Chapter-1

Measures of location

Mean → gives central tendency of data.
(Average) → determined by adding all the data points in a population & then dividing total by the no. of points.

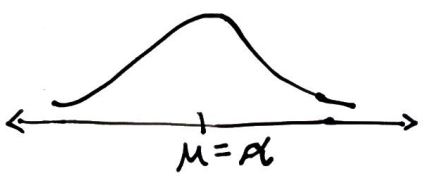
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median → middle value once the data is sorted (ordered) from small to largest.
(Middle)

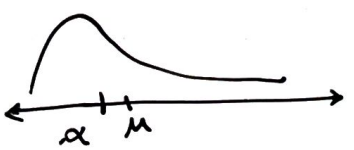
$$\tilde{x} = \begin{cases} \text{single mid value} & \text{if } n \% 2 != 0 \\ \text{avg of } \left(\frac{n}{2}^{th} \text{ \& } \frac{n}{2}^{th}+1\right) & \text{if } n \% 2 == 0 \end{cases}$$

(odd) $\left(\frac{n+1}{2}\right)^{th}$

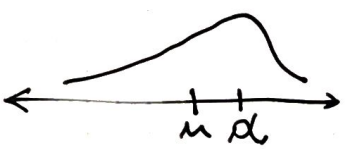
(even) $\left(\frac{(n/2)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}\right)$



Symmetric



positive skew
α < μ



negative skew
μ < α

μ = mean median = $\tilde{\mu}$ or α (here)

Other measures: Quartiles, Percentiles & trimmed means

10% trimmed mean for eg. is mean calc. by eliminating smallest 10% & largest 10%

Measures of variability

→ gives deviation from above computed central values.
We need to sometime find out the variability in data.

We cannot simply add the difference of all values from mean since,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

!!

∴, Other ways are:

- do absolute value $|x_i - \bar{x}|$ Summation
  But this isn't a good idea theoretically
- Square summation then root of result
  - ↳ Pros: solves our problem as well as many theoretical problems
  - Cons → Works crap with really skewed data.

★ So, sample variance (s^2) is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

★ Sample standard deviation $= \sqrt{s^2}$

Standard deviation is σ^2 for variance, & σ for standard deviation

$$\sigma^2 = \sum_{i=1}^{N} (x_i - M)^2 / N$$

Note
($n-1$ is statistical correction to compensate if data is extremely ~~narrow~~ symmetrically skewed.)

if observation is within 1.5 fs from closest fourth it's mild
farther than 1.5 fs → outlier
farther than 3 fs → extreme.

Ch2- Probability

→ finding out likelyhood of an outcome amidst multiple outcomes of an event

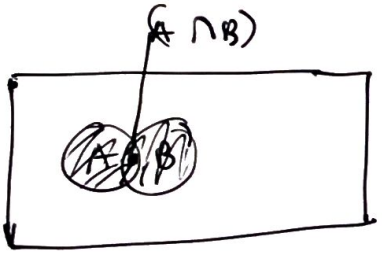
Sample Space (S)

→ set of all possible outcomes of an experiment

Event

→ any collection of outcomes contained in the sample space S

Simple if exactly 1 outcome
Compound if multiple outcomes that comprise



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

✡ Product Rule:

1st element of ordered pair can be chosen in n_1 ways & for each of these n_1 ways, second element can be chosen in n_2 ways, then no. of pairs is $n_1 * n_2$

✡ Permutation & Combination

↓ ordered ↓ unordered.

P → no. of permutations of size k that can be formed from n individuals or objects

↳ unordered subset is called Combination.

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

$${}^nC_r = \frac{n!}{n!\ n-r!}$$

Conditional Probability:

↳ probability of A, given event B has occured

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

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

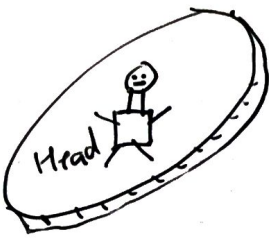
Bayes theorem:

let $A_1 \dots A_k$ be mutually exclusive events.

for any event B, $P(B) = P(B|A_1)P(A_1) + \dots P(B|A_n)P(A_n)$

$$\boxed{P(B) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i)}$$

Binomial Variable



lets consider an unfair coin,

$P(H) = 0.6$

$P(T) = 1 - 0.6 = 0.4$

X = # of heads after 10 flips of my coin.

(A) These sort of problems can be <u>categorised</u> if they follow certain rules:

1) The outcome of each trial can be classified either as a success or failure
2) Each trial is independent of the others
3) There is a fixed number of trials
4) The probability of success on each trial remains constant.

If these conditions are fulfilled we can employ a <u>Binomial variable</u> to solve our problem.

* How to Recognize a binomial Variable you ask?
Ans: Well the process is simple. you ask intelligent questions about the conditions in exhibit (A).

<u>illustration</u> if we are keeping track of # of kings in a deck by checking if the card is king or not,

i) if we check the card then remove it from deck, we can say ~~it's~~ ~~not~~ we can't employ a binomial variable.

(since after each trial, we change the no. of cards in deck, & hence affect(change) the probability of each new card about to be drawn. ~~the~~ each experiment & trial is now dependent on previous outcome))

trial 1

$P(\text{success}) = \frac{4}{52}$

trial 2

$P(\text{success} \mid \text{success in } T_1) = \frac{3}{51}$

* Why we use binomial variables??

we can say the distribution is binomial, useful inferences.

i) [Sum of a bunch of independent trials]

⋮

* 10% Rule?

if sample ≤ 10% of population

≈ independence

Let's consider an example

X = # of H from flipping coin 5 times

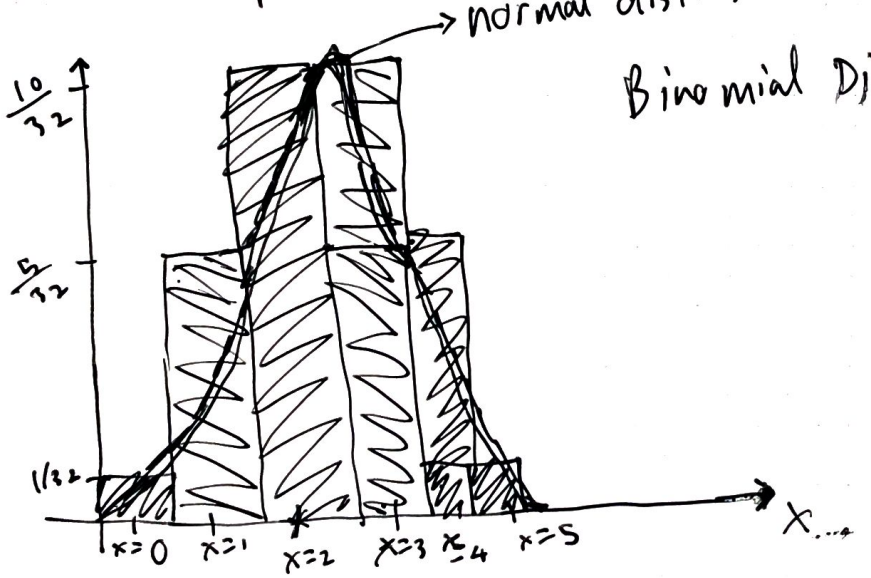
possible no. of outcome from 5 flips $= 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^5 = 32$

$P(x=0) = \frac{1}{32} = \frac{{}^5C_0}{32}$ $\quad P(x=2) = \frac{{}^5C_2}{32} = \frac{10}{32}$

$P(x=1) = \frac{5}{32} = \frac{{}^5C_1}{32}$ $\quad P(x=3) = \frac{{}^5C_3}{32} = \frac{10}{32}$

$P(x=4) = \frac{{}^5C_4}{32} = \frac{5}{32}$ $\quad P(x=5) = \frac{{}^5C_5}{32} = \frac{1}{32}$

If we plot this



* Binomial distribution is discrete version of normal distribution

more observations will give out closer & closer discrete versions of the normal distributions

i.e. more x values, smoother the bar graph.

Calculating Probability

If I'm playing basketball,

Prob (Score) = 70% or 0.7

Prob (miss) = 30% or 0.3

P (exactly 2 scores in 6 attempts)

S S M M M M

$(0.7)(0.7)(0.3)(0.3)(0.3)(0.3) = (0.7)^2(0.3)^4$ → this combination prob

Similarly,

M S M S M M

$(0.3)(0.7)(0.3)(0.7)(0.3)(0.3) = (0.7)^2(0.3)^4$ → this combination prob

⋮

We can see many ways are to do this but it's $(0.7)^2(0.3)^4$ nonetheless

added to however many times it happens

?

$\binom{6}{2}$

∴ total prob of P (exactly 2 scores in 6 attempts)

$= \binom{6}{2} 0.7^2 0.3^4$

or $\binom{\text{total attempts}}{\text{favorable outcomes}} \times$ (Chance of favorable) $\times$ (Chance of fav out) $\times$ (Chance of fail)

Hypergeometric Distribution:

Eg: 6 docs, 19 nurses attend a small conference.

all 25 names in a hat, 5 names picked randomly without replacement

what is P(4 docs, 1 nurse selected)?

sol → trials not independent

Binomial dist not relevant

lets see ☺

P(4 docs, 1 nurse)

$$\frac{\binom{6}{4}\binom{19}{1}}{\binom{25}{5}} = \frac{15 \cdot 19}{53130} = 0.00536$$

not easy to do with complex stuff.

☆ Had the sampling been done with replacement, using binomial distribution, would be appropriate.

$$P(4 \text{ docs}) = \binom{5}{4}\left(\frac{6}{25}\right)^4\left(1-\frac{6}{25}\right)^{5-4}$$

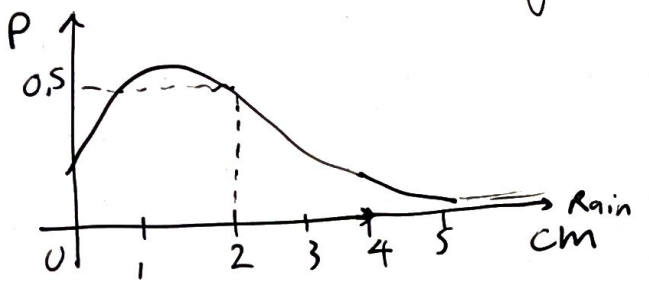
$= 0.0126$ with replacement!

without replacement 0.00536

Probability density function Section 4.1

Continuous RVs need these to express data.

eg Y = exact amount of rain tomorrow (cm)



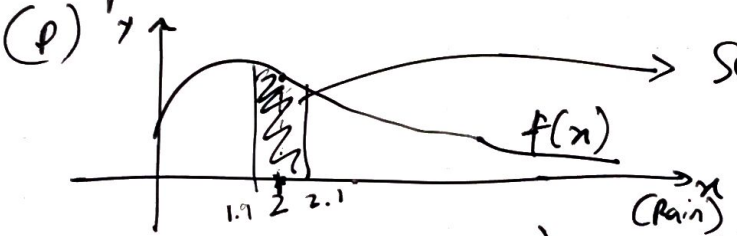
$P(Y=2) = ?$

~~$P(Y=2) = 0.5$~~ Wrong!

we can't be that accurate!

So, we make bounds for Y's value like

$$\text{★} \quad P(|Y-2|<0.1) \Rightarrow P(1.9<Y<2.1)$$



So, we need to measure area under this curve (shaded)

$$P(1.9<Y<2.1) = \int_{1.9}^{2.1} f(x)\,dx$$

★ also, $\int_0^{\infty} f(x)\,dx = 1$

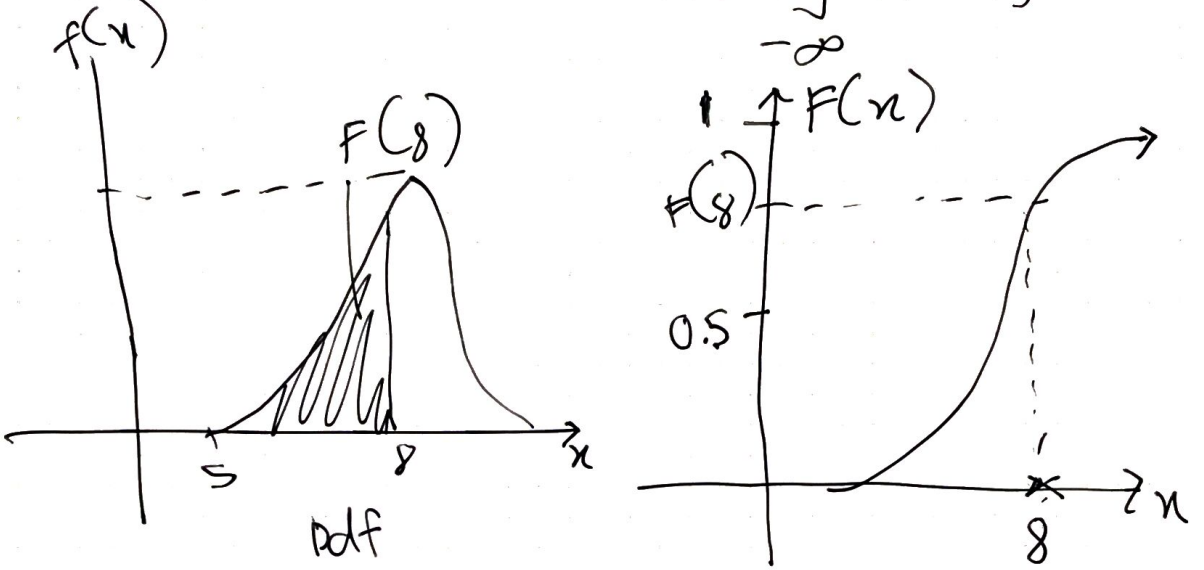
DEF: A continuous RV X is said to have a uniform distribution on the interval $[A,B]$ if pdf of x is

$$f(x;A,B) = \begin{cases} \frac{1}{B-A} & A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

4.2 Cumulative Distributive Functions & EV

CDF is just $F(x)$

$$F(x) = P(X \le x) = \int_{-\infty}^{\infty} f(y)\,dy$$



$$P(x > a) = 1 - F(a)$$

$$P(a \le X \le b) = F(b) - F(a)$$

$$\boxed{F'(x) = f(x)}$$

Percentiles

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\,dy$$

$(100\cdot p)^{th}$ percentile $\rightsquigarrow \eta(p)$

$$\mu_x = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx$$

$$\sigma_x^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx$$

$$\sigma_x = \sqrt{V(X)} \qquad = E[(X-\mu)^2]$$

$$V(X) = E(X^2) - [E(X)]^2$$

Normal Distribution

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

if $\mu=0, \sigma=1$, => standard normal distribution

& we have standard normal RV => Z

pdf of $Z = f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}}$

cdf of z $p(Z \le z) = \int_{-\infty}^{x} f(y;0,1)\,dy$

which is denoted by

$\Phi(z)$

$p\{a \le Z \le b\} = \Phi(b) - \Phi(a)$

& $\Phi(\infty) = 1$

$Z = \frac{X-\mu}{\sigma}$ by $p(Z \le z) = \int_{-\infty}^{z} \frac{1}{2\pi} e^{\frac{-y^2}{2}}\,dy = \phi(z$

$E[X] = \mu$ & $V[X] = \sigma^2$

Gamma Distribution & Exponential Distribution

Expo for param $\lambda > 0$ if pdf is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \frac{1}{\lambda} \qquad V(X) = \frac{1}{\lambda^2}$$

Gamma

for $\alpha > 0$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

properties

1) $\alpha > 0$ $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
2) any int n $\Gamma(n) = (n-1)!$
3) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

if $\alpha > 0$, $\beta > 0$, if pdf(x) =

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & \text{if } x \\ 0 & \text{otherwise} \end{cases}$$

- $f(x;\alpha,\beta)$ is standard gamma

$$E(X) = \mu = \alpha\beta$$

$$V(X) = \sigma^2 = \alpha\beta^2$$

in standard gamma dist,

$$F(x;\alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy$$

X be a gamma dist,

$$P(X \leq x) = F(x;\alpha,\beta) = F\left(\frac{x}{\beta};\alpha\right]$$

CHI SQ distribution

v = int, greater 0

if pdf of x is gamma density with $\alpha = v/2$ & $\beta = 2$

pdf of chi-sq rv is

$$f(x;v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} x^{\frac{v}{2}-1} e^{\frac{x}{2}} & \text{if } x \\ 0 & \text{otherwise} \end{cases}$$

WEIBULL DIST

if $\alpha > 0$, $\beta > 0$ if pdf of x is

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} & \text{if } x \geq \\ 0 & \text{otherwise.} \end{cases}$$

we got weibull dist

cdf of weibull is

$$F(x;\alpha,\beta) = \begin{cases} 1 - e^{\left(\frac{-x}{\beta}\right)^{\infty}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

LOG NORMAL DIST

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sigma x \sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} & \text{if } x \\ 0 & \text{otherwise} \end{cases}$$

we have lognormal dist.

$$E(x) = e^{\mu+\frac{\sigma^2}{2}} \quad V(x) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

$$F(x;\mu,\sigma) = p(X \leq x) = p(Y \leq Y) = p(Z \leq z)$$

$$= p\left(Z \geq \frac{\ln x - \mu}{\sigma}\right) = \phi\left(\frac{\ln x - \mu}{\sigma}\right)$$