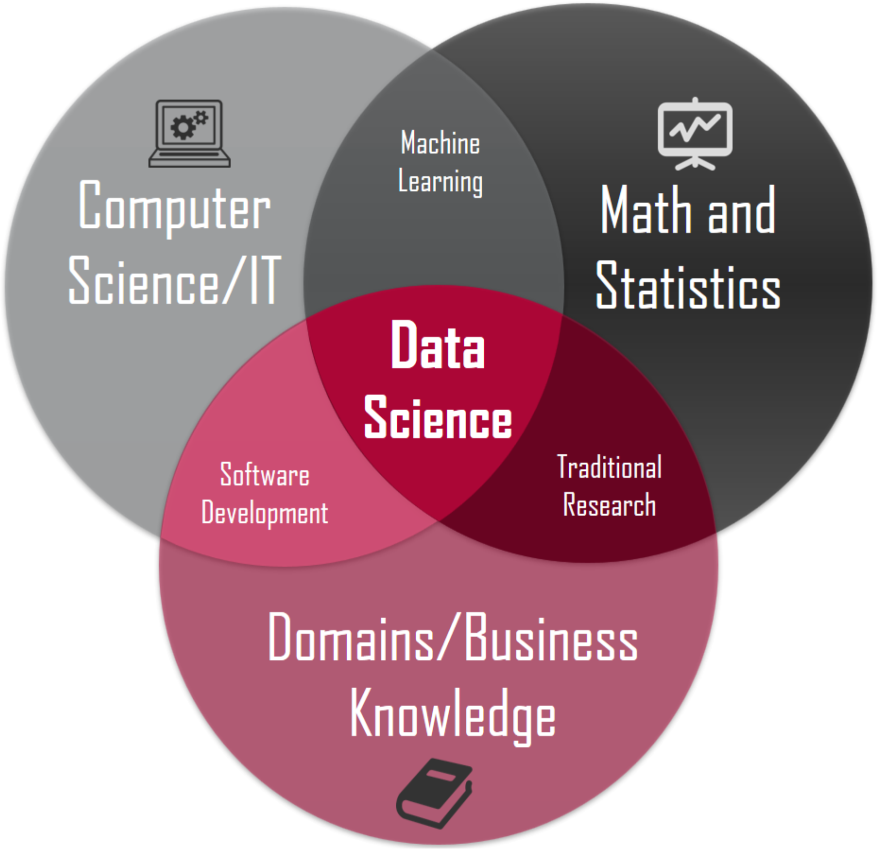
EXPLORATORY DATA ANALYSIS

Sonpvh

RECAP


Computer Science/IT

Machine Learning

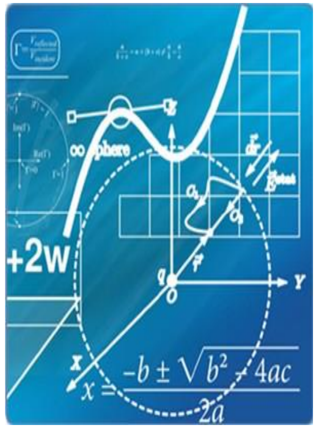
Math and Statistics

Data Science

Software Development

Traditional Research

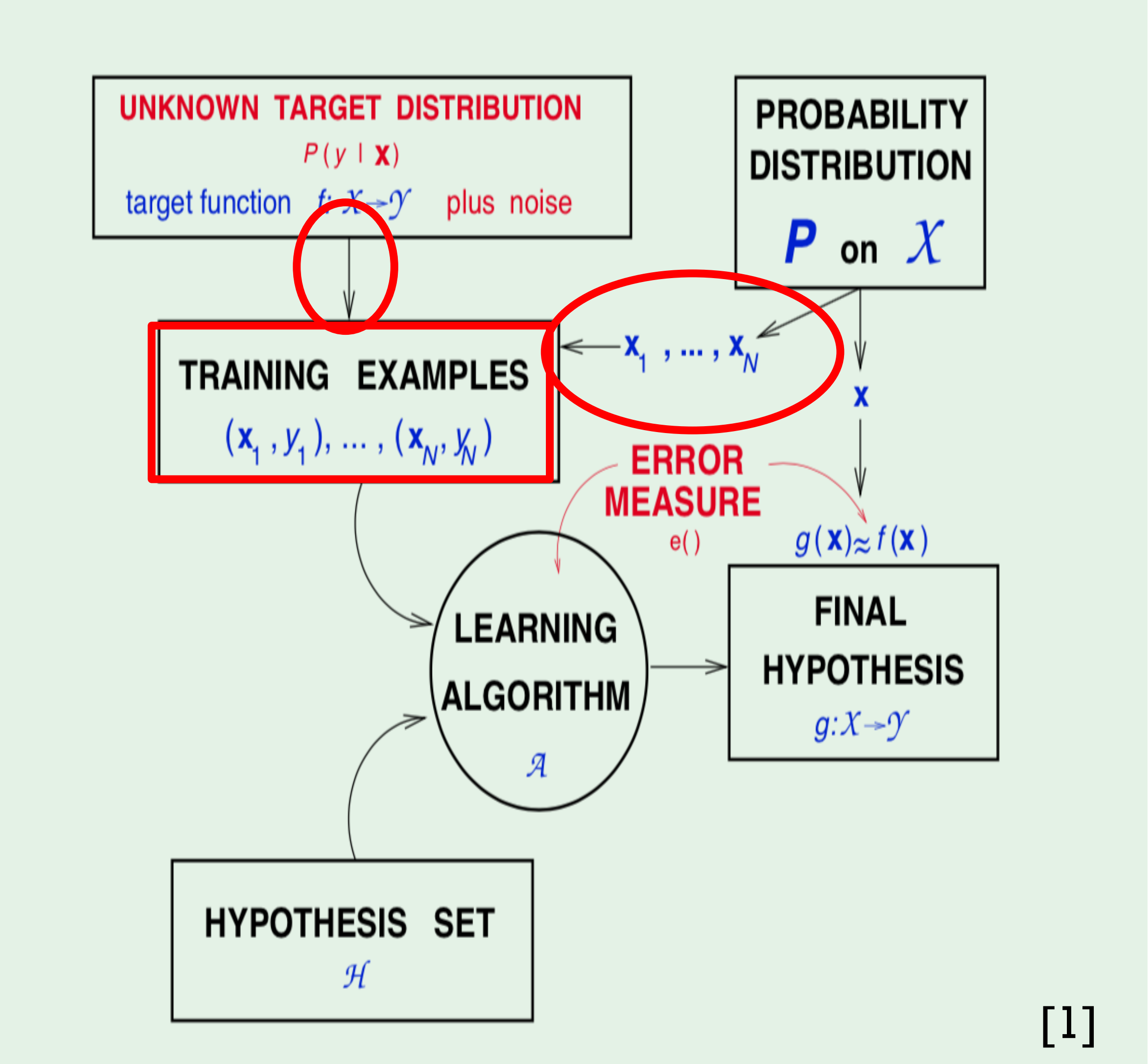
Domains/Business Knowledge

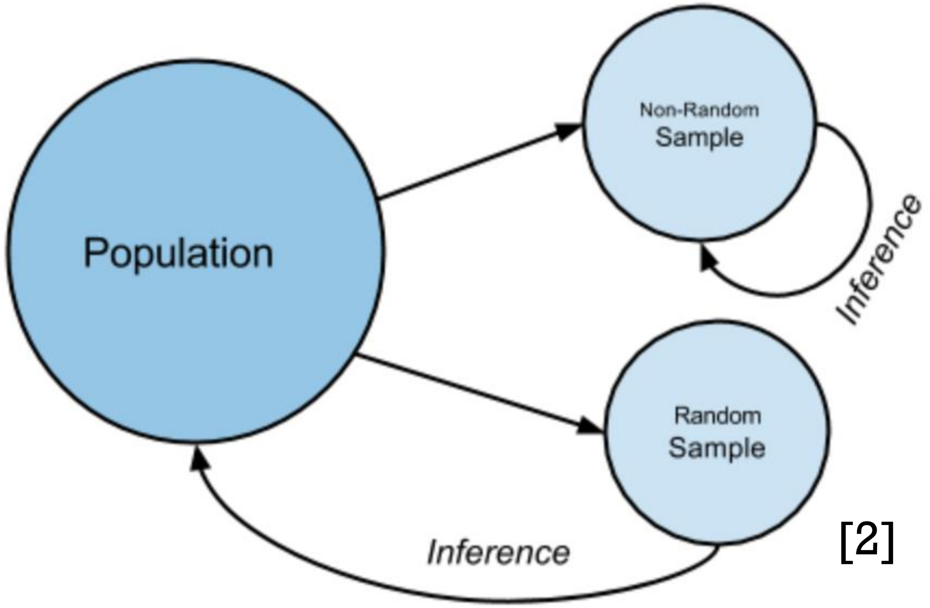


OUTLINE

1. Sampling
2. Exploratory Data Analysis - EDA
3. STAT 101
4. Data Visualization
5. Missing value
6. Outlier
7. Anomaly detection

1. SAMPLING


UNKNOWN TARGET DISTRIBUTION
$P(y \mid \mathbf{x})$
target function $f: \mathcal{X} \to \mathcal{Y}$ plus noise
PROBABILITY DISTRIBUTION
P on $\mathcal{X}$
$\mathbf{x}_1, \dots, \mathbf{x}_N$
$\mathbf{x}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$
ERROR MEASURE
e( )
$g(\mathbf{x}) \approx f(\mathbf{x})$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g: \mathcal{X} \to \mathcal{Y}$
HYPOTHESIS SET
$\mathcal{H}$
[1]



Representation

- **Randomness**: Each member of the larger population has an equal chance of being chosen. [4]
- **Large enough:** Depends on the precise degree of confidence required for making an inference

1. SAMPLING

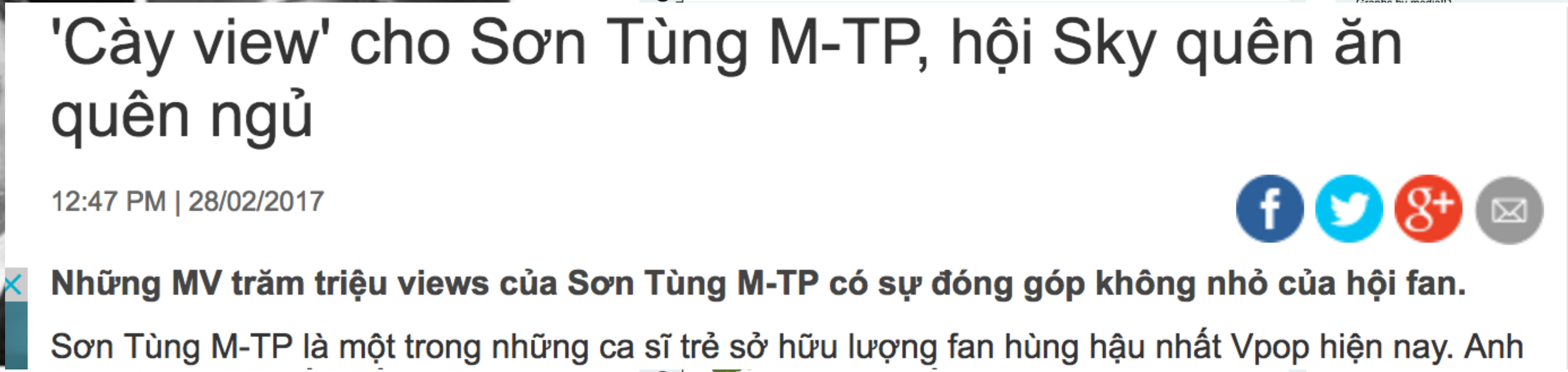
Type of Sources:

- Primary Sources: Collect by yourself for a specific purpose
- Secondary Sources: Collect by someone else, some other purpose. VHLSS, PAPI, SME …

Problems of Sampling:

- Biased sampling
- Noise
- Missing data, errors logs, ...
- …

1. SAMPLING



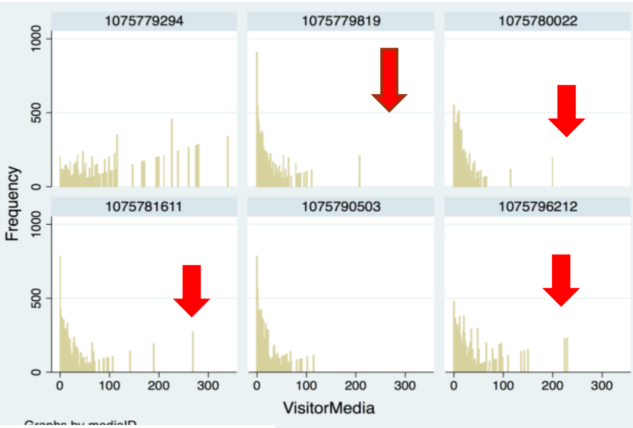
'Cày view' cho Sơn Tùng M-TP, hội Sky quên ăn quên ngủ

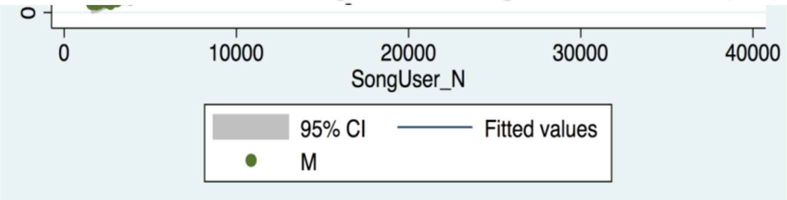
12:47 PM | 28/02/2017

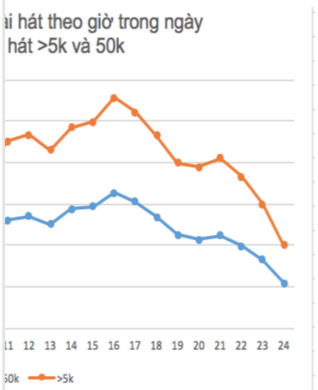
Những MV trăm triệu views của Sơn Tùng M-TP có sự đóng góp không nhỏ của hội fan.

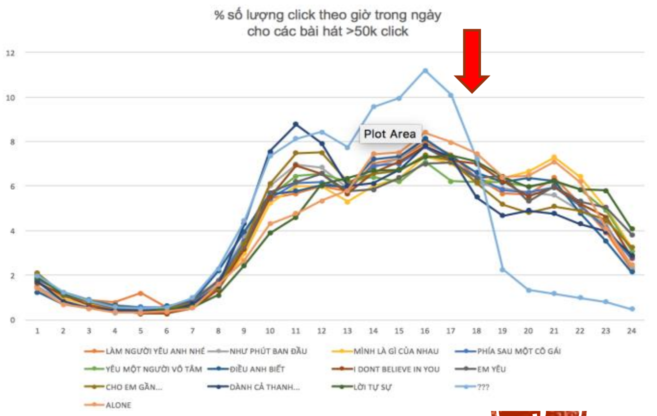
Sơn Tùng M-TP là một trong những ca sĩ trẻ sở hữu lượng fan hùng hậu nhất Vpop hiện nay. Anh

1936 - Franklin D. Roosevelt vs Alf Landon [3]

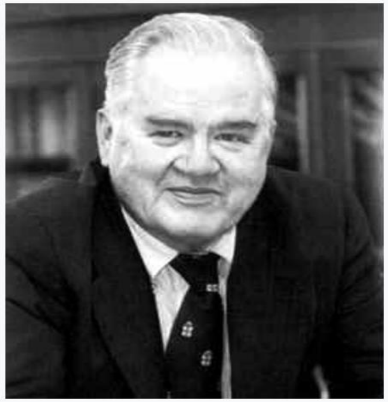


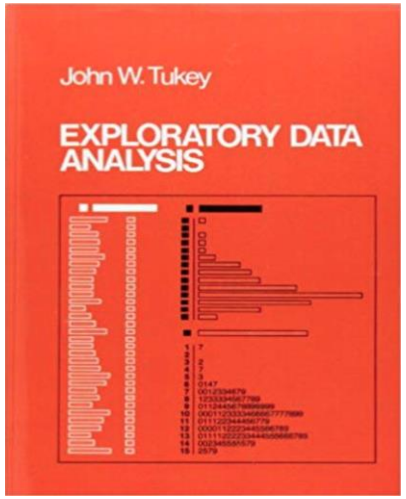






2. EXPLORATORY DATA ANALYSIS – EDA





- “**Too much emphasis in** statistics was placed on **statistical hypothesis testing**…, more emphasis **needed to be placed on using data to suggest hypotheses** to test” Turkey - 1977 [6]

- "**Procedures** for analyzing data, **techniques for interpreting** the results of such procedures, ways of **planning the gathering of data** to make its **analysis easier, more precise or more accurate**, and all the machinery and results of (mathematical) statistics which apply to analyzing data." Turkey - 1961 [5]

- The idea of EDA encouraged the development of statistical computing: S, S-PLUS, R.

- Appling to data science and big data analysis

2. EDA VS HYPOTHESIS TEST

- **Traditional hypothesis testing** designed to verify a **priori hypotheses** about relations between variables

- **Exploratory Data Analysis (EDA)** is used to **identify systematic relations** between variables when there are **no a priori expectations** as to the nature of those relations.

[8]

- From **Business-Driven** to **Data-Driven**

2. EDA – "UNDERSTANDING ABOUT DATA"

1. Uncover underlying structure
2. Detect outliers and anomalies, missing, mistakes
3. Maximize insight into a data set
4. Extract important variables
5. Determine optimal factor settings
6. Test underlying assumptions
7. Develop parsimonious models

[7]

2. EDA - TECHNIQUES

1. **Data quantitative measurements**
   - Univariable
   - Mutilvariable
2. **Data visualization**

3. STAT 101: VARIABLES AND TYPE

1. Qualitative (category)
    1. **Binary** – where there are two choices, e.g. Male and Female;
    2. **Ordinal** – where the names imply levels with hierarchy or order of preference, e.g. level of education
    3. **Nominal** – where no hierarchy is implied, e.g. political party affiliation.
2. Quantitative
    1. **Discrete** (number of students in class)
    2. **Continuous** (amount of milk in a gallon)

3. STAT 101: PLOT

1. **Graphs for a Categorical Variable**
   1. Pie Chart: percentile
   2. Bar Chart: many categories
   3. …

1. **Graphs for a Single Quantitative Variable**
   1. Dot Plot
   2. Frequency Histogram and Relative Frequency Histogram
   3. Stem-and-Leaf Diagram
   4. Time Plot
   5. Boxplot or Box-and-Whisker Plot
   6. …

3. STAT 101: CENTRAL TENDENCY [9]

1. **Measures of Central Tendency**
   1. Mean : not resistant
   2. Median
   3. Mode
   4. Trimmed Mean: (solve outlier)
      - Care about mistakenly recorded

3. STAT 101: CENTRAL TENDENCY [9]

Series: 95, 78, 69, 91, 82, 76, 76, 86, 88, 80

- Mean = 82.1
- Median = 81
- Trimmed Mean:
- (69), 76, 76, 78, 80, 82, 86, 88, 91, (95)

  The 10% trimmed mean = 82.13
- How about: 950, 78, 69, 91, 82, 76, 76, 86, 88, 80

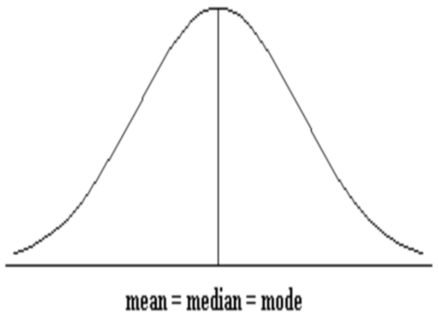
Error series: 95, 78, 69, 9, 82, 76, 76, 86, 88, 80

- Mean = 73.9
- Median =79
- (9), 69, 76, 76, 78, 80, 82, 86, 88, (95)

  The 10% trimmed mean = 79. 38

3. STAT 101: SKEWNESS [9]

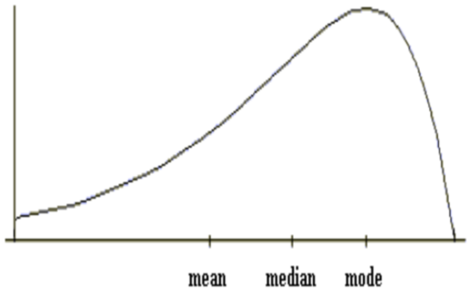
2. Skewness



The above distribution is symmetric.

2. Skewed Left

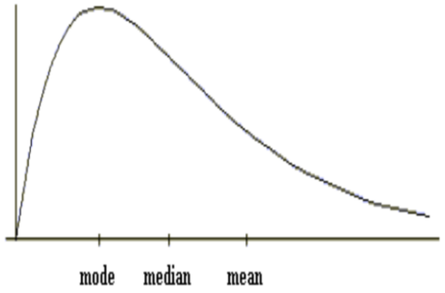
Mean to the left of the median, long tail on the left.



The above distribution is skewed to the left.

3. Skewed Right

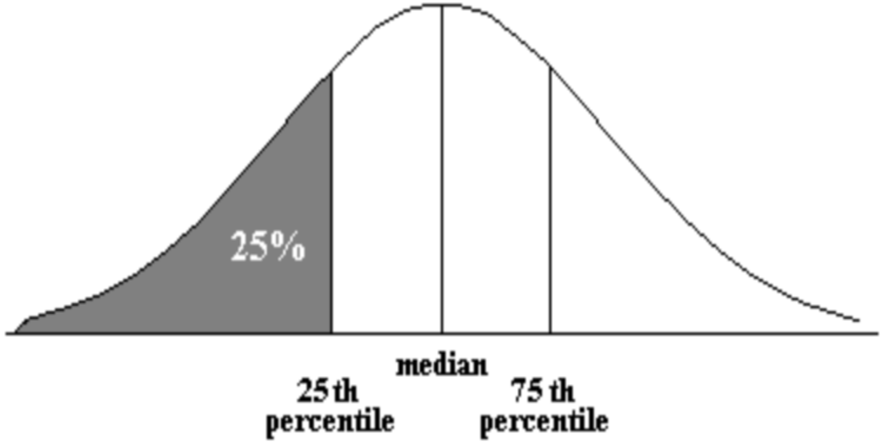
Mean to the right of the median, long tail on the right.



The above distribution is skewed to the right.

3. STAT 101: MEASURES OF VARIABILITY [9]

1. Range **(affected by extreme values)**
2. Interquartile Range (IQR): Q_3 - Q_1 (don’t affected by extreme values)



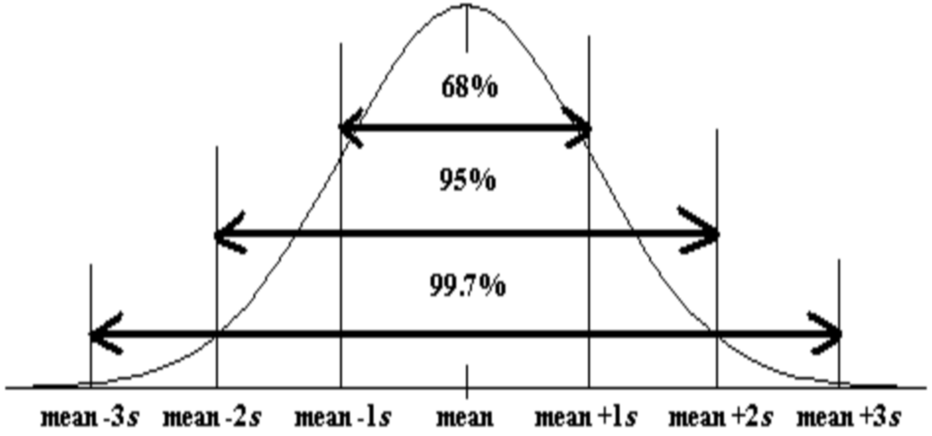
3. STAT 101: MEASURES OF VARIABILITY [9]

3. Variance and Standard Deviation

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1} \longrightarrow \sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{N}$$

Sample

Population



- Add constant => sd not change, multi constant => sd * constant
- Why sample variance divide n-1 [10]
- Range ≈ $4s$

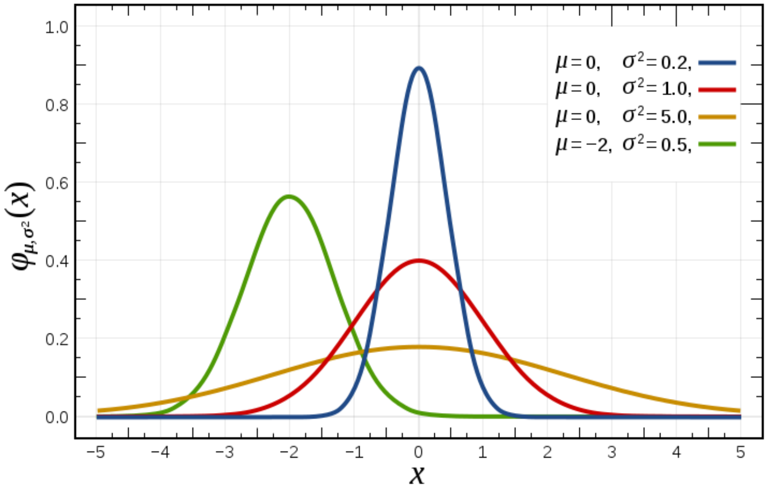
  Approximate value of $s \approx \frac{range}{4}$

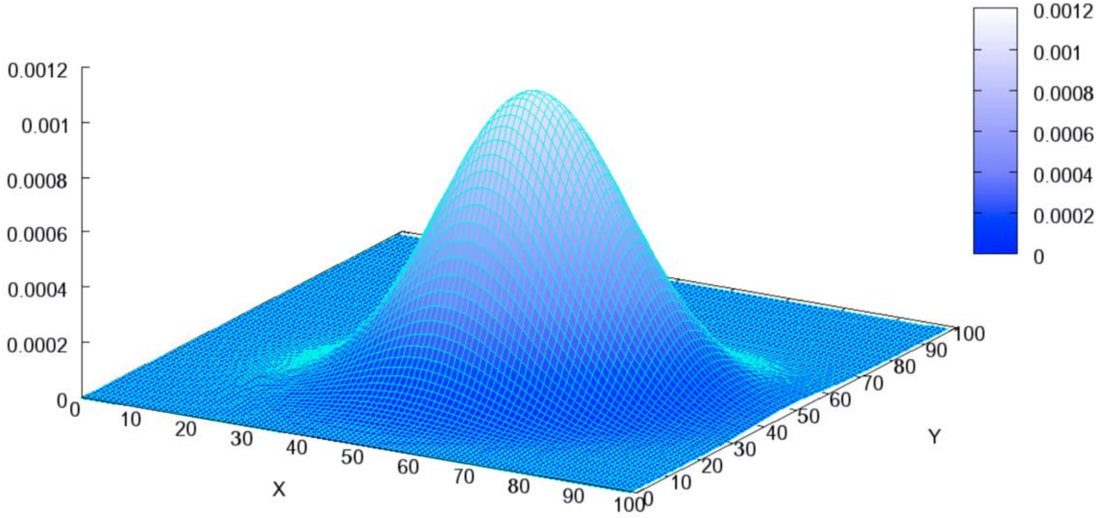
3. STAT 101: MEASURES OF VARIABILITY [9]

4. Coefficient of Variation:
   - CV = Standard Deviation / Mean
   - Compare dispersion from 2 or more distinct population
5. Zscore
   - Z = (observed value – mean) / SD

3. STAT 101: PROBABILITY DISTRIBUTION

1. Continues variable - Normal distribution – Multivariable Normal Distribution

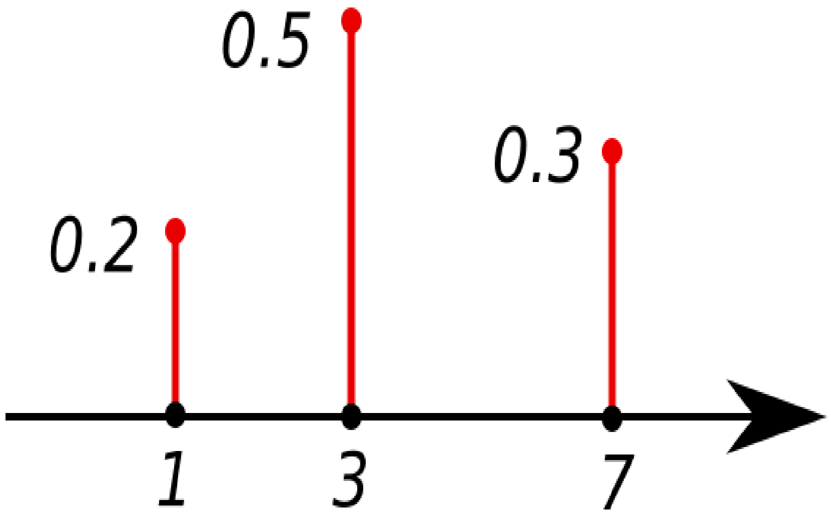


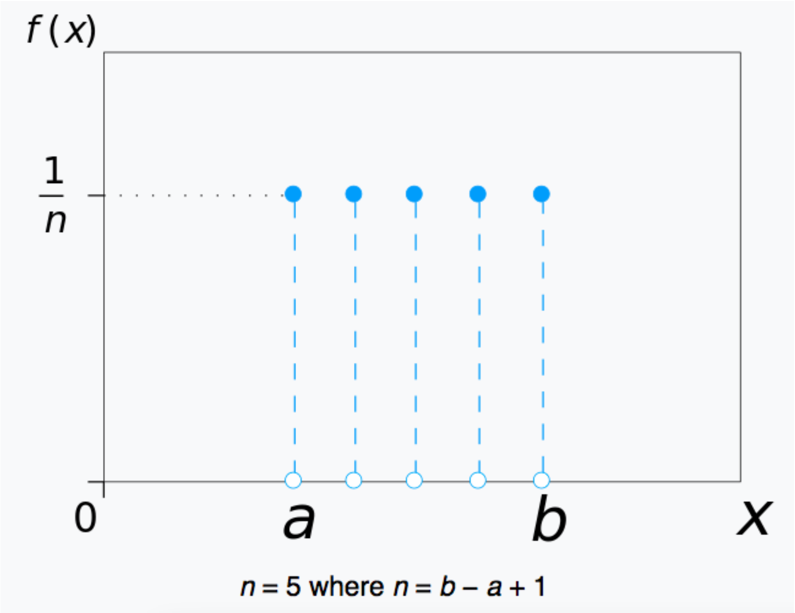


Carl Friedrich Gauss
(1777 – 1855)

3. STAT 101: PROBABILITY DISTRIBUTION

Discrete distribution – Multinominal distribution



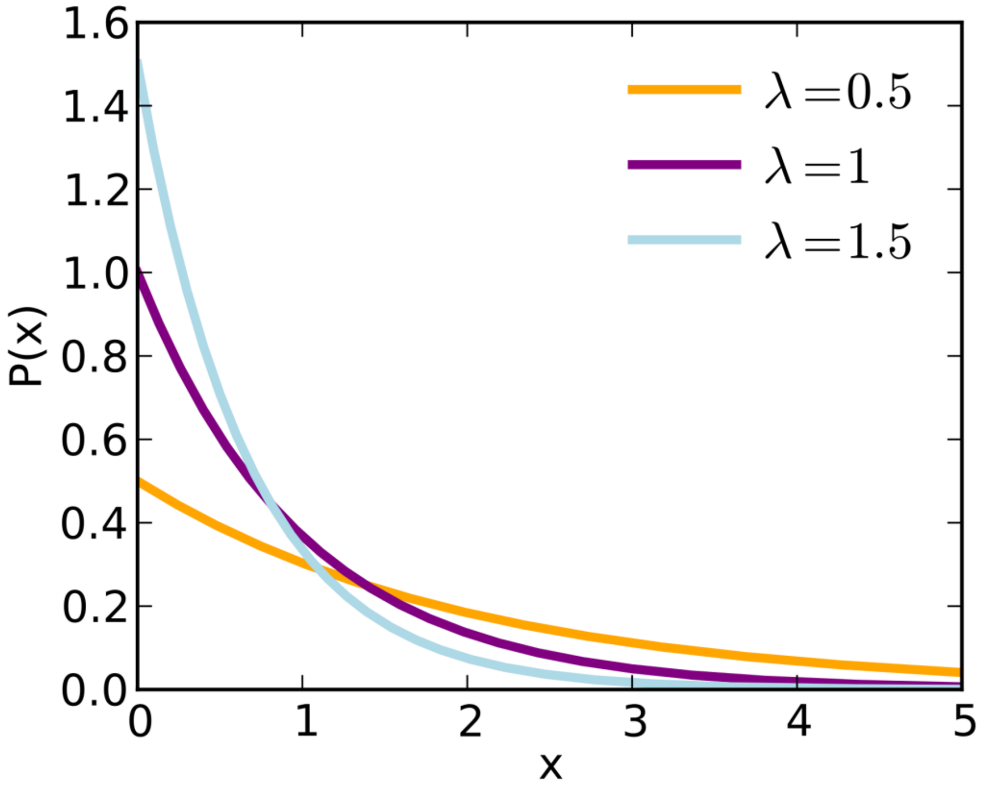


Probability mass function

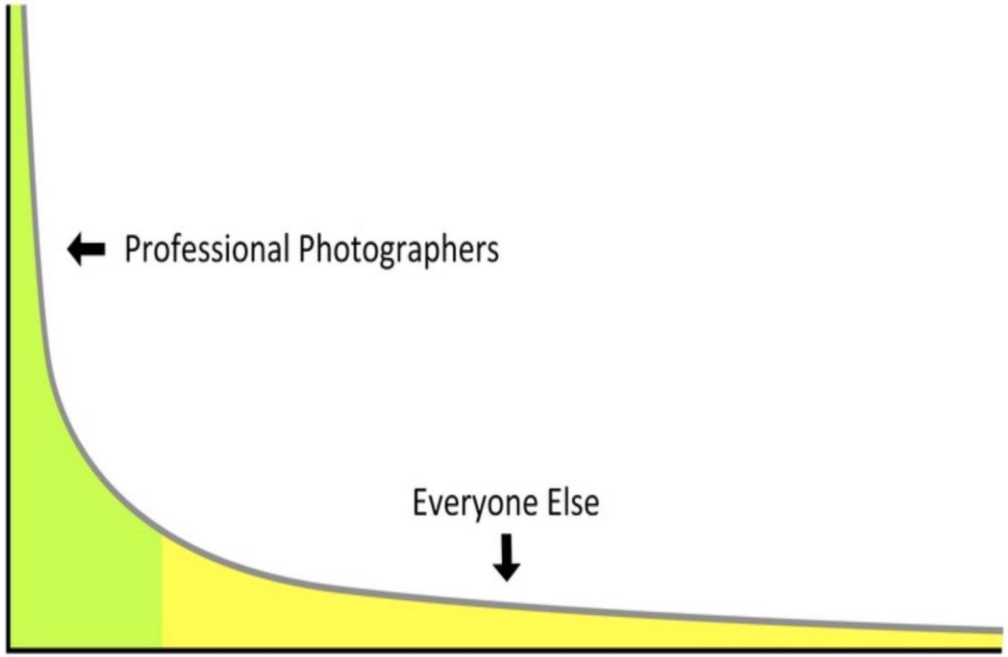
3. STAT 101: PROBABILITY DISTRIBUTION

3. Exponential Family

Probability density function $P(x) = \lambda e^{-\lambda x}$

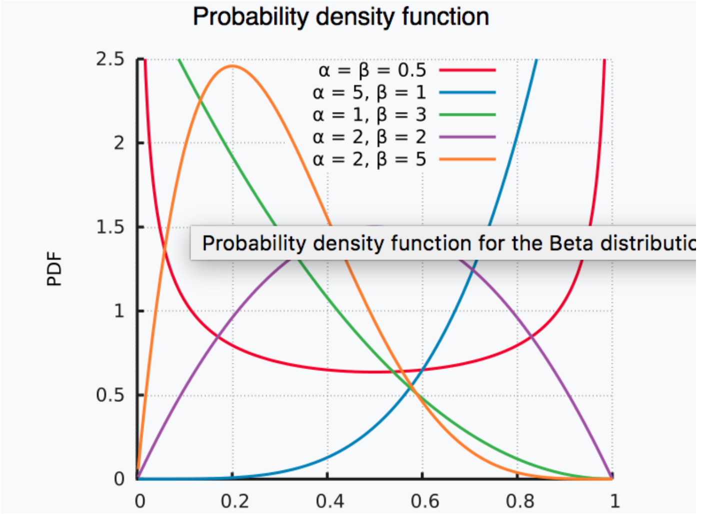


LONG TAIL

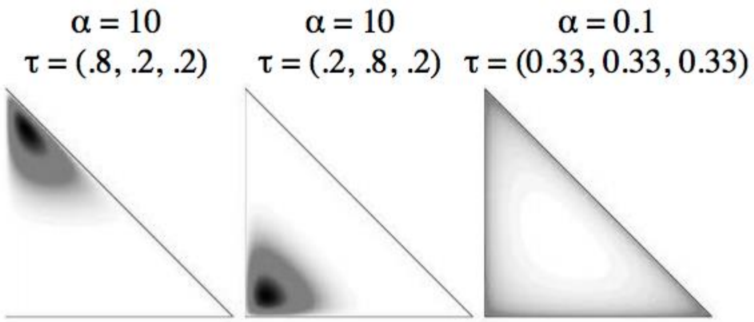


3. STAT 101: PROBABILITY DISTRIBUTION

2. Binary variable - Beta distribution – Dirichlet distribution



$$f(x;\alpha,\beta) = \text{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1}$$



$$f(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

Peter Gustav Lejeune Dirichlet (1777 – 1855)

3. STAT 101: MULTIVARIABLE - CORRELATION

1. Type of variable
   1. Non-category – Non-category
   2. Non-category – category
   3. Category – category
2. Analysis
   1. Pearson correlation
   2. Ttest
   3. ANOVA
   4. Regression analysis
   5. …

4. DATA VISUALIZATION

- Show in R

- http://jkunst.com/highcharter/hchart.html
- https://www.r-graph-gallery.com/ridgeline-plot/
- https://www.youtube.com/watch?v=e2w-kOVHNQ4

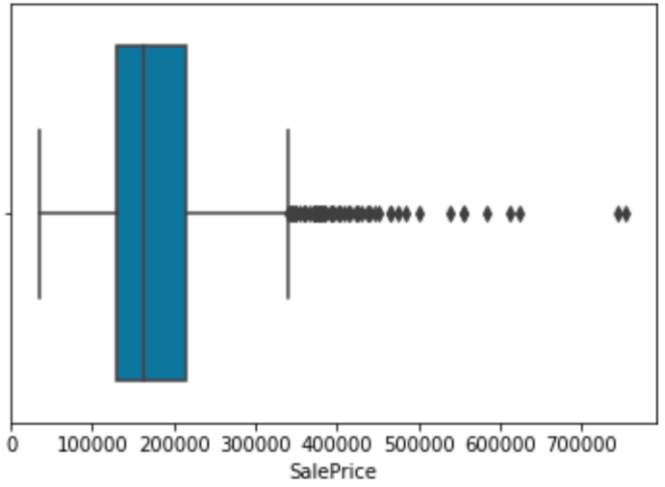
5. MISSING VALUE [11]

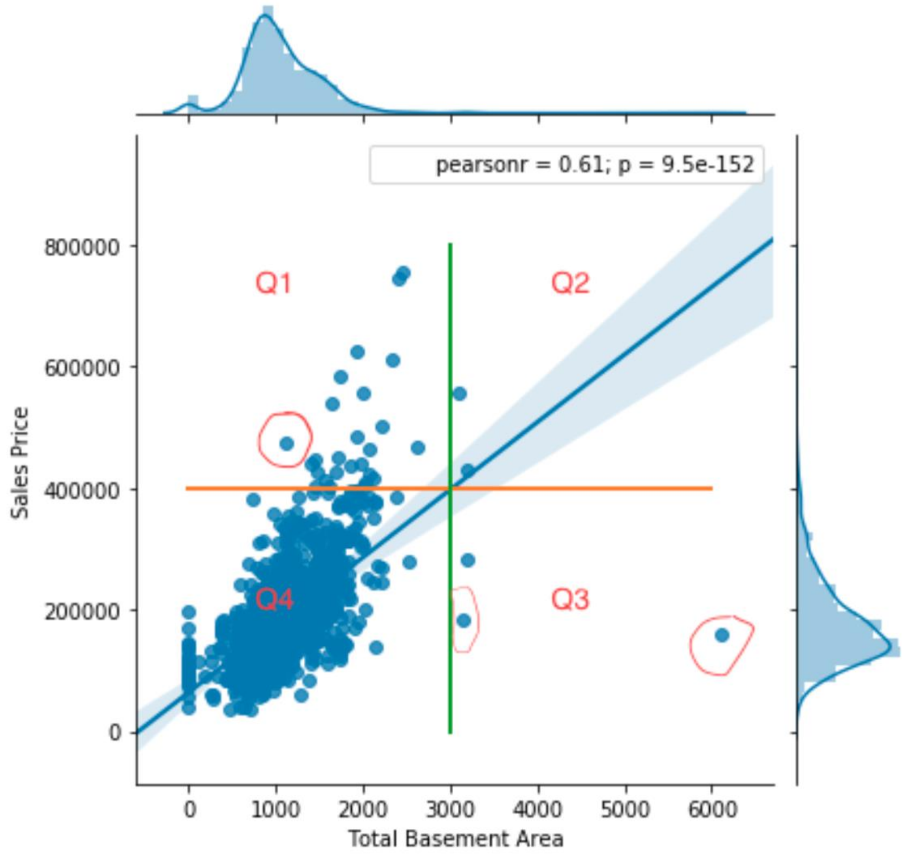
- Ignore missing value
- Back-fill or forward-fill
- Replace with mean/median/mode/cluster mean …
- Assigning An Unique Category
- Predict the missing value
- …

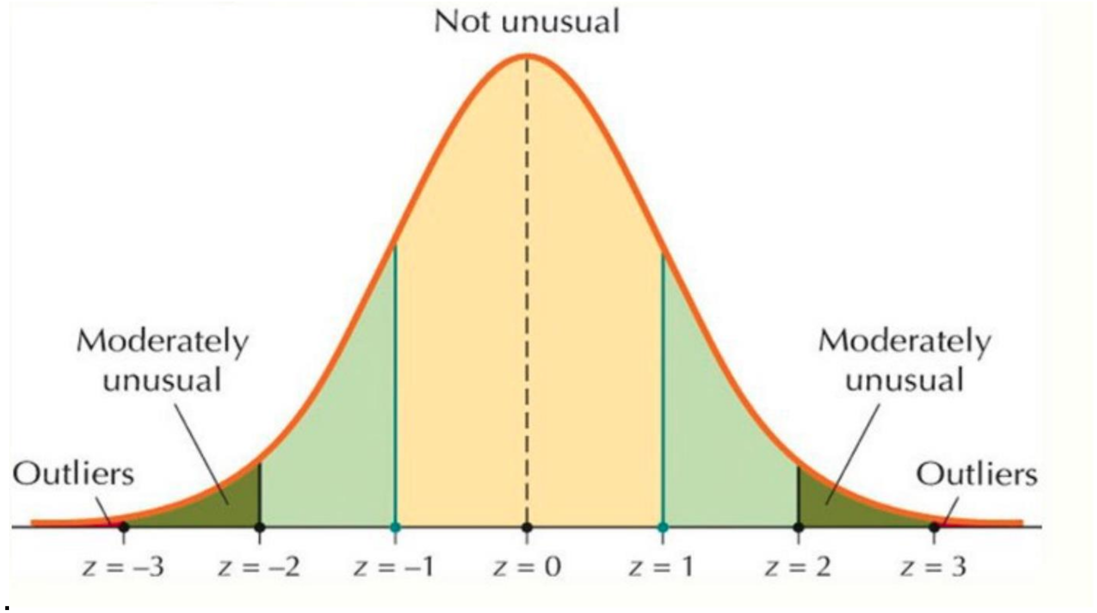
6. OUTLIER

TYPES:

1. Univariate Outlier
2. Multivariate Outlier

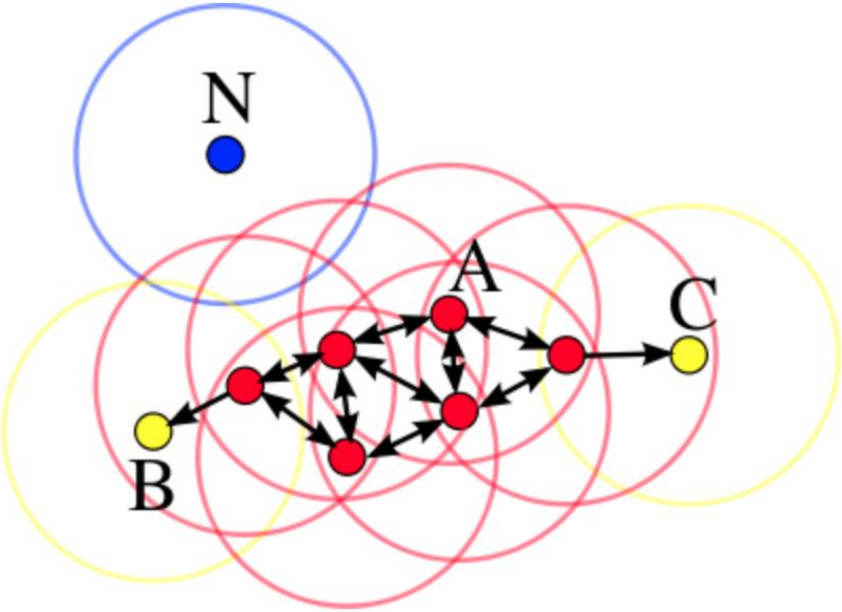




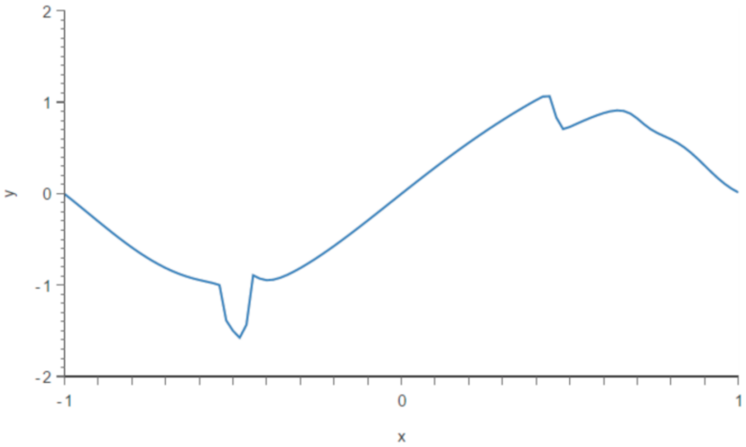


[12,13]

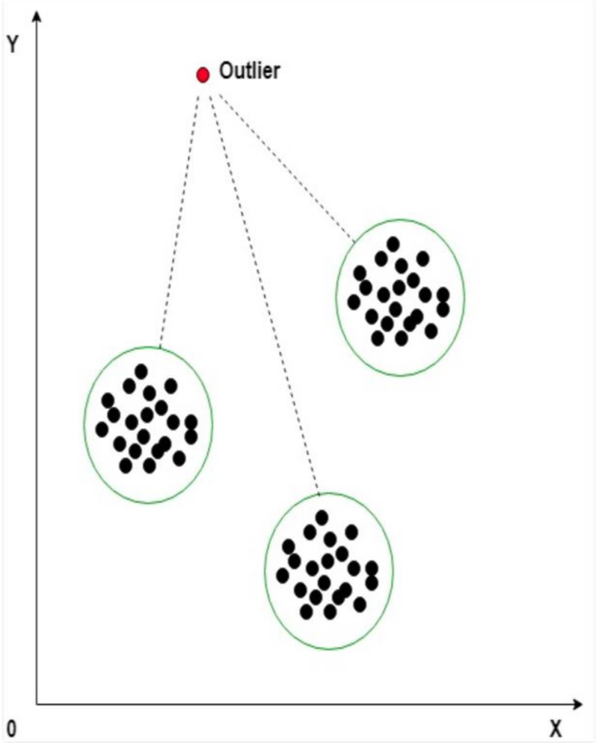
6. OUTLIER



DBScan



Minkowski error



KMean

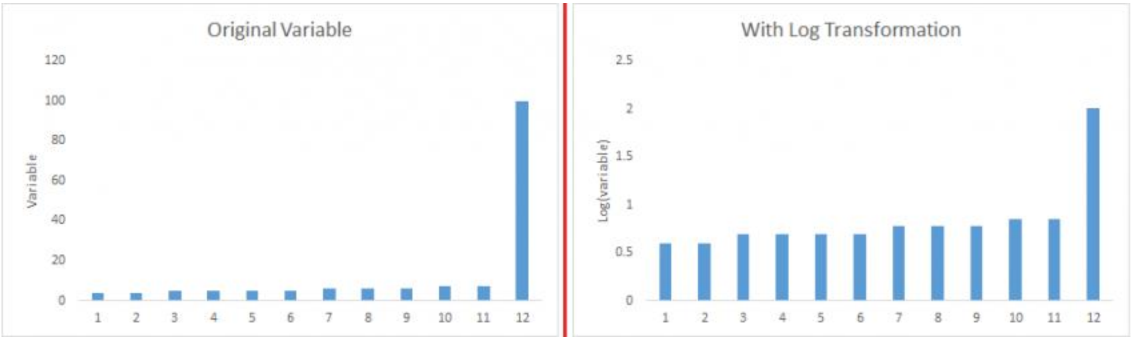
6. OUTLIER

CAUSE

1. Data entry errors (human errors)
2. Measurement errors (instrument errors)
3. Experimental errors
4. Intentional
5. Data processing errors
6. Sampling errors
7. Natural
8. **…**

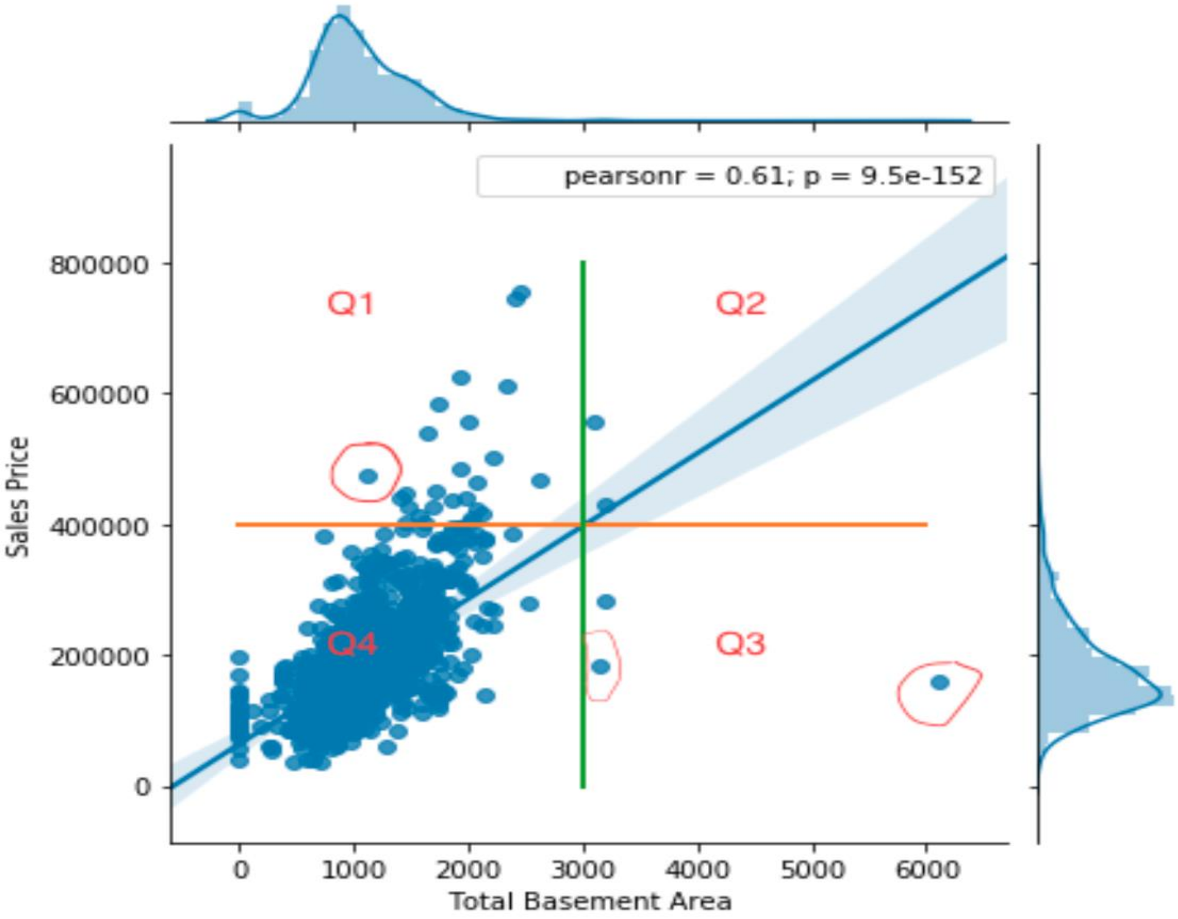
6. OUTLIER

1. Transforming and binning values
2. Deleting observations:
3. Imputing: max, min …
4. Treat Outliers separately
5. Detect error from systems
6. …


Original Variable

With Log Transformation

7. ANOMALY DETECTION



ace

REFERENCES:

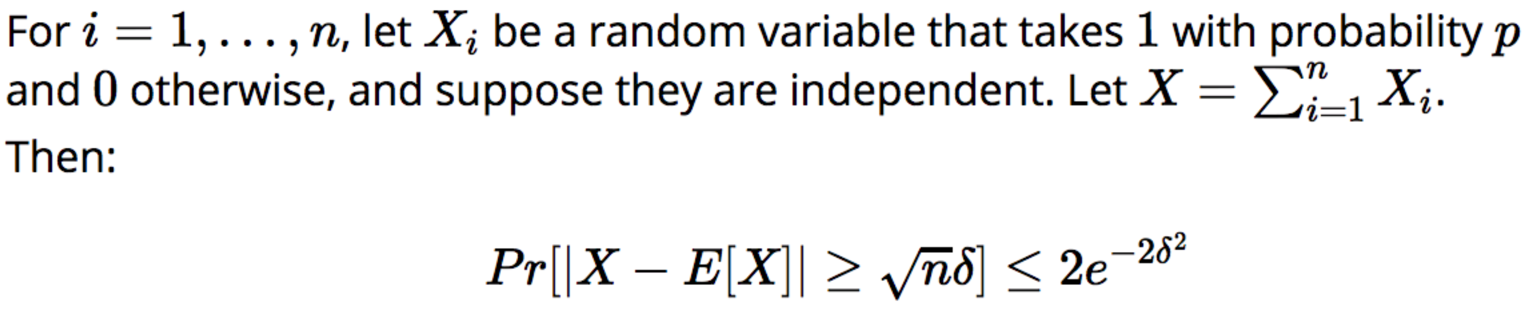
1. Learning from data – Yaser S. Abu, Malik Madon-Ismal, Hsan Tien Lin – 2012
2. Applied statistics course – Penstate University – STAT 500 - https://newonlinecourses.science.psu.edu/stat500/node/111/
3. https://www.thesociologicalcinema.com/videos/biased-sampling-in-predicting-a-presidential-election
4. *Definition taken from Valerie J. Easton and John H. McColl's Statistics Glossary v1.1*
5. The Future of Data Analysis – John Tukey – 1961
6. Exploratory Data Analysis – John Tukey – 1977
7. https://www.itl.nist.gov/div898/handbook/ - chapter 1. Explore
8. http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Common/DataMining/ExploratoryDataAnalysisEDAandDataMiningTechniques
9. https://newonlinecourses.science.psu.edu/stat500/node/12/
10. http://www.visiondummy.com/2014/03/divide-variance-n-1/

REFERENCES:

11. https://www.analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/
12. https://medium.com/datadriveninvestor/unboxing-outliers-in-machine-learning-d43fe40d88a6
13. https://www.kdnuggets.com/2018/08/make-machine-learning-models-robust-outliers.html

LESSON 1.2: COLLECTING THE DATA

Margin of Error: How many sample we need ask?



For $i = 1, \ldots, n$, let X_i be a random variable that takes 1 with probability p and 0 otherwise, and suppose they are independent. Let $X = \sum_{i=1}^{n} X_i$. Then:

$$Pr[|X - E[X]| \geq \sqrt{n}\delta] \leq 2e^{-2\delta^2}$$

Chernoff-Hoeffding bound