Probability and Statistics (PROB)

Lecture 1

Monday. 22nd Jan, 24

Statistics
- Descriptive Statistics – Describe + summarise + tabulation + no calculations
- Inferential Statistics – inferring something + calculations

Population: Objects in a specific category that is under consideration
Sample: Small "representative" part of the population.
Observation: Numerical record of information.
Dataset, Variable, Constant, Group Data, Ungroup Data

Variable
- Quantitative variable – measurable, age, weight, time etc.
  - (countable) Discrete
  - Continuous (uncountable) e.g. 5/4, 1.2
- Qualitative variable – non-measurable, numerical like religion.

Errors
- Systematic error / cummulative / biased error – personal involvement
- Random error / unbiased error – not personal involvement

Study Design
- Time Series data
- Cross-sectional data

Measurement scales:
1) Nominal scale – making groups / categories / classes of data
2) Ordinal scale – nominal + ordered data
3) Interval scale – constant interval size, "no natural zero", 0 means something e.g. temperature
4) Ratio scale – "true-zero-point", 0 means nothing e.g. bank balance.

Elements: Data being collected of an object

Probability and Statistics
Wednesday

Lecture 2
24th Jan, 24

Measures of Central Tendency or Averages: [Values given in sequence/symmetry eg: 5, 10, 15, 20, 25, 30, 35 ...]

→ Arithmetic mean or mean ($\bar{x} = \frac{\Sigma x}{n}$ — sample mean, sample size; $\mu = \frac{\Sigma x}{N}$ (Mu) — population, population size)
→ Geometric mean X
→ Harmonic mean X
→ Median – data should be sorted, divide into two equal parts, for odd: middle, even: $\frac{\text{middle two}}{2}$
→ Mode – when see trend a more frequent. (unimodal, bimodal, multimodal)

* If extreme values in dataset like 5, 10, 15, 20, 10000, then use median instead of mean to get more accurate answer.

The marks obtained by 9 students are given as 45, 32, 37, 46, 39, 36, 41, 48, 36. Calculate arithmetic mean.

$\bar{x} = \frac{\Sigma x}{n} = \frac{45+32+37+46+39+36+41+48+36}{9} = \frac{360}{9} = 40.$

Let the values be 1, 5, 3, 9, 10000. Calculate median.
First sort: 1, 3, 5, 9, 10000. Now, the median is 5, since middle value and is odd.

~~Quartiles, Deciles, & Percentiles~~

Quantiles:
Quartiles — divide in 4 parts
Deciles — divide in 10 parts
Percentiles — divide in 100 parts.

* First sort the dataset.

Q_1 Q_2 Q_3

Q1: Lower Quartile / First Quartile.
Q2: Second Quartile / Median
Q3: Upper Quartile / Third Quartile

To find Percentile:
Step 1: Arrange the data in ascending order.
Step 2: Compute the index i, $\boxed{i = \left(\frac{p}{100}\right) n}$ where n = no of observations
p = percentile of interest

Step 3:

a) If i is not an integer, round-up. The next integer greater than i denotes the position of the p^{th} percentile.

b) If i is an integer, the p^{th} percentile is the average of the values in position i and i+1.

Q: 85^{th} Percentile? P_{85}?

The values are 3310, 3355, 3450, 3480, 3480, 3480, 3520, 3540, 3550, 3650, 3730, 3825

The data is sorted.

$i = \left(\frac{85}{100}\right) \times 12 = 10.2$ a) (not integer, so round up - 11).

Position at 11th is 3730 so $P_{85} = 3730$.

b) If it were 10, then take average of 3650 and 3730

Lower Quartile $= \frac{25}{100} \times 12 = 3$ $P_{25} = \frac{3450 + 3480}{2}$ $Q_1 = 3465$

Second Method for Calculating Quartiles:

Let 7, 8, 11, 13, 8, 12 — Even

First, sort 7, 8, 8, 11, 12, 13.

Divide in 2 halves.

7 8 8 | 11 12 13

$Q_1 = 8$ $Q_3 = 12$

$Q_2 = \frac{8+11}{2} = 10$

Let 7, 8, 11, 13, 15, 18, 19, 20, 22

1) Sorted

2) Apart from median, divide in two parts.

7 8 11 13 | 18 19 20 22

Median $= \frac{8+11}{2}$ Median $= \frac{19+20}{2}$

$Q_1 = 9.5$ $Q_3 = 19.5$

Probability & Statistics — Lecture 3

Monday — 29th Jan, 2024

Trimmed Mean:

Trim some values from the dataset and then find mean. Remove a certain percentage.

- If total trimmed = 10%, then 5% from each side (lower & upper).
- If trimmed = 10%, then 10% from each side.

Question: The following measurements were recorded for the ~~drying~~ drying time in hours of a certain brand of latex paint.

3.4, 2.5, 4.8, 2.9, 3.6, 2.8, 3.3, 5.6, 3.7, 2.8, 4.4, 4.0, 5.2, 3.0, 4.8.

Assume that the measurements are a simple random sample.

a) Sample size of above sample? $n = 15$

b) Compute the 20% trimmed mean for this dataset?

2.5, 2.8, 2.8, | 2.9, 3.0, 3.3, 3.4, 3.6, 3.7, 4.0, 4.4, 4.8 | 4.8, 5.2, 5.6

$$\bar{x}_{tr20} = \frac{2.9 + 3.0 + 3.3 + \dots + 4.8}{9}$$

$$= 3.67$$

Measures of Dispersion:

Absolute dispersion (same & unit; squared) / Relative dispersion (unitless for comparison)

A:	48	52	60	60	60	68	72	Mean = Median = Mode = 60
B:	0	10	60	60	60	110	120	Mean = Median = Mode = 60

Interquartile Range:

Absolute: IQ Range $= Q_3 - Q_1$

- We do not use all values, so not accurate. We do not consider below Q_1 & above Q_3.
- use when outliers.

Relative: Coefficient of Quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Variation: Variance:

Absolute S^2 = Variance = $\frac{\sum (x - \bar{x})^2}{n-1}$ ($\bar{x}$: Sample mean) & Population variation $= \frac{\sum (x-\mu)^2}{N}$ (σ^2: sigma; μ: Population mean)

- most reliable cause consider all values. — Squared units.

Standard Deviation

Absolute

$$S = S.D = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$ — Same units.

Q: An engineer is interested in a pH meter. Data are collected on the meter by measuring the pH of unwanted substance. The sample of size 10 is taken with results given by

7.07, 7.00, 7.10, 6.97, 7.00, 7.03, 7.01, 7.01, 6.98, 7.08

$\bar{x} = 7.025$

$n = 10$

Variance = ?

S.D = ?

$(x-\bar{x})^2$
$(7.07-7.025)^2 = 0.002025$
0.000625
0.005625
0.003025
0.000625
0.000025
0.000225
0.000225
0.002025
0.003025
$\Sigma (x-\bar{x})^2 = 0.017451$

$$\text{Variance} = \frac{0.017451}{9} = 0.00194$$

$$S.D = \sqrt{0.00194} = 0.044$$

Variance of Standard Deviation

Relative · coefficient of variation (C.V)

$$C.V = \frac{SD}{\bar{x}} \times 100$$

C.V ↑ variation ↑ consistency ↓

Group A	Group B
$\bar{x} = 18.4$	$\bar{x} = 211$
$S = 7.66$	$S = 78.38$
$CV = \frac{7.66}{18.4} \times 100$	$C.V = \frac{78.38}{211} \times 100$
$= 41.6\%$	$= 37.15\%$

More consistency less variability

Probability & Statistics — Lecture 4

Wednesday — 31st Jan, 2024

Five Number Summary:

1. Minimum value
2. Q_1
3. Q_2 = Median
4. Q_3
5. Maximum

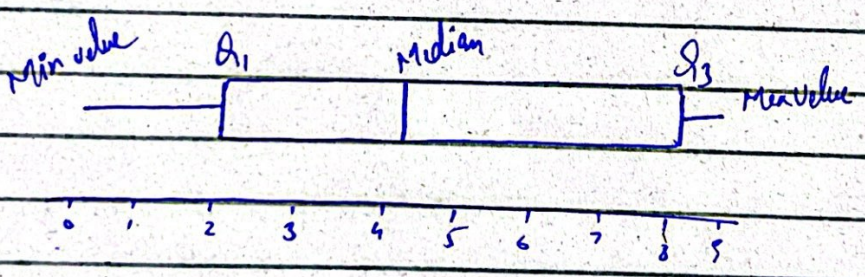
Q: 0, 2, 5, 2, 0, 4, 4, 8, 5, 8, 8

0, 0, 2, 2, 4, (4), 5, 8, 8, 8, 8

Min = 0 Median = 4 $Q_3 = 8$.

Max = 9 $Q_1 = 2$

Box and Whisker Plot:



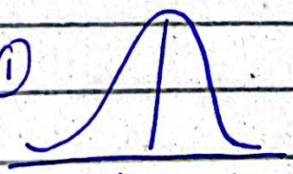
Outlier Detection:

Upper fence = $Q_3 + 1.5(Q_3 - Q_1)$ $= 8 + 1.5(6)$ $= 17$.

Lower fence = $Q_1 - 1.5(Q_3 - Q_1)$ $= 2 - 1.5(6)$ $= -7$.

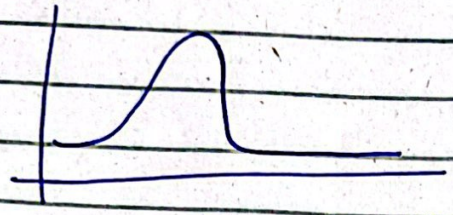
No outliers cause no values 17 and above or -7 and below are present.

Skewness:

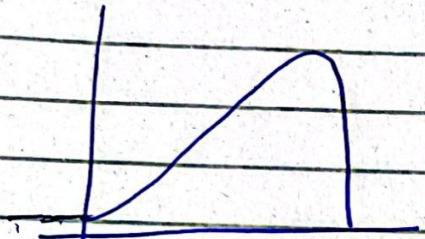


Symmetrical distribution.

Mean = Median = Mode

②  (+vely skewed distribution).

Mean > Median > Mode

③  (-vely skewed distribution).

Mean < Median < Mode.

If $Q_3 - Q_2 = Q_2 - Q_1$ — Symmetrical distribution
If $Q_3 - Q_2 > Q_2 - Q_1$ — +vely skewed
If $Q_3 - Q_2 < Q_2 - Q_1$ — -vely skewed

Co-efficient of Skewness:

Pearson's coefficient of Skewness = $S_k = \frac{\text{Mean} - \text{Mode}}{S.D} = \frac{3(\text{Mean} - \text{Median})}{SD}$

Ranges from -3 to +3.
- -3 (-vely skewed)
- 0 (symmetrical)
- +3 (+vely skewed)

* Mean - Mode = 3(Mean - Median)
* Mode = 3Median - 2Mean

} Empirical relation b/w mean, median & mode

Bowley's Co-efficient of Skewness

$S_k = \frac{Q_1 + Q_3 - 2\,\text{Median}}{Q_3 - Q_1}$

-1 to +1

-ve skewed . +ve skewed

Graphical Representation:

1- Dot-Plot
2- Bar Chart
3- Pie-chart

Lecture 5
07/02/24

Wednesday Online Class

Frequency Distribution:

1) $k = 1 + 3.3 * \log_5 N$ — observations
Number of classes — Estimate not necessary.

* Number of classes should be $>=5$ and $<=20$. If 7.1, round-up to 8.

2) Range of data.

3) Interval/Class width. Range / no. of classes.
Width = Range / No. of classes.

⇒ Class Limits & Class Boundary — $\frac{UCL + LCL}{2}$ — Mid-Point
upper class limit

Relative Frequency: $\frac{\text{Frequency}}{\text{Total frequency}}$

Cummulative Frequency: Sums everything upto that level.

Percent Frequency: Relative Frequency x 100.

Probability & Statistics Continued... Lecture 5

Wednesday 07/02/24

Histogram:

If intervals are diffant, do not take frequency but frequency density.

$FD = \frac{Freq}{CL}$

Monday Lecture 6

12/02/24

Grouped Data:

$Mean = \frac{\Sigma fx}{\Sigma f}$ — Mid point $= \frac{7350}{60} = 122.5$

Weights	f	x	fx	fx^2
65-84	8	$\frac{65+84}{2} = 74.5$	670.5	49952.25
85-104	10	94.5	945	89302.50
105-124	17	114.5	1946.5	⋮
125-144	10	134.5	1345	⋮
145-164	5	154.5	772.5	⋮
165-184	4	174.5	698	⋮
185-204	5	194.5	972.5	⋮
	60		7350	973335.00

Variance: $\frac{\Sigma f(x-\bar{x})^2}{n} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2$

Standard deviation: $\sqrt{\frac{\Sigma f(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$

$= \sqrt{\frac{973335}{60} - \left(\frac{7350}{60}\right)^2}$

$= 34.87$ grams.

Probability & Statistics

Lecture 6

Monday

12/02/24

Probability

Sample Space: Complete list of all possible outcomes.

S.S = {1, 2, 3, 4, 5, 6} ← Sample point

S.S = {H, T}

Experiment: A planned activity or process.

Outcomes: Results obtained from experiment.

Trial: A single performance of an experiment.

Event: Outcome of interest

Mutually Exclusive Events:

→ Disjoint events (Cannot occur together).

→ No common points.

Not Mutually Exclusive Events:

→ Common points (Can occur together).

Exhaustive Events:

S.S = {1, 2, 3, 4, 5, 6}

A: Even number occurs = {2, 4, 6} They are mutually exclusive

B: Odd number occurs = {1, 3, 5}

A ∪ B = S.S so, exhaustive events.

Equally Likely Events:

Chances of occuring is same.

Probability & Statistics — Lecture 7

Wednesday — Probability — 14/02/24

Counting Sample Points:
→ Rule of Multiplication
→ Permutations
→ Combinations

Permutation: $^nP_r = \frac{n!}{(n-r)!}$

Permutation when objects are not distinct.

$$\frac{n!}{p!q!\times\cdots}$$

Combination:

1 — A B C D

ABC ACD BCD ABD
ACB
BAC
BCA
CAB
CBA

$^nC_r = \frac{^nP_r}{r!}$ (order does not matter)

$= \frac{n!}{(n-r)!r!}$

Q: A team of 5 people which must contain three men & 2 women is chosen from 8 men & 7 women.

$^8C_3 \times {}^7C_2 = 1176$ ways

Q: A three person committee is to be formed from 4 person list. How many sample points are associated with this experiment?

$^4C_3 = 4$ ways.

A: Favourable outcomes / Total outcomes

$P(A) = \frac{10}{40}$

$P(B) = \frac{30}{40}$

1) $0 \leq P(A) \leq 1$
2) Sum of probabilities is 1.

S.S = {1, 2, 3, 4, 5, 6}

A: 6 ~~outcomes~~ occurence.

$P(A) = \frac{1}{6}$

Addition law of Probability:

If A & B are mutually exclusive events.

$P(A \text{ or } B) = P(A) + P(B)$

$P(A \cup B) = P(A) + P(B)$

⇒ If A and B are not mutually exclusive events.

$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$

Q: Two fair dice are thrown. A prize is won if total score on two rows is 4 or if each individual score is over 4.

a) S.S = {(1,1), (1,2), (1,3) .....
(2,1) (2,2) (2,3) .....
⋮ ⋮ ⋮ }

A: Total on two rolls is 4 = {(1,3) (2,2) (3,1)} = 3/36

B: Ind. score is over 4. = {(5,5) (5,6) (6,6)} = 4/36

(Mutuelly Exclusive)

b) A prize is won if total score on two rolls is 10 or if each individual score is over 4.

C: Total score is 10 = {(4,6) (5,5) (6,4)} = 3/36 = 1/12

$P(\text{won}) = P(B \cup C) = \{(5,5)\} = \frac{1}{36} = P(A) + P(B) - P(A \cap B)$

$= \frac{4}{36} + \frac{3}{36} - \frac{1}{36}$

$= \frac{6}{36} = \frac{1}{6}$

Probability & Statistics — Lecture 8

Monday — 19/02/24

Q: Probability of girls & boys & ~~Engineer~~ who wear glasses.

W: selected is women

M: Selected is men

G: wear glasses

G': Don't wear glasses.

	W	M	Total
G	4	2	6
G'	3	11	14
	7	13	20.

$P(W \text{ or } G) = ?$

$P(W) = 7/20$

$P(G) = 6/20$

$P(W \cap G) = 4/20$

$P(W \cup G) = P(W) + P(G) - P(W \cap G)$

$= 7/20 + 6/20 - 4/20$

$= 9/20.$

Joint Probability Table:

	W	M	
G	4/20	2/20	6/20
G'	3/20	11/20	14/20
	7/20	13/20	1

Conditional Probability:

$S = \{1, 2, 3, 4, 5, 6\}$ A: occurrence of 6

$P(A) = 1/6$

B: Shown even no. of die

$S:S = \{2, 4, 6\}$

$P(A|B) = 1/3$ * Reduces sample size

Given that.

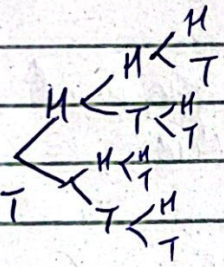
Q: Consider a class of 30 students, where 17 are girls 13 are boys. Suppose 5 of the girls and 6 are left-handed. If a student selected at random is a girl. What is probability that student is left handed?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

	G	B	
L	5	6	11
R	12	7	19
	17	13	30

$$P(L|G) = \frac{P(L \cap G)}{P(G)} = \frac{5/30}{17/30} = \frac{5}{30} \times \frac{30}{17} = \frac{5}{17}$$

Tree Diagram



S.S: {HHH, HHT ~~ ---}

Dependent / Independent Variable:

Multiplication Rule

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$P(A \text{ and } B) = P(A) \cdot P(B|A)$ — Dependent events

$P(A \cap B) = P(A) \cdot P(B)$ — Independent events.

Q: Suppose a jar contains 7 red & 4 white balls. 2 balls are selected w/o replacement. What is probability that a) Both are red b) First white second c) Both same colour.

a) $\frac{7}{11} \times \frac{6}{10}$

b) $\frac{4}{11} \times \frac{7}{10}$

$P(W_1) \cdot P(R_2|W_1)$

$= 4/11 \times 7/10$

Without replacement - dependent

c) $P(R_1) \cdot P(R_2|R) + P(W_1) \cdot P(W_2|W_1)$

$= \frac{7}{11} \cdot \frac{6}{10} + \frac{4}{11} \cdot \frac{3}{10}$

$= \frac{27}{55}$

Probability & Statistics — Lecture 09

Wednesday — 21/02/24

Q. Suppose a jar contains 7 red, 4 white balls. 2 balls selected with replacement. What is the probability that second ball is red?

$P(R_2) = P(R_1 \cap R_2) + P(W_1 \cap R_2)$

$= P(R_1) \cdot P(R_2) + P(W_1) \cdot P(R_2)$

$= \frac{7}{11} \cdot \frac{7}{11} + \frac{4}{11} \cdot \frac{7}{11}$

$= 7/11$

Multiplication Rule:

Dependent events

$P(A \cap B) = P(A) \cdot P(B/A)$

P(A and B)

Independant Events:

$P(A \cap B) = P(A) \cdot P(B)$

Q: In a carnival game, a contestant has to first spin a fair coin & then roll a dice whose faces are numbered 1-6. The contestant wins a prize if coin shows head & dice shows score below 3. Find the probability that the contestant wins a prize.

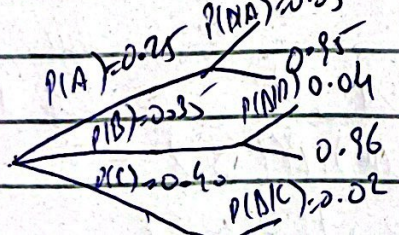
P(Prize win) = P(Coin shows head & Dice below 3) — Independant

$P(A \cap B) = P(A) \cdot P(B)$

$= \frac{1}{2} \cdot \frac{2}{6} = \frac{1}{6}$

Bayes Theorem

Q. In a bolt factory machines A, B, C manufactures 25, 35, and 40% of the total outcome respectively. Of their output 5, 4, 2% are defective bolts. A bolt is selected at random & found to be defective. WITPT the bolt came from machine C.



$P(C/D) = \frac{P(C \cap D)}{P(D)}$

$P(D) = P(D \cap A) + P(D \cap B) + P(D \cap C)$

$= P(A) \cdot P(D|A) + P(B) \cdot P(D|B) + P(C) \cdot P(D|C)$

$= (0.25)(0.05) + (0.35)(0.04) + (0.40)(0.02)$

$= 0.0125 + 0.014 + 0.008$

≈ 0.0345

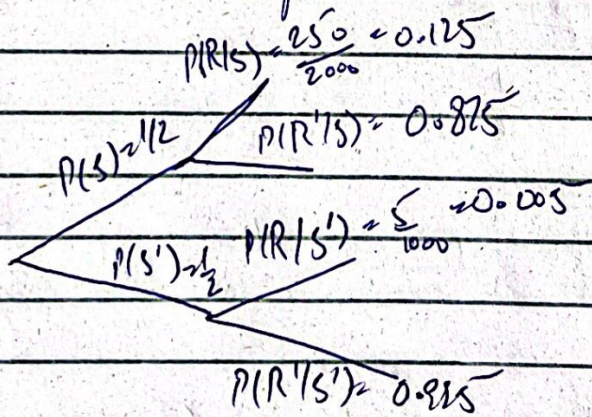
$P(C|D) = \frac{P(C) \cdot P(D|C)}{P(D)} = \frac{(0.40)(0.02)}{0.0345} = 0.232$

Spam

Bayesian Filter

Q. Suppose that we have found that word "Rolex" occurs in 250 of 2000 messages known to be spam and 5 of 1000 messages is known not to be spam. Estimate TPT the incoming message containing the word Rolex is spam. Assuming that it is equally likely that an incoming message is spam or not spam. If our threshold for rejecting message as spam is 0.9 will we reject such messages? How?

S: Spam msg S': Non spam R: Rolex R': Rolex not occurs.



$P(S|R) = \frac{P(S \cap R)}{P(R)} = \frac{P(S) \cdot P(R|S)}{0.065} = \frac{\frac{1}{2} \cdot 0.125}{0.065} = 0.962$

Rejected cause threshold > 0.9

$= P(R \cap S) + P(R \cap S')$

$= P(S) \cdot P(R|S) + P(R) \cdot P(R|S')$

$= (0.5)(0.125) + (0.5)(0.005)$

≈ 0.065