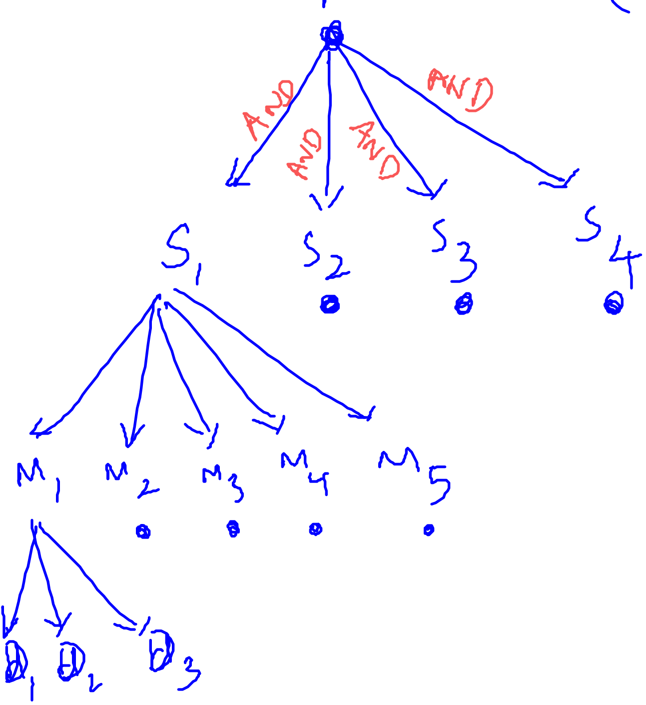
# Lecture - 2

How to count?

E.g. 3 restaurants, 4 different soups, 5 main courses and 3 desserts.

The role of the words AND / OR

$R_1$ (or) $R2$ (or) $R3$




One possible choice: $\underline{R_1 S_1 M_1 D_1}$

Another possible choice: $\underline{R_1 S_1 M_1 D_2}$

⋮

| Restaurant | Soup | Main course | Dessert |
|---|---|---|---|
| 3 | 4 | 5 | 3 |

$$3 * 4 * 5 * 3$$

$3$ → $R1$ (or) $R2$ (or) $R3$ $= 1+1+1 = 3$

No. of choices $= (1+1+1) * (1+1+1+1)$

$* (1+1+1+1+1)$

$* (1+1+1) = 180$

Sampling (Selection) process

How many ways can we select $r$ objects out of $n$ distinct objects?

$S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$

Select $\boxed{3}_{r}$ out of $\boxed{10}_{n}$ students for the following:

| | | E.g | |
|---|---|---|---|
| Scenario-1<br>$10 * 10 * 10 = 10^3$ | To answer 3 labeled questions such that any student can be selected more than once | $S_1, S_2, S_3$<br>$S_1, S_1, S_3$<br>$S_2, S_1, S_3$<br>⋮ | with replacement and order matters |
| Scenario-2<br>$\frac{10 * 9 * 8 * 7 * 6 * \cdots * 1}{7 * 6 * \cdots * 1} = \frac{10!}{7!} = \frac{10!}{(10-3)!} = {}^{10}P_3$ | To answer 3 labeled questions such that any student should be selected once only. | $S_1, S_2, S_3$<br>$\cancel{S_1, S_1, S_3}$<br>$S_2, S_1, S_3$<br>⋮ | without replacement and order matters. |

| | | e.g. | |
|---|---|---|---|
| Scenario-3 $\frac{10!}{3!(10-3)!} = {}^{10}C_3$ | to answer 3 unlabeled questions such that any student can be selected once only. | $S_1 S_2 S_3$ ~~$S_2 S_3 S_1$~~ ~~$S_1 S_1 S_2$~~ | without replacement and order does not matter |
| Scenario-4 | to answer 3 unlabeled questions such that any student can be selected more than once | $S_1 S_2 S_3$ $S_1 S_1 S_2$ ~~$S_1 S_2 S_1$~~ ⋮ | with replacement and order does not matter |

$\boxed{\text{Answer for Scenario-3}}$ AND $\boxed{\text{no. of ways to Arrange the 3 students}}$

$= \boxed{\text{Answer for Scenario-2}}$

$\boxed{\text{Answer for Scenario-3}}$

E.g.

$\boxed{S_1 \quad S_2 \quad S_3}$

$S_1\ S_2\ S_3$

$S_1\ S_3\ S_2$

$S_2\ S_1\ S_3$

$S_2\ S_3\ S_1$

$S_3\ S_1\ S_2$

$S_3\ S_2\ S_1$

$\boxed{S_1 \quad S_4 \quad S_5}$

⋮

$$\boxed{\text{Answer for Scenario - 3}} * \boxed{\text{no. of ways to arrange 3 distinct students}} = \frac{10!}{(10-3)!}$$

$\Downarrow$

Same as

$$\boxed{\text{The no. of ways to select 3 students out of 3 distinct students without replacement and order matters.}} = \frac{3!}{(3-3)!} = 3!$$

Answer for scenario-3 $* \; 3!$ = Answer for scenario-2

$$= \frac{10!}{(10-3)!}$$

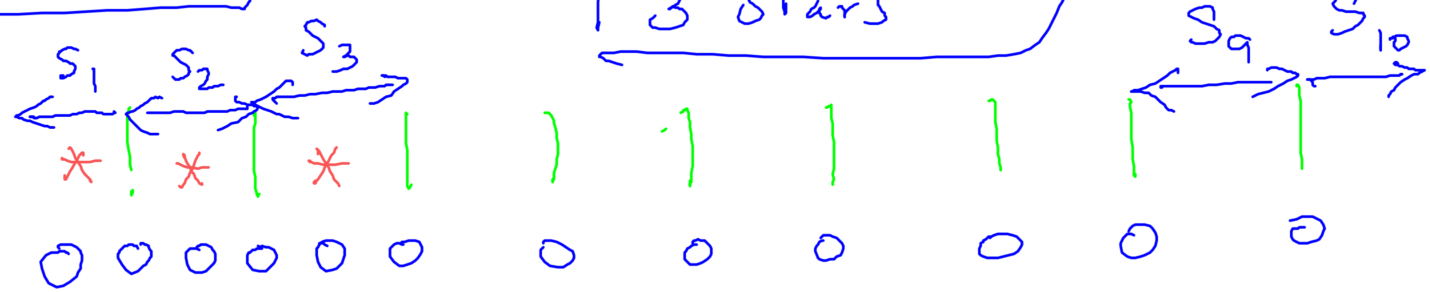
$\Rightarrow$ Answer for scenario-3 $= \dfrac{10!}{3!\,(10-3)!} = \begin{cases} {}^{10}C_3 \\ \dbinom{10}{3} \end{cases}$

For scenario-4

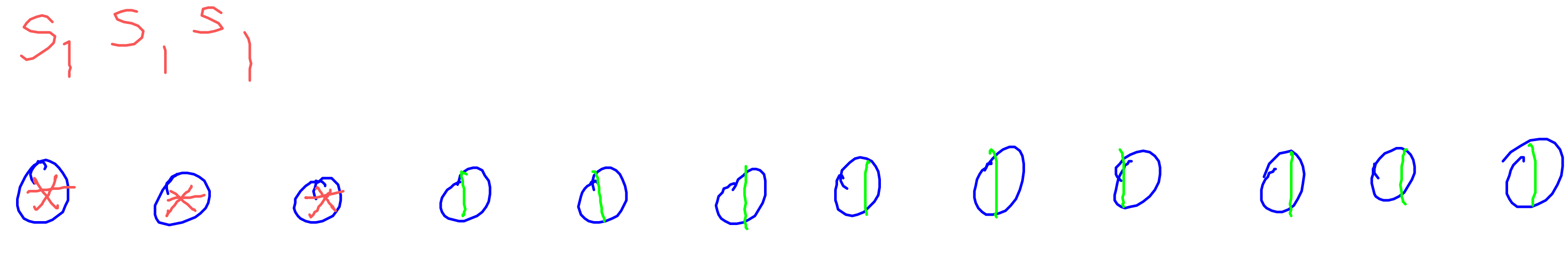
9 = (10-1) vertical bars

3 stars

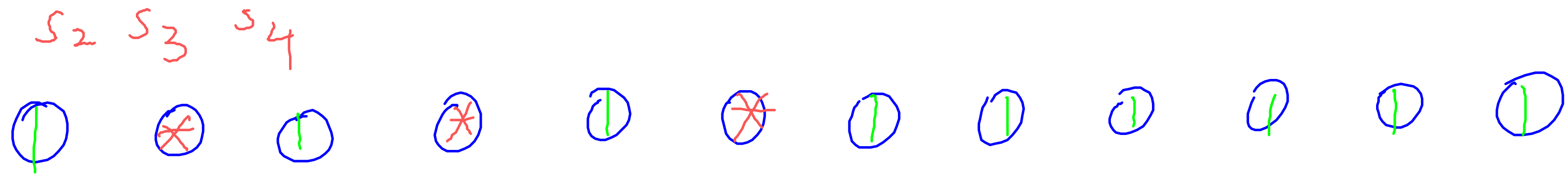
12 slots



E.g.




E.g.




= No. of ways to select 3 slots out of 12 slots $\left(\underbrace{10-1}_{\text{bars}} + \underbrace{3}_{\text{stars}}\right)$ without replacement and order does not matter $= {}^{12}C_3$

$$= {}^{(10-1+3)}C_3$$

**Problem** 4 passengers → 4 origins, 4 destinations

How many distinct (legal) routes are possible?

One possible (legal) route: $o_1\ o_2\ o_3\ o_4\ d_1\ d_2\ d_3\ d_4$

One possible (illegal) route: $d_1\ o_1\ o_2\ d_2\ o_3\ d_3\ o_4\ d_4$

Legal route $\boxed{o_1\ d_1\ o_2\ d_2\ o_3\ d_3\ o_4\ d_4}$ ⟷

All possible routes derived from one legal route

| |
|---|
| $o_1\ d_1\ o_2\ d_2\ o_3\ d_3\ o_4\ d_4$ |
| $d_1\ o_1\ o_2\ d_2\ o_3\ d_3\ o_4\ d_4$ |
| $o_1\ d_1\ d_2\ o_2\ o_3\ d_3\ o_4\ d_4$ |
| ⋮ |

$\boxed{\text{Total no. of legal routes}} * \boxed{\;?\;} = \boxed{\text{Total no. of routes possible}}$

We want to calculate this.

How many ways can we arrange

$o_1\ o_2\ o_3\ o_4\ d_1\ d_2\ d_3\ d_4$ ?

is same as No. of ways to select 8 out of 8 objects without replacement and order matters $= \dfrac{8!}{(8-8)!} = 8!$

How many ways can we distribute 4 distinct objects into 2 distinct bins? $= 2^4$

$$n^r = \begin{cases} \bullet \text{ no. of ways to select } r \text{ out of } n \text{ distinct objects with replacement and order matter (Scenario - 1)} \\ \bullet \text{ no. of ways to distribute } r \text{ distinct objects into } n \text{ distinct bins} \end{cases}$$

4 passengers

Total no. of legal routes $= \dfrac{8!}{2^4} = \dfrac{(2 \times 4)!}{2^4}$

$(2n)! > 2^n$

2 passengers

$2^2$

$o_1 \, d_1 \, o_2 \, d_2$

$o_1 \, d_1 \, o_2 \, d_2$

$d_1 \, o_1 \, o_2 \, d_2$

$o_1 \, d_1 \, d_2 \, o_2$

$d_1 \, o_1 \, d_2 \, o_2$

$o_1 \, o_2 \, d_1 \, d_2$

$o_1 \, o_2 \, d_1 \, d_2$

$d_1 \, o_2 \, o_1 \, d_2$

$o_1 \, d_2 \, d_1 \, o_2$

$d_1 \, d_2 \, o_1 \, o_2$

4

$= 4!$

$\frac{4!}{2!}$

$o_2 \, o_1 \, d_1 \, d_2 \longrightarrow d_2 \, d_1 \, o_1 \, o_2$

Total no. of legal routes with 4 passengers

$$= \frac{(2 * 4)!}{2^4}$$

Another approach

$N(K)$ = no. of legal routes possible with $K$ passengers.

$$N(1) = 1$$

claim $$N(K+1) = N(K) * \left[ \binom{2K+1}{1} + \binom{2K+1}{2} \right]$$

# Lab Notes APS BDA

(1) Random experiment

(2) Trial (simulating the random experiment once)

(3) Outcome of a random experiment

(4) Sample space corresponding to a random experiment

(5) Sampling space corresponding to a random experiment

(6) Event corresponding to a random experiment

Eg. Random experiment of tossing 2 fair coins

Sample space $S = \{(H, H), (H, T), (T, H), (T, T)\}$

Sampling space $S = \{H, T\}$

No. of outcomes in the sample space = no. of ways to select 2 objects out of the sampling space $S$ with replacement and order matters

$n^r = 2^2$

E.g. Random experiment of tossing 2 fair coins

order matters

$$S = \{(H,H), (H,T), (T,H), (T,T)\}$$

$\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

$$s = \{H, T\}$$

Sampling with replacement order matters

$$= 2^2$$

Order does not matter

$$S = \{(H,H), (H,T), (T,T)\}$$

$\frac{1}{4}$ $\frac{1}{2}$ $\frac{1}{4}$

$$s = \{H, T\}$$

Sampling with replacement order does not matter

$$= \binom{n+r-1}{r} = \binom{2+2-1}{2} = \binom{3}{2} = 3$$

Eg. Random experiment of rolling two fair dice

order matters

Order does not matter

$\binom{n+r-1}{r} = \binom{6+2-1}{2} = 21$

Sampling 2 out of 6 objects with replacement, order does not matter

$$S = \left\{ \begin{matrix} (1,1), (1,2), \ldots, (1,6) \\ (2,1), (2,2), \ldots, (2,6) \\ \vdots \\ (6,1), (6,2), \ldots (6,6) \end{matrix} \right\}$$

$$S = \left\{ \begin{matrix} (1,1), (1,2), \ldots (1,6) \\ (2,2), \ldots (2,6) \\ \ddots \\ (6,6) \end{matrix} \right\}$$

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

Selecting 2 out of 6 objects with replacement and order matters $= 6^2$

## Event

Random experiment of tossing 2 fair coins

$$S = \{(H,H), (H,T), (T,H), (T,T)\}$$

- Events are subsets of the sample space $S$

- Define the events
$$\begin{cases} E = \text{getting at least one heads} \quad (\geq 1) \\ F = \text{getting at least one tail} \\ G = \text{getting at most one tail} \quad (\leq 1) \end{cases}$$

$$E = \{(H,H), (H,T), (T,H)\}, \quad G = \{(H,T), (T,H), (H,H)\}$$

$$F = \{(T,T), (H,T), (T,H)\}$$

E.g.

Sample space $\Rightarrow$ $S = \{ (0g), (0f), (0s), (1g), (1f), (1s) \}$

Insured

(or)

Not insured

$\boxed{\text{AND}}$

Wound condition

good

(or)

fair

(or)

serious

$2 * 3 = 6$

Is the sampling space $S = \{0, 1, g, f, s\}$ ?

$\boxed{\text{Two Sampling Spaces}}$ $\qquad\qquad$ $\boxed{\text{Sample Space}}$

$S_1 = \{0, 1\}$ , $S_2 = \{g, f, s\}$

$$S = \left\{ \begin{matrix} (0g), (0f), (0s) \\ (1g), (1f), (1s) \end{matrix} \right\}$$

$\binom{2}{1} * \binom{3}{1} = 6$

A = Event that the patient is in serious condition

$= \{ (0s), (1s) \}$

$F$ = event that patient is uninsured and serious

$= \{ os \}$

$B$ = event that patient is uninsured

$= \{ og, of, os \}$

---

Eg. $E$ = event that patient is serious

$F$ = event that patient is uninsured

$$E = \{ \underline{(0S)}, \underline{(1S)} \}.$$

$$F = \{ (0g), (0f), (0S) \}$$

<u>Complimentary events</u>

$$\underline{E}^{c} = \text{NOT}\,(\text{event that patient is serious})$$

$$= \text{event that patient is not serious}$$

$$= \{ (0g), (0f), (1g), (1f) \}$$

# Compound Events

(1) Union

$$E \cup E^c = \{ (0s), (1s), (0g), (0f), (1g), (1f) \} \quad S \text{ (Sample Space)}$$

$E$ has outcomes, $\cup$ OR, $E^c$ has outcomes

(2) Intersection

$$E \cap E^c = \{ \varnothing \}$$

$\cap$ AND

$B^c$ = event that patient is insured.

$A$ = event that patient is serious

$$\boxed{B^c \cup A} = \{(1g),(1f),(1s)\} \cup \{(0s),(1s)\}$$

$$= \{(1g),(1f),(1s),(0s)\}$$

Event that patient is insured or serious

## Complement of compound events

$E$ = I have pizza

$F$ = I have curd rice

$E \cup F$ = I have pizza or curd rice

$(E \cup F)^c$ = I do not have pizza and I do not have curd rice

$$\boxed{(E \cup F)^c = E^c \cap F^c}$$

$E^c \Downarrow$ I do not have pizza

$F^c \Downarrow$ I do not have curd rice

$$(E \cap F)^c = E^c \cup F^c$$

---

**Digression** Negate this statement:

"If all rich people are happy then all poor people are sad".

| Word | Negation of the word |
|---|---|
| (100) All | $\rightarrow$ Not all (0-99) |
| (0) None | Some / at least one (1-100) |

E.g. All students who went to coaching did well on the entrance test

negate

Negation

Equivalent

Not all students who went to coaching did well on the entrance test

At least one student who went to coaching did not do well on the exam

## Mutually exclusive events

E.g. $S = \left\{ \begin{array}{l} (0g), (0f), (0s) \\ (1g), (1f), (1s) \end{array} \right\}$ Sample space

### Three events

$A$ = patient in good condition $= \{0g, 1g\}$

$B$ = patient in fair condition $= \{0f, 1f\}$

$C$ = patient in serious condition $= \{0s, 1s\}$

No outcomes are common to $A, B, C$

## Not mutually exclusive events

$E$ = patient is insured $= \{ IG, IF, IS \}$

$F$ = patient is serious $= \{ OS, IS \}$

$E \cap F$ = outcomes that are common to $E$ and $F$

$= \{ IS \}$, the box is not empty

Now we go back to A, B, C.

- $A \cap B = \{\phi\}$, $A \cap C = \{\phi\}$

  and $B \cap C = \{\phi\}$.

- $A \cap B \cap C = \{\phi\}$

A and B are mutually exclusive
A and C are mutually exclusive
B and C are mutually exclusive

A, B, C are also mutually exclusive

$$A \cup B \cup C = \left\{ \begin{matrix} og, & ig, \\ of, & if, \\ os, & is \end{matrix} \right\} = \text{Sample Space } S$$

$\Downarrow$

$A, B, C$ are mutually exclusive

and collectively exhaustive

$\Downarrow$

The union of the events gives us the

Sample space

All outcomes are equally likely intuitively

$S = \{1, 2, 3, 4, 5, 6\}$, Sample Space for rolling a 6-sided dice.

likelihood $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$

$$S = \left\{ \begin{matrix} 0g, 1g, \\ 0f, 1f, \\ 0s, 1s \end{matrix} \right\}$$

, Sample Space for the hospital administration problem

All outcomes are not equally likely intuitively

Assigning likelihood is a bit tricky.

Eg. Random experiment of tossing a single fair coin.

Theory

$$S = \{H, T\}$$

$$S = \{H, T\}$$

Sample 1 out of 2 objects without replacement and order does not matter

no. of ways $= \binom{2}{1} = 2$

Simulation

$$S = \{H, T\}$$

Sampling with replacement

# What is probability? → Theoretical

- A measure of how likely an event is to happen
- $S = \{(H,H), (H,T), (T,H), (T,T)\}$

  Likelihood $\quad 1/4 \quad 1/4 \quad 1/4 \quad 1/4$
- $E$ = event of getting at least one heads

$$= \{(H,T), (T,H), (H,H)\}$$

$$\quad 1/4 \quad 1/4 \quad 1/4$$

$P(E)$ = how likely is event $E$ going to happen

= how likely is an outcome from event $E$ going to happen $= 1/4 + 1/4 + 1/4$

$P(E) = 3/4$ (Remember, we said the outcomes in $S$ are equally likely).

## Frequency based (simulation) definition of probability

$$P(E) := \lim_{n_{simulations} \to \infty} \frac{n(E)}{n_{simulations}}$$

no. of times event E happened

no. of trials that are simulated

| Sampling space | Sample space |
| --- | --- |
| $S = (H, T)$ | $S = ((H,H), (H,T), (T,H), (T,T))$ |

with replacement, order matters

val

data

Simulated Data =

| | | | | | |
| --- | --- | --- | --- | --- | --- |
| H | T | T | H | H | T |
| T | H | T | H | T | H |

E.g. Random experiment of rolling a pair of fair die.

Sampling Space $S = \{1, 2, 3, 4, 5, 6\}$

Sampling 2 out of 6 objects with replacement and order matters

no. of ways $= 6^2 = 36$

Sample space

likelihood

$$S = \left\{ \begin{array}{cccc} \tfrac{1}{36} & \tfrac{1}{36} & \cdots & \tfrac{1}{36} \\ (1,1), & (1,2), & \cdots, & (1,6) \\ \tfrac{1}{36} & \tfrac{1}{36} & \cdots & \tfrac{1}{36} \\ (2,1), & (2,2), & \cdots, & (2,6) \\ \vdots & & & \\ \tfrac{1}{36} & \tfrac{1}{36} & & \tfrac{1}{36} \\ (6,1), & (6,2), & \cdots, & (6,6) \end{array} \right\}$$

no. of outcomes in S

E = event that the sum of the rolls is at least 7

$$= \left\{ \begin{array}{l} \text{likelihood} \quad \tfrac{1}{36} \quad \tfrac{1}{36} \quad \tfrac{1}{36} \quad \tfrac{1}{36} \quad \tfrac{1}{36} \quad \tfrac{1}{36} \\ (1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1) \\ (2,6),\ (3,5),\ (4,4),\ (5,3),\ (6,2) \\ \vdots \\ (6,6) \end{array} \right\}$$

$$P(E) = \frac{1}{36} * \text{no. of outcomes in event } E$$

$$= \frac{1}{36} * 21 = \frac{21}{36}$$

E.g. Faulty-worker problem

Who could have possibly broken the 5 machines

Sampling space $S = \{w_1, w_2, w_3, w_4, w_5\}$

Sampling 5 out of 5 objects with replacement and order matters.

no. of ways $= 5^5$

Sample space

All outcomes equally likely

$$S = \left\{ \begin{array}{l} (w_1, w_1, w_1, w_1, w_1), \\ (w_1, w_2, w_3, w_4, w_5) \\ \vdots \\ (w_3, w_5, w_5, w_5, w_3) \end{array} \right\}$$

no. of outcomes in the sample space

$E$ = event that a specific worker breaks at least 4 machines

Probability of interest = $P$(any specific worker breaks at least 4 machines)

$$P(E) = \frac{1}{5^5} * \text{no. of outcomes in } E$$

More than two compound events

$$(A \cup B \cup C)^c = A^c \cap B^c \cap C^c$$

$$(A \cap B \cap C)^c = A^c \cup B^c \cup C^c$$

---

Road map for sampling problems

$$\text{Sampling space} \xrightarrow{\text{build}} \text{Sample space}$$

How is the sample space built from the sampling space?

**with replacement**

E.g. Selecting 3 out of 10 icecream flavors where the flavors can repeat

$S = \{$ Vanilla, Strawberry, ..., Pista $\}$

$S = \{$ (vanilla, strawberry, choco), ⋮ $\}$

**without replacement**

E.g. Selecting 3 out of 7 days to have the probability lectures (one lecture per day atmost)

$S = \{M, T, W, Th, F, S, Su\}$

$S = \{(MWF), (MWS) \ldots\ldots\}$

Outcomes

$(f_1\ f_2\ f_3)$

$(f_3\ f_2\ f_1)$

$(f_1\ f_1\ f_1)$

Ice-cream Sample space

order matters

~~$(f_1\ f_2\ f_3)$~~

~~$(f_1\ f_1\ f_1)$~~

~~Order does not matter~~

**E.g.** Random experiment of tossing a pair of fair dice. What is the probability that the sum on the dice is atleast 7?

$S = \{1, 2, 3, 4, 5, 6\}$

likelihood

order matters

$$S = \begin{Bmatrix} (1,1) & (1,2) & \cdots & (1,6) \\ \tfrac{1}{36} & \tfrac{1}{36} & \cdots & \tfrac{1}{36} \\ (2,1) & (2,2) & \cdots & (2,6) \\ \tfrac{1}{36} & \tfrac{1}{36} & \cdots & \tfrac{1}{36} \\ \vdots & & & \\ (6,1) & (6,2) & \cdots & (6,6) \\ \tfrac{1}{36} & \tfrac{1}{36} & \cdots & \tfrac{1}{36} \end{Bmatrix}$$

No. of outcomes

$$= 6^2 = 36$$

order does not matter

$$S = \begin{Bmatrix} \tfrac{1}{36} & \tfrac{2}{36} & \cdots & \tfrac{2}{36} \\ (1,1) & (1,2) & \cdots & (1,6) \\ & (2,2) & \cdots & (2,6) \\ & & \ddots & \vdots \\ & & & (6,6) \\ & & & \tfrac{1}{6} \end{Bmatrix}$$

likelihood

No. of outcomes

$$= \binom{6+2-1}{2} = 21$$

Event of interest (sum $\geq 7$)

(i) <u>for order matters</u>, $E = \{ (1,6)^{\,1/36},\ (2,5)^{\,1/36}, (2,6)^{\,1/36},\ (3,4)^{\,1/36}\ (3,5)^{\,1/36}\ (3,6)^{\,1/36},\ (4,3)^{\,1/36}, (4,4)^{\,1/36}, (4,5)^{\,1/36}\ (4,6)^{\,1/36},\ (5,2), (5,3)\ (5,4)\ (5,5)\ (5,6),\ (6,1)\ (6,2)\ (6,3)\ (6,4)\ (6,5)\ (6,6) \}$

likelihood

$$P(E) = \frac{1}{36} * \text{no. of outcomes in } E$$

$$= \frac{1}{36} * 21 = \frac{21}{36}$$

(2) <u>for order does not matter</u>, $E = \{ (1,6),\ (2,5), (2,6),\ (3,4)\ (3,5)\ (3,6),\ (4,5)\ (4,6),\ (5,6),\ (6,6) \}$

$$P(E) = \frac{1}{36} * 3 + \frac{2}{36} * 9 = \frac{21}{36}$$

If sampling (from the sampling space) space is with replacement, then use order matters to build the sample space

E.g.

sampling space

$S = \{M, T, W, Th, F, S, Su\}$

no. of outcomes $= {}_7P_3 = \frac{7!}{(7-3)!} = 210$

order matters

$S = \{ (M\,T\,W), (M\,T, Th), \ldots (F, S, Su) \}$

$\frac{1}{210}$ $\frac{1}{210}$ $\frac{1}{210}$

order does not matter

$S = \{ (M\,T\,W), \cancel{(W\,T\,M)}, \ldots (F, S, Su) \}$

$\frac{1}{35}$ $\frac{1}{35}$ $\frac{1}{35}$

no of outcomes $= \binom{7}{3} = 35$

Event of interest $E$ = outcomes corresponding to lectures scheduled on a weekday

$$P(E) = \begin{cases} \boxed{\text{Order matters : } \frac{1}{210} * \text{no. of outcomes in } E} \\ \boxed{\text{Order does not matter} = \frac{1}{35} * \text{no. of outcomes in } E} \end{cases}$$

No. of outcomes in E =

Order matters

no. of ways to select 3 days out of 5 days (weekdays) without replacement and order matter

$= {}^5P_3$

Order does not matter

no. of ways to select 3 days out of 5 days (weekdays) without replacement and order does not matter $= {}^5C_3$

---

P(E) =

order matters

$\frac{1}{{}^7P_3} * {}^5P_3$

order does not matter

$\frac{1}{{}^7C_3} * {}^5C_3$

- $\dfrac{1}{{}_7P_3} * {}_5P_3 = \dfrac{1}{\dfrac{7!}{(7-3)!}} * \dfrac{5!}{(5-3)!}$

- $\dfrac{1}{{}_7C_3} * {}_5C_3 = \dfrac{1}{\dfrac{7!}{\cancel{3!}\,(7-3)!}} * \dfrac{5!}{\cancel{3!}\,(5-3)!}$

Equal

When sampling is with replacement and order is not specified naturally, use order matters to build the sample space because this makes the outcomes in the sample space equally likely.

When sampling is without replacement and order is not specified naturally, use order does not matter as it simplifies the calculation.

E.g. Ten fair coins are tossed. What is the probability that we get exactly three heads?

Sampling space $= \{H, T\}$ $\xrightarrow[\text{order matters}]{\text{with replacement}}$ Sample space $= \left\{ \begin{array}{c} \overset{1/2^{10}}{(HTHTHTHTHT)} \\ \vdots \\ \underset{1/2^{10}}{(HHHHHHHHHH)} \end{array} \right\}$

no. of outcomes in $S$ = no. of ways to select 10 objects out of 2 with replacement and order matters $= 2^{10}$

Event $E$ = event of getting three heads

$$= \left\{ \begin{array}{l} (H\ H\ H\ T\ T\ \overset{1/2^{10}}{T}\ T\ T\ T\ T) \\ (H, T, H, T, H, \overset{1/2^{10}}{T}, T\ T\ T\ T) \\ \vdots \end{array} \right\}$$

no. of outcomes in event $E$ = no. of ways to select 3 out of 10 slots to put an H (without replacement and order does not matter)

$$= {}_{10}C_3$$

$$P(E) = \frac{1}{2^{10}} * {}^{10}C_3$$

$$= {}^{10}C_3 * \left(\frac{1}{2}\right)^3 * \left(1 - \frac{1}{2}\right)^{10-3}$$

no. of coins tossed

no. of heads

likelihood that a coin shows up heads

likelihood that a coin shows up tails

# Bus-ridership analysis

Model:

(1) At each stop, each passenger is likely to get off the bus with a 20% chance.

(2) At every stop there is a 50% / 40% / 10% chance of 0/1/2 passengers getting on board.

(3) Bus never gets full; new passengers at any stop can always be accommodated

(4) Bus is empty when it arrives at the first stop.

- How likely is it for the bus to be empty after visiting the 10th stop?

Simulation

- How likely is it that no passengers board the bus at the first three stops?

$P(B_1=0 \cap B_2=0 \cap B_3=0)$

$L_i$ = random no. of passengers getting off at the ith stop

$B_i$ = random number of passengers boarding at the ith stop.

$$P\left(L_2 = 0\right)$$

$$= P\Big(\underline{L_2 = 0 \cap B_1 = 0} \cup \underline{L_2 = 0 \cap B_1 = 1}$$

$$\cup\ \underline{L_2 = 0 \cap B_1 = 2}\Big)$$

E.g. Consider rolling a pair of fair dice

Event B = first roll is even = $\{(2,1),(2,2),\cancel{(2,3)},(2,6)$
$(4,1)\ (4,2)\ \ldots\ (4,6)$
$(6,1)\ (6,2),\ \ldots\ (6,6)\}$

Event A = sum of the rolls is atleast seven
$=\{(1,6),(2,5),(2,6),(3,4),(3,5)$
$(3,6)\ \ldots\}$

Probabilities $P(A \cap B)$ compound event $=\{(2,5),(4,3),\ldots\}$

$P(A \mid B)$ given

conditional event.

A conditioned on B

A given B

The difference between an 'AND' and a 'conditional' probality?

A = event that a person is a smoker

B = event that a person has cancer

Person is a smoker and has cancer

$P(A \cap B) =$

$P(A | B) =$

Person is a smoker given that he/she has cancer.

Eg. Random experiment of rolling a single fair die

$$S = \{ \overset{1/6}{1}, \overset{1/6}{2}, \overset{1/6}{3}, \overset{1/6}{4}, \overset{1/6}{5}, \overset{1/6}{6} \}$$

→ likelihoods (equally likely outcomes)

Events of interest:

A = roll is at least equal to 3

B = roll is even

$A = \{3, \underline{4}, 5, \underline{6}\}$, $B = \{2, \underline{4}, \underline{6}\}$

$$A \cap B = \{\overset{1/6}{4}, \overset{1/6}{6}\}$$

$$\begin{cases} A = \{\overset{1/6}{3}, \overset{1/6}{4}, \overset{1/6}{5}, \overset{1/6}{6}\} \\ B = \{\overset{1/6}{2}, \overset{1/6}{4}, \overset{1/6}{6}\} \end{cases}$$

$$P(A \cap B) = \frac{1}{6} \times 2 = \frac{1}{3}$$

$P(A|B)$

$$S|_B = \{\overset{1/3}{2}, \overset{1/3}{4}, \overset{1/3}{6}\}$$

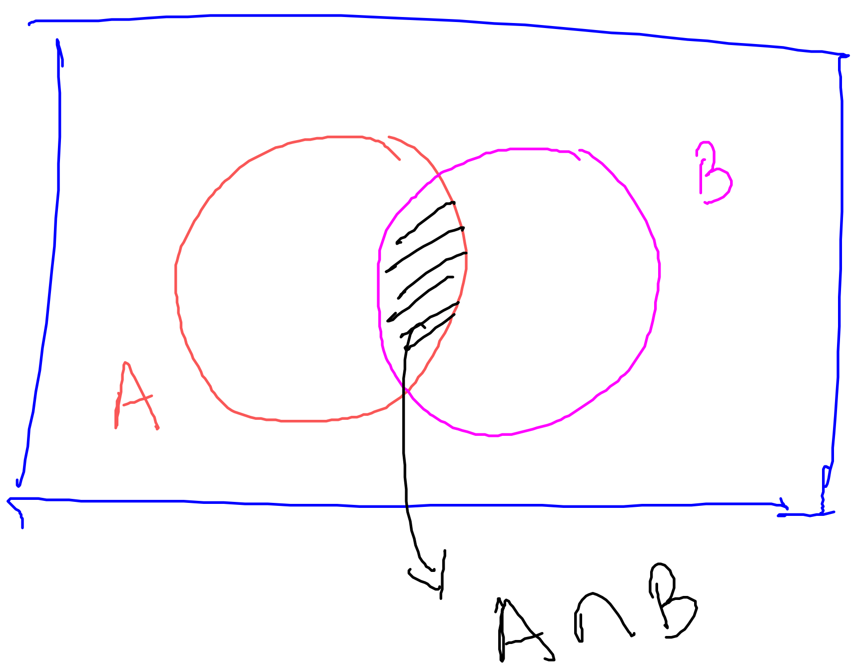
$$A|B = \{\overset{1/3}{4}, \overset{1/3}{6}\}$$

$$P(A|B) = \frac{1}{3} + \frac{1}{3} = 2/3$$

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

## Conditional probability

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



Random experiment: roll two fair dice

$$S = \left\{ \begin{matrix} \overset{1/36}{(1,1)}, & \cdots & \overset{1/36}{(1,6)} \\ \overset{1/36}{(2,1)}, & \cdots & ,\overset{1/36}{(2,6)} \\ \overset{1/36}{(6,1)}, & \cdots & ,\overset{1/36}{(6,6)} \end{matrix} \right\} = 36 \text{ outcomes}$$

$$A = \text{Sum is at least } 7 = \left\{ \begin{matrix} \overset{1/36}{(1,6)} \\ \overset{1/36}{(2,5)}, \overset{1/36}{(2,6)} \\ \overset{1/36}{(3,4)}, \overset{1/36}{(3,5)}, \overset{1/36}{(3,6)} \\ \overset{1/36}{(4,3)}\ \overset{1/36}{(4,4)}, \overset{1/36}{(4,5)}, \overset{1/36}{(4,6)} \\ \vdots \end{matrix} \right\} \leftarrow 21 \text{ outcomes}$$

$$B = \text{first roll even} = \left\{ \begin{matrix} \overset{1/36}{(2,1)}, & \cdots & \overset{1/36}{(2,6)} \\ \overset{1/36}{(4,1)}, & \cdots, & \overset{1/36}{(4,6)} \\ \overset{1/36}{(6,1)}, & \cdots & \overset{1/36}{(6,6)} \end{matrix} \right\} \rightarrow 18 \text{ outcomes}$$

$$A \cap B = \left\{ \begin{matrix} \overset{1/36}{(2,5)}\ \overset{1/36}{(2,6)} \\ \overset{1/36}{(4,3)}, \overset{1/36}{(4,4)}, \overset{1/36}{(4,5)}, \overset{1/36}{(4,6)} \\ \overset{1/36}{(6,1)}, \overset{1/36}{(6,2)} \cdots \overset{1/36}{(6,6)} \end{matrix} \right\} \longrightarrow 12 \text{ outcomes}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{36} * 12}{\frac{1}{36} * 18}$$

$$= \frac{1/3}{1/2} = 2/3 \approx 0.66\ldots$$

---

Conditional probability and Bayes' formula

A = a person is a smoker

B = a person has cancer

$P(A \cap B)$

$P(A|B)$

$P(B|A)$

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

(definition of conditional probability)

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

(the same)

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

$$\parallel \qquad\qquad \parallel$$

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

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

$$\Rightarrow \boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}$$

First version of Bayes' formula

Note that $B$ = person has cancer

= (Person has cancer AND is smoker)

(OR)

(Person has cancer AND is not smoker)

Law of Total Probability

$$\boxed{\begin{aligned} B &= \underline{(B\cap A)}\ \underline{\underline{\cup}}\ \underline{(B\cap A^c)} \\ P(B) &= P(B\cap A) + P(B\cap A^c) \end{aligned}}$$

We use this in Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B\cap A) + P(B\cap A^c)}$$

$$P(A|B)$$

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

$$= \frac{P(B|A)\,P(A)}{P(B\cap A) + P(B\cap A^c)}$$

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

# Cellphone signal problem

→ observe 2 bars on the cellphone

$$P(\overbrace{\text{person is at some location } (i,j)}^{A} \mid \overbrace{\text{observe 2 bars}}^{B})$$

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$= \frac{P(\overbrace{\text{observe 2 bars}}^{B} \mid \overbrace{\text{Person is at } (i,j)}^{A})\,P(\overbrace{\text{Person is at } (i,j)}^{A})}{P(\underbrace{\text{observe 2 bars}}_{B})}$$

$$\frac{P(\text{observe 2 bars of signal} \mid \text{user is at position } (i,j)) * P(\text{user is at position } (i,j))}{P(\text{observe 2 bars of signal})}$$

$$= \frac{b_{ij} * a_{ij}}{P\left(\begin{array}{c}\text{observe 2 bars } \underline{\text{AND}} \text{ user at } (1,1) \\ \text{(OR)} \\ \text{observe 2 bars } \underline{\text{AND}} \text{ user at } (4,4)\end{array}\right)}$$

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$= \frac{P(B \mid A)\,P(A)}{P(B \cap A) + P(B \cap A^c)}$$

$$= \frac{P(B \mid A)\,p(A)}{P(B \mid A)\,P(A) + P(B \mid A^c)\,P(A^c)}$$

$$= \frac{b_{ij} * a_{ij}}{\left( \begin{array}{c} P(\overbrace{\text{observe 2 bars}}^{B} \text{ AND } \overbrace{\text{user at } (1,1)}^{A}) \\ + \\ P(\text{observe 2 bars AND user at } (1,2)) \\ + \\ \vdots \\ P(\overbrace{\text{observe 2 bars}}^{B} \text{ AND } \overbrace{\text{user at } (4,4)}^{A^c}) \end{array} \right)}$$

$$= \frac{b_{ij} * a_{ij}}{\left( \begin{array}{c} \underbrace{P(\overbrace{\text{observe 2 bars}}^{B} \mid \overbrace{\text{user at } (1,1)}^{A})}_{b_{11}} * \underbrace{P(\overbrace{\text{user at } (1,1)}^{A})}_{a_{11}} \\ + \\ \underbrace{P(\text{observe 2 bars} \mid \text{user at } (1,2))}_{b_{12}} * \underbrace{P(\text{user at } (1,2))}_{a_{12}} \\ + \\ \vdots \\ \underbrace{P(\text{observe 2 bars} \mid \text{user at } (4,4))}_{b_{44}} * \underbrace{P(\text{user at } (4,4))}_{a_{44}} \end{array} \right)}$$

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

$\Rightarrow P(A \cap B) = P(A|B)\,P(B)$

Definition of conditional probability

$$P\left(\text{user at } (i,j) \,\middle|\, 2 \text{ bars signal}\right) = \frac{b_{ij}\, a_{ij}}{b_{11} a_{11} + b_{12} a_{12} + \cdots + b_{44} a_{44}}$$

E.g.

- Say 60% of all email is spam
- 90% of spam emails have a forged header
- 20% of non-spam emails have a forged header
- Event E = email contains a forged header
- Event F = email is spam

What is $P(F|E)$?

$$P(\underbrace{F}_{\text{email is spam}} \mid \underbrace{E}_{\text{forged header}}) = \frac{P(\underbrace{E}_{\text{forged header}} \mid \underbrace{F}_{\text{spam}}) \times P(\overbrace{F}^{\text{spam}})}{P(\underbrace{E}_{\text{forged header}})}$$

$$= \frac{0.9 \times 0.6}{P\left(E \cap F \;\cup\; \underline{E \cap F^c}\right)}$$

$$P(F|E) = \frac{\overbrace{P(E|F)\,P(F)}^{0.9 \times 0.6}}{P(E\cap F) + P(E\cap F^c)}$$

$$\boxed{P(F|E) = \frac{P(E|F)\,P(F)}{P(E|F)\,P(F) + P(\underbrace{E}_{\text{forged header}}|\underbrace{F^c}_{\text{not spam}})\,P(\underbrace{F^c}_{\text{not spam}})}}$$

$$P(\overset{\text{spam}}{F}|\overset{\text{forged header}}{E}) = \frac{0.9 \times 0.6}{0.9 \times 0.6 + 0.2 \times (1-0.6)} \approx 0.87$$

E.g.

- 18.5% of people are smokers $(S)$
- 34.05% of people among smokers have cancer $(C \mid S)$
- 10.86% of people among non-smokers also have cancer $(C \mid S^c)$

$$P(\overset{S}{\text{smoker}} \mid \overset{C}{\text{cancer}}) = \, ?$$

# HIV Testing

- A test is 98% effective at detecting HIV $P(E|F) = 0.98$
  - However, test has a "false positive" rate of 1% $= P(E|F^c)$
  - 0.5% of US population has HIV $P(F) = 0.005$
  - Let E = you test positive for HIV with this test
  - Let F = you actually have HIV
  - What is P(F | E)?
- Solution:

$$P(F \mid E) = \frac{\overset{0.98}{P(E|F)}\,\overset{0.005}{P(F)}}{\underset{?}{P(E)}} = \frac{\overset{0.98}{P(E|F)}\,\overset{0.005}{P(F)}}{\underset{0.98}{P(E|F)}\,\underset{0.005}{P(F)} + \underset{0.01}{P(E|F^c)}\,P(F^c)}$$

$P(F^c) = 1 - P(F) = 1 - 0.005 = 0.995$

$P(F \mid E) = 0.33$

Positive test results cannot be trusted

# Why it's Still Good to Get Tested

| | HIV + | HIV – |
|---|---|---|
| **Test +** | 0.98 = P( $E \mid F$ ) | 0.01 = P( $E \mid F^c$ ) |
| **Test –** | 0.02 = P( $E^c \mid F$ ) | 0.99 = P( $E^c \mid F^c$ ) |

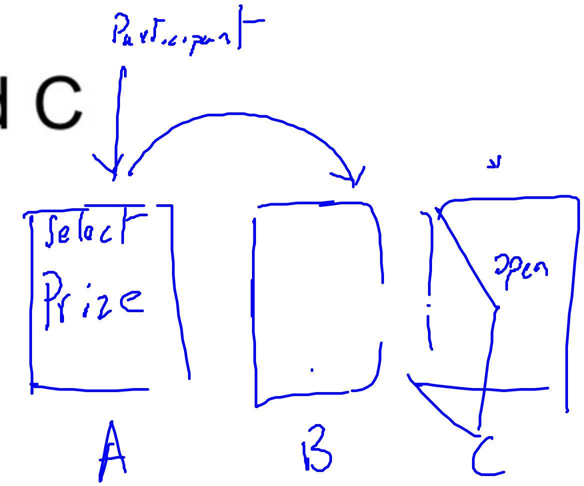
- Let $E^c$ = you test <u>negative</u> for HIV with this test
- Let F = you actually have HIV
- What is $P(F \mid E^c)$?

$P(F \mid E^c) =$ close to 0 = ?

$P(F^c \mid E^c) =$ close to 1

# Let's Make a Deal

- Game show with 3 doors: A, B, and C




  - Behind one door is prize (equally likely to be any door)
  - Behind other two doors is nothing
  - We choose a door
  - Then host opens 1 of other 2 doors, revealing nothing
  - We are given option to change to other door
- Should we?
  - Note: If we don't switch, P(win) = 1/3 (random)

# Let’s Make a Deal

- Without loss of generality, say we pick A
  - P(A is winner) =
    - Host opens either or , we by switching
    - P(win | A is winner, picked A, switched) =
  - P(B is winner) =
    - Host <u>must</u> open (can’t open A and can’t reveal prize in B)
    - So, by switching, we switch to and
    - P(win | B is winner, picked A, switched) =
  - P(C is winner) =
    - Host <u>must</u> open (can’t open A and can’t reveal prize in C)
    - So, by switching, we switch to and
    - P(win | C is winner, picked A, switched) =
  - Should always switch!
    - P(win | picked A, switched) =

E.g. A person is known to have atleast one girl child of the two children. What is the probability that the other child is a boy?

Simulate

$S = c('b', 'g')$

$p = c(0.5, 0.5)$

Simulated Data = B B G . . . . .
B G B

Filtered Simulated Data

Solve using a theoretical approach

using a reduced Sample Space

Conditional probability and Bayes formula

$P(\text{smoker} \mid \text{cancer})$

Fraction of smokers among those who have cancer

Probability that a random who is known to have cancer is a smoker.

$$S = \{BB, BG, GB, GG\}$$

$$S|_{A|G} = \{\overset{1/3}{BG}, \overset{1/3}{GB}, \overset{1/3}{GG}\}$$

$A|G$: at least one girl $\Rightarrow E|_{A|G} = \{\overset{1/3}{BG}, \overset{1/3}{GB}\}$

$$P(\underline{E} \mid \cancel{A} \mid \sigma) = \frac{2}{3}$$

event that
there is a brother-sister
pair

---

Recall definition of conditional probability

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

A: Smoker  B: Cancer

$\Downarrow$ definition of conditional probability

$\Downarrow$ 1st version of Bayes' formula

Second version of Bayes' formula using the law of total probability

$$\boxed{\begin{aligned} P(A \cap B) &= P(A|B)\,P(B) \\ P(B \cap A) &= P(B|A)\,P(A) \end{aligned}}$$

AND probabilities    Conditional Probabilities

$P(B) = P(\text{cancer})$

$= P(\text{Cancer AND Smoker OR Cancer AND NOT Smoker})$

Mutually exclusive

$= P(\text{Cancer AND Smoker}) + P(\text{Cancer AND NOT Smoker})$

---

$S = \{ \overset{1/4}{BB}, \overset{1/4}{BG}, \overset{1/4}{GB}, \overset{1/4}{GG} \}$, $A|G = \{ \overset{1/4}{BG}, \overset{1/4}{GB}, \overset{1/4}{GG} \}$

$$P\{1B1G \mid A1G\} = \frac{P(1B1G \text{ and } A1G)}{P(A1G)}$$

Event of a boy-girl pair (for $1B1G$)

Event of atleast one girl (for $A1G$)

$$= \frac{P(\{\overset{1/4}{BG}, \overset{1/4}{GB}\})}{P(\{\underset{1/4}{BG}, \underset{1/4}{GB}, \underset{1/4}{GG}\})} = \frac{2/4}{3/4} = 2/3$$

$$A1G = \{\underline{BG}, \underline{GB}, GG\}$$

$$1B1G = \{\underline{BG}, \underline{GB}\}$$

$$A1G \text{ and } 1B1G = \{BG, GB\}$$

# Independent Events

Recall conditional probability

$A$ = person is a smoker

$B$ = person is a male

$P(A)$ = prior probability

$P(A|B)$ = posterior probability

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

For e.g. in Manipal (student community)

$P(A) = 0.1$ (Prior probability)

$P(A|B) = 0.1$, this means the evidence $B$ (person is a male) adds no information to the prior probability,

If $P(A|B) = P(A)$, then $A$ and $B$ are independent events

- In this case, the events that the person is a smoker and the person is male are independent.

- $$\underbrace{\boxed{P(A|B) = P(A)}}_{\text{Independent events}} = \underbrace{\boxed{\frac{P(A \cap B)}{P(B)}}}_{\text{Conditional probability}}$$

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

$$\Rightarrow \boxed{P(A \cap B) = P(A)\,P(B) \text{ for independent events } A \text{ and } B}$$

Recall Generic formula for two events A and B

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

Generic formula for three events A, B, and C

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

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

(Generic formula)

A, B are mutually exclusive

$A\cap B = \phi$

$P(A\cap B) = 0$

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

'OR' becomes an addition when events are mutually exclusive

A, B are independent

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

$= P(A|B)P(B)$ a

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

'AND' becomes a multiplication when events are independent

Suppose A and B are mutually exclusive events:

Which of the following is true?

(1) A and B are independent

(2) A and B are dependent

Rolling a die and tossing a coin

$$S = \left\{ \begin{array}{ccccc} \overset{1/12}{(1H)}, & \overset{1/12}{(2H)}, & \ldots & \overset{1/12}{(6,H)} \\ \underset{1/12}{(1T)}, & \underset{1/12}{(2T)}, & \ldots & \underset{1/12}{(6,T)} \end{array} \right\}$$

$$A = \left\{ \underset{1/12}{(2H)}, \underset{1/12}{(4H)}, \underset{1/12}{(6H)} \right\}, \quad B = \left\{ \underset{1/12}{(1H)}, \underset{1/12}{(3H)}, \underset{1/12}{(5H)} \right\}$$

A and B are mutually exclusive (but not collectively exhaustive)

Prior

What is $P(A)$? $= \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{1}{4}$

What is $P(A|B)$? $= 0$ Posterior

B has influenced A $\Rightarrow$ A and B are dependent.

# Conditional probability with multiple conditioning

$$P(\overset{A}{\text{Smoker}} \mid \underbrace{\overset{B}{\text{Cancer}}, \overset{C}{\text{l.t. 20 yrs old}}}_{\text{AND}})$$

$$= \frac{P(\overset{B}{\text{cancer}} \mid \overset{A}{\text{smoker}}, \text{l.t. 20 yrs old}) \times P(\overset{A}{\text{Smoker}} \mid \underset{C}{\text{l.t. 20 yrs old}})}{P(\overset{B}{\text{cancer}} \mid \overset{C}{\text{l.t. 20 yrs old}})}$$

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

$C$ is simply an additional conditioning event.

# Law of total probability with multiple conditioning

## Recall insurance problem

P(accident)

$$= P\left(\begin{array}{c}\text{accident AND good driver} \\ \text{(OR)} \\ \text{accident AND avg. driver} \\ \text{(OR)} \\ \text{accident AND bad driver}\end{array}\right)$$

$$\underbrace{P(\text{accident}}_{A} \text{ AND } \underbrace{\text{good driver})}_{B}$$

$$\boxed{\begin{aligned} &P(A \cap B) \\ &= P(A|B)\,P(B) \end{aligned}}$$

$$+\, P(\text{accident AND avg. driver})$$

$$+\, P(\text{accident AND bad. driver})$$

$$= P(\underbrace{\text{accident}}_{A} \mid \underbrace{\text{good driver}}_{B}) \times P(\text{good driver})$$

$$+\, P(\text{accident} \mid \text{avg. driver}) \times P(\text{avg-driver})$$

$$+\, P(\text{accident} \mid \text{bad. driver}) \times P(\text{bad driver})$$

E.g.

$$P(\overset{A}{\text{Smoker}} \mid \overset{B}{\text{Cancer}})$$

$$= P(\text{Smoker AND l.t. 20 yrs old} \mid \text{Cancer})$$

$$+ P(\text{Smoker AND g.t. 20 yrs old} \mid \text{Cancer})$$

Law of total probability with additional conditioning

# The Monty Hall Problem (3-doors problem)

A B C

- $P(\text{win} \mid \text{choose a door initially (door-A), switch})$

$$= P\left(\text{win AND prize-A} \;\text{OR}\; \text{win AND prize-B} \;\text{OR}\; \text{win AND prize-C} \;\middle|\; \text{choose a door initially, switch}\right)$$

$$= P\left(\underline{\text{win AND prize-A}} \;\middle|\; \underbrace{\text{choose a door initially, switch}}_{(\text{door A})}\right)$$

$$+ P\left(\underline{\text{win AND Prize-B}} \;\middle|\; \underbrace{\text{choose a door initially, switch}}_{(\text{door A})}\right)$$

$$+ P\left(\underline{\text{win AND Prize-C}} \;\middle|\; \underbrace{\text{choose a door initially, switch}}_{(\text{door A})}\right)$$

$$= \underline{P\left(\text{win} \mid \text{Prize-A, choose A, switch}\right)}_{=0} * \underline{P\left(\text{Prize-A} \mid \cancel{\text{choose A, switch}}\right)}_{=1/3}$$

$$+ \underline{P\left(\text{win} \mid \text{Prize-B, choose A, switch}\right)}_{=1} * \underline{P\left(\text{Prize-B} \mid \cancel{\text{choose A, switch}}\right)}_{=1/3}$$

$$+ \underline{P\left(\text{win} \mid \text{prize-C, choose A, switch}\right)}_{=1} * \underline{P\left(\text{Prize-C} \mid \cancel{\text{choose A, switch}}\right)}_{=1/3}$$

$$= \frac{2}{3}$$

# Simulation

What happens in one trial?

Host decides which door to put the prize in: A

Participant selects a door A

Based on what the participant selects, host opens an empty door. B

Switch (or) not switch 1

Simulated Data

| | |
|---|---|
| A | A |
| A | A |
| B | C |
| 1 | 1 |

{B, C}

| | | | |
|---|---|---|---|
| Where does the host put the prize? | c('A', 'B', 'C') c(1/3, 1/3, 1/3) | A | A |
| Which door the participant selects? | c('A', 'B', 'c') c(1/3, 1/3, 1/3) | A | B |
| Which door the host opens? | c('B', 'C') c(1/2, 1/2) | B | c('c') c(1) C |
| Whether the participant switches or not? | c(0, 1) c(1/2, 1/2) | 1 | 0 |

Simulated Data =

| | | |
|---|---|---|
| A | A | . . . . . . . |
| A | B | |
| B | C | |
| 1 | 0 | |

# Negation (hashing problem)

**6 strings** $\quad s_1 \quad s_2 \quad s_3 \quad s_4 \quad s_5 \quad s_6$

**5 buckets** $\quad \boxed{p_1} \quad \boxed{p_2} \quad \boxed{p_3} \quad \boxed{p_4} \quad \boxed{p_5}$

$\qquad\qquad\quad\; b_1 \quad\; b_2 \quad\; b_3 \quad\; b_4 \quad\; b_5$

$E$ = All of buckets 1-4 have at least one string hashed to them

$$P(E) = 1 - P(E^c)$$

Complement of event $E$

| | Word | Negation |
|---|---|---|
| (1) | All (100) | Not all (0-99) |
| (2) | None (0) | Some / at least one (1-100) |

$E^c$ = Not all of buckets 1-4 have at least one string hashed to them.

= At least one of buckets 1-4 have none strings hashed to them.

$E^c$ = At least one of buckets 1-4 have no string hashed to them

$\boxed{F_i = \text{event that the ith bucket has at least one string hashed to it.}}$

$$E = F_1 \cap F_2 \cap F_3 \cap F_4$$

$F_1, F_2, F_3, F_4$ are not independent

$$P(E) = P(\overset{A}{\underline{F_1}} \cap \overset{B}{\underline{F_2 \cap F_3 \cap F_4}})$$

$$\overset{?}{=} \cancel{P(F_1) * P(F_2) * P(F_3) * P(F_4)}$$

$$
\begin{aligned}
P(E) &= 1 - P(E^c) \\
&= 1 - P\left(\left(F_1 \cap F_2 \cap F_3 \cap F_4\right)^c\right) \\
&= 1 - P\left(F_1^c \cup F_2^c \cup F_3^c \cup F_4^c\right) \\
&= 1 - \Big\{ P(F_1^c) + P(F_2^c) + P(F_3^c) + P(F_4^c) \\
&\qquad - P(F_1^c \cap F_2^c) - P(F_1 \cap F_3^c) - P(F_1^c \cap F_4^c) \\
&\qquad - P(F_2^c \cap F_3^c) - P(F_2^c \cap F_4^c) - P(F_3^c \cap F_4^c) \\
&\qquad + P(F_1^c \cap F_2^c \cap F_3^c) + P(F_1^c \cap F_3^c \cap F_4^c) + \cdots \Big\}
\end{aligned}
$$

$$\underline{P(F_1^c)} = (1-P_1)^6$$

$$P(\underline{F_1^c} \cap \underline{F_2^c}) \neq P(F_1^c) \times P(F_2^c)$$

$$\boxed{\begin{array}{c}\text{No string is hashed to bucket-1} \\ \text{AND} \\ \text{No string is hashed to bucket-2}\end{array}}$$

= String-1 is not hashed to bucket-1 <u>or</u> bucket-2

AND

→ String-2 is not hashed to bucket-1 <u>or</u> bucket-2

⋮

$$\Rightarrow P(F_1^c \cap F_2^c) = P\left(\begin{array}{l}\text{String-1 is not hashed} \\ \text{to bucket-1 or bucket-2}\end{array}\right) * P\left(\begin{array}{l}\text{String-2 is not hashed} \\ \text{to bucket-1 or bucket-2}\end{array}\right) * \cdots * P\left(\begin{array}{l}\text{String-6 is not hashed to} \\ \text{bucket-1 or bucket-2}\end{array}\right)$$

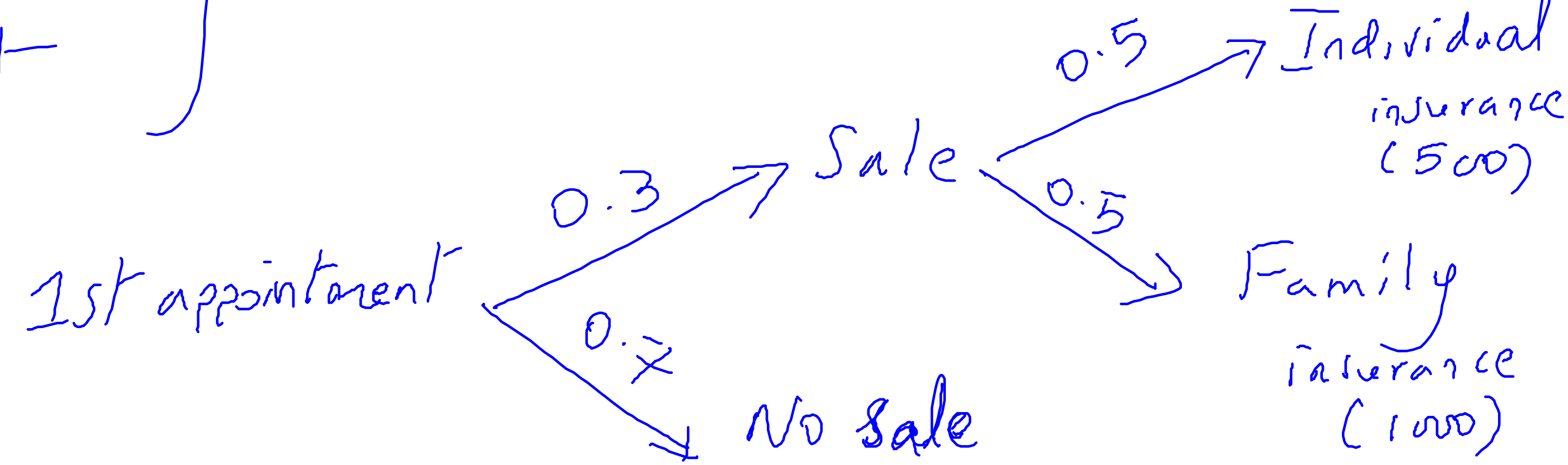
$$= \left[1 - (P_1 + P_2)\right] * \left[1 - (P_1 + P_2)\right] * \cdots * \left[1 - (P_1 + P_2)\right]$$

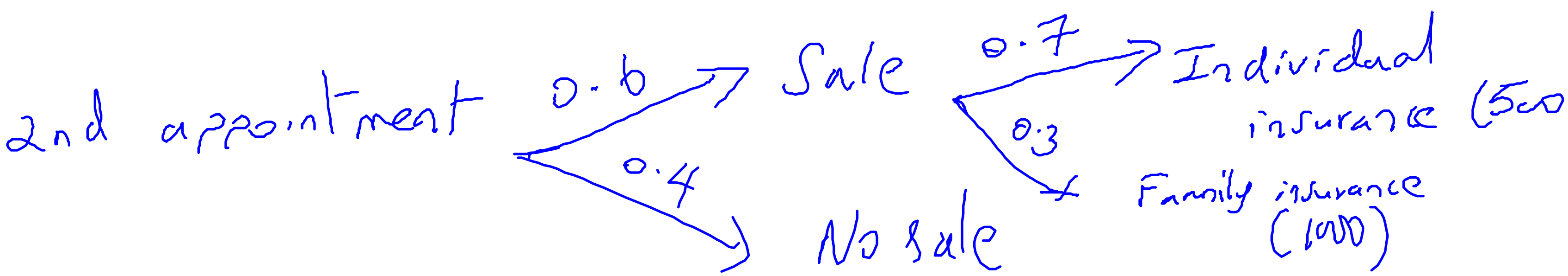
$$= \left[1 - (P_1 + P_2)\right]^6$$

$$P(F_1^c \cap F_2^c \cap F_3^c) = \left[1-(p_1+p_2+p_3)\right]^6$$

---

{Insurance Agent} has 2 appointments in a day



- How much the insurance agent earns in a day is random

$$\text{day is random} \longrightarrow \left.\begin{matrix}0\\500\\1000\\1500\\2000\end{matrix}\right\}\text{Possible earnings}$$

- $X$ = agent's earning per day (random variable)

- $P(X=0) = P(\text{agent earning nothing over a day})$

  $P(X=500) = P(\text{agent earns Rs-500 over a day})$

- $X = 0$ is an event

  $X = 500$ is another event

- E.g. $P(X=0) = P\left(\begin{array}{c}\text{Appointment-1 no sale} \\ \text{AND} \\ \text{Appointment-2 no sale}\end{array}\right)$

Assuming that the appointments are independent

$$\Rightarrow P(X=0) = P(\text{Appointment-1 no sale})$$
$$* P(\text{Appointment-2 no sale})$$
$$= 0.7 * 0.4 = 0.28$$

$$P(x = 500)$$

$= P\big($Appointment-1 is sale AND sells individual insurance

AND Appointment-2 is no sale

(OR)

Appointment-1 is no sale AND Appointment-2 is sale

AND sells individual insurance$\big)$

Mutually

exclusive

Independent

$= P\big($Appointment-1 is sale AND sells individual insurance

AND Appointment-2 is no sale$\big)$

$+ P\big($Appointment-1 is no sale AND Appointment-2 is sale

AND sells individual insurance$\big)$

$$= P(\overset{B}{\text{Appointment-1 is sale}} \text{ AND } \overset{A}{\text{sells individual insurance}})$$

$$* \, P(\text{Appointment-2 is no sale})$$

$$+ \, P(\text{Appointment-1 is no sale})$$

$$* \, P(\overset{B}{\text{Appointment-2 is sale}} \text{ AND } \overset{A}{\text{sells individual insurance}})$$

conditional probability definition

$$= P(\overset{A}{\text{Sells individual insurance}} \mid \overset{B}{\text{Appointment-1 is sale}})$$

$$* \, P(\overset{B}{\text{Appointment-1 is sale}}) * P(\text{Appointment-2 is no sale})$$

$+ P(\text{Appointment-1 is no sale})$

$$\times P(\underbrace{\text{Sells individual}}_{A}\ \text{insurance} \mid \underbrace{\text{appointment-2 is sale}}_{B})$$

$$\times P(\underbrace{\text{Appointment-2 is sale}}_{B})$$

$$= 0.5 \times 0.3 \times 0.4 + 0.7 \times 0.7 \times 0.6$$

# Expected value of a random variable

E.g. Insurance problem

$$E[X] = 0 * P(X=0) + 500 * P(X=500) + \dots + 2000 * P(X=2000)$$

$$E[X] \approx 0 * \frac{\text{no. of times } 0 \text{ appears}}{N \text{ simulations}} + 500 * \frac{\text{no. of times } 500 \text{ appears}}{N \text{ simulations}} + \dots$$

Population size

no. of successes

## Scenario-1

$\boxed{10}$ balls in total ($\boxed{4}$ white, $\boxed{6}$ black)

no. of failures

Sample size

- draw $\boxed{4}$ balls with replacement

drawing a white ball is treated success

- Probability that $\boxed{2}$ balls are white

total no. of successes we are interested in

- $X$ = total no. of successes

0
1
2
3
4

- $P(X = 2)$

$$P(X=2) = P\left(\begin{array}{c}\text{2 successes in a sample of} \\ \text{size 4 drawn from a population} \\ \text{of size 10 with 4 successes}\end{array}\right)$$

$$\text{Sampling space} = \{w_1, w_2, w_3, w_4, b_1, b_2, b_3, b_4, b_5, b_6\}$$

$$\text{Probability} = \left\{\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right\}$$

with replacement, order matters (no of outcomes $= 10^4$)

$$S = \left\{\begin{array}{ll} \frac{1}{10^4} & \boxed{w_1\ w_1\ w_1\ w_1} \\ \frac{1}{10^4} & \boxed{w_1\ w_2\ b_1\ b_2} \\ \frac{1}{10^4} & \boxed{w_1\ b_1\ w_2\ b_2} \\ & \vdots \end{array}\right\}$$

| slot-1 | slot-2 | slot-3 | slot-4 |
|---|---|---|---|

$$E = \left\{\begin{array}{ll} \frac{1}{10^4} & \boxed{w_1 \mid w_2 \mid b_1 \mid b_2} \\ \frac{1}{10^4} & \boxed{w_1\ b_3\ w_1\ b_4} \\ \frac{1}{10^4} & \boxed{w_1\ b_3\ b_4\ w_3} \\ & \vdots \end{array}\right\}$$

$4c_2$

Select 2 out of 4 slots without replacement and order does not matter

AND

$4^2$

Select 2 out of 4 white balls with replacement and order matters

AND

select 2 out of 6 black balls WR and OM.

$6^2$

$$P\{\text{outcome} = (w_1\ w_1\ w_1\ w_1)\} = P(w_1 \underset{\text{AND}}{} w_1 \underset{\text{AND}}{} w_1 \underset{\text{AND}}{} w_1)$$

$$= P(w_1) * P(w_1) * P(w_1) * P(w_1)$$

$$\frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} = \frac{1}{10^4}$$

$$P(E) = \frac{1}{10^4} * \boxed{\text{no. of outcomes in } E}$$

$$= \frac{\text{no. of outcomes in } E}{\text{no. of outcomes in } S}$$

no. of outcomes in $E = \underline{{}^4C_2} * \underline{4}^2 * \underline{6}^2$

$$P(E) = \frac{{}^4C_2 * 4^2 * 6^2}{10^4}$$

$$= {}^4C_2 * \left(\frac{4}{10}\right)^2 * \left(\frac{6}{10}\right)^2$$

$$= {}^4C_2 * \left(\frac{4}{10}\right)^2 * \left(1 - \frac{4}{10}\right)^2$$

$$P(x=2) = {}^4C_2 * \left(\boxed{\frac{4}{10}}\right)^{2} * \left(\boxed{1-\frac{4}{10}}\right)^{4-2}$$

Success fraction in the population

failure fraction in the population

Abstraction

$$S = \{w,\ b\}$$

$$p = \left\{\frac{4}{10},\ \frac{6}{10}\right\}$$

$$S = \left\{ \begin{matrix} \boxed{w\ w\ w\ w} \\ \boxed{w\ b\ w\ b} \\ \boxed{b\ w\ w\ w} \\ \vdots \end{matrix} \right\}$$

$$P(wwbb) = P(w) * P(w) * P(b) * P(b) = \frac{4}{10} * \frac{4}{10} * \frac{6}{10} * \frac{6}{10}$$

$$E = \left\{ \begin{matrix} \boxed{w\,|\,w\,|\,b\,|\,b} \\ w\ b\ w\ b \\ b\ w\ w\ b \\ \vdots \end{matrix} \right\}$$

$$P(E) = {}^4C_2 * \frac{4}{10} * \frac{4}{10} * \frac{6}{10} * \frac{6}{10}$$

$$P(E) \neq \frac{\text{no. of outcomes in } E}{\text{no. of outcomes in } S}$$

because the outcomes in the sample space are not equally likely.

$$P(wwww) = P(w) * P(w) * P(w) * P(w) = \frac{4}{10} * \frac{4}{10} * \frac{4}{10} * \frac{4}{10}$$

$$P(bbbb) = \left(\frac{6}{10}\right)^4$$

# Binomial experiment

- Sampling from the population happens with replacement. Therefore, the success probability in each draw is the same. The draws are independent.

$X$ = no. of successes in a sample of size $n$

with success probability $p = \dfrac{n_{success}}{n_{success} + n_{failure}} = \dfrac{4}{4+6}$

- $X$ is a binomial random variable
- $X \sim Bin(n, p)$, E.g. $X \sim Bin(n=10, p=0.4)$

$n$: Sample size / no. of trials / no. of draws

$p$: Success probability / proportion

Parameters $= n, p$ have to be specified.

- $P(X = j) \Leftrightarrow$ dbinom$(j, n, p)$

E.g. dbinom$(2, 4, 4/(4+6))$

$$P(X=j) = {}^nC_j\, p^j (1-p)^{n-j}$$

$X \sim Bin(n, p)$

$j = 0, 1, 2, \ldots n$

E.g. $4C_2 \left(\frac{4}{10}\right)^2 \left(1-\frac{4}{10}\right)^{4-2}$

for the 4 white balls and 6 black balls problem

$n = 4$

$p = 4/(4+6)$, $p = 0.4$

dbinom(2, n, p)

What if sampling is without replacement?

- Same setup: 4 white, 6 black balls
- Draw a sample of size = 4
- We are in the probability of 2 white balls (2 successes)

$$S = \{w_1, w_2, w_3, w_4, b_1, b_2, b_3, b_4, b_5, b_6\}$$

$$S = \left\{ \begin{array}{ll} \boxed{w_1\ w_2\ w_3\ w_4} & 1/{}^{10}C_4 \\ \boxed{w_1\ b_1\ w_2\ b_2} & 1/{}^{10}C_4 \\ \boxed{b_1\ b_2\ b_3\ b_4} & 1/{}^{10}C_4 \\ \vdots & \end{array} \right\}$$

Sampling 4 out of 10 balls without replacement and order does not matter

$$= \boxed{{}^{10}C_4 \text{ ways}}$$

$$P(w_1 \text{ AND } w_2 \text{ AND } w_3 \text{ AND } w_4)$$

$$P(\text{first ball is } w_1) = 1/10, \quad P(\text{second ball is } w_2 \mid \text{first ball is } w_1)$$

$$= \frac{1}{9}$$

$$\frac{1}{9} = \boxed{P(\text{Second ball is } w_2 \mid \text{first ball is } w_1)}$$

$$\stackrel{?}{=} \boxed{P(\text{Second ball ball is } w_2)}$$

Choose 2 out of 4 white balls without replacement order does not matter

$\boxed{4C_2} * \boxed{6C_2}$ outcomes here

choose 2 out of 6 black balls without replacement order does not matter

$$\big\| \; / 9$$

$$E = \left\{ \begin{array}{ll} \frac{1}{10C_4} & \boxed{w_1 \mid w_2 \mid b_1 \mid b_2} \\ \frac{1}{10C_4} & \boxed{w_1 \mid b_2 \mid w_3 \mid b_4} \\ \vdots & \quad \vdots \end{array} \right\}$$

$$P(E) = \frac{1}{10C_4} * \text{no of outcomes in } E$$

$$P(E) = \frac{\text{no. of outcomes in } E}{\text{no. of outcomes in } S} = \frac{{}^{4}C_2 * {}^{6}C_2}{{}^{10}C_4}$$

$$P(X = 2) = \frac{{}^{4}C_{2} * {}^{6}C_{2}}{{}^{10}C_{4}}$$

no. of successes in the population

no. of failure in the population

Sample size

Population size

still counts the no. of successes

- $X$ is a hypergeometric random variable

- $\boxed{X \sim \text{HyperGeom}(n_{success}, n_{failure}, n)}$

- Parameters $\left\{ \begin{array}{l} n_{success} = \text{no. of successes in population} \\ n_{failure} = \text{no. of failures in population} \\ n = \text{sample size} \end{array} \right.$

$\Downarrow$

have to be specified

- $\boxed{P(X = j) = \dfrac{\binom{n_{success}}{j}\binom{n_{failure}}{n-j}}{\binom{n_{success} + n_{failure}}{n}}}$

In R, we calculate $P(X=2)$ where

$X \sim \text{Hypgeom}(n_{success}=4,\ n_{failure}=6,\ n=4)$

nsuccess = 4

nfailure = 6

n = 4

dhyper(2, nsuccess, nfailure, n)

Special case of a negative binomial random variable

- $X \sim NegBin(r, p)$ is the same as sample size

- X counts the no. of trials until the $r^{th}$ success

  where success probability is p (Sampling is with replacement)

- Special case where $r = 1$

- $X \sim NegBin(r=1, p)$ counts the no. of trials

  until the 1st success

- $X = 1, 2, 3, \ldots\ldots$

- $P(X=j) = \binom{j-1}{r-1} p^r (1-p)^{j-r}$

$X \sim$ NegBin$(r, p)$

dnbinom calculates this

- If $r=1$, $X \sim \text{NegBin}(r=1, p)$

$$P(X=j) = \binom{j-1}{1-1} p^1 (1-p)^{j-1}$$

$(j = 1, 2, 3, \ldots)$

$$= p(1-p)^{j-1} = (1-p)^{j-1} p$$

- $X$ is called a geometric variable

- $X \sim \text{Geom}(p)$

---

Special case of a binomial random variable

- $X \sim \text{Bin}(n, p)$

- $P(X = j) = \binom{n}{j} p^j (1-p)^{n-j}$

$(j = 0, 1, \ldots, n)$

dbinom calculates this

- Special case when $n = 1$ • $X \sim \text{Bin}(n=1, p)$

• $P(X=j) = \binom{1}{j} p^{j} (1-p)^{1-j}$

$j = 0, 1$

$j = 0$

0 Success

$$\binom{1}{0} p^{j} (1-p)^{1-0} = 1-p$$

$j = 1$

1 Success

$$\binom{1}{1} p^{1} (1-p)^{1-1} = p$$

• $P(X=j) = p^{j} (1-p)^{1-j}$

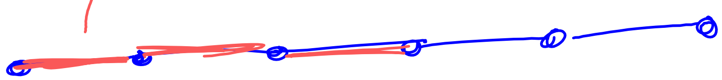
• $X$ is called a Bernoulli random variable

- $X \sim \text{Ber}(p)$

```
Usual average = (0 + 500 + 1000 + 1500 + 2000) / 5

= (1/5)* 0 + (1/5)*500 + (1/5)*1000 + (1/5)*1500 + (1/5)*2000

E[X] = weighted average of the values that the random variable can take

= 0.27*0 + 0.28*500 + 0.315*1000 + 0.09*1500 + 0.045*2000 = 675

= expected value, mean value, average value of the random variable

approximately = (# of times 0 appeared / nsimulations)*0 +
                (# of times 500 appeared / nsimulations)*500 +
                (# of times 1000 appeared / nsimulations)*1000 +
                (# of times 1500 appeared / nsimulations)*1500 +
                (# of times 2000 appeared / nsimulations)*2000

                = [the sum of all simulated (realized) values] / nsimulations

                = long term average of the realizations of the random variable.

Expected value of a binomila random variable X~(n, p) is E[X] = np.

What about a negative binomila random variable X~NegBin(r, p)?
r = 5, p = 0.4, E[X] = 7.5
r = 4, p = 0.2, E[X] = 16

|--|--|--|------------------|--| = 1/60 seconds

In the 1st millisecond, we have n = 1e09 photos strike the pixel. The probability that a photon will register information is
1e-10. If success is defined as a photon regustering information, how many photons register information in the 1st millisecond?

X ~ Bin(n = 1e09, p = 1e-10).

P(X = j) = nC_j * p^j *(1-p)^(n-j), here n >> p, sample size >> success probability

lambda = n * p = 1e09 * 1e-10 = 0.1 => p = lambda / n

P(X = j) = nC_j * (lambda / n)^j *(1-(lambda/n))^(n-j) is approximately equal to (e^(-lambda)*lambda^j)/j!

Under this condition, X is called the Poisson random variable.

X ~ Poi(lambda = 0.1)
```

# Poisson random variable

$X = Bin(n, p)$, $n >> p$



Spatial interval 10km

$n = 10^3$, $p = 10^{-3}$

$X$ = no. of accidents over a 10km interval

Time interval $10^{-3}$ s

$n = 10^6$, $p = 10^{-7}$

$X$ = no. of photons that strike a pixel every millisecond

Time interval 100 years

$n = 10^3$, $p = 10^{-2}$

$X$ = no. of volcanic eruptions over 100 years

$X$ is a from a binomial distribution with parameters $n$ and $p$ $\Longleftrightarrow$ $X \sim Bin(n, p)$

In the first case (accident scenario $X \sim Bin(n = 10^3, p = 10^{-3})$)

- population size $= 10^7$
- Sample size $= 10^3$
- no. of white balls (accident) $= 10^4$
- no. of black balls (no accident) $= 10^7 - 10^4$

$$P = \frac{n_W}{n_W + n_B}$$

$$= 10^{-3} n_W \times 10^7$$

Still counts the no. of successes

$$\boxed{\begin{gathered} X \sim \text{Poisson}(\lambda) \\ X \sim \text{Poi}(\lambda) \end{gathered}}$$

## Three conditions

(1) $n >> p$

(2) What happens in one interval does not affect what happens in another (events in each interval are independent)

(3) On an average, there are $\lambda$ occurences over each interval.

$\lambda = np$

$\lambda$ may be given directly or indirectly

On an average there are 10 accidents every 1000 kilometer

- $\lambda = \dfrac{10}{1000} = \dfrac{1}{100} \dfrac{\text{accidents}}{\text{km}}$

$\lambda = \dfrac{1}{100} \dfrac{\text{accidents}}{\text{km}} = \dfrac{1}{10} \boxed{\dfrac{\text{accidents}}{10 \text{ Km}}}$

- $n = 10^3$
- $p = 10^{-3}$
- 10 Km segments

$\lambda = np$

$= 1 \boxed{\dfrac{\text{accident}}{10 \text{ km}}}$

$X \sim Poi\left(\lambda = 2 \frac{\text{students}}{\text{min}}\right)$

E.g. Students arrive in a lecture hall at the rate of 2 per min. The no. of students is Poisson distributed.

(1) What is the probability that in the next 15 seconds, 1 student arrives?

(2) What is the probability that in the next 2 minutes, atmost 6 students arrive?

$$\lambda = 2\ \frac{\text{student}}{\text{minute}} = \frac{\frac{2}{4}\ \text{students}}{15\ \text{seconds}}$$

$$= 2\ \frac{\text{student}}{60\ \text{seconds}} = \frac{2\ \text{students}}{4 * 15\ \text{seconds}}$$

$$\lambda = 0.5\ \frac{\text{students}}{15\ \text{seconds}}, \quad X \sim Poi\left(\lambda = 0.5\ \frac{\text{students}}{15\ \text{seconds}}\right)$$

In R

dpois(1, 0.5)

Lambda

$$P(X = 1) = \frac{e^{-\lambda} \lambda^{j}}{j!} = \frac{e^{-0.5} * (0.5)^{1}}{1!}$$

j

$$\lambda = 2\ \frac{\text{students}}{1\ \text{minute}} = 4\ \frac{\text{students}}{2\ \text{minutes}}$$

$$X \sim \text{Poi}\left(\lambda = 4\ \frac{\text{students}}{2\ \text{minutes}}\right)$$

lambda

In R ppois(6, 4)

$$P(X \le 6)$$

$$= P(\underline{X=0} \text{ or } \underline{X=1} \text{ or } \underline{X=2} \text{ or } \dots \text{ or } \underline{X=6})$$

$$= P(\underline{x=0}) + P(\underline{x=1}) + \dots + P(\underline{x=6})$$

$j \qquad j \qquad j$

$$= \frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \dots + \frac{e^{-4}4^6}{6!}$$

$$\frac{e^{-\lambda}\lambda^j}{j!}$$

# Salesman Problem

$$X = \begin{cases} 0 \\ 500 \\ 1000 \\ 1500 \\ 2000 \end{cases}$$

$P_X(0) = P(X=0)$

$P_X(500) = P(X=500)$

$P_X(1000) = P(X=1000)$

$P_X(1500) = P(X=1500)$

$P_X(2000) = P(X=2000)$

- What is the expected earning of the salesman?

One possible way to do this is a simple average:

$$\frac{0 + 500 + 1000 + 1500 + 2000}{5}$$

$$= \frac{1}{5} * 0 + \frac{1}{5} * 500 + \frac{1}{5} * 1000 + \frac{1}{5} * 1500 + \frac{1}{5} * 2000$$

Equal weights assigned to all possible values of X.

= Weighted sum with equal weights (1/5)

- The weights should reflect the importance of each possible value that X can take

modified weights

$$P(X=0) * 0 + P(X=500) * 500 + P(X=1000) * 1000 + P(X=1500) * 1500 + P(X=2000) * 2000$$

Weighted sum with unequal weights

$$E[X] = 0 * P(X=0) + 500 * P(X=500) + 1000 * P(X=1000) + 1500 * P(X=1500) + 2000 * P(X=2000) = 675$$

$P(X=0) \approx \dfrac{\text{no. of times 0 appears}}{\text{nsimulations}}$

$E[X]$: Expected value of $X$

675: Expected earnings of the salesman is Rs. 675

$$\approx 0 * \frac{\text{no. of times 0 appears}}{\text{nsimulations}} + 500 * \frac{\text{no. of times 500 appears}}{\text{nsimulations}} + \ldots + 2000 * \frac{\text{no. of times 2000 appears}}{\text{nsimulations}}$$

Realizations of $X$

$$E[X] \simeq \frac{\text{Sum of the simulated values}}{n \text{ simulations}}$$

Total no. of simulations

Expected Value

Theoretical definition

$$E[X] \simeq \boxed{\text{Long term average of the realizations of } X}$$

Large no. of simulations

Simulation interpretation

$$= \sum_x x P_X(x)$$

$$= 0 * P_X(0) + 500 * P_X(500) + \dots + 2000 * P_X(2000)$$

$$P_X(0) = P(X=0)$$

- Build another random variable using X (random earning of the salesman)

$E[X]$ = expected earnings is not a random variable but a fixed number

$= 675$

$X - E[X] = \begin{cases} \text{is a random variable, } X - 675 \\ \quad \hookrightarrow \text{is random} \\ \text{Deviation of the random earning from the expected earning} \end{cases}$

$$X - E[x] = \begin{cases} 0 - 675 = -675 & P(x=0) \\ 500 - 675 = -175 & P(x=500) \\ 1000 - 675 = 325 & P(x=1000) \\ 1500 - 675 = 825 & P(x=1500) \\ 2000 - 675 = 1325 & P(x=2000) \end{cases}$$

$$= X - 675$$

$$(X - E[x])^2 = (X - 675)^2$$

- Squared deviation of the salesman's random earnings from the expected earnings
- Also a random variable

$(X - E[X])^2$ = Squared deviation of $X$ from its expected value $E[X]$

$$\downarrow$$

$$E\Big[(X - \underbrace{E[X]}_{675})^2\Big] =$$

$X \in \{0, 500, 1000, 1500, 2000\}$

<u>Theoretical definition</u>

$$(0-675)^2 * P(X=0) + (500-675)^2 * P(X=500) + \cdots + (2000-675)^2 * P(X=2000)$$

<u>Simulation interpretation</u>

Long term average of the realizations of the squared deviations definition

- $\mathrm{Var}[X] \overset{\uparrow}{=} E\left[\left(X - E[X]\right)^2\right]$ = a single number

units of $\mathrm{Var}[X]$ =

- $E\left[\left(X - E[X]\right)^2\right] = \left\{ \left[\text{simulatedData} - \text{mean}(\text{simulatedData})\right]^2 \right.$

$(X - E[X])^2$ approximated by mean(simulatedData)

Simulate many times

## Units

- $X$ = random earning of the salesman = Rupees
- $E[X]$ = expected earning = Rs. 675
- $X - E[X]$ = deviation from the expected earning = Rupees
- $(X - E[X])^2$ = squared deviation from the expected earning = Rupees$^2$

- $E\left[(X - E[X])^2\right] = \text{Var}[X]$

= expected squared deviation from the expected value = $\text{Rupees}^2$

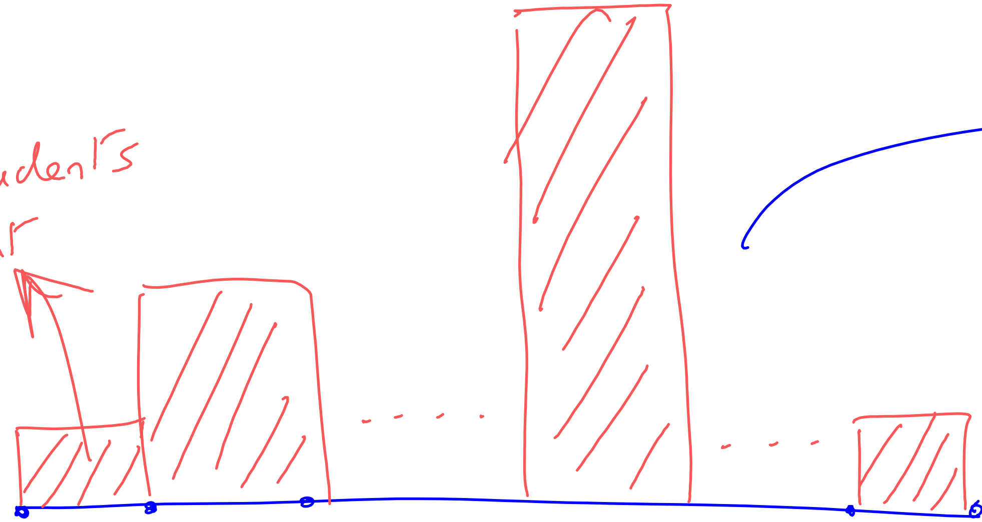
- $\sqrt{E\left[(X - E[X])^2\right]}$ = expected deviation from the expected value

= Rupees.

$\|$

$SD[X]$, standard deviation of $X$

no. of student's whose height is in the range 138-140cm



138cm 140cm 142cm . . - - - 200cm 202cm

bin width = 2cm

Frequency histogram

Approximate Probability histogram

Relative frequency histogram

Density histogram

A continuous random variable, like a random student's height → $X$ can possibly take infinitely many values.

$$P(X = \text{specific value}) = 0$$

- We started with defining probability as a relative frequency
- Relative frequency histogram $\Rightarrow$ bar heights are approximations of probabilities.

Frequency (Absolute count) $\Longrightarrow$ Relative frequency (proportional count) $\Longrightarrow$ Density (relative frequency divided by bin width)

$$\text{Density between 170 cm and 172 cm} = \frac{\boxed{\text{Relative frequency between 170 - 172 cm}}}{\boxed{\text{Bin width} = 2\text{ cm}}}$$

Density between 170-172 cm

$$= \frac{\text{Approximate probability that a random student's height is in between 170-172 cm}}{2\ \text{cm}}$$

$$= \frac{0.098}{2} \ \frac{\text{approximate probability 170cm-172cm}}{\text{cm}}$$

$$= \boxed{0.049} \ \frac{\text{approximate probability}}{\text{cm}}$$

probability density over 170cm-172cm

available in the range 170cm-172cm

$$0.049 = \text{probability per cm available over the interval } 170\text{cm} - 172\text{cm}$$

$$= \text{probability density available over the interval } 170\text{cm} - 172\text{cm}$$

For a continuous random variable $X$ :

(i) CDF (cumulative distribution function)

$$F_X(x) = P(X \leq x)$$

Eg. $F_X(170) = P(X \leq 170)$

$F_X(170) = $
- area of the density histogram bars to the left of 170
- area under the density curve to the left of 170

formula

(2) PDF (probability density function)

$f_X(x) = $ Probability density available at the input value $x$

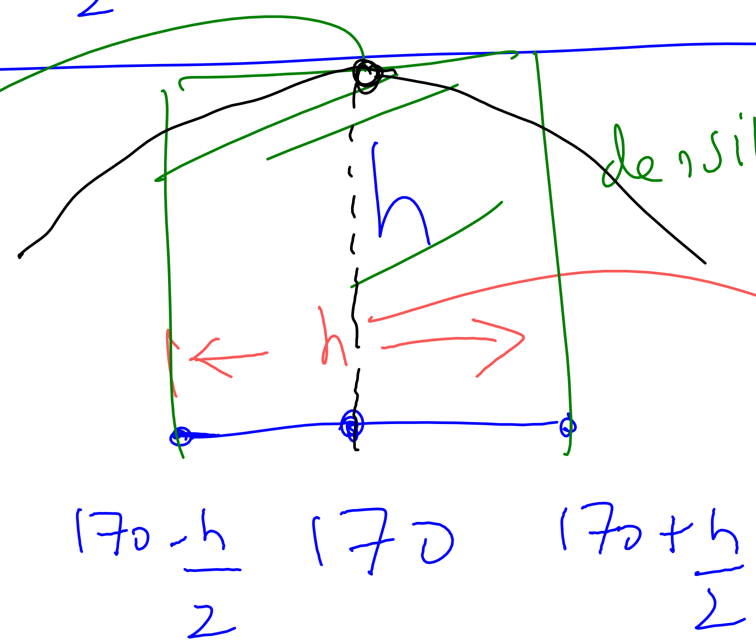
E.g. $f_X(170)$ = probability density (probability per cm) available around 170 cm.

PDF

$$f_X(170) \approx P\left(170-\frac{h}{2} \le X \le 170+\frac{h}{2}\right)$$

E.g. $h = 0.0001$

$\approx 0.05$




a very small value

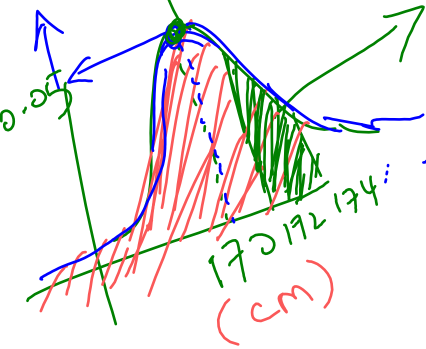
CDF

$$F_X(170) = P(X \leq 170)$$

$= $ Area under the PDF $f_X(x)$ for values of $x$ less than $170$.

dnorm(170, mean = 170, sd = 8) = $f_X(170)$ = PDF at 170cm

Likelihood at 170cm




$P(172 \le X \le 174)$ = Area under the PDF $f_X(x)$ for $x$ in between 172 and 174 = $\underbrace{\text{pnorm}(174, \text{mean} = 170, \text{sd} = 8)}_{F_X(174)} - \underbrace{\text{pnorm}(172, \text{mean} = 170, \text{sd} = 8)}_{F_X(172)}$

## Normal random variable

mean

standard deviation

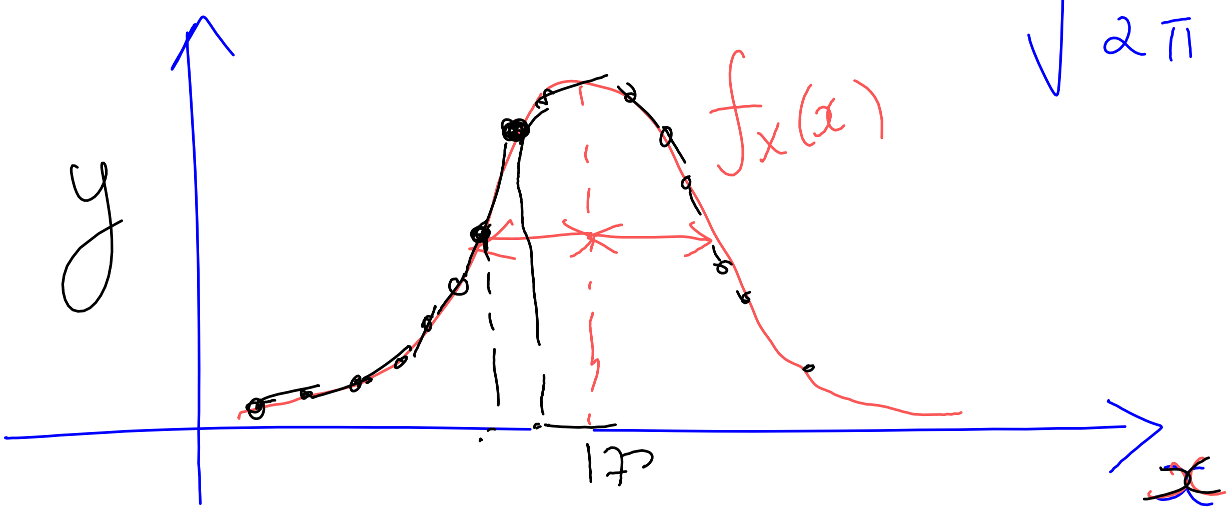
$$X \sim N(\mu, \sigma)$$

E.g. $X \sim N(\mu = 170\text{ cm}, \sigma = 8\text{ cm})$

PDF $f_X(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

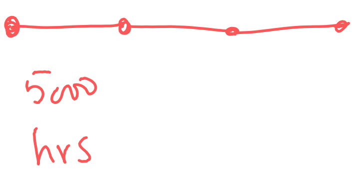
E.g. Case $f_X(x) = \dfrac{1}{\sqrt{2\pi 8^2}} e^{-\frac{1}{2}\left(\frac{x-170}{8}\right)^2}$

$X \sim N(\mu = 170\,cm, \sigma = 8\,cm)$



# Recall Poisson random variable

- Discrete random variable
- Counts the number of successes (no. of events) over an interval (time or space) of interest
- Only one parameter $\lambda$ = average rate at which an event occurs. E.g. $\lambda = \dfrac{1 \text{ failure}}{\boxed{5000 \text{ hours}}}$




(on an average a laptop fails once every 5000 hours)

- $P(X = 3)$ = probability that in the next time interval (5000 hrs), we have 3 failures

- $P(X=3)$   $x$

Pen & Paper

$$\frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-1}\,1^{3}}{3!} = \frac{e}{3!}$$

R function

dpois(3, lambda = 1)

- $X \sim Poi\left(\lambda = 1 \ \frac{\text{failure}}{5000\text{ hrs}}\right)$, i.e., $X$ counts the random number of failures over a time interval (5000 hours)

- Under the hood $X \sim Poi(\lambda = 1)$   $\lambda = np$

is the same as $X \sim Bin(n = 10^{6}, p = 10^{-6})$

Where $n >> p$

- $\lambda = 1 \dfrac{\text{failure}}{5000 \text{ hrs}} = \dfrac{1}{5000} \boxed{\dfrac{\text{failure}}{\text{hour}}}$

↳ unit

Time interval of interest = 1 hour

- $P(X = 3)$ = probability that in the next time interval (1 hour) we have 3 failures

- Using R, dpois(3, lambda = 1/5000)

500 hrs $\qquad$ 1 hour $\qquad$ 1 year

- $\lambda = 1 \dfrac{\text{failure}}{500 \text{ hours}} = \dfrac{1}{500} \dfrac{\text{failures}}{\text{hour}} = \dfrac{365 \times 24}{500} \dfrac{\text{failures}}{\text{year}}$

$X \sim Poi(\lambda) \Rightarrow P(X=3)$

- Now define $T$ as a random variable that measures the time until the next occurrence of the event.

- $T$ is a continuous random variable. When the no. of occurrences follow a Poisson process (i.e. $\lambda$ is given) we say that $T$ is an exponential random variable with the same 'rate' parameter $\lambda$. We write

$$T \sim Exp(\lambda) \; ; \; \text{e.g. } T \sim Exp\left(\lambda = 1 \frac{\text{failure}}{500 \text{ hrs}}\right)$$

unit for time

$T$ is measuring the time until the next laptop failure.

→ measured in 500 hours

* $P(T < 2)$ = probability that the next event (failure of laptop) happens within 2 (500 hrs) = 1000 hours

- $T \sim Exp\left(\lambda = \frac{1}{5000} \frac{\text{failure}}{\text{hour}}\right)$

unit for time = 1 hour

$P(T < 2)$ = probability that the next event (failure of a laptop) will occur within the next 2 hours

$\lambda = \frac{1}{5000} \frac{\text{failure}}{\text{hour}}$

- PDF of $T \sim Exp(\lambda)$

$f_T(t)$ = probability density (likelihood)

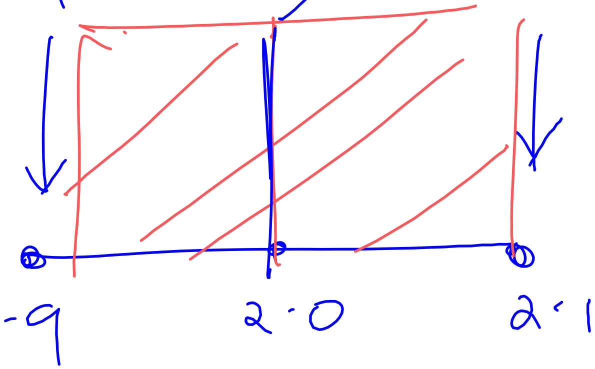
= probability per unit interval available at $t$

= probability per hour available at $t$.

- For e.g. $f_T(2)$ = probability per unit hour available at $t = 2$.

- $P(1.9 < T \le 2.1) \approx \underbrace{f_T(2) \times 0.2}_{\text{approximate probability}}$

PDF



- $P(1.99 \le T \le 2.01) \approx f_T(2) * 0.02$

$P(1.999 \le T \le 2.001) \approx f_T(2) * 0.002$

$\Downarrow$

Keep continuing, but approximations will get better as the interval gets smaller

- $P\left(2-\frac{h}{2} \le T \le 2+\frac{h}{2}\right) \approx f_T(2) * h$

$h$ is very very small $\Rightarrow$ $\boxed{f_T(2) \approx \frac{P\left(2-\frac{h}{2} \le T \le 2+\frac{h}{2}\right)}{h}}$

- PDF for $T \sim Exp(\lambda)$

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Curve that hugs the density histogram

<u>E.g.</u> $T \sim Exp\left(\lambda = \frac{1}{500} \text{ failures}/\text{hour}\right)$

$$f_T(t) = \begin{cases} \frac{1}{500} e^{-\frac{1}{500} t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

# Sessional–2 Practice Problems, January 7, 2022

1. On an average, a laptop fails once every 5000 hours. Suppose you use your laptop 5 hours each day. What is the probability that your laptop will last for more than 4 years?

**Solution:** The first step in such problems is to identify $\lambda$, which is the average rate at which an event occurs. We have

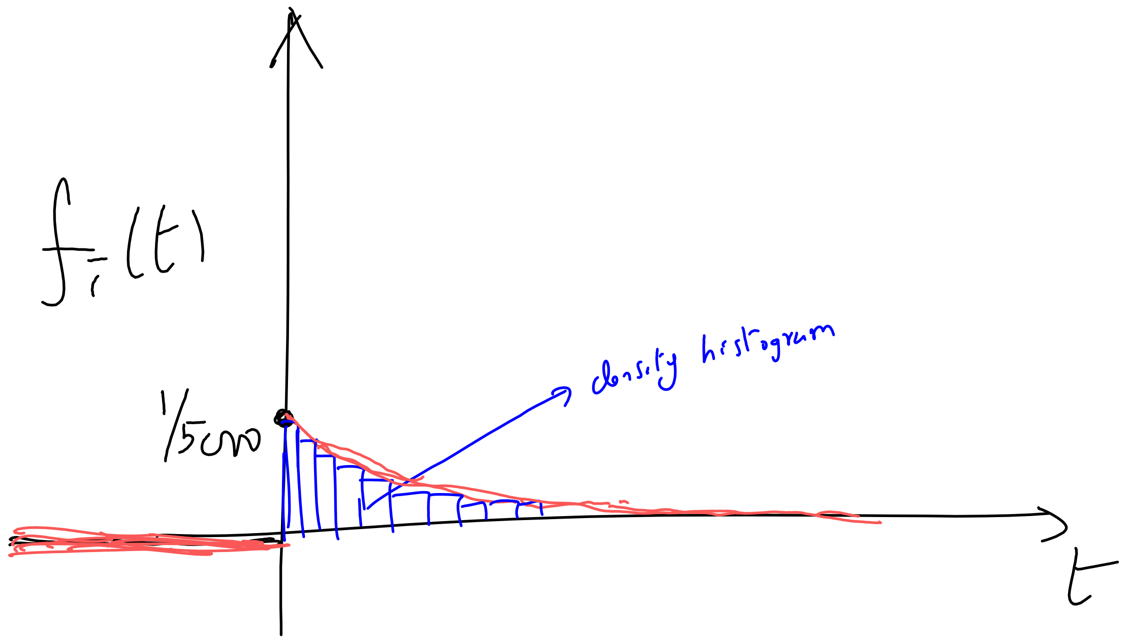
$$\lambda = \begin{cases} 1 \dfrac{\text{failure}}{5000 \text{ hours}}, \\[2ex] \dfrac{1}{5000} \dfrac{\text{failure}}{\text{hour}}, \\[2ex] \dfrac{365\times 24}{5000} \dfrac{\text{failure}}{\text{year}}. \end{cases}$$

Now $X \sim \text{Poi}(\lambda)$ in each case would model (that is, count) the number of failures in the next 5000 hours, the number of failures in the next hour, and the number of failures in the next year, respectively. For example, suppose we consider $X \sim \text{Poi}(\lambda = \frac{365\times 24}{5000} \frac{\text{failure}}{\text{year}})$, then $P(X = 2)$ means the probability that we have exactly 2 failures in the next $\underline{\text{year}}$, which is equal to

$$\frac{e^{-\left(\underbrace{\frac{365\times 24}{5000} \frac{\text{failure}}{\text{year}}}_{\lambda}\right)} \times \left(\underbrace{\frac{365\times 24}{5000}}_{\lambda}\right)^2}{2!},$$

or using R, `dpois(2, lambda = (365*24)/5000)`. Note that the underlined units for time matches with that of the units for time in the denominator of $\lambda$.

Now define $T$ as the time until next failure. For example, if $T \sim \text{Exp}(\lambda = \frac{365\times 24}{5000} \frac{\text{failure}}{\text{year}})$, then $P(T < 4)$ means the probability that the next failure happens within 4 $\underline{\text{years}}$; note that the underlined units for time matches with that of the units for time in the denominator of $\lambda$. On the other hand, if $T \sim \text{Exp}(\lambda = \frac{1}{5000} \frac{\text{failure}}{\text{hour}})$, then $P(T < 4)$ means the probability that the next failure happens within 4 $\underline{\text{hours}}$; note again that the underlined units for time matches with that of the units for time in the denominator of $\lambda$. You can use any variation of the definition of the random variable $T$ to solve for the probability that your laptop will last more than 4 years if you use the laptop 5 hours per day. Of course, in each case, you have to decide what is $t$ in the expression $P(T > t)$. For example if you use $T \sim \text{Exp}(\lambda = \frac{1}{5000} \frac{\text{failure}}{\text{hour}})$, then we need $P(T > t)$, where $t$ would correspond to number of hours over 4 years if you use your laptop for 5 hours per day. You can solve it in using R using `1-pexp(?, rate = ?)`. I will leave it to you to fill in the question marks.




## Laptop Screen Problem

Time until 1st failure

- $T_A \sim N(\mu = 20, \sigma = 4)$

  $\mu = 20 \Downarrow$ (100 hours), $\sigma = 4 \Downarrow$ (100 hours)

- $T_B \sim \text{Exp}\left(\lambda = \frac{1}{20} \; \frac{\text{failures}}{100 \text{ hours}}\right)$

$$P(\overset{A}{\text{A manufactured}} \mid \overset{B}{T > 18})$$

Bayes' theorem

$$= \frac{P(\overset{B}{T > 18} \mid \overset{A}{\text{A manufactured}}) \times P(\overset{A}{\text{A manufactured}})}{\boxed{P(\overset{B}{T > 18})}}$$

use law of total probability

$$= \frac{P(T_A > 18) \times P(\text{A manufactured})}{P(T > 18 \text{ AND A manufactured OR } T > 18 \text{ AND B manufactured})}$$

$$= \frac{P(T_A > 18) * P(A \text{ manufactured})}{P(\overset{A}{T} > 18 \text{ AND } A \text{ } \overset{B}{\text{manufactured}}) + P(T > 18 \text{ AND } B \text{ manufactured})}$$

$$= \frac{\boxed{P(T_A > 18)} * P(A \text{ manufactured})}{P(\overset{A}{T} > 18 \mid A \text{ } \overset{B}{\text{manufactured}}) \overset{A}{*} P(A \text{ manufactured}) + P(T > 18 \mid B \text{ manufactured}) * P(B \text{ manufactured})}$$

$$= \frac{P(T_A > 18) * P(A \text{ manufactured})}{P(T_A > 18) * P(A \text{ manufactured}) + P(T_B > 18) * P(B \text{ manufactured})}$$

area under the PDF greater than 18

$T_A \sim N(\mu = 20, \sigma = 4)$

$$= \frac{P(T_A > 18) * 0.5}{P(T_A > 18) * 0.5 + P(T_B > 18) * 0.5}$$

$T_B \sim Exp(\lambda = \frac{1}{20})$

$$= \frac{\text{pnorm}(18, \text{mean} = 20, \text{sd} = 4, \text{lower.tail} = \text{FALSE}) * 0.5}{\boxed{\phantom{x}} + \text{pexp}(18, \text{rate} = 1/20, \text{lower.tail} = \text{FALSE})}$$

Sindata = rnorm(1e4, mean = 20, SD = 4)

mean(Sindata > 18)

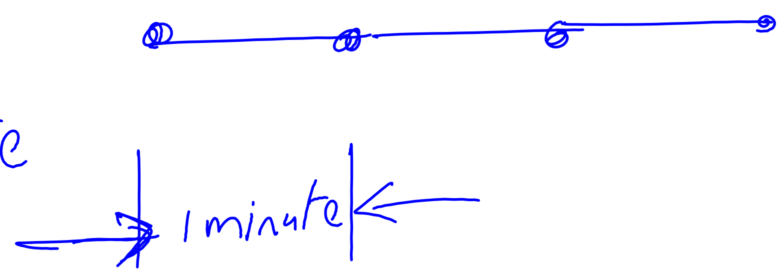
# Telephone Problem

$$\frac{10 \text{ minutes}}{\text{call}} \Longleftrightarrow \frac{1}{10} \frac{\text{calls}}{\text{minute}} \rightarrow \lambda \text{ (rate)}$$

Time at bottom

$$\Longleftrightarrow \frac{1 \text{ call}}{10 \text{ minutes}} \rightarrow \lambda$$

$$\lambda = \frac{1}{10} \frac{\text{calls}}{\text{minute}}$$



$$T \sim Exp\left(\lambda = \frac{1}{10} \frac{\text{calls}}{\text{minute}}\right)$$

memoryless property

$$P(T > 20 \text{ minutes}) = \text{Pexp}(20, \text{rate} = 1/10, \text{lower.tail} = \text{FALSE})$$

$$\boxed{P(T > 10 + 20 \mid T > 10)} \overset{?}{=} \boxed{P(T > 20)}$$

$$= \boxed{P(\underbrace{T > 30}_{A} \mid \underbrace{T > 10}_{B})}$$

definition of conditional probability

$$= \boxed{\frac{P(T > 30 \text{ AND } T > 10)}{P(T > 10)}}$$

$$P(T > 20)$$

$$\boxed{P(T > m + 20 \mid T > m) = P(T > 20)}$$

memoryless property of an exponential random variable

Rainfall problem

Rainfall for a particular year

$R \sim N(\mu = 40 \text{ inches}, \sigma = 4 \text{ inches})$

$R_1 \sim N(\mu = 40, \sigma = 4)$

$R_2 \sim N(\mu = 40, \sigma = 4)$

$\vdots \qquad \vdots$

$R_j \sim N(\mu = 40, \sigma = 4)$

Success probability

$p = P(R > 50)$

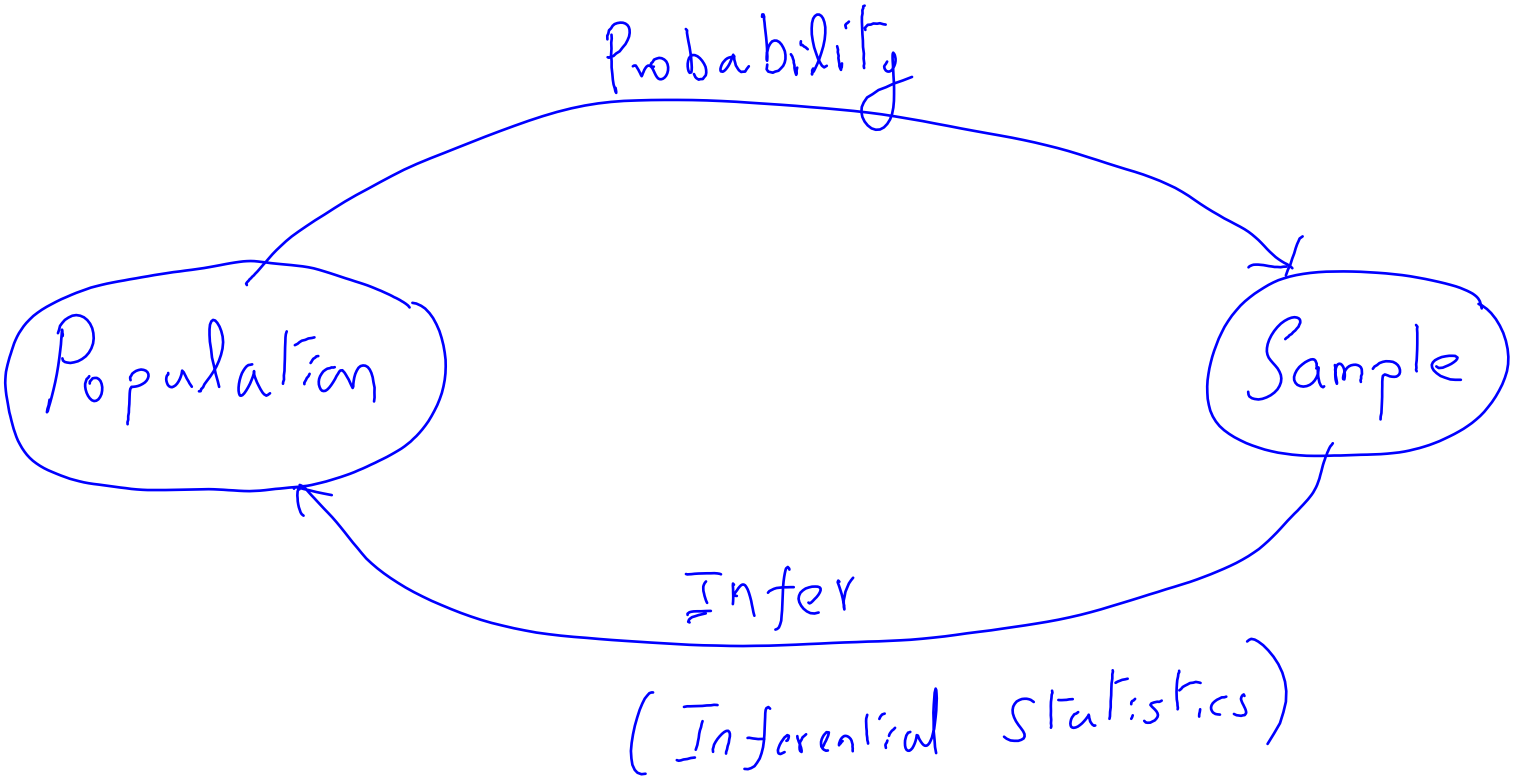
Success number $r = 1$ st success

$$\boxed{\begin{array}{l} p, \ r, \ X \sim \text{NegBin}(r, p) \\ \ \| \quad \ \ \| \\ P(R>50) \ \ 1 \end{array}}$$

$$P(X > 10) = 1 - P(X \le 10)$$

$$\Downarrow$$

$$1 - \text{pnbinom}(?, r, p)$$




E.g. Population = entire undergraduate population in Manipal

Sample = a certain number of (random) undergraduate students in Manipal.

$$\sqrt{\frac{1}{10^6}\left[\begin{array}{l}(* - \text{mean height})^2 \\ + \\ (* - \text{mean height})^2 \\ + \\ (* - \quad\cdot\quad)^2 \\ + \\ * - \cdot \\ * \quad \cdot \quad \cdot \end{array}\right]}$$

$$\Downarrow$$

Standard deviation

$= \sqrt{\text{Variance}}$

= The square root of the mean of the deviations-squared from the mean itself

Variance

= The mean of the deviations-squared from the mean itself.

- If the student's heights (in the population) are normally distributed, $X \sim N(\mu, \sigma)$, then approximately 68% of student's heights in the population will fall between $\mu - 1 * \sigma$ and $\mu + 1 * \sigma$.

- In other words, $P(\mu - 1 * \sigma \leq X \leq \mu + 1 * \sigma)$

$$\approx 0.68$$

- $P(\mu - 2 * \sigma \leq X \leq \mu + 2 * \sigma) \approx 0.95$

- $P(\mu - 3*\sigma \leq X \leq \mu + 3*\sigma) = 0.997$

$\boxed{\text{68 - 95 - 99.7 rule}}$ $\Rightarrow$ applicable only when the underlying quantity follows a normal distribution

$$\bar{X}_{10} = \frac{X_1 + X_2 + \cdots + X_{10}}{10}$$

1st random student's height

Sample mean

## 1st Simulation

$$\bar{x}_{10}^{(1)} = \frac{x_1^{(1)} + x_2^{(1)} + \cdots + x_{10}^{(1)}}{10}$$

## General Simulation

$$\bar{x}_{10} = \frac{x_1 + x_2 + \cdots + x_{10}}{10}$$

$\boxed{\text{Random Variable}}$

$\hookrightarrow \bar{X}_{10}$

$$\bar{X}_{10} \to \bar{x}_{10}^{(1)}$$

$$\bar{X}_{10} \to \bar{x}_{10}^{(2)}$$

$$\bar{X}_{10} \to \vdots$$

$\bar{X}_{10}$ = sample mean (random variable)

has its own distribution

$$\boxed{\bar{X}_{10} \sim N(\mu = 170, \sigma = 2.6)}$$

$$\bar{X}_{10} = \frac{X_1 + X_2 + \cdots + X_{10}}{10}$$

$$E\left[\bar{X}_{10}\right] = E\left[\frac{X_1 + X_2 + \cdots + X_{10}}{10}\right]$$

$\Downarrow$

Expected value of sample mean

$\Downarrow$ Expectation is linear

$$\frac{1}{10} \times \left( E[X_1] + E[X_2] + \cdots + E[X_n] \right)$$

Expected values of a sum of random variables = Sum of the expected values of the random variables

$$\underbrace{E\left[\alpha H + \beta W\right]}_{\text{Expected BMI}} = \alpha \underbrace{E[H]}_{\text{Expected height}} + \beta \underbrace{E[W]}_{\text{Expected weight}}$$

$H$: Height

$W$: Weight

Sample (IID)

Independent and Identically Distributed Random Variables

$$X_1 \sim N(\underset{=170}{\mu}, \underset{=8}{\sigma})$$

$$X_2 \sim N(\mu = 170, \sigma = 8)$$

$$\vdots$$

$$X_{10} \sim N(\mu = 170, \sigma = 8)$$

Population

$$X \sim N(\underset{170}{\underset{\parallel}{\mu}}, \underset{8}{\underset{\parallel}{\sigma}})$$

$$\mathrm{Var}[\underline{\text{Random variable}}] = E\left[\left(\text{Random variable} - E[\text{Random variable}]\right)^2\right]$$

$$E[\bar{X}_{10}] = \frac{1}{10} * \left(\underline{E[X_1]} + E[X_2] + \cdots + E[X_{10}]\right)$$

This is a random variable

$$= \frac{1}{10} * \left(\underline{170} + \underline{170} + \cdots + \underline{170}\right)$$

$$= 170 = \mu \quad (\text{Population mean})$$

IID

$$\mathrm{Var}[\bar{X}_{10}] = \mathrm{Var}\left[\frac{1}{10} * (X_1 + X_2 + \cdots + X_{10})\right]$$

$$\| \quad \sigma^2/n = \frac{1}{10^2}\left[\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \cdots + \mathrm{Var}[X_{10}]\right]$$

$$= \frac{1}{10^2}\left[8^2 + 8^2 + \cdots + 8^2\right] = \frac{1}{10} \times 8^2$$

$$SD[\bar{X}_{10}] = \sqrt{Var[\bar{X}_{10}]} = \sqrt{\frac{1}{10} * 8^2} = \frac{1}{\sqrt{10}} * 8$$

$\sigma/\sqrt{n}$

$\|$

Sample mean is a random variable

What will be $SD[\bar{X}_{100}] = ?$

$$= \frac{X_1 + X_2 + \cdots + X_{100}}{100}$$

Sample size $= 100$

$$= \frac{1}{\sqrt{100}} * 8$$

In general, if sample size $= n$, the $SD[\bar{X}_n] = \frac{1}{\sqrt{n}} * \sigma$

$$\bar{X}_{10} \sim N\left(\mu = 170, \ \sigma = \frac{1}{\sqrt{10}} * 8\right)$$

In general $\bar{X}_n \sim N\left(\mu, \frac{1}{\sqrt{n}}\sigma\right)$

As $n$ increases the simulated! (or) realized sample mean values will be close to the population mean $\mu$

Central limit theorem: If $X_1, X_2, \ldots, X_n$ are (I.I.D) random variables (Normal, exponential, Beta, etc.)

Sample mean $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$, where $\begin{cases} \mu \text{ is the } E[X_1], \ldots E[X_n] \\ \sigma \text{ is the } SD[X_1], \ldots SD[X_n] \end{cases}$

$$P\left(160 < X < 180\right), \text{ given } X \sim N(\mu = 170, \sigma = 8)$$

$$= P\left(\frac{160 - E[X]}{SD[X]} < \frac{X - E[X]}{SD[X]} < \frac{180 - E[X]}{SD[X]}\right)$$

$$= P\left(\frac{160-170}{8} < Z < \frac{180-170}{8}\right)$$

$$\boxed{Z \sim N(\mu = 0, \sigma = 1)}$$

= Probability that a standard normal random variable is in between $-\frac{10}{8}$ and $\frac{10}{8}$

Recall that for a standard normal random variable $Z$, $z_\alpha$ is the value such that

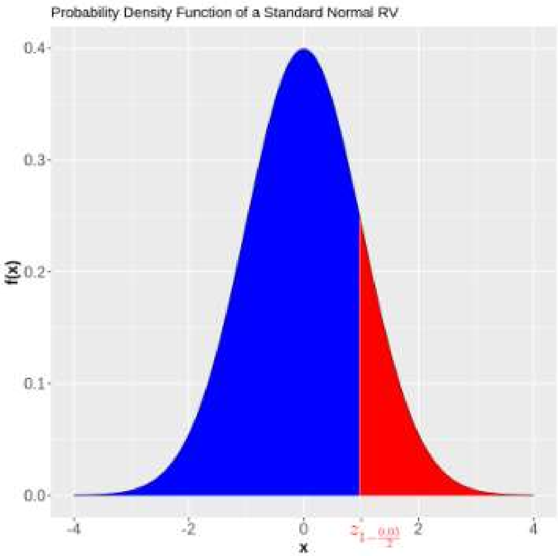
$$P(Z \leq z_\alpha) = \alpha.$$

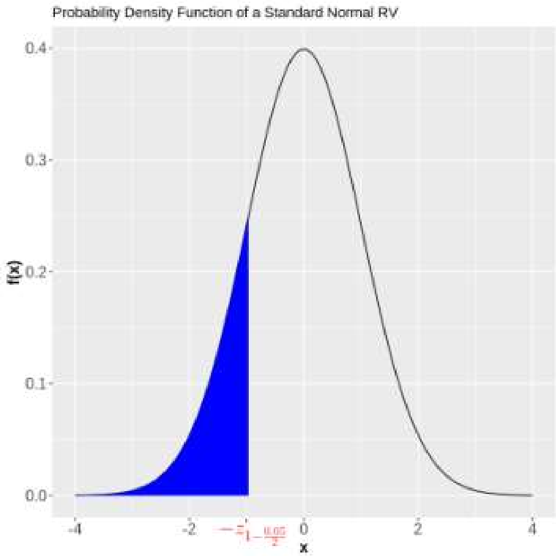
The value $z_\alpha$ is called the $\alpha$-th quantile or the $(100\alpha)$-th percentile of the standard normal distribution.

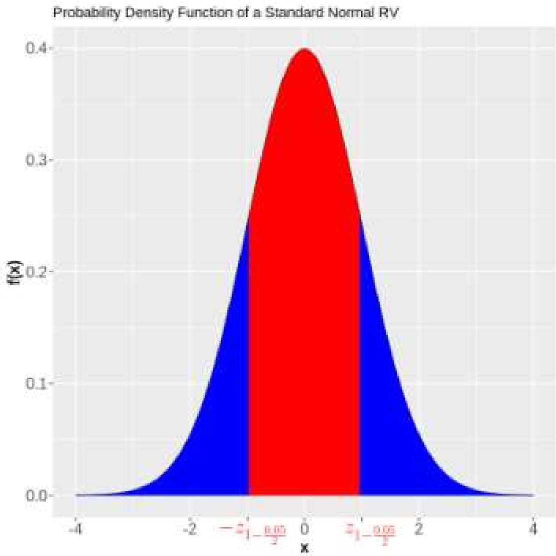
For example, the $.95$-th quantile or the $95$-th percentile of the standard normal distribution is a value denoted as $z_{0.95}$ such that

$$P(Z <= z_{0.95}) = 0.95.$$

Write the areas of the shaded regions in the figures below in terms of 0.05. Recall that the area under the whole curve is the probability $P(-\infty < Z < \infty) = 1$.




Suppose $X \sim \mathrm{N}(\mu = 75, \sigma = 10)$ represents the weight of a randomly chosen individual in a population. Assume that we do not know the population mean and standard deviation. All we can do is to sample individuals from the population. We know that we can look at a small random sample of, say of sample size $= n$, individuals and approximate the population mean using the sample mean

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We note that the sample mean $\bar{X}_n$ is a random variable; every time we choose a random sample of $n$ individuals, we will get a different sample mean. Suppose we want to be 95% confident that the sample mean we got is a good enough approximation of the population mean. For this purpose we calculate a so-called 95% confidence interval (CI) as follows by setting $1 - \alpha = 0.95$:

$$95\% \text{ CI} = \left[\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_n}{\sqrt{n}} \quad \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_n}{\sqrt{n}}\right],$$

where $\hat{\sigma}_n$ is the sample standard deviation. Note that this interval changes every simulation.

The 95% confidence interval (using $1-\alpha = 0.95$, where $\alpha$ represents the diffidence level and $1-\alpha$ represents the confidence level) for a standard normal random variable $Z \sim N(\mu=0, \sigma=1)$

is $$\left[-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}\right] = \left[\underbrace{-Z_{0.975}}_{\text{negative of the 0.975th quantile}}, \underbrace{Z_{0.975}}_{\text{positive of the 0.975th quantile}}\right]$$

- We want to be confident about our sample mean

$$\text{Sample mean} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \text{ a random variable}$$

- $$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - E[\bar{X}_n]}{SD[\bar{X}_n]} \sim N(0, 1)$$

$E[\bar{X}_n] =$

$Z$

Standardizing the random variable $\bar{X}_n$, the sample mean

A standard normal random variable

$$P\left(-z_{1-\frac{\alpha}{2}} \le \underbrace{Z}_{N(0,1)} \le z_{1-\frac{\alpha}{2}}\right) = \underbrace{1-\alpha}_{\text{confidence level}}$$

$\Downarrow$ difference level

$$P\left(-z_{1-\frac{\alpha}{2}} \le \frac{\bar{X}_n - E[\bar{X}_n]}{SD[\bar{X}_n]} \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow P\left(-z_{1-\frac{\alpha}{2}} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\mu$ ⇓ population mean

$$\Rightarrow P\left(\text{The random interval } \left[\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] \text{ contains } \mu\right)$$

$$= 1-\alpha$$

If $1-\alpha = 0.95 \Rightarrow P\left(\text{The random interval } \left[\bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}}\right] \text{ contains } \mu\right) = 0.95$

95% confidence interval for the population parameter $\mu$ = $\left[ \bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \ \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}} \right]$

What about $\sigma$? We don't know it!

$\sigma$ = population standard deviation

$\approx$ sample standard deviation = $\hat{\sigma}$

95% CI

$$= \left[ \bar{X}_n - 1.96 \frac{\hat{\sigma}}{\sqrt{n}} , \ \bar{X}_n + 1.96 \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

mean(data) - qnorm(1 - alpha/2, mean = 0, sd = 1) * sd(data)/sqrt(length(data))

E.g. [169.8337, 172.2816] is the 95% CI that Senthil obtains