

### exploratory data analysis

types of variables: **Categorical** (nominal no order ex color of eyes or ordinal order ex. lvl of education variables)/  
**Numerical** variables: discrete and continuous variables

numerical summaries: **quantile**: value that proportion  $p$  the data is smaller than  
 $Q(p)$  and  $1-p$  bigger  
first quantile  $Q1: p=0.25$ , **median**  $Q2: p=0.5$  and third quantile:  $p=0.75$   $Q3$ ,  
**IQR** is the interquartile range =  $Q3-Q1$  contains 50% of the data  
Formula for the rank is  $p(n-1)+1$  if not integer extrapolate with 2 values between with weight

measures of center: **MODE**: most frequent value  
**MEDIAN**:  $Q(0.50)$ /  
**MEAN**: average,  $\text{tot}/n$   
if **unimodal** and **symetric** distribution mean=median, right skewed mode < median < mean

variance and sd

Graphics: **pies**,  
**barplots** (frequency or rf, any order, specific categories ex faculties),  
**contingency tables** (2 or + categorical variables),  
**mosaic plot** (translation of CT, if aligned, independent),  
**frequency table** (numerical variable,  $f$ ,  $rf$ =proportion, cumulative  $f$ , cumulative  $rf$ , densities  $rf/\text{amplitude}$ , order),  
**hitograms** (translation of FT, area proportional to class frequency = density, numerical variables, order needed, size can be an interval no precise value as bp),  
**BOXPLOT** (IQR and  $1.5 \cdot \text{IQR}$ , put median, LB, UB),  
**QQ-plot** (compare two distribution theoretical and empirical, if  $45^\circ$  same distribution)

### Statistical inference

simpson paradox: heterogeneous sources: divide to more homogeneous subgroups: ex by major because could bias the proportion :  
**controlling for the confounding factor** men chose the easiest program whereas women chose the more difficult to enter:  
the solution is to use a weighted average of the admission rates

sampling the population: **population**: what we want to analyse, want to find the population's parameters, these are true and fixed values but usually unknown

**sample**: what we have, piece of the population chosen randomly, parameters are random variables, should be as large as possible to limit bias, sample have incomplete information, if finite population without replacement of sample can affect results

point estimation: **estimators** an estimator is a parameter calculated with the sample. it tries to estimate the true parameter of the population it is a random variables and parameter are fixed but unknown within a certain certitude: **confidence intervals**

Estimator: to estimate a parameter and its uncertainty: ex:  $\mu$ , the more sampling, the more precise because variance decreases with  $N$  large concentrated distribution around true value



By Nathaliemayor

Not published yet.

Last updated 11th November, 2018.

Page 1 of 2.

Sponsored by **Readability-Score.com**

Measure your website readability!

<https://readability-score.com>

### Statistical inference (cont)

central limit thm when we sum random variables from the same distribution:  
sum/n= new variable that follow a normal distribution when n is large special case for proportion (binomial)

estimating variance if x follow a normal distribution, follos khi 2 distribution with n-1 degrees of freedom similar to variance estimation  
s<sup>2</sup> and s<sup>~2</sup>,

confidence intervals from central limit thm: C is a certain value for with prob of (1-a) that the estimator is in the interval, small alfa, bigger interval, not exactly 95/100 but around value, prob, if normal distribution use student distribution so modify CI to be more precise,

for proportions :  
 $\hat{p}$  estimate mean

for median

for variance

for the difference of means when 0 is not in the interval: significant difference

theory of estimation depends on situation, can evaluate the quality of estimator, good one has nu bias,



By **Nathaliemayor**

[cheatography.com/nathaliemayor/](https://cheatography.com/nathaliemayor/)

Not published yet.

Last updated 11th November, 2018.

Page 2 of 2.

Sponsored by **Readability-Score.com**

Measure your website readability!

<https://readability-score.com>