  - bay for numdeic needs :1 bind and conver to cate, 2 prob. model p(x|y=i)
  - discrininat analysi : LDA taelss measure of importanace and 好计算
  - lda believes the covarince matrix same for groups e covariance matrix:
  - **LOGIST TIC** 查查

- ## evalue model

  - accuracy
  - confusion matrix 看图
  - sebsitivity = percent of 1s cirreckt classfie
  - bagging resample records
  - rf bag+ resample variables
  - boosting requires more care and tune.: give weights to the record with large residual
  - regularization : add penalty term besed on number of parameters in model to avoid overffting

  - ## Unsupervised learning

  - principal component: lienar combination of the predictor vaeibels 查查

  - http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/multivariate/principal-components-and-factor-analysis/what-is-pca/

  - loading : weights that trasnform the predictors into the compoonents

  - ## (Redirected from Ridge regression)

  - lasso set coe === zero , while ridge not

  - least absolute shrinkage and selection operator

```
# bayesian :
- http://uc-r.github.io/naive_bayes
his is primarily because what is usually needed is not a propensity (exact posterior
probability) for each record that is accurate in absolute terms but just a reasonably
```

```
 accurate rank ordering of propensities.


#z Gini index: Mainly used with tree-based methods and commonly referred to as a meas
ure of purity where a small value indicates that a node contains predominantly observ
ations from a single class. Objective: minimize




# A simple way to detect collinearity is to look at the correlation matrix of the pre
dictors. An element of this matrix that is large in absolute value indicates a pair o
f hig


# Leave-one-out cross-validation

- Filter methods are generally used as a preprocessing step. The selection of feature
s is independent of any machine learning algorithms. Instead, features are selected o
n the basis of their scores in various statistical tests for their correlation with t
he outcome variable. The correlation is a subjective term here. For basic guidance, y
ou can refer to the following table for defining correlation co-efficients.

fs1

Pearson’s Correlation: It is used as a measure for quantifying linear dependence betw
een two continuous variables X and Y. Its value varies from -1 to +1. Pearson’s corre
lation is given as:
fs2

LDA: Linear discriminant analysis is used to find a linear combination of features th
at characterizes or separates two or more classes (or levels) of a categorical variab
le.
ANOVA: ANOVA stands for Analysis of variance. It is similar to LDA except for the fac
t that it is operated using one or more categorical independent features and one cont
inuous dependent feature. It provides a statistical test of whether the means of seve
ral groups are equal or not.
Chi-Square: It is a is a statistical test applied to the groups of categorical featur
es to evaluate the likelihood of correlation or association between them using their
frequency distribution.


看看 如何 logistc odds之类的
 sensititivy and specificity
 knn
```

```
# model tranfer
- https://newonlinecourses.science.psu.edu/stat501/node/320/


# Variance Inflation Factors (VIF)


#chi
A chi-squared test, also written as χ2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

# PCA和LDA（linear discriminant analysis）都可以用来减少feature。PCA保留variation最大的feature，LDA保留对于结果最容易进行分类的feature。




My name is JIASHU MIAO and im currently a third year student at ucla double majors in math of computation and staistics. These two majors are a actually a great overlap which contains knowledge of mathematics, statistics and programming and scripting langugaes. The combined study occupies necessary basics for data science and I also get many past experince working as an intern or lab assistant in the fields of data analysis in different indusstris like healthcare, biomedicine, educatoinal it company and financial service. I like data science and never feel that involves so many topics.

RAM



# database optimization
- probr index
- rettrive relevant data
- getting rid of corredlate subs
- avoid coding loops
```