Week-3

Data Science for Engineers (NPTEL)

TA:

Saladi Pravallika,

Prime Ministers Research Fellow

PhD Scholar, IISc Bangalore.

1. Intro to Random Variables & how they are characterised by prob. measures & prob density f^ns.

2. How parameters of the density f^n is estimated (estimation)
   Hypothesis Testing (for decision making)

→ Session is being recorded

→ 3-5 min break every half hour & 10 min break every hour.

→ If you would like me to cover more topics in next session please let me know at the end.

→ Be respectful during the session, I will not hesitate to remove you from the session

Phenomenon → Deterministic
Phenomenon → Stochastic

Characterization of Random Phenomenon

Sources of Error → Model Error
Sources of Error → Measurement Error

Types of Ph. → Random
Types of Ph. → Continous

* Important Terms

① Sample Space – Set of all possible outcomes.
Eg: 2 coins toss

Consider the experiment in which a coin is tossed repeatedly until a head comes up.
Describe the sample space $\{H, TH, TTH, TTTH, \ldots\}$

② Event – Subset of sample set
Eg: Having H/T first

③ Probability Measure – f^n that assigns a real value to every outcomes.

$P(A)$

$P(\text{Head}) = \frac{1}{2}$
$P(\text{Tail}) = \frac{1}{2}$

$\frac{1}{2} + \frac{1}{2} = 1$ → Sample Space (Head + Tail)

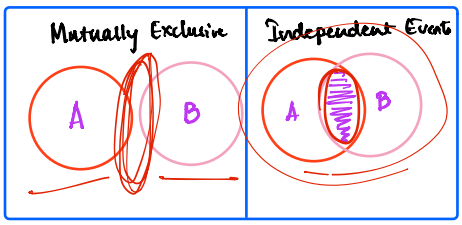
H & T → 50% 50%

$\frac{50}{100} = \frac{1}{2}$

$0 \le P(A) \le 1$
$P(S) = 1$
$P(A \cup B) = P(A) + P(B)$ – mutually exclusive

S → Sample Space

$P(A \cap B) = P(A) * P(B)$
↳ Independent (no influence on each other's occurence)



A → likes dogs
B → likes cats
→ likes dogs & cats

Venn Diagrams

Prob. w.r.t frequency.

$$\boxed{P(A) = \frac{N_A}{N}}$$

$A \rightarrow$ Event
$N_A \rightarrow$ # Event occurred
$N \rightarrow$ # Outcomes

2 coins
$\hookrightarrow \{HH, HT, TH, TT\}$

$P(HH) = \frac{1}{4}$

Question:

Two coins are tossed 500 times, and we get:

(i) Two heads: 105 times

(ii) One head: 275 times

(iii) No head: 120 times

Find the probability of each event to occur.

Also check the sum of probabilities.

$N = 500$

(i) $N_{HH} = 105$

$P(HH) = \frac{N_{HH}}{N} = \frac{105}{500} = 0.21$

(ii) $N_H = 275$

$P(H) = \frac{N_H}{N} = \frac{275}{500} = 0.55$

(iii) $N_{\sim H} = 120$

$P(\sim H) = \frac{N_{\sim H}}{N} = \frac{120}{500} = 0.24$

$1.00 = 1$

① $0 \le P \le 1$

② $P(S) = 1$

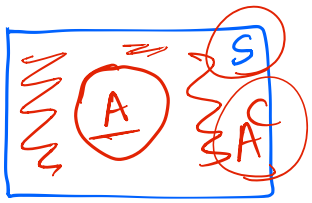
$\Sigma P = 1$

$A + A^C = 1$

$\boxed{A^C = 1 - A}$

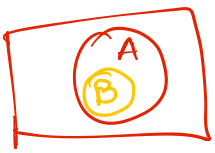
* Rules

① $A^C = 1 - A$

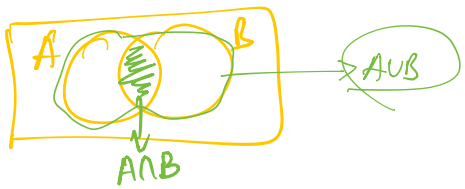


Venn Diagram

② If $B \subseteq A$, $P(B) < P(A)$



③ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



* Conditional Probability

A, B → not independent
outcomes influence each other.
i.e., we want to know the prob. of one given that we know the occurance of another

Eg: - 6 Red Balls
- 3 Blue Balls
- 1 green ball

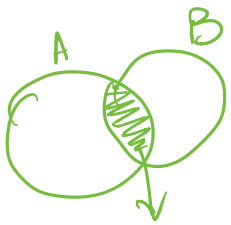
} N = 10

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \; ; \; P(B) > 0$$

B → known
A → unknown

A → 1st ball is red
B → 2nd is not green

$P(A) = \frac{6}{10}$ $\quad P(B) = \frac{9}{10}$ $\quad P(A \cap B) = \frac{\text{Red Balls}}{\text{All Balls}} = \frac{6}{10}$



Alt, $P(\text{green}) = \frac{1}{10}$

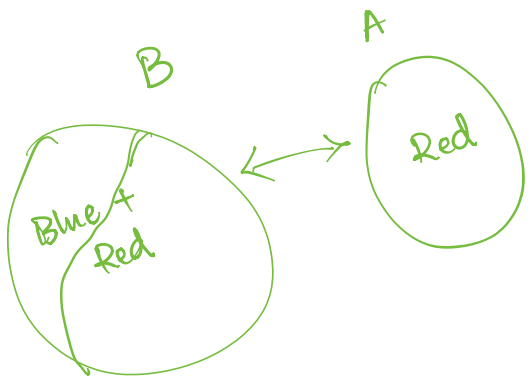
$P(\sim\text{green}) = 1 - P(\text{green})$
$= 1 - \frac{1}{10}$
$= 9/10$

① $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{6/10}{9/10} = \frac{2}{3}$

② $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{6/10}{6/10} = 1$

Bayes formula → $P(A|B)\,P(B) = P(B|A)\,P(A)$

$$P(A) = P(A|B)\,P(B) + P(A|B^c)\,P(B^c)$$



Problem (R-Code)

5 Red Balls
3 Green Balls } $N = 8$

2 Ball are drawn at rnd.

find the prob. that both are green.

$$\underset{\downarrow \atop 3/8}{\underline{G}} \quad \underset{\downarrow \atop 2/7}{\underline{G}}$$

$$P(GG) = \frac{3}{\cancel{8}_{4}} \times \frac{\cancel{2}}{7} = \frac{3}{28}$$

* Random Variable

RV → f^n that maps outcomes to numeric values.

Eg: H → 1/2 = 0.5
T → 1/2 = 0.5

RV → Discrete
RV → Continous

Probability Distribution → Prob. Mass f^n (PMF) – for discrete RV
Probability Distribution → Prob. Density f^n (PDF) – for conti. RV.

PD → to understand RV, we shld know prob. distribution.
It tells us the likelihood of each possible value or range of values.

why do we use prob distributions?
Sometimes we have so many outcomes it is not possible to track all, along with their probabilities. In such cases, we use a mathematic f^n to represent the outcomes with their prob.

PMF

Assigns probability to every outcome in Sample Space.

$P(x=0) = 0.5$
$P(x=1) = 0.5$

Given $\{H, T\} \rightarrow \{0, 1\}$

This is used for discrete R.V.

PDF

Assigns probability to every interval on a real line.

$F(a < x < b) = \int_a^b f(x)\,dx$

$F(b) = \int_{-\infty}^{b} f(x)\,dx$

Area under the curve

F(b) → Cumulative density f^n.

Examples (PMF)

1. Let X be a random variable, and P(X=x) is the PMF given by,

X	0	1	2	3	4	5	6	7
P(X=x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

1. Determine the value of k
2. Find the probability (i) $P(X \le 6)$, (ii) $P(3 < x \le 6)$

Hint: (i) Sum of all prob = 1.
(ii) $0 \le P \le 1$

1

$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$

$9k + 10k^2 = 1$

$10k^2 + 9k - 1 = 0$

$10k^2 + 10k - k - 1 = 0$

$10k(k+1) - 1(k+1) = 0$

$(k+1)(10k-1) = 0$

$k + 1 = 0$

$\boxed{k = -1}$ ✗

$10k - 1 = 0$

$\boxed{k = 1/10}$ ✓

$\boxed{0 \le P \le 1}$

2. (i) $P(X \le 6)$

$P(0) + P(1) + \dots + P(6)$

Rewrite,

$P(X \le 6) = 1 - P(7)$

$= 1 - 7k^2 - k$

$= 1 - 7(0.1)^2 - 0.1$

$= 1 - 7(0.01) - 0.1$

$= 0.9 - 0.07$

$= 0.83$

(ii) $P(3 < X \le 6) = P(4) + P(5) + P(6)$

$= 3k + k^2 + 2k^2$

$= 3k + 3k^2$

$= 3(0.1 + 0.01)$

$= 3(0.11)$

$= 0.33$ //

② PDF

Let X be a continuous random variable with the PDF given by:

$$f(x) = \begin{cases} x & ; \quad 0 < x < 1 \\ 2-x & ; \quad 1 < x < 2 \\ 0 & ; \quad x > 2 \end{cases}$$

Find $P(0.5 < x < 1.5)$.

Hint : $F(a < x < b) = \int_a^b f(x) \cdot dx$

$a = 0.5$
$b = 1.5$

$$F(0.5 < x < 1.5) = \int_{0.5}^{1.5} f(x) \cdot dx.$$

$$= \int_{0.5}^{1} f(x) \cdot dx + \int_{1}^{1.5} f(x) \cdot dx$$

$$= \int_{0.5}^{1} x \cdot dx + \int_{1}^{1.5} (2-x)\, dx$$

$$= \left[\frac{x^2}{2}\right]_{0.5}^{1} + \left[2x - \frac{x^2}{2}\right]_{1}^{1.5}$$

$$\frac{\frac{3}{2} \times \frac{3}{2}}{2}$$

$$= \left(\frac{1}{2} - \frac{1}{8}\right) + \left(3 - \frac{9}{8} - 2 + \frac{1}{2}\right)$$

$$\frac{9}{8}$$

$$= \frac{1}{2} - \frac{1}{8} + 3 - \frac{9}{8} - 2 + \frac{1}{2}$$

$$= 1 + 1 - \frac{1}{8}(1+9)$$

$$= 2 - \frac{10}{8}$$

$$= \frac{16-10}{8} = \frac{6}{8} = \frac{3}{4}$$

$$= 0.75$$

Binomial Mass function (for discrete)

↓ 2 ↳ outcome
Ex: Success or Failure

→ If I toss coin n times & 'k' times it is a head with a prob. 'p'.

then,

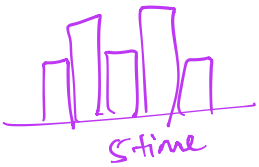
$$f(x=k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

If 'n' ↑ → gaussian dist.

! → factorial

$3! = 1 \times 2 \times 3 = 6$

$5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$



$\boxed{\text{Bootstraping in R.}}$

If a coin is tossed 5 times, find the probability of: (R-code)

(a) Exactly 2 heads

$p = 1/2$

(b) At least 4 heads.

$n = 5 \quad p = 1/2$

(a) $k = 2.$

$$f(x=2) = \frac{5!}{2!\,3!} \left(\frac{1}{2}\right)^2 \left(1-\frac{1}{2}\right)^3$$

$$= \frac{5 \times \cancel{4}^{2} \times \cancel{3} \times \cancel{2} \times 1}{1 \times \cancel{2} \times 1 \times \cancel{2} \times \cancel{3}} \left(\frac{1}{2}\right)^5$$

$$= 10 \times \frac{1}{2^5} = \frac{10}{32} = \frac{5}{16}$$

(b) $f(x=4) + f(x=5)$

Gaussian / Normal Density Function (for continous)

↳ to characterize random errors.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

PDF $\xrightarrow{\text{characterized}}$ μ, σ (parameters)

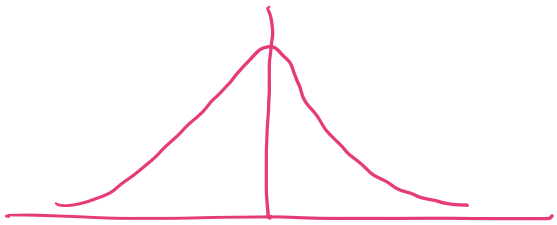
Std. normal dist : $\mu = 0$, $\sigma = 1$

Gaussian density f^n : $\mu = 0$, $\sigma = 1$

Variance $= \sigma^2$

S.d : $\sqrt{var}$

$= \sigma$



Calculate the probability density function of Gaussian distribution using the following data. $x = 2$, $u = 5$ and $sd = 3$

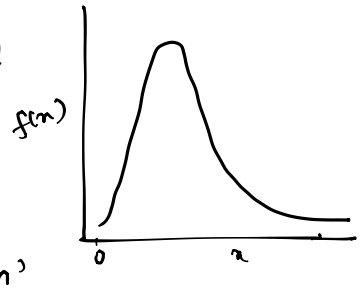
μ σ

(R Code)

— R Code —

Chi-square density function

The chi-square distribution results when v independent variables with standard normal distributions are squared and summed.



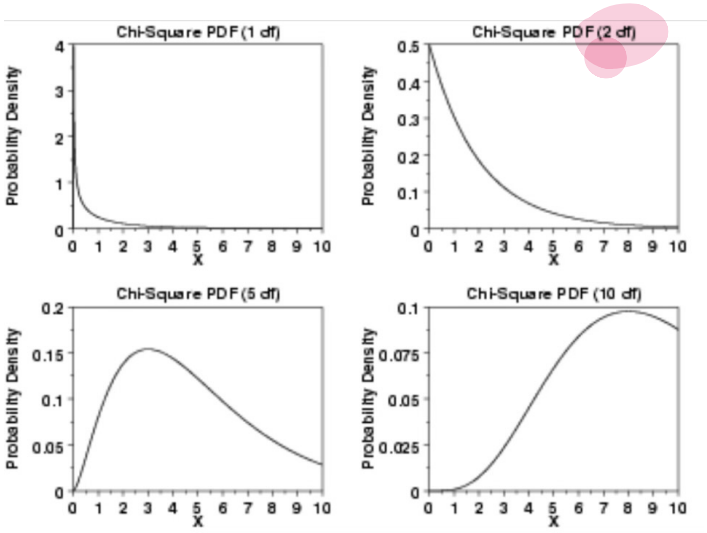
$$f(x) = \frac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}\, x^{n/2-1}\, e^{-x/2}$$

density characterized by 'n'

$\Gamma \rightarrow$ Gamma f^n

$$\Gamma(\alpha) = \int_0^\infty t^{a-1} e^{-t} \cdot dt$$

$n \rightarrow$ degrees of freedom.



Moments of PDF

Moments $\longrightarrow$ certain quantitative measures related to shape of f^n's graph

In PDF, 1st Moment $\rightarrow$ expected value
2nd " $\rightarrow$ variance (σ^2)
3rd " $\rightarrow$ Skewness (measure of asymmetry)

$$E[x^k] = \int_{-\infty}^{\infty} x^k f(x)\,dx \quad \text{— cont.}$$

$$E[x^k] = \sum_{i=1}^{N} x_i^k p(x_i)$$

Mean (μ) = $E[x]$
Variance (σ^2) = $E[(x-\mu)^2]$
$= E[x^2] - \mu^2$
S.d = σ

Symbolically, $x \sim \mathcal{N}(\mu, \sigma^2)$

Sample Statistics

Population → set of all possible outcomes of a rnd experiment characterized by $f(x)$

Sample set → finite set of observations

Inference → Conclusion

→ Measures of Central Tendency.

* Represents single value of the entire distribution.

CT
- → Mean
  - → Geometric Mean
  - → Weighted Mean
  - → Harmonic Mean
- → Median
- → Mode

$2 \times \text{Mean} + \text{Mode} = 3 * \text{Median}$

Question

38, 22, 23, 40, 29, 30, 26, 24, 33, 29, 5, 33, 23, 33

—— R Code ——

- Represents spread of sample set
  - Sample variance : $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
    - Unbiased estimate of population variance : $\mathrm{E}[s^2] = \sigma^2$
    - Standard deviation is sqrt of variance
  - Mean absolute deviation : $\bar{d} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$
  - Range : $R = x_{max} - x_{min}$
  - Eg. Sample heights of 20 cherry trees

s^2 = 70.5132 and 212.25 with outlier

s = 8.392 (population std used for generating numbers was 10)

MAD = 6.85 and 9.5 with outlier

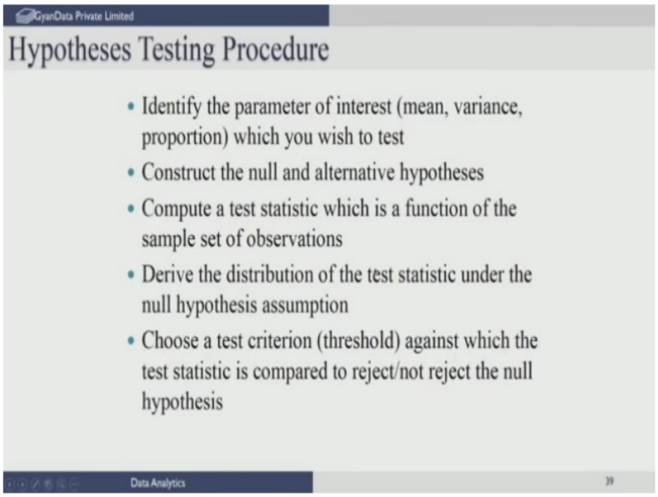
Range = 83 - 55 = 28

Hypothesis Testing

Parameters : mean / variance, etc

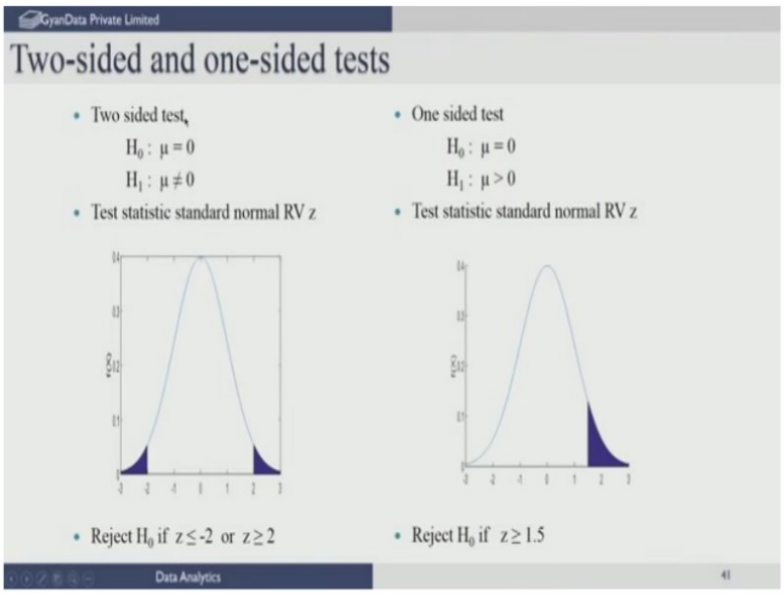
Null Hypothesis (H_0) : $\mu = \mu_0$

Alternative Hypothesis : $\mu < \mu_0$



Slide for NPTEL Lecture

Tests :



Hypothesis Testing

Higher values
What is the prob. of finding some value in a large sample

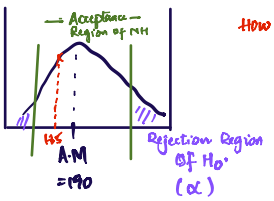
NULL HYPOTHESIS (H_0): The idea that what I currently see is true.
→ sample of value is same as population.

Alternate HYPOTHESIS (H_1): Opp of NULL Hypo.

Errors : Type I Error: When you reject H_0 when it is true.
(α) → significance of a test

Type II Error: When you accept H_0 when it is not true.
(β) (θ_1) When you reject H_1 when it is true.

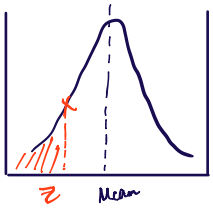
($1-\beta$) → Power of a test



How far away is 135 from A.M?

What is the prob. of rejecting?
↳ α.

$\boxed{\text{Fisher}}$ ⇒ $\alpha = \{0.05, 0.01, 0.001\}$ → both sides sum.
usually these values in biomedical
→ with statistical significance



$$z = \frac{\bar{x} - \mu_{pop}}{\sigma/\sqrt{n}}$$

n → # samples to calculate $\bar{x}$.

σ → S.D of population

Gossett : when n is small, sample distribution is close to normal but not normal. And called it as T- distribution.

Advantage of T : finding S.D of population is always difficult to find. In such cases, you can merge t-dist. and S.D lack

$$t = \frac{\bar{x} - \mu_{pop}}{s/\sqrt{n}}$$

s → Sample S.D

"t-statistic"
↳ depends on t_{n-1}, α

$n-1$ → degree of freedom