17/02/2021

MA2040 : Probability, Stochastic Process & Statistics.

* INSTRUCTORS : Prof. S. H. Kulkarni
  shk@iitpkd.ac.in
  Prof. Ashok Kumar
  ashokm@iitpkd.ac.in

* TAs : Abhinav Thakur
  142002002@smail.iitpkd.ac.in
  Shreee Ganesha Sharma M S
  142002015@smail.iitpkd.ac.in

* SLOTS : G1 Wed 1130-1230 hrs
  T4 Thu 1400-1530 hrs

- Evaluation: 1 assignment (10 Marks)
  4 Quizzes (60 Marks)
  1 endsem exam (30 Marks)

⚠ Quizzes ~ 8 Mar, 26 Mar, 19 Apr, 7 May (0900-0945 hrs)

* Textbook : Introduction to Probability & Statistics for Engineers & Scientists
  – Sheldon. M. Ross
  5th ed

- Probability, Random Variables & Stochastic Processes
  – Athanasios Papoulis & S. Unnikrishna Pillai

⚠ Take own notes.

⚠ Discuss actively in the forum.
Freely ask doubts, keep your shyness (& ego) aside :)

- Basic steps in Statistics

a) Data Collection (& cleaning)
Outlier removal, standardization, ...

b) Representing the data
Visualization – plot, histogram, ...

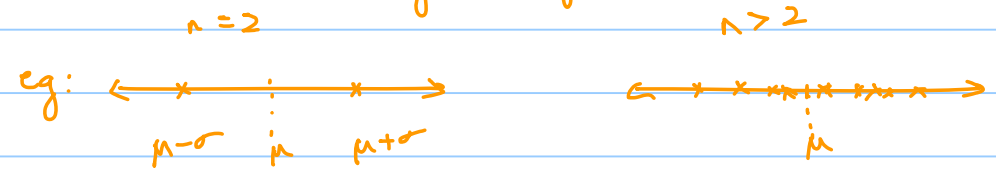
c) Summarize the data
Central tendency, dispersion,
* Skewness, * Kurtosis

d) Conclusion

⚠ Given mean, $\mu = 5.4$, std deviation, $\sigma = 1.2$ } for a dataset X,

→ what can you say about the distribution of the data (distro)

→ Significance of sample size n of X in making inference about X



17/02/2021 ↑ ↓ 18/02/2021

LECTURE 2

a) DATA COLLECTION

- eg: Clinical testing of efficacy of a vaccine.

Design of experiments } N individuals/volunteers.

Treatment group (expt. treatment) | Control group (Placebo)

* Some objectives.

Neither doctors, nor volunteers know placebo/vaccine.

N → N/2 treatment group — admin the vaccine
N → N/2 control group.

After incubation period, check on "patients".

Expectation:

- More people in control group are infected.

⚠ •

CONFUSION MATRIX

Test \ Disease	present	absent
detected	TP	FP
not detected	FN	TN

https://en.wikipedia.org/wiki/False_positives_and_false_negatives

⚠ What's a good N? 100 / 1000 / 10000 ...
to make good decision.
Will find out in due course.

- Minimize N
- Still have good testing accuracy, efficiency, & confidence in inference.

F2 score etc.

5) REPRESENTING / DESCRIBING THE DATA

- MATLAB (optional, useful for assignment).

++ Can use MATLAB online using institute mail-id @

matlab.mathworks.com.

- Setup: 30 students choose an integer b/w 1 & 10. (frequently used). denoted by X.

$$X = [1, 2, 2, \dots, 10]$$

a) Frequency table

discrete data, $x_i \in \{1, 2, \dots, 10\}$

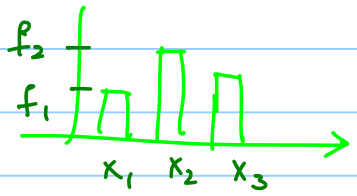
# written, x_i	1	2	3	4	5	6	7	8	9	10	Total
# of students, f_i	1	3	3	2	1	4	9	3	3	1	30 (N)

frequency.

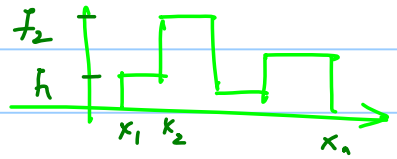
Plot

https://en.wikipedia.org/wiki/Histogram

b) Bar chart
- for discrete observations



c) Histogram.
- Usually for grouped/continuous intervals.



- Normalized / relative frequency = proportion of students
- $f_{i(rel)} = f_i / \text{length}(X)$
- $f_i(rel) \in [0, 1] \ \forall \ i$
- area under this curve is 1 ⟶ Similar to a probability mass / density function.

- relative frequency is analogous to probability.

d) - pie-chart is a helpful tool / appropriate for this.
  - ↳ area of sector } ∝ frequency of observation.
    angle of sector }

e) * Stem and leaf plot.

↳ Helps keep record of the data in a concise manner.

Say all observations are in lakhs or in fractions.

eg1.

phone #s, pincodes, ...

9886784591, 9901425926, ...

9886907843, 9482569380, ...

Stem	Leaf
990	1425926
988	6784591, 6907843
948	2569380

airtel
vodafone
BSNL

first 3 digits of phone #

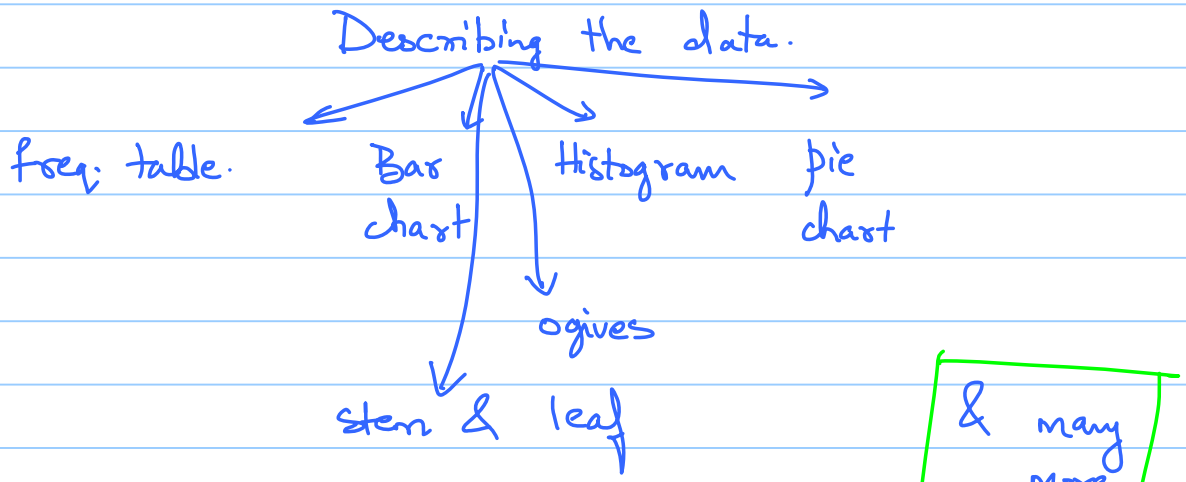
rest of the observation.

eg2.

0.0011, 0.0025, 0.0019,
0.0042, 0.0069, ...

Stem	leaf
0.001	1, 9
0.002	5
0.004	2
0.006	9

++ Stem & leaf plot helps us group/form clusters based on the stem.



18/02/2021

Textbook: Chapter 1
Chapter 2 till section 2.2 page 14.

24/02/2021

LECTURE 3

Summarizing the data.

++ Inference @ a glance.
++ Helps compare datasets.

Some measures - mean, median, mode, sd, variance, ...

a) Measures of Central tendency.
↳ The value (need not be an observation in our dataset)
  ↳ around which the dataset is spread.
  ↳ One value that "represents" the data.

given $\{x_i\}_{i \in I_n}$ // $I_n := \{1, 2, 3, \ldots, n\}$

Sample Mean.

$$\bar{X} = \frac{\sum_i x_i}{n}$$

- arithmetic mean
- ++ mathematically treatable
- ++ simple to use, intuitive
- -- Susceptible to changes with extreme values.
- for frequency table, $\{(x_i : f_i)\}_{i=1}^{\ell}$ (x_i appears f_i times)

$$\bar{X} = \frac{\sum_i x_i}{n} = \frac{x_1 f_1 + x_2 f_2 + \cdots}{f_1 + f_2 + \cdots}$$

$$\bar{X} = \frac{\sum_{i}^{\ell} f_i x_i}{\sum_i f_i}$$

There are ℓ distinct observations

* $$\bar{X} = \sum_{i=1}^{\ell} x_i \left(\frac{f_i}{\sum f_i} \right)$$ → relative frequency as weight.
→ = EMPIRICAL probability/ proportion of times x_i appears in the data set.

* as $n \to \infty$ $f_i/n \longrightarrow$ (theoretical) probability

Other means: averages

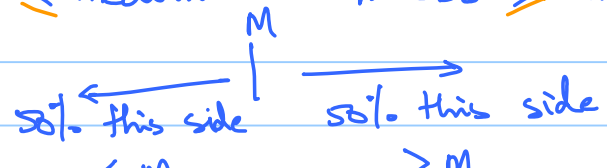
a) Geometric mean, $\bar{X} := \left(\prod_{i=1}^{\ell} x_i^{f_i} \right)^{1/n}$
b) Harmonic mean, $\bar{X} := \left(\frac{\sum x_i^{-1}}{n} \right)^{-1}$
c) ⋮

Sample median

→ A locational / positional measure.

→ divides the dataset into 2 halves of equal # of observations.

→ # obs ≤ median = # obs ≥ median.



ROBUST

→ Not unique ++ less sensitive to outliers.

Procedures to find Median

(I) • sort & find the "middle" element.

↳ given $x_1, x_2, \ldots, x_n$ arrange as

$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ ( ) denotes indices may get permuted & we're using order statistics.

↳ if n is odd:

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

else if n is even:

$$\text{median} = \left[\frac{\left(\frac{n}{2}\right)^{th} \text{obs} + \left(\frac{n}{2}+1\right)^{th} \text{obs}}{2}\right]$$

→ // need not be in $\{x_i\}$.

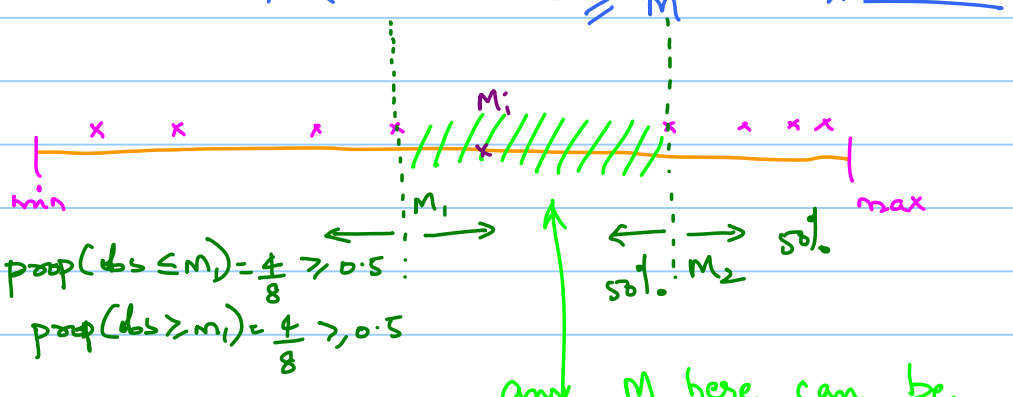
→ // Need not be unique.

Definition

M is 'a' median of $\{x_i\}_i$ if

a) proportion of observations ≤ M is atleast 50 %

b) proportion of observations ≥ M is atleast 50 %



⚠ Image proc, salt & pepper noise.

⚠ Median income, salary than mean.

LECTURE 4

Percentiles → Another positional measure.

Let $0 \le p \le 1$

then $100p^{th}$ percentile is a value m_p, such that

$$\frac{\#\{obs \le m_p\}}{n} \ge p \quad \&$$

$$\frac{\#\{obs \ge m_p\}}{n} \ge 1-p$$

- $m_{1/2}$ (m @ $p=0.5$) is the median.
- $m_{1/4}, m_{2/4}, m_{3/4}$ are called **Quartiles**. I, II & III quartile or Q_1, Q_2, Q_3 respectively.
- they are equivalent to $25^{th}, 50^{th}$ & 75^{th} percentiles resp.

Mode
- observation ($\in \{x_i\}$) with the largest frequency.

Empirical relation
- $3\text{Median} \simeq 2\text{Mean} + 1\text{Mode}$ (practically, for many datasets)

- Property of mean,

given $\{x_i\}_{i \in I_n}$, $\bar{x}$

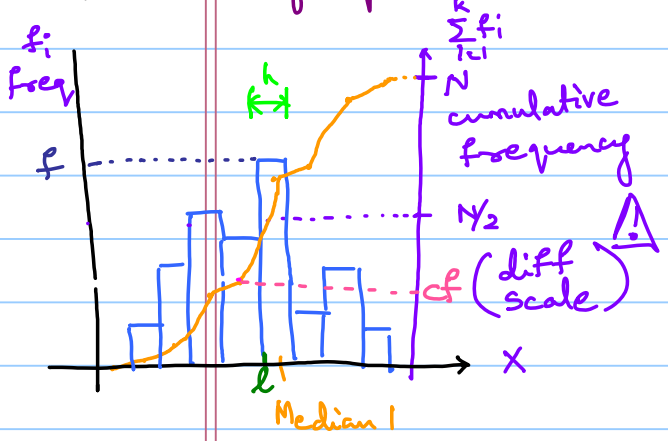
$$x_i \longmapsto ax_i + b = y_i$$

$$\Downarrow$$

$$* \quad \bar{x} \longmapsto a\bar{x} + b = \bar{y}$$

- change in origin & scale affects the mean.
- linear if $b = 0$

Median for grouped data.



Interpolated median

$$M := l + \frac{h}{f}\left(\frac{N}{2} - cf\right)$$

- l: lower limit of median class
- f: freq of median class
- h: width of median class
- cf: cum freq previous to median class

https://en.wikipedia.org/wiki/Median

b) Measures of deviation / spread / variability (of the data from a point X'. If $X' = \text{mean}$, we call it central moments.).
$X' = \bar{X}$

(i) Root Mean Square deviation.

→ root (mean (square (deviation))))

$x_i - x'$

$(x_i - x')^2$

$\frac{1}{n}\sum_i (x_i - x')^2$

$\text{rms}(x) := \sqrt{\frac{1}{n}\sum_i (x_i - x')^2}$ ← this is a function of x'

⚠ check for
$x' = 0$
$x' = \bar{x}$ → minimizes rms (x)

Doesn't have the units of X

(ii) Variance.

$$V(x) := \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 =: \text{Var}(x) \text{ , } \sigma_x^2 \text{ etc}$$

(iii) Standard deviation := positive square root of V(x)

$$\sigma_x = |\sqrt{\text{Var}(x)}|$$
$$= \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

(iv) Sample Standard deviation.

$$s_x := \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

$$= \sqrt{\frac{n}{n-1}} \cdot \sigma_x$$

$\lim_{n\to\infty} s_x = \sigma_x$

But why $n-1$?

⚠ Bias of an estimator

Unbiased estimator

(v) Mean Absolute deviation. from 0 / Mean / arb. point (x')

$$\text{MAD}(x) = \frac{1}{n}\sum_i |x_i - x'|$$ // L_1 norm

↑ Median minimizes MAD(x') !!

++ Not differentiable, mathematically less treatable.

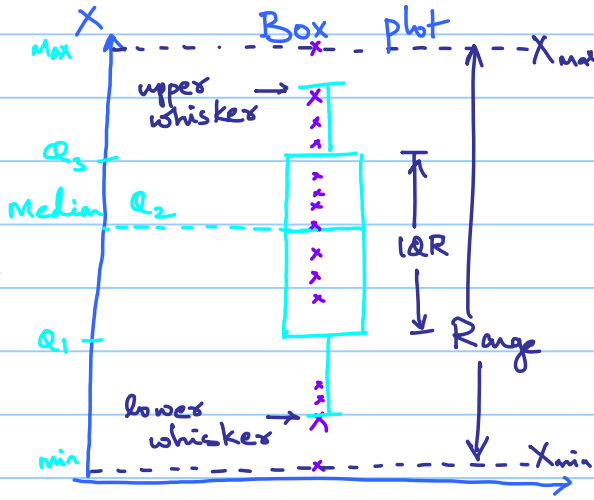
-- Requires no multiplication, less complex.

?

LECTURE 5

* Range := $X_{max} - X_{min}$
  → sample range ≤ population range.

* Interquartile range (IQR)
  $IQR := Q_3 - Q_1$



Outliers :- elements X_i such that
$X_i > Q_3 + 1.5\,IQR$ or
$X_i < Q_1 - 1.5\,IQR$

! Chebyshev's inequality (in statistics)

Statement: At least $100\left(1 - \frac{1}{k^2}\right)\%$ of data points lie within k std. deviations from the mean. $(k > 1)$

$$\frac{\#\{x_i \mid |x_i - \bar{x}| \le k s_x\}}{n} \ge 1 - \frac{1}{k^2}$$

or

$$\frac{\#\{x_i \mid |x_i - \bar{x}| \ge k s_x\}}{n} \le \frac{1}{k^2}$$

$\bar{x} - 3s_x$ → contain ≥ 75% data. ← $\bar{x} + 3s_x$

→ contains at least 88.9% data

++ Empirical relationship, true for all kinds of data.
{ $s_x \neq 0$ is not a constant sequence }

⚠ → Caveats of Chebyshev's inequality in terms of probability, continuous Random Variables, integrability,

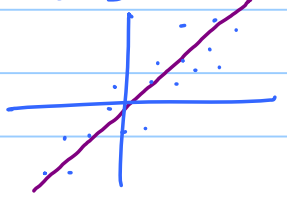
-- Chebyshev's bound is loose, better bounds are there. (Chernoff, Markov, ...)

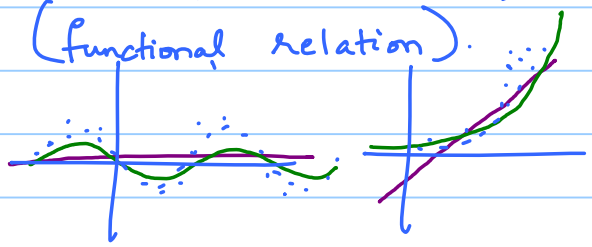
• Multivariate dataset.
eg in 2D $\{(x_i, y_i)\}_{i \in I_n}$ ordered pairs. ▷ points (x_i, y_i) in 2D

CORRELATION
measure of LINEAR relationship
$Y = aX + b$
for some a & b

⊂ Subset of

all kinds of relations b/w 2 variables.
eg: $Y = e^X$, $Y = \log X$,
X = time, Y = acceleration,
(functional relation).

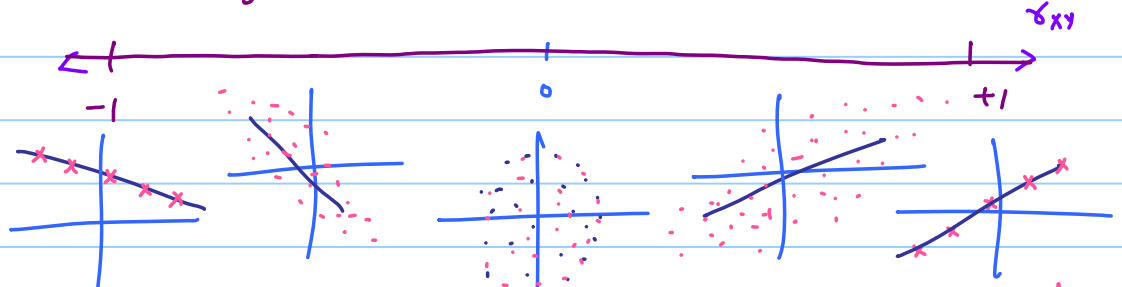




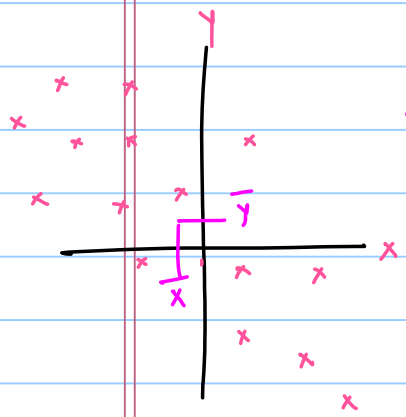
Regression: finding this function → actual relation – $Y = f(X)$
linear relation –

r_{XY} { Correlation → sign
Coefficient → magnitude

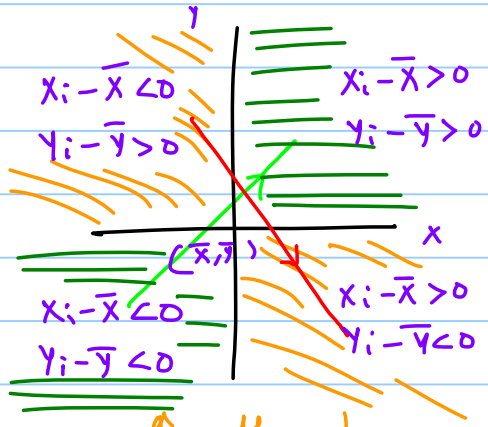
we want a measure like this
$r_{XY} \in [-1, 1]$



Correlation
perfect negative
negative
Zero
positive
perfect positive



without loss of generality → translate origin to the mean of data



- $Cov(X,Y) := \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

↳ as one ↑, other ↓ $(X_i - \bar{X})(Y_i - \bar{Y}) < 0$

$(X_i - \bar{X})(Y_i - \bar{Y}) > 0$ ↳ as one ↑, other ↑

-- X & Y can be heterogeneous. { if X_i = mass of i^{th} body,
⚠ units of XY, unintuitive. { Y_i = force on i^{th} body

$Cov(X,Y)$ has the dimension of force (N) or $[MLT^{-2}]$

- Karl Pearson's Correlation coefficient, $r_{XY} = \frac{Cov(X,Y)}{S_X \cdot S_Y}$

++ r_{XY} is normalized, unitless. ++ Compare datasets ✓

ROUGH ⚠

gray areas

!? With an inner product processes L_2 norm is easy

$\pm$ $X = [x_1^T \; x_2^T \; \cdots \; x_n^T]$ vector, one instance.

$X = [x_1 \; x_2 \; \cdots \; x_n]_{d \times n}$ matrix

$$XX^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & & & \\ \vdots & & & \\ x_{1d} & & & x_{nd} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & & & \\ \vdots & & & \\ x_{n1} & & & x_{nd} \end{bmatrix}$$

$X = (x_1, x_2, \cdots, x_n)^T$

$Y = (y_1, y_2, \cdots, y_n)^T$

$$Var(X) = \frac{1}{n-1} \sum_i (x_i - \bar{X})^2 = \frac{1}{n-1} \sum_i (x_i^2 + \bar{X}^2 - 2x_i\bar{X}) = \frac{1}{n-1} \{ X^TX + n\bar{X}^2 - 2\bar{X} \cdot n\bar{X} \}$$

$$= \frac{1}{n-1} (X^TX - n\bar{X}^2)$$

$\rightarrow \frac{(X-\bar{X})^T(X-\bar{X})}{n-1}$

$$Var(Y) := \frac{1}{n-1} (Y-\bar{Y})^T(Y-\bar{Y})$$

$$Cov(X,Y) := \frac{1}{n-1} (X-\bar{X})^T(Y-\bar{Y})$$

$$r_{XY} = \frac{(X-\bar{X})^T(Y-\bar{Y})}{\sqrt{(X-\bar{X})(X-\bar{X})}\sqrt{(Y-\bar{Y})^T(Y-\bar{Y})}} \quad \{ \div \frac{n-1}{n-1} \}$$

- Prob Set →

bijections
functions
injections

→ SRSWoR

SRSWR