Sampling Process

- Select (sample) r objects out of n distinct objects
- Why do we want to sample?
- How do we sample?
- $r = 3$ students out of $n = 28$ students

	With replacement	Without replacement
Order matters	Select 3 out of 28 students to answer 3 labeled questions with more than one question per student allowed	Select 3 out of 28 students to answer 3 labeled questions with exactly one question per student
Order does not matter	Select 3 out of 28 students to answer 3 unlabeled questions with more than one question per student allowed	Select 3 out of 28 students to answer 3 unlabeled questions with exactly one question per student

$S_1 S_2 S_3$, $S_1 S_3 S_2$, $S_1 S_1 S_1$, ⋮

$S_1 S_2 S_3$, ~~$S_1 S_3 S_2$~~, $S_1 S_1 S_1$

$S_1 S_2 S_3$, $S_1 S_3 S_2$, ~~$S_1 S_1 S_1$~~

$S_1 S_2 S_3$, ~~$S_1 S_3 S_2$~~, ~~$S_1 S_1 S_1$~~

$r=3$ out of $n=28$ students (with replacement, order matters)

In counting, the word
(1) AND corresponds to $*$
(2) OR corresponds to $+$

$$28 * 28 * 28 = 28^3 = n^r$$

$r=3$ out of $n=28$ students (without replacement, order matters)

$$28*27*26 = \frac{28\times 27\times 26 \times 25 \times 24 * \cdots * 1}{25 * 24 * \cdots * 1} = \frac{28!}{25!} = \frac{28!}{(28-3)!}$$

$$= {}_{28}P_3 \quad (\text{no. of 3-permutations of 28 objects})$$

$$= {}_nP_r$$

$r=3$ out of $n=28$ students (without replacement, order does not matter)

$S_1\, S_2\, S_3 \Rightarrow$ $S_1 S_2 S_3$, $S_1 S_3 S_2$, $S_2 S_1 S_3$, $S_2 S_3 S_1$, $S_3 S_1 S_2$, $S_3 S_2 S_1$

$S_1\, S_2\, S_4 \Rightarrow$ ⋮

$\Rightarrow$ no. of selections without replacement and order does not matter
$*$
no. of arrangements that each selection results in
$=$
no. of selections without replacement and order matters

$\Rightarrow$ no. of selections without replacement and order doesn't matter $=$ $\dfrac{\text{no. of selections without replacement and order matters}}{\text{no. of arrangements that each selection on the left results in}}$

$$= \frac{{}_{28}P_3}{{}_3P_3} = \frac{28!/25!}{3!/(3-3)!}$$

$0! = 1$

$$= \frac{28!}{3!\,(28-3)!} = {}_{28}C_3 = \binom{28}{3}$$

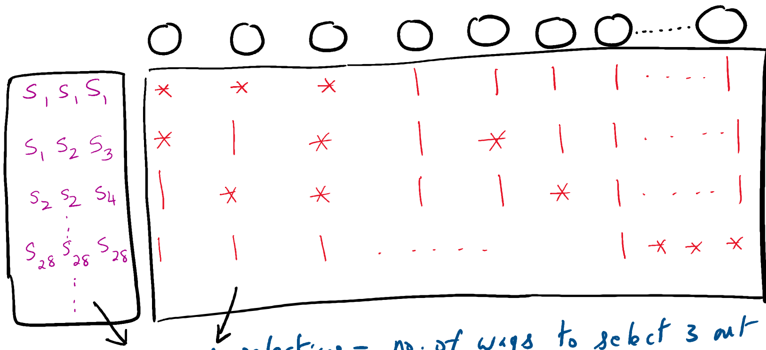
$$= {}_nC_r = \binom{n}{r}$$

$$= \frac{28!}{3!\,(28-3)!} = {}^{28}C_3 = \binom{28}{3}$$

$$= {}^{n}C_r = \binom{n}{r}$$

$r = 3$ out of $n = 28$ students (with replacement, order does not matter)

Draw $(\overset{27 \text{ bars}}{28-1}) + \overset{3 \text{ stars}}{3}$ circles (or) slots



no. of selections = no. of ways to select 3 out of 30 slots without replacement and order does not matter to put the stars

$$= {}^{30}C_3 = {}^{(28+3-1)}C_3 = {}^{(n+r-1)}C_r$$

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

· Recall $\binom{n}{r} = {}^{n}C_r \Rightarrow$ binomial coefficient

= no. of ways to select r objects out of n distinct objects without replacement and order does not matter

= no. of ways to select n objects out of n distinct objects without replacement and order doesn't matter such that r objects fall into the 1st group and the remaining $n-r$ into the 2nd group

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

· Consider this scenario: from 10 students, I want to pick 3 for literature survey, 4 for coding, and 3 for documentation. How many ways can we do this?

$$\underset{\text{Literature Survey}}{\binom{10}{3}} * \underset{\text{coding}}{\binom{7}{4}} * \underset{\text{Documentation}}{\binom{3}{3}}$$

Literature Survey and coding and Documentation

$$= \frac{10!}{3!\,\cancel{(10-3)!}} * \frac{\cancel{7!}}{4!\,\cancel{(7-4)!}} * \frac{\cancel{3!}}{3!\,(3-3)!} = \frac{\overset{n}{10!}}{3!\,4!\,3!}$$

$$= \frac{10!}{3!\,\cancel{(10-3)}!} * \frac{7!}{4!\,\cancel{(7-4)}!} * \frac{3!}{3!\,\underbrace{(3-3)!}_{0!=1}} = \frac{10!}{3!\,4!\,3!}$$

$r_1 = 3$ $r_2 = 4$ $r_3 = 3$

Compare the above with $\binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!\,7!}$

$10! \to n$, $r_1 = 3$, $r_2 = 4$

$$\binom{n}{r} = \binom{10}{3} = \frac{10!}{3!\,7!} \Rightarrow \boxed{\text{binomial coefficient}}$$

$$\binom{n}{r_1, r_2, r_3} = \binom{10}{3,4,3} = \frac{10!}{3!\,4!\,3!} \Rightarrow \boxed{\text{multinomial coefficient}}$$

Ideas towards Probability

E.g. $\boxed{\text{Consider tossing 2 fair coins}}$ → Random experiment

Possible outcomes in one trial

one simulation

Sample space $S = \left\{ \underset{1/4}{(HH)}, \underset{1/4}{(HT)}, \underset{1/4}{(TH)}, \underset{1/4}{(TT)} \right\}$ → outcomes are equally likely.

How did we build the sample space?

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

No. of outcomes in the sample space S = no. of ways to select 2 objects from 2 objects (in the sampling space) with replacement and order matters $= 2^2 = 4$

E.g. Emergency hospital administrator
- patient is in good/fair/serious condition
- patient is insured/not insured

0.1×0.6 $\quad 0.1 \times 0.4$

Sample space $S = \left\{ \begin{matrix} g0, & g1 \\ 0.2 \times 0.6 & 0.2 \times 0.4 \\ f0, & f1 \\ 0.7 \times 0.6 & 0.7 \times 0.4 \\ s0, & s1 \end{matrix} \right\}$ Sample space with not equally likely outcomes.

Sampling spaces
- $\mathcal{S}_1 = \{g, f, s\}$ (10% 20% 70%)
- $\mathcal{S}_2 = \{0, 1\}$ (60% 40%)

- Load the csv file emergency.csv onto a dataframe dfPatient (read_csv())
- What are the features? (str(dfPatient))
- How many levels in the two features? (contrasts())
- How likely is it for the next patient to be in

- How many levels in the two features? (contrasts())
- How likely is it for the next patient to be in (1) good (2) fair (3) serious conditions?
- How likely is it for the next patient to be (1) insured (2) not insured?
- Given that the next patient is insured, how likely is it for them to be in a serious condition?
- Given that the next patient is in a serious condition, how likely is it for them to be insured?

Encoding a categorical column

Condition → (1) good, (2) $\boxed{\text{fair}}$, (3) serious

	Label encoding (as factor)	Dummy encoding (building models) Condition good	Condition serious
(1) good	2	1	0
(2) fair	1	0	0
(3) serious	3	0	1

one-hot encoding (building models)

	Condition fair	condition good	Condition serious
good	0	1	0
fair	1	0	0
serious	0	0	1

Revisit the 5-judges problem:

Sample space $S = \left\{ \begin{array}{l} \underset{\cancel{1/32}}{(00000)}, \underset{\cancel{1/32}}{(11110)}, \ldots \\ \underset{\cancel{1/32}}{(11111)}, \underset{\cancel{1/32}}{(00001)} \ldots \end{array} \right\}$

Are these outcomes equally likely or not? Intuitively not equally likely.

Sampling space $s = \{1, 0\}$

- no. of outcomes in the sample space = no. of ways to select $r=5$ objects from $n=2$ objects with replacement and order matters $= n^r = 2^5 = 32$ outcomes
- How to calculate the likelihood values for the outcomes in the sample space? E.g. outcome $(1\,1\,1\,1\,0) \Rightarrow$ Likelihood $= 0.95 * 0.95 * 0.9 * 0.9 * 0.8$
- The hospital administrator Scenario:

0.1×0.6

$\times 0.8$

- The hospital administrator Scenario:

Sample space $S = \left\{ \begin{array}{cc} \overset{0.1\times 0.4}{(g1)}, & \overset{0.1\times 0.6}{(g0)} \\ \overset{0.2\times 0.4}{(f1)}, & \overset{0.2\times 0.6}{(f0)} \\ \overset{0.7\times 0.4}{(s1)}, & \overset{0.7\times 0.6}{(s0)} \end{array} \right\}$

Sampling spaces

$S_1 = \{ \overset{0.1}{g}, \overset{0.2}{f}, \overset{0.7}{s} \}$

$S_2 = \{ \overset{0.4}{1}, \overset{0.6}{0} \}$

Events of interest

Medical supervisor $E_1 = \{ \overset{0.28}{(s1)}, \overset{0.42}{(s0)} \}$

How likely is it for the next random patient to be in a serious condition?

$\Updownarrow$

Fraction of patients who were serious in the dataset

Insurance supervisor $E_2 = \{ \overset{0.04}{(g1)}, \overset{0.08}{(f1)}, \overset{0.28}{(s1)} \}$

Compound Events

E.g. E_1 AND E_2 $= E_1 \cap E_2 = \begin{cases} \text{patient is serious AND insured} \\ \{ \overset{0.28}{(s1)} \} \end{cases}$

E_1 OR E_2 $= E_1 \cup E_2 = \begin{cases} \text{patient is serious OR insured} \\ \{ \overset{0.28}{(s1)}, \overset{0.42}{(s0)}, \overset{0.04}{(g1)}, \overset{0.08}{(f1)} \} \end{cases}$

$$n(E_1 \cup E_2) = n(E_1) + n(E_2) - n(E_1 \cap E_2)$$

Common-sense Rules

(1) $E_1 \cap E_2 = E_2 \cap E_1$

(2) $E_1 \cup E_2 = E_2 \cup E_1$

(3) $(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$

(4) $(E_1 \cap E_2) \cap E_3 = E_1 \cap (E_2 \cap E_3)$

(5) $(E_1 \cap E_2) \cup E_3 = (E_1 \cup E_3) \cap (E_2 \cup E_3)$

Complement of an event E

$E_1^c = \begin{cases} \text{Patient is not serious} \\ \{(g1), (g0), (f1), (f0)\} \end{cases}$

↳ outcomes in the sample space S that do not contain outcomes in E_1

De Morgan's laws

$(E_1 \cap E_2)^c$ = negation of "patient is serious and insured"

= patient is not serious Or patient is not insured

$E^c \cup E^c$

$= \text{patient is not serious Or patient is not insured}$

$= E_1^c \cup E_2^c$

$(E_1 \cup E_2)^c = \text{negation of "patient is serious or insured"}$

$= \text{patient is not serious And patient is not insured}$

$= E_1^c \cap E_2^c$

$$(E_1 \cup E_2 \cup E_3 \cup E_4)^c = E_1^c \cap E_2^c \cap E_3^c \cap E_4^c$$

Practical definition of Probability

$$P(\text{Event}) = \begin{cases} \text{(1) Simulation} & \text{Long run relative frequency } \lim_{n \to \infty} \frac{n(E)}{n} \\ \text{(2) Pen \& Paper} & \text{Sample space } S = \{ \dots \} \end{cases}$$

sampling space $S = \{ \dots \}$

subset

$E = \{ \dots \}$

Axioms of Probability

Axiom-1 $\quad 0 \le P(E) \le 1$

Axiom-2 $\quad P(S) = 1$

Axiom-3 $\quad$ For mutually exclusive events $E_1, E_2, E_3, \dots$

$$P(E_1 \cup E_2 \cup E_3 \cup \dots) = P(E_1) + P(E_2) + P(E_3) + \dots$$

E.g. Random experiment: roll a pair of fair dice

E_1 = rolls sum up to an odd number	E_2 = rolls are the same	E_3 = rolls sum up to an even number
$E_1 = \{(12),(14),(16),(21),(23),(25), \dots, (61),(63),(65)\}$	$E_2 = \{(11),(22), \dots, (55),(66)\}$	$E_3 = \{(11)\ (13)\ (15)\ (22)\ (24)\ (26)\ \dots\ (62)\ (64)\ (66)\}$

Are E_1 and E_2 mutually exclusive? Yes

Are E_1 and E_3 mutually exclusive? Yes

but E_2 and E_3 are not mutually exclusive

E_1 and E_3 are mutually exclusive and collectively exhaustive

$E_1 \cap E_3 = \{\emptyset\}$ $\qquad E_1 \cup E_3 = S$

E_1 and E_3 are mutually exclusive and collectively exhaustive

$E_1 \cap E_3 = \{\phi\}$ $\qquad$ $E_1 \cup E_3 = S$

$\{$but $E_1 \cup E_2 \subset S\}$

Propositions

(1) $P(E \cup E^c) = P(S) = 1$

$\Downarrow$

$P(E) + P(E^c) = 1 \Rightarrow P(E^c) = 1 - P(E)$

(2) $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

(3) $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_1 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$

(4) $E_1 \subset E_2 \Rightarrow P(E_1) \leq P(E_2)$

E.g. Coin toss problem (toss coin 10 times, probability of getting 3 heads) with replacement, order matters (select 10 out of 2 objects)

$$\text{Sample space } S = \left\{ \begin{array}{ll} \overset{1/2^{10}}{(HHH\cdots H)}, & \overset{1/2^{10}}{(HTHT\cdots HT)} \\ \overset{1/2^{10}}{(THTH\cdots TH)} & \cdots \\ \cdots & \overset{1/2^{10}}{(TTT\cdots T)} \end{array} \right\}$$

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

Sampling with replacement → order does not matter → outcomes not equally likely
→ order matters → outcomes are equally likely

Sampling without replacement → order does not matter → outcomes are equally likely
→ order matters → outcomes are equally likely

$$\text{Event } E = \left\{ \begin{array}{ll} \overset{1/2^{10}}{(HHHTT\cdots T)}, & \overset{1/2^{10}}{(HTHTHTT\cdots T)} \\ \vdots & \\ \cdots & \overset{1/2^{10}}{(TTT\cdots THHH)} \end{array} \right\}$$

$P(E) = \frac{1}{2^{10}} * \text{no. of outcomes in } E = \frac{\text{no. of outcomes in } E}{\text{no. of outcomes in } S}$

$$P(E) = \frac{1}{2^{10}} * \underbrace{\text{no. of outcomes in } E}_{\substack{\text{no. of ways to select 3}\\\text{out of 10 locations (or slots)}\\\text{without replacement and order does not matter} = {}^{10}C_3}} = \frac{\text{no. of outcomes in } E}{\text{no. of outcomes in } S}$$

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

success probability

no. of trials

no. of successes

failure probability

E.g. Lecture scheduling problem

Sampling space $\mathcal{S} = \{$'Mo', 'Tu', 'We', 'Th', 'Fr', 'Sa', 'Su'$\}$

select 3 out of 7

without replacement, order does not matter

Sample Space $S = \{ \overset{1/{}^7C_3}{(Mo, Tu, We)}, \overset{1/{}^7C_3}{(Mo, Tu, Sa)}, \ldots, \overset{1/{}^7C_3}{(Fr, Sa, Su)} \}$

Event $E = \{ \overset{1/{}^7C_3}{(Mo, Tu, We)}, \overset{1/{}^7C_3}{(Mo, Tu, Th)}, \ldots, \overset{1/{}^7C_3}{(Fr, We, Mo)} \}$

$$P(E) = \frac{1}{{}^7C_3} * \underbrace{\text{no. of outcomes in } E}_{\substack{\text{Select 3 out of 5}\\\text{weekdays without replacement}\\\text{and order does not matter}\\= {}^5C_3}} = \frac{{}^5C_3}{{}^7C_3}$$

Conditional Probability

E.g. Random experiment of rolling a die

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

... at least $5 = \{ \overset{1/6}{5}, \overset{1/6}{6} \}$

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

Events
- E = event that the roll is at least 5 $= \{\overset{1/6}{5}, \overset{1/6}{6}\}$
- F = event that the roll is even $= \{\overset{1/6}{2}, \overset{1/6}{4}, \overset{1/6}{6}\}$

$P(E)$
- Sample space-way of calculating $= 1/6 + 1/6 = 1/3$
- Frequency-way (simulation) of calculating ≈ 0.33
  = fraction of times at least 5 shows up
  in, say, 10^6 simulations of the random experiment

Quick recap on compound events
- $E \cap F = \{\overset{1/6}{6}\}$
- $E \cup F = \{\overset{1/6}{2}, \overset{1/6}{4}, \overset{1/6}{5}, \overset{1/6}{6}\}$

Conditional event $E|F$ = E given F is known to have happened

$P(E|F)$ = Conditional probability of E happening given that F has happened

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

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

reduced sample space
(or) conditional sample space

$E|F = \{\overset{1/3}{6}\} \Rightarrow \boxed{P(E|F) = 1/3}$

$$\boxed{P(E) = 1/3 \quad , P(E|F) = 1/3} = \frac{P(E\cap F)}{P(F)} = \frac{1/6}{1/2} = 1/3$$

Now consider
- E = roll is at least 3
- F = roll is even

$\Rightarrow$
- $P(E) =$
- $P(E|F) =$

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

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

$P(E) = 4 * \frac{1}{6} = \frac{2}{3}$

$E \cap F = \{\overset{1/6}{4}, \overset{1/6}{6}\}$

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

$E|F = \{\overset{1/3}{4}, \overset{1/3}{6}\}$

$P(E|F) = \frac{2}{3} = \frac{P(E\cap F)}{P(F)} = \frac{2/6}{3/6} = 2/3$

Now consider this
- E = roll is atmost 3
- F = roll is even

(F = roll is even

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

$E = \{ \overset{1/6}{1}, \overset{1/6}{2}, \overset{1/6}{3} \}, F = \{ \overset{1/6}{2}, \overset{1/6}{4}, \overset{1/6}{6} \}$

$P(E) = 3 * 1/6 = 1/2$

$E \cap F = \{ \overset{1/6}{2} \}$

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

$E | F = \{ \overset{1/3}{2} \}$

$P(E|F) = 1/3 = \dfrac{P(E \cap F)}{P(F)} = \dfrac{1/6}{1/2} = 1/3$

E.g. I have a medical kit to detect HIV

Test kit company
- The test is 98% effective in detecting HIV ? $\boxed{P(E|F)}$? $P(F|E)$
  True positive rate
- The False positive rate is very small = 1% $P(E|F^c)$

Policy maker
- The percentage of people in the entire population with HIV = 0.5% $= P(F)$

E = event that a person tests positive for HIV with the test kit

F = event that a person actually has HIV

$$\boxed{\text{The conditional probability formula:} \quad P(E|F) = \frac{P(E \cap F)}{P(F)}}$$

True for all sample spaces: equally likely or not

In the context of probabilistic models in ML, we have

$$\underbrace{\boxed{P(\text{model} \mid \text{data})}}_{\text{Posterior}} = \frac{\underbrace{\boxed{P(\text{data} \mid \text{model})}}_{\text{likelihood}} \underbrace{\boxed{P(\text{model})}}_{\text{Prior}}}{\boxed{P(\text{data})}}$$

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \Rightarrow P(E \cap F) = P(E|F)P(F)$$

the same

will also be equal

Now we can write $P(F \cap E) = P(F|E)P(E)$

$$\Rightarrow P(E|F)P(F) = P(F|E)P(E)$$

$$\Rightarrow \boxed{P(E|F) = \frac{P(F|E)\,P(E)}{P(F)}}$$

This is the (first version of) the Bayes' formula

$P(F)$

This is the (first version of) the Bayes' formula

Back to the HIV problem

Does it make sense to buy the test kits in bulk and distribute to the population?

- Recall that the medical company says $\begin{cases} P(E|F) = 0.98 \\ P(E|F^c) = 0.01 \end{cases}$ True positive rate (TPR)
- The deciding body in the govt. knows that $P(F) = 0.005$
- The probability of interest for the deciding body

$$= P\left(\begin{matrix}\text{Person has} \\ \text{HIV}\end{matrix} \,\middle|\, \text{Test is positive}\right) = P(F|E)$$

would love this to be 1

$$\Rightarrow P(F|E) = \frac{\overset{0.98}{P(E|F)}\,\overset{0.005}{P(F)}}{P(E) \rightarrow ?}$$

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

$$\begin{matrix}\text{Person tests} \\ \text{positive}\end{matrix} = \begin{bmatrix}\text{Person tests positive} \\ \text{AND} \\ \text{Person has HIV}\end{bmatrix} \text{ OR } \begin{bmatrix}\text{Person tests positive} \\ \text{AND} \\ \text{Person does not have HIV}\end{bmatrix}$$

Mutually exclusive and collectively exhaustive

Recall $\begin{cases} \text{for generic events } A, B,\ P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ \text{for mutually exclusive events } A, B,\ P(A \cap B) = 0 \\ \Rightarrow P(A \cup B) = P(A) + P(B) \end{cases}$

$$\Rightarrow \boxed{P(E) = P(E \cap F) + P(E \cap F^c)}$$

Law of total Probability

$$\Rightarrow P(F|E) = \frac{P(E|F)\,P(F)}{P(E \cap F) + P(E \cap F^c)}$$

$P(E \cap F) = P(E|F)P(F)$

$\Updownarrow$

$P(E|F) = \dfrac{P(E \cap F)}{P(F)}$

$$= \frac{\underset{0.98}{P(E|F)}\,\underset{0.005}{P(F)}}{\underset{0.98}{P(E|F)}\,\underset{0.005}{P(F)} + \underset{0.01}{P(E|F^c)}\,\underset{1-P(F)=1-0.005}{P(F^c)}}$$

$$\Rightarrow \boxed{P(F|E) \approx 0.33 = 33\%}$$

$$\Rightarrow \boxed{P(F|E) \approx 0.33 = 33\%}$$

Govt. decides not to procure the kit in bulk.

But as an individual, is it good to get tested?

$$P\left(\begin{matrix}\text{Person does}\\ \text{not have HIV}\end{matrix} \middle| \text{Test result negative}\right) = P(F^c | E^c)$$

$\boxed{\text{The complement of the conditional event } F|E \text{ is } F^c|E}$

$$P(\underbrace{F^c|E^c}_{\text{Complement events}}) + P(\underbrace{F|E^c}) = \begin{cases} P(E^c)\ ? \\ 1\ ?\ \checkmark \end{cases}$$

$$P(F^c) + P(F) = 1$$

$$\Rightarrow P(F^c|E^c) = 1 - P(F|E^c) \rightarrow \frac{P(E^c|F)P(F)}{P(E^c)}$$

$$0.005 = \frac{P(E^c|F)P(F)}{P(E^c \cap F) + P(E^c \cap F^c)}$$

$$= 1 - \frac{P(E^c|F)\,P(F)}{\underset{1-0.98}{P(E^c|F)}\,\underset{0.005}{P(F)} + \underset{\substack{1-P(E|F^c)\\ =1-0.01}}{P(E^c|F^c)}\,\underset{1-0.005}{P(F^c)}}$$

$$1 - P(E|F) = 1 - 0.98$$

$$= 99.8\%$$

$$\underset{(TPR)}{\text{True positive rate}} = P\left(\begin{matrix}\text{Test}\\ \text{positive}\end{matrix} \middle| \text{has the disease}\right) = P(E|F)$$

$$\underset{(FNR)}{\text{False negative rate}} = P\left(\begin{matrix}\text{Test}\\ \text{negative}\end{matrix} \middle| \text{has the disease}\right) = P(E^c|F)$$

$$\Rightarrow TPR + FNR = 1 \Rightarrow \boxed{TPR = 1 - FNR}$$

$$\underset{(TNR)}{\text{True negative rate}} = P\left(\begin{matrix}\text{Test}\\ \text{negative}\end{matrix} \middle| \begin{matrix}\text{does not have}\\ \text{the disease}\end{matrix}\right) = P(E^c|F^c)$$

$$\underset{(FPR)}{\text{False positive rate}} \quad P\left(\begin{matrix}\text{Test}\\ \text{positive}\end{matrix} \middle| \begin{matrix}\text{does not have}\\ \text{the disease}\end{matrix}\right) = P(E|F^c)$$

$$\Rightarrow TNR + FPR = 1 \Rightarrow \boxed{TNR = 1 - FPR}$$

Suppose we assume that 5% of people are drug-users. A test is 95% accurate, which means that if a person is a drug-user then the test result is positive 95% of the time; and if the person is not a drug-user then the test result is negative 95% of the time. Let us define the following events:

U = person is a drug user,

P = test positive.

(d) $P(U) = 0.05$ ✓

TPR $P(P|U) = 0.95$ ✓

TNR $P(P^c|U^c) = 0.95$

(a) Explain what the following quantities mean:

- $P(U \mid P^c)$
- $P(P \mid U^c)$
- $P(P \,\text{and}\, U)$
- $P(U^c \mid P)$

$$P(U|P) = \frac{P(P|U)\,P(U)}{P(P)} = \frac{0.95 * 0.05}{P(P|U)P(U) + P(P|U^c)P(U^c)}$$

$P(P \cap U) + P(P \cap U^c) \leftarrow P(P)$

$$= \frac{0.95 * 0.05}{0.95 * 0.05 + (1-0.95)(1-0.05)} = 0.5$$

(b) What quantity above would be of interest if we are concerned about making a wrong accusation?

(c) What quantity above would be of interest if we are concerned about high safety standard when recruiting some one for a job?

(d) A randomly chosen person tests positive. Is the individual highly likely to be a drug-user? Show your reasoning clearly.

Independent events

- Conditional probability $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

- Events A and B are (statistically) independent if

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

$$\Rightarrow \boxed{P(A \cap B) = P(A) * P(B)}$$

AND becomes a multiplication w.r.t. probability only if the connected events are independent

We say events A and B are dependent

if $P(A \cap B) = P(A|B) * P(B)$

10. **[Coding]** Your cell phone is constantly trying to keep track of where you are. At any given point in time, for all nearby locations, your phone stores a probability that you are in that location. Right now your phone believes that you are in one of 16 different locations arranged in a grid with the following probabilities (see the figure on the left):

i = 1 DS Lab

j = 1

Prior Belief of Location

P =

0.05	0.10	0.05	0.05
0.05	0.10	0.05	0.05
0.05	0.05	0.10	0.05
0.05	0.05	0.10	0.05

Prior matrix

P(Observe two bars of signal | Location)

L =

0.75	0.95	0.75	0.05
0.05	0.75	0.95	0.75
0.01	0.05	0.75	0.95
0.01	0.01	0.05	0.75

Likelihood matrix

$$P(\text{Person is at } (i,j) \mid \text{observed 2 bars}) = \frac{P(\text{observe 2 bars} \mid \text{Person is at } (i,j)) * P(\text{Person is at } (i,j))}{P(\text{observe 2 bars})}$$

0.75

0.05

?

Bayes' formula

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

$P(E|F) = P(F|E)\,P(E)$

Your phone connects to a known cell tower and records two bars of signal. For each grid location L_i you know the probability of observing two bars from this particular tower, given that the cell phone is in location L_i (see the figure on the right). That value is based on knowledge of the dynamics of this particular cell tower and stochasticity of signal strength.

Example: the highlighted cell on the left figure means that you believed there was a 0.05 probability that the user was in the bottom right grid cell prior to observing the cell tower signal. The highlighted cell on the right figure means that you think the probability of observing two bars, given the user was in the bottom right grid cell, is 0.75.

For each of the 16 location positions, calculate the new probability that the user is in each location given the cell tower observation. Write a program to calculate the probabilities. The matrices are provided on the website on the problem set #2 page. Report the probabilities of

signal. The highlighted cell on the right figure means that you think the probability of observing two bars, given the user was in the bottom right grid cell, is 0.75.

For each of the 16 location positions, calculate the new probability that the user is in each location given the cell tower observation. Write a program to calculate the probabilities. The matrices are provided on the website on the problem set #2 page. Report the probabilities of all 16 cells and write a short explanation of your program. The grid in the left figure is stored in a file called "prior.csv" the grid in the right figure is stored in a file called "conditional.csv"

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

$$P(E|F) = \frac{P(F|E)\,P(E)}{P(F)}$$

$$P(\text{Event-1}|\text{Event-2}) = \frac{P(\text{Event-2}|\text{Event-1})\,P(\text{Event-1})}{P(\text{Event-2})}$$

(1,1) = AIML Lab

$$\underbrace{P\left(\text{Person is at }(1,1) \,\middle|\, \text{observe 2 bars of signal}\right)}_{\text{Posterior}}$$

$$= \frac{\overbrace{P\left(\text{observe 2 bars of signal} \,\middle|\, \text{Person is at }(1,1)\right)}^{\text{likelihood}} * \overbrace{P\left(\text{Person is at }(1,1)\right)}^{\text{Prior}}}{P(\text{observe 2 bars of signal})}$$

Law of total Probability

$$\left\{ \text{Observe 2 bars of signal} = \left[\begin{matrix}\text{observe 2 bars of signal}\\ \text{AND}\\ \text{Person is at }(1,1)\end{matrix}\right] \text{ OR } \left[\begin{matrix}\text{observe 2 bars of signal}\\ \text{AND}\\ \text{Person is not at }(1,1)\end{matrix}\right] \right.$$

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

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

$$= P(E|F)P(F) + P(E|F^c)P(F^c)$$

Recall

$$\boxed{\begin{matrix} P(E|F) = \dfrac{P(E \cap F)}{P(F)} \\ P(E \cap F) = P(E|F)P(F) \end{matrix}}$$

$$= P\left(\text{Observe 2 bars of signal} \,\middle|\, \text{person at }(1,1)\right) * P\left(\text{Person at }(1,1)\right)$$

$$+ P\left(\text{observe 2 bars of signal} \,\middle|\, \text{Person not at }(1,1)\right) * P\left(\text{person not at }(1,1)\right)$$

$$P\left(\overset{A}{\text{Person is at }(1,1)} \,\middle|\, \overset{B}{\text{observe 2 bars of signal}}\right)$$

$$= \frac{P\left(\overset{B}{\text{observe 2 bars of Signal}} \,\middle|\, \overset{A}{\text{Person at }(1,1)}\right) * P\left(\overset{A}{\text{Person at }(1,1)}\right)}{P\left(\overset{B}{\text{observe 2 bars of signal}} \,\middle|\, \overset{A}{\text{Person at }(1,1)}\right) * P\left(\overset{A}{\text{Person at }(1,1)}\right) + P\left(\overset{B}{\text{observe 2 bars of signal}} \,\middle|\, \overset{A^c}{\text{Person not at }(1,1)}\right) * P\left(\overset{A^c}{\text{Person not at }(1,1)}\right)}$$

.25

(bars of signal) (1,1)) ((1,1)) signal | at (1,1))

$$= \frac{0.75 * 0.05}{0.75 * 0.05 + ?}$$

$= 1 - 0.05$

$$P\left(\begin{matrix}\text{observe 2} \\ \text{bars of signal}\end{matrix} \;\middle|\; \begin{matrix}\text{Person not} \\ \text{at } (1,1)\end{matrix}\right) = \frac{P\left(\begin{matrix}\text{observe 2 bars of signal} \\ \text{AND} \\ \text{Person not at } (1,1)\end{matrix}\right)}{}$$

$$P\left(\begin{matrix}\text{Observe 2 bars of signal AND person at } (1,2) \\ \text{OR} \\ \text{observe 2 bars of signal AND person at } (1,3) \\ \text{OR} \\ \vdots \\ \text{observe 2 bars of signal AND person at } (4,4)\end{matrix}\right)$$

mutually exclusive

$$\begin{aligned} = \; & P(\text{observe 2 bars of signal AND Person at } (1,2)) \\ & + P(\text{observe 2 bars of signal AND person at } (1,3)) \\ & + \vdots \\ & + P(\text{observe 2 bars of signal AND person at } (4,4)) \end{aligned}$$

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

$$\begin{aligned} = \; & P(\text{observe 2 bars of signal} \mid \text{Person at } (1,2)) * P(\text{Person at } (1,2)) \\ & + P(\text{observe 2 bars of signal} \mid \text{Person at } (1,3)) * P(\text{Person at } (1,3)) \\ & + \vdots \\ & + P(\text{observe 2 bars of signal} \mid \text{Person at } (4,4)) * P(\text{Person at } (4,4)) \end{aligned}$$

$$P\left(\begin{matrix}\text{Person at} \\ (1,1)\end{matrix} \;\middle|\; \begin{matrix}\text{Observe 2} \\ \text{bars of signal}\end{matrix}\right)$$

$$\underset{0.05}{P_{11}} * \underset{0.75}{L_{11}}$$

$$= \frac{P_{11} * L_{11}}{P_{11} * L_{11} + P_{12} * L_{12} + P_{13} * L_{13} + \cdots + P_{44} * L_{44}}$$

$$P\left(\begin{array}{c}\text{Person at}\\(i,j)\end{array} \middle| \begin{array}{c}\text{observe 2}\\\text{bars of signal}\end{array}\right) = \frac{P_{ij} * L_{ij}}{\sum_{i,j=1}^{4} P_{ij} * L_{ij}}$$

Digression

Hadamard product of matrices

E.g. $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \circledast \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$

$$\begin{bmatrix} 1*7 & 2*8 & 3*9 \\ 4*10 & 5*11 & 6*12 \end{bmatrix}$$

. Suppose that an insurance company classifies people into one of three classes: good risks, average risks, and bad risks. The company's records indicate that the probabilities that given good-,average-,and bad-risk persons will be involved in an accident over a 1-year span are, respectively, 0.05, 0.15, and 0.30. Suppose 20% of the population is a good risk, 50% an average risk, and 30% a bad risk.

(a) What proportion of people have accidents in a fixed year?
Hint: *law of total probability followed by Bayes' theorem.*

(b) If policyholder A had no accidents in 1997, what is the probability that he or she is a good or average risk?

$P\left(\begin{array}{c}\text{good}\\\text{or}\\\text{avg}\end{array} \middle| \text{No accident}\right)$

Proportion / fraction / Percentage } Probability from a frequency (data) perspective $\Rightarrow P\left(\begin{array}{c}\text{next random person would}\\\text{have met with an accident}\end{array}\right)$

$$P(\text{accident}) = P\left(\begin{bmatrix}\text{accident}\\\text{AND}\\\text{good}\end{bmatrix} \text{ OR } \begin{bmatrix}\text{accident}\\\text{AND}\\\text{average}\end{bmatrix} \text{ OR } \begin{bmatrix}\text{accident}\\\text{AND}\\\text{bad}\end{bmatrix}\right)$$

Law of total probability

$$= P(\text{accident AND good}) + P(\text{accident AND average}) + P(\text{accident AND bad})$$

$P(\text{good} \cancel{|} \text{accident})$

$P(\text{good } \cancel{\text{and}} \text{ accident})$

$$= P(\text{accident} \mid \text{good}) * P(\text{good})$$

0.05, 0.2, 0.15, 0.5

$S = \{1, 2, 3, 4, 5, 6\}$ (each $\tfrac{1}{6}$)

A = at most 3 $= \{1, 2, 3\}$

B = odd $= \{1, 3, 5\}$

C = even $= \{2, 4, 6\}$

D = at least 5

$$= P(\text{accident} \mid \text{good}) * P(\text{good})$$
$$+ P(\text{accident} \mid \text{average}) * P(\text{average})$$
$$+ P(\text{accident} \mid \text{bad}) * P(\text{bad})$$

0.15, 0.5, 0.3, 0.3

C = even $= \{2, 4, 6\}$ 1/6 1/6 1/6

D = at least 5 $= \{5, 6\}$ 1/6 1/6

$A \cap D = \{\phi\}$

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

mutually exclusive

Mutually exclusive and collectively exhaustive

(b) $P(\text{good or average} \mid \text{No accidents})$

$$= P(\text{good} \mid \text{No accidents}) + P(\text{average} \mid \text{No accidents})$$

Bayes' formula

$$= \frac{P(\text{No accident} \mid \text{good}) * P(\text{good})}{P(\text{No accident})} + \frac{P(\text{No accident} \mid \text{average}) * P(\text{average})}{P(\text{No accident})}$$

$1 - P(\text{accident} \mid \text{good}) = 1 - 0.05$

0.2

$1 - P(\text{accident}) = 1 - 0.175$

Two events A and B are independent if $\boxed{P(A \mid B) = P(A)}$

$P(\text{Smoker}) = P(\text{Smoker} \mid \text{Female})$ (In MIT)

$P(\text{Smoker}) \neq \underbrace{P(\text{smoker} \mid \text{Female})}_{=0}$ (In rural karnataka)

For independent events A, B, we have $\begin{cases} P(A \cap B) = P(A \mid B) P(B) \\ P(A \cap B) = P(A) * P(B) \end{cases}$

A, B mutually exclusive $\Rightarrow$ A, B are dependent

A, B independent $\Rightarrow$ A, B mutually exclusive?

$\Downarrow$

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

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

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?

$P(A|B)$, $P(A|B,C)$

$P(\text{Win} \mid \underbrace{\text{Choose } A, \text{switch}}_{\text{Compound event}})$

Choose A AND Switch

$= P\left(\boxed{\text{Win AND Prize in } A} \text{ OR } \boxed{\text{Win AND Prize in } B} \text{ or } \boxed{\text{Win AND Prize in } C} \mid \text{choose } A, \text{switch}\right)$

$= P(\text{Win AND Prize in } A \mid \text{choose } A, \text{switch})$

$+ P(\text{Win AND Prize in } B \mid \text{choose } A, \text{switch})$

$+ P(\text{Win AND prize in } C \mid \text{choose } A, \text{switch})$

$P(\text{Prize in } A \mid \text{choose } A, \text{switch}) = P(\text{Prize in } A) = 1/3$

$P(\text{win} \mid \text{Prize in } A, \text{choose } A, \text{switch}) * P(\text{Prize in } A \mid \text{choose } A, \text{switch})$ $= 0$, $= 1/3$

$+$

$= P(\text{win} \mid \text{Prize in } B, \text{choose } A, \text{switch}) * P(\text{Prize in } B \mid \text{choose } A, \text{switch})$ $= 1$, $= 1/3$

$+$

$P(\text{win} \mid \text{Prize in } C, \text{choose } A, \text{switch}) * P(\text{Prize in } C \mid \text{choose } A, \text{switch})$ $= 1$, $= 1/3$

$= 2/3$

20 September 2023
02:40 PM

Without replacement and order does not matter

A circus act needs to select a juggler, a clown and an acrobat from a group of 10 performers. How many different choices of circus artists are possible if:

(a) There are no restrictions. $^{10}C_3 = \frac{10!}{3!(10-3)!}$ = Leave it as such
(b) Two of them (the 10 artists) will not perform together.
(c) Two of them will perform together or not at all.
(d) One of them must be in the act.
(e) One of them can only perform as an acrobat.

(b) Both do not perform OR {one of them performs as juggler or clown or acrobat}

$^8C_3 + {}^2C_1 * [{}^8C_2 + {}^8C_2 + {}^8C_2]$

(c) Two of them perform together OR not at all

$3C. * 8C. + \quad 8C$

(c) Two of them perform together OR not at all

$${}^3C_2 * {}^8C_1 + {}^8C_3$$

(d) $3C_1 * 9C_2$

e) One of them performs as an acrobat (or) that one does not perform at all

$$9C_2 + 9C_3$$

[9 points] [TLO 1.1, CO 1] A student taking a test has to select 10 out of 20 questions. How many different choices does she have if:

(a) there are no other restrictions? $20C_{10}$

(b) she has to answer exactly 2 of the last 5? $5C_2 * 15C_8$

(c) she has to answer at least 3 of the first 5? 3 of first 5 or 4 of first 5 ... or 5 of first 5

$$= 5C_3 * 15C_7 + 5C_4 * 15C_6 + 5C_5 * 15C_5$$

Eleven soccer players are to be divided into 4 functional groups: 3 forwards, 3 midfields, 4 defenses, and 1 goalie. There are only 2 people who can play goalie. Both of these two players can play any other position. Of the remaining 9, 4 can play only forward or midfield; the other 5 can play only defense or midfield. We want to calculate the number of possible ways to divide the team into the 4 functional groups. Follow the hint below and get to the answer:

Select 1 goalie out of 2 in $\binom{?}{?}$ ways **AND** $*$ $\left\{\begin{array}{l}\text{Remaining goalie plays defense} \\ \textbf{OR} + \\ \text{Remaining goalie plays midfield} \\ \textbf{OR} + \\ \text{Remaining goalie plays forward}\end{array}\right.$

$2C_1$

$5C_3 * 4C_3 * 3C_3$

3 defense ✓
3 midfields
3 forwards

4 defense
2 midfields
3 forwards

4 defense
3 midfields
2 forwards

Kara has ten books which she wants to place in her (single shelf) bookshelf arranged by subject. She has 4 probability books, 3 optimization books, 2 decision analysis books and 1 economic book. In how many ways can she arrange her books?

$\boxed{PPPP}$ $\boxed{OOO}$ $\boxed{DD}$ $\boxed{E}$

$$\boxed{4!} * (4! * 3! * 2! * 1!)$$

Mr. Brown needs to take 1 tablet of type A and 1 tablet of type B together on a regular basis. One tablet of type A corresponds to a 1 mg dosage, and so does 1 tablet of type B. He keeps these two types of tablets in two separately labeled bottles as they cannot be differentiated easily. One day, on a buisness trip, Mr. Brown brought 10 tablets of type A and 10 tablets of type B.. Unfortunately, he drops the bottles and breaks them. He does not have the time to go to a pharmacy to buy a new set of tablets but he needs to take his required dosage of both tablets A and B. The safe dosage that he needs for both tablets A and B is given by

$$0.9\,\text{mg} \leq \text{safe dosage} \leq 1.1\,\text{mg}.$$

Taking either an excess or a shortage of the required intake will result in serious health issues.

(a) Suppose that after investigating the broken bottles, Mr. Brown finds 2 tablets that are still intact in the bottle for tablet A. The other 18 tablets are found to be mixed in a pile. Is it better for him to take one known tablet from the bottle and one from the pile, or take two tablets from the pile? Answer this by calculating the respective probabilities that he will not have any serious health issues for both options.

(b) Suppose that after investigating the broken bottles, Mr. Brown finds that the tablets are all mixed up. What is the probability that he will not have any serious health issues if he randomly picks 2 tablets?

Bottle for type-A

$\boxed{A_1, A_2}\ \boxed{A_3, \dots, A_{10}}$ = 10 type-A tablets

$\boxed{B_1, B_2, \dots, B_{10}}$ = 10 type-B tablets

outcomes are equally likely

Sample space = $\left\{ \begin{array}{l} \overset{1/36}{(A_1, A_3)}, \overset{1/36}{(A_1, A_4)} \dots \overset{1/36}{(A_1, A_{10})} \\ (A_1, B_1) \dots (A_1, B_{10}) \\ (A_2, A_3), (A_2, A_4) - \overset{1/36}{\dots} (A_2, A_{10}) \\ (A_2, B_1) \dots (A_2, B_{10}) \end{array} \right\}$

$2C_1 * 18C_1$

Sampling Space

$S_1 = \{A_1, A_2\}$, $S_2 = \{A_3, \dots A_{10}, B_1, B_2, \dots B_{10}\}$

$2C_1 * 10C_1$

$E = \left\{ \begin{array}{l} (A_1, B_1), (A_1, B_2) \dots (A_1, B_{10}) \\ (A_2, B_1), (A_2, B_2) \dots (A_2, B_{10}) \end{array} \right\}$, $n(E) = 2C_1 * 10C_1$

$p(E) = \frac{1}{36} * n(E) = \frac{1}{36} * 20 = 0.56$

(b) Sample space = $\left\{ \begin{array}{l} (A_1, A_2), (A_1, A_3) \dots (A_1, A_{10}) \\ (A_1, B_1), (A_1, B_2) \dots (A_1, B_{10}) \\ \vdots \\ (B_{10}, B_1), (B_{10}, B_2) \dots (B_{10}, B_9) \end{array} \right\}$

$20C_2$

Sampling space $S = \{A_1, A_2, \dots, A_{10}, B_1, B_2, \dots, B_{10}\}$

Event $E = \left\{ \begin{array}{l} (A_1, B_1) \dots (A_1, B_{10}) \\ (A_2, B_1) \dots (A_2, B_{10}) \\ \vdots \\ (A_{10}, B_1) \dots (A_{10}, B_{10}) \end{array} \right\}$

$\Rightarrow P(E)$

$= \frac{1}{n(S)} * n(E)$

$= \frac{10C_1 * 10C_1}{20C_2}$

$= 0.53$

Say in Silicon Valley, 36% of engineers program in Java and 24% of the engineers who program in Java also program in C++. Furthermore, 33% of engineers program in C++.

(a) What is the probability that a randomly selected engineer programs in Java and C++?

(b) What is the probability that a randomly selected engineer who is known to program using C++ also programs in Java?

$P(\text{Java}) = 0.36$ $\quad P(\text{Java AND C++}) \stackrel{?}{=} 0.24$

$P(\text{C++}) = 0.33$ $\quad P(\text{C++} \mid \text{Java}) \stackrel{?}{=} 0.24$ ✓

$$\text{(a) } P(\text{Java AND C++}) = P(\text{Java} \mid \text{C++})\, P(\text{C++})$$
$$= P(\text{C++} \mid \text{Java})\, p(\text{Java})$$
$$= 0.24 * 0.36$$

$$\text{(b) } P(\text{Java} \mid \text{C++}) = \frac{P(\text{Java AND C++})}{P(\text{C++})} = \frac{0.24}{0.33}$$

In a certain community, 36 percent of the families own a dog and 22 percent of the families that own a dog also own a cat. In addition, 30 percent of the families own a cat.

(a) What is the probability that a randomly selected family owns a dog? 0.36 or 36%

(b) What the probability that a randomly selected family owns a cat? 0.3 or 30%.

(c) What probability does 22 percent in the problem statement indicate? $P(\text{Cat} \mid \text{Dog}) = 0.$ (10)

(d) What is the probability that a randomly selected family owns both a dog and a cat? $P(\text{Cat AND Dog}) = P(\text{cat} \mid \text{Dog})\, P(\text{Dog}) = 0.22 * 0.36$

(e) What is the the conditional probability that a randomly selected family owns a dog given that it owns a cat? $P(\text{Dog} \mid \text{cat}) = \frac{P(\text{Dog AND Cat})}{P(\text{cat})} = \frac{0.22 * 0.36}{0.3}$

$P(\text{Dog}) = 0.36$, $P(\text{Cat} \mid \text{Dog}) = 0.22$, $P(\text{cat}) = 0.$

)

$22 = 22'$

3

A doctor assumes that a patient has one of three diseases d_1, d_2, or d_3. Before any test, he assumes an equal probability for each disease. He carries out a test that will be positive with probability 0.8 if the patient has d_1, 0.6 if he has disease d_2, and 0.4 if he has disease d_3. Given that the outcome of the test was positive, what probabilities should the doctor now assign to the three possible diseases?

d_1 = Malaria

d_2 = Dengue

d_3 = Cold

$\Rightarrow$ $P(d_1) = 1/3$, $P(d_2) = 1/3$, $P(d_3) =$

$P(+ve \mid d_1) = 0.8$

$P(+ve \mid d_2) = 0.6$

$P(+ve \mid d_3) = 0.4$

		Test
M	5000	$\cancel{4}$
D	3000	$\times$
C	2000	$\times$

$$P(\overset{A}{d_1} \mid \overset{B}{+ve}) = \frac{P(d_1 \text{ AND } +ve)}{P(+ve)} = \frac{P(+ve \mid d_1)^{0.8}\, P(d_1)^{1/3}}{P(+ve)\,?}$$

$$= \frac{P(+ve \mid d_1)\, P(d_1)}{P(+ve \text{ AND } d_1 \text{ OR } +ve \text{ AND } d_2 \text{ OR } +ve \text{ AND } d_3)}$$

$P(+ve \text{ AND } d_1)$ + $P(+ve\ \cancel{\text{AND}}\ d_2)$ + $P(+ve \text{ AND } d_3)$

$$= \frac{P(+ve \mid d_1)\, P(d_1)}{P(+ve \mid d_1)P(d_1) + P(+ve \mid d_2)P(d_2) + P(+ve \mid d_3)P(d_3)}$$

- Given that test is +ve, what is the probability that it is d_2 or d_3?

$P(d_2 \text{ or } d_3 \mid +ve)$ $\nearrow$ $P(d_2 \mid +ve) + P(d_3 \mid +ve) + P(d_1 \mid +ve) = 1$

$\searrow$ $1 - P(d_1 \mid +ve)$

$^1/_3$

ve

00

$$P\left(d_2 \text{ or } d_3 \mid +ve\right) \rightarrow 1 - P(d_1 \mid +ve)$$

- Given that test is +ve, what is the probability that is not ~~a~~ dengue or

De Morgan's Formulas $\begin{cases} (A \cap B)^c = A^c \cup B^c \\ (A \cup B)^c = A^c \cap B^c \end{cases}$

$$P\left((d_2 \cup d_3)^c \mid +ve\right) = P\left(d_2^c \cap d_3^c \mid +ve\right) = P(d$$

- Complement $\rightarrow P(A) + P(A^c) = 1$

  $\rightarrow P(A \mid B) + P(A^c \mid B) = 1$

In a city, half of the days have some rain. The weather forecaster is correct 2/3 of the time, i.e., the probability that it rains, given that the forecaster has predicted rain and the probability that it does not rain, given that she has predicted that it won't rain are both equal to 2/3. When rain is forecast, Mr. X takes his umbrella. When rain is not forecast, he takes it with probability 1/3.

(a) Calculate the probability that Mr. X has no umbrella given that it rains.

(b) Calculate the probability that he brings his umbrella given that it doesn't rain.

$$P\left(\text{Rain} \mid \text{Forecast rain}\right) = \frac{2}{3}$$

$$P\left(\text{No Rain} \mid \text{Forecast No Rain}\right) = \frac{2}{3}$$

$\rightarrow$ Forecaster is correct

$$P\left(\text{Umbrella} \mid \text{Forecast Rain}\right)$$

$$P\left(\text{Umbrella} \mid \text{Forecast no Rain}\right)$$

$$P\left(\text{Umbrella}^c \mid \text{Rains}\right) = 1 - P\left(\text{Umbrella} \mid \text{Rains}\right)$$

r cold.

, (+ve)

$= 1$

$= 1/3$

A salesman has scheduled two appointments to sell software, one in the morning and another one in the afternoon. There are two software editions available: the base edition costing Rs. 5000 and the premium edition costing Rs. 10000. His morning appointments typically lead to a sale with a 30% chance while the afternoon ones typically lead to a sale with a 60% chance independent of what happened in the morning. If the morning appointment ends up in sale, the salesman has a 70% chance of selling the premium edition and if the afternoon appointment ends up in a sale, he is equally likely to sell either of the editions. Let X be the random variable representing the total Rupee value of sales. What are the different values that X can take? Calculate the probability that X takes the value 5000? Use the preliminary steps below to calculate that probability.

Solution: The event $X = 5000$ is equivalent to the following:

$$P\left(\begin{array}{c}\text{morning appointment ends up in sale } \underline{\text{and}} \text{ sell standard model } \underline{\text{and}} \text{ afternoon appointment no sale} \\ \underline{\text{or}} \\ \text{morning appointment no sale } \underline{\text{and}} \text{ afternoon appointment ends up in sale } \underline{\text{and}} \text{ sell standard model}\end{array}\right)$$

We can now calculate the probability $P(X = 5000)$ as follows:

$$P(X=5000) = \begin{cases} P\left(\text{morning appointment ends up in sale } \underline{\text{and}} \text{ sell standard model } \underline{\text{and}} \text{ afternoon appointment no sale}\right) \\ \underbrace{+}_{\text{why?}} \quad \text{mutually exclusive} \\ P\left(\text{morning appointment no sale } \underline{\text{and}} \text{ afternoon appointment ends up in sale } \underline{\text{and}} \text{ sell standard model}\right) \end{cases}$$

$$= \begin{cases} P\left(\text{morning appointment ends up in sale } \underline{\text{and}} \text{ sell standard model}\right) \underbrace{\times}_{\text{why?}} P\left(\text{afternoon appointment no sale}\right) \quad \text{independent} \\ \underbrace{+}_{\text{why?}} \\ P\left(\text{morning appointment no sale}\right) \underbrace{\times}_{\text{why?}} P\left(\text{afternoon appointment ends up in sale } \underline{\text{and}} \text{ sell standard model}\right). \quad \text{independent} \end{cases}$$

B A B A

Recall
$$\begin{cases} A, B \text{ mutually exclusive} \Rightarrow P(A \cup B) = P(A) + P(B) \\ A, B \text{ independent} \Rightarrow P(A \cap B) = P(A) * P(B) \end{cases}$$

$$P(\overset{A}{\text{Standard}} \mid \overset{B}{\text{morning sale}}) * P(\overset{B}{\text{morning sale}}) * P(\text{afternoon no sale})$$
$$+$$
$$P(\text{morning no sale}) * P(\text{Standard} \mid \text{afternoon sale}) * P(\text{afternoon sale})$$

$$= 0.3 * 0.3 * (1 - 0.6) + (1 - 0.3) * 0.5 * 0.6$$

X = Random earnings over a day → 0, 5000, 10000, 15000, 20000

j	P(X=j)
0	0.28
5000	0.24
10000	
15000	
20000	

$$\begin{array}{ccc} & \text{No sale} & \text{Sale} \\ \text{Morning} & \multicolumn{2}{c}{P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}} \\ \text{Afternoon} & & \end{array}$$

probability matrix

$P[1,]$ = morning probabilities

$$\begin{array}{ccc} & \text{Base} & \text{Premium} \\ \text{morning} & 0.3 & 0.7 \end{array}$$

$$P = \begin{array}{l} \\ \text{afternoon} \end{array} \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}$$

Probability Mass Function (PMF)

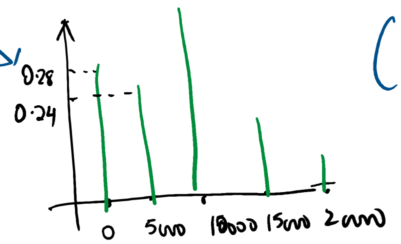
$$\underbrace{P_X}_{\text{name of the function}} (\underbrace{\text{input}}_{\text{Possible values that } X \text{ can take}}) = \underbrace{\text{Output}}_{P(X = \text{input})}$$

E.g. For the Salesman problem, $P_X(5000) = P(X = 5000)$

X → Discrete random variable

x	$P_X(x)$
0	$0.7 * 0.4$
5000	0.24
1000	*
1500	*
2000	*

(Tabular information)



(Visual information)

Plot of the PMF

Long run average of the simulated/realized values of a random variable = expected value of that variable

expected value → Average, Mean

$$\frac{\text{Sum (Simulated Data)}}{n\text{simulations}} = \text{expected earning of the salesman}$$

$$\Rightarrow E[X] = \frac{0 * \overset{\text{How many times 0 appeared in Simulated Data}}{n_0} + 5000 * n_{5000} + 1000 * n_{1000} + 1500 * n_{1500} + 2000 * n_{2000}}{n\text{simulations}}$$

$$\Rightarrow E[X] = \frac{0 * n_0 + 500 * n_{500} + 1000 * n_{1000} + 1500 * n_{1500}}{n \text{ simulations}}$$

$$= 0 * \frac{n_0}{n \text{ simulations}} + 500 * \frac{n_{500}}{n \text{ simulations}} + \cdots + 2000 * \frac{n_{2000}}{n \text{ simulations}}$$

$\approx P(X=0)$ $\quad$ $\approx P(X=500)$ $\quad$ $\approx P(X=2000)$

$P_X(0)$ $\quad$ $P_X(500)$ $\quad$ $P_X(2000)$

$$\Rightarrow E[X] = 0 * P_X(0) + 500 * P_X(500) + \cdots + 2000 * P_X(2000)$$

$$\boxed{E[X] = \sum_{x} x P_X(x)}$$

0, 500, 1000, 1500, 2000

Connection between Probs & stats and linear algebra

Probs & Stats	Linear Algebra
Random variable X	Vector $x = \begin{bmatrix} 500 \\ 1500 \\ 0 \\ 500 \\ \vdots \end{bmatrix}$ $\Rightarrow n$ components
	realized version of X
$E[X] = \sum_x x P_X(x)$	$E[X] \approx (\mathbb{1}^T x)/n = \text{avg}(x)$
$Var[X] = E[(X - E[X])^2]$	Variance of x
$= E[X^2 + (E[X])^2 - 2XE[X]]$	$Var[X] \approx$ Exercise
$= E[X^2] + E[E[X]^2] + E[-2XE[X]]$	

Problem

(1) We have a box with 10 balls (4 white and 6 black).

Binomial — We randomly pick 5 balls with replacement. What is the probability that we get 3 white balls?

$$P(X=3) = \frac{{}^5C_3 * 4^3 * 6^2}{10^5} = {}^5C_3 \left(\frac{4}{10}\right)^3 * ($$

(2) Now consider that the sampling is without replacement.

Hypergeometric — What is the probability that we get 3 white balls?

$$P(X=3) = \frac{{}^4C_3 * {}^6C_2}{{}^{10}C_5}$$

(1) Sampling Space $= \{ w_1, w_2, w_3, w_4, b_1, b_2, b_3, b_4, b_5, b_6 \}$

5 out of 10 with replacement, order matters

$$\text{Sample Space} = \left\{ \begin{array}{cc} \overset{1/10^5}{(w_1 w_2 w_3 w_4 b_1)}, & \overset{1/10^5}{(b_1 w_1 w_2 w_3 w_4)} \\ \vdots & \vdots \\ \overset{1/10^5}{(b_1 b_1 b_3 b_1 b_1)} & \overset{1/10^5}{(b_2 b_3 b_4 b_5 b_6)} \end{array} \right\}$$

outcomes here are equally likely

$\Rightarrow n(S) =$ no. of ways to select 5 out of 10 objects with replacement and order matters

$= 10^5$

$$\text{Event } E = \left\{ \begin{array}{l} \overset{1/10^5}{(w_1 w_1 w_1 b_1 b_1)}, \overset{1/10^5}{(w_1 w_2 w_3 b_1 b_2)} \\ \overset{1/10^5}{(b_1 b_2 w_1 w_2 w_3)}, \dots \\ \dots \end{array} \right\}$$

$w_1 w_2 w_3 \quad b_1 b_2$ → $w_1 w_2 w_3 b_1 b_2$, $b_1 b_2 w_1 w_2 w_3$, $w_1 b_1 w_2 b_2 w_3$

$$P(E) = \frac{1}{10^5} * n(E) = \frac{n(E)}{n(S)} = \frac{{}^5C_3 * 4^3 * 6^2}{10^5}$$

5 slots $\Rightarrow$ $\overset{S_1}{\bigcirc} \quad \overset{S_2}{\bigcirc} \quad \overset{S_3}{\bigcirc} \quad \overset{S_4}{\bigcirc} \quad \overset{S_5}{\bigcirc}$

- Is this selection of slots valid: $S_1 S_1 S_1$?

Not valid as one slot gets exactly one ball

$\Rightarrow$ slots are selected without replacement.

$\left(1-\frac{4}{10}\right)^{5-3}$

$\Rightarrow$ slots are selected without replacement.

- Is this selection of slots $s_1 s_2 s_3$ different from $s_1 s_3 s_2$? They are not different, so order does not matter.

$$\Rightarrow P(E) = \frac{{}^5C_3 * 4^3 * 6^2}{10^5}$$

$$= {}^5C_3 * \left(\frac{4}{10}\right)^3 * \left(\frac{6}{10}\right)^2$$

$$= {}^5C_3 * \left(\frac{4}{10}\right)^3 * \left(1-\frac{4}{10}\right)^{5-3}$$

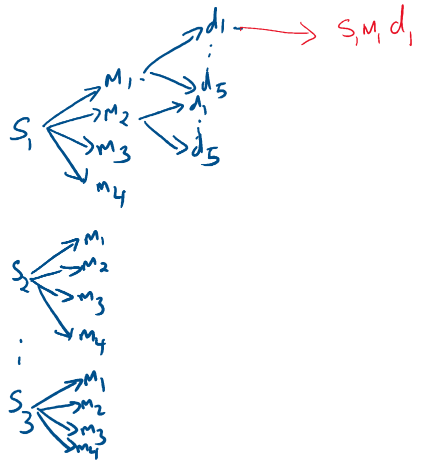
success probability per trial

no. of trials / sample size

no. of successes (white = success, black = failure)

Detour back to counting

3 soups, 4 main courses, 5 desserts



$3 * 4 * 5$ different meal combinations

- We calculated the probability of getting $\boxed{3}$ successes (3 white balls) in $\boxed{5}$ trials (or) in a sample of size 10 with a success probability of $\boxed{4/10}$ per trial

one possible no. of successes

| the trials are independent.

of size 10 with a success probability of 4/10,

and the trials are independent.

Success failure

- X = no. of successes in n independent trials with a constant success probability p per trial

$= 0, 1, 2, \dots, n$

$X \sim Bin(n, p)$

fixed parameters

In the balls example

$X = 0, 1, 2, 3, 4, 5$

$P(X=j) = 5c_j \left(\frac{4}{10}\right)^j \left(1-\frac{4}{10}\right)^{5-j}$

X is a Binomial Random variable with parameters n and p

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

0 to n

$5c_3 (0.4)^3 (0.6)^2$

SSSFF
SFSSF
FFSSS

$X \sim Bin(n,p)$

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

0 to n

E.g. A box with 10 balls (4 white, 6 black), Sample a ball with replacement

S_1 = 1st draw is white (Success)

F_1 = 1st draw is black (failure)

S_2 = 2nd draw is white (Success)

$P(S_1) = 4/10$, $P(S_2) = 4/10$ $\Rightarrow$ Success probabilities are the same across the trials

Are the 1st and 2nd trials independent

$P(S_2|S_1) \overset{?}{=} P(S_2) \Rightarrow S_1$ and S_2 are independent

Are the 1st and 2nd trials independent

$$P(S_2|S_1) \stackrel{?}{=} P(S_2) \Rightarrow S_2 \text{ and } S_1 \text{ are independent}$$

$$\underset{4/10}{\parallel} \quad = \quad \underset{4/10}{\parallel}$$

What if sampling is without replacement?

$P(S_1) = 4/10$, $P(S_2) = P(S_2 \text{ and } S_1 \text{ OR } S_2 \text{ and } F_1)$

$$= P(\overset{A}{S_2} \text{ and } \overset{B}{S_1}) + P(S_2 \text{ and } F_1)$$

$$= P(S_2|S_1)P(S_1) + P(S_2|F_1)P(F_1)$$

$$= \frac{3 \times 4}{9 \quad 10} + \frac{4}{9} \times \frac{6}{10} = \frac{4}{10}$$

Success probabilities across the trials are the same

$$P(S_2|S_1) \stackrel{?}{=} P(S_2) \Rightarrow S_2 \text{ and } S_1 \text{ are not independent}$$

$$\frac{3}{9} \neq \frac{4}{10}$$

but they are dependent

E.g. I have 10 oil drilling machines that I'm going to rent out. From the past experience, I know that each machine has a 40% chance of hitting oil independent of other machines. The number of machines that will hit oil is random.

$$X \sim Bin(n = 10, p = 0.4)$$

$$P(X = j) = {}^{10}C_j (0.4)^j (1 - 0.4)^{10-j}, \quad j = 0, 1, \cdots, 10$$

E.g. I run a company that manufactures batteries. The production defective rate in my factory is 1%. I package the batteries into packs of 10 and sell them

The production defective rate in my ...

I package the batteries into packs of 10 and sell them with the claim that if the customer finds more than 1 defective battery in a pack of 10 batteries, they will get a 2x refund. Aishwarya buys 3 packs of batteries from me. What is the probability that I will end up giving only one 2x refund to Aishwarya

Discrete random variables for counting

- draw a sample of size $\boxed{n}$ from a finite population (or) perform $\boxed{n}$ trials
- Each trial is either a success or failure
- Sampling is
  - with replacement $\Rightarrow$ trials are independent
  - without replacement $\Rightarrow$ trials are dependent
- But in both sampling scenarios, success probability $\boxed{p}$ per trial is the same across all trials

given as a fraction (or) as no. of successes in the population

$$p = \frac{n_{success}}{n_{success} + n_{failure}}$$

- We count the number of successes denoted as X
- Sampling is
  - with replacement $\rightarrow \begin{cases} X \sim Bin(n, p) \\ P(X=j) = {}^nC_j \, p^j (1-p)^{n-j} \end{cases}$

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

    $p = \frac{n_{success}}{n_{success} + n_{failure}}$
  - without replacement $\rightarrow \begin{cases} X \sim Hypgeom(n_{success}, n_{failure}, n) \\ P(X=j) = \dfrac{\binom{n_{success}}{j}\binom{n_{failure}}{n-j}}{\binom{n_{success}+n_{failure}}{n}} \end{cases}$

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

- Probability mass function of X

$$\underset{\text{Input}}{P_X(j)} = \underbrace{P(X=j)}_{\text{output}} = \begin{cases} {}^nC_j \, p^j (1-p)^{n-j} & \text{for } X \sim Bin(n, p) \\ \dfrac{\binom{n_{success}}{j}\binom{n_{failure}}{n-j}}{\binom{n_{success}+n_{failure}}{n}} & \text{for } X \sim Hypgeom(n_{success}, n_{failure}, n) \end{cases}$$

Example: approximately 42% of people have type O blood. On a given day in a blood bank, 120 people arrive to donate blood. What is the probability that 30 of those 120 people have type O blood?

$X \sim \text{Hypgeom}(n_{success} = 420{,}000, n_{failure} = 580{,}000, n = 120)$

$P(X = 30) = ?$

How to identify if it is a binomial or a hypergeometric scenario

Hypergeometric $\longrightarrow$ finite population, small sample from that population, no. of successes and failures in the population are explicitly made available

[10 points] [TLO 2.2 CO 1] You run APS Mobile Company that offers cell phones with three storage capacities: Small (64GB), Medium (128GB), and Large (256GB). Based on your market research, you believe that the three sizes will be ordered by a potential customer with probabilities 0.3, 0.5, and 0.2, respectively, for each order independently of the other orders. For each of the following, identify the correct random variable with the associated parameters clearly shown. Using the parameters, calculate the expected value of each random variable. For example,

$$\boxed{X \sim \text{Bin}(n = 12, p = 0.2)} \text{ and } E[X] = np = 12 \times 0.2.$$

X = no. of Large phone orders
$X \sim \text{Bin}(n = 100, p = 0.2)$
$P(X = j)$, $j = 0, 1, \cdots, 100$

(a) The number of Large phone orders out of the next one hundred phone orders.

(b) The number of phones that will be ordered until six Medium phones are ordered.

(c) Suppose that 40 of the previous 100 orders were for Small phones. You randomly choose 25 of those 100 orders and would like to know the number of Small phone orders in them.

X = no. of Small phone orders
$X \sim \text{Hypgeom}(n_{success} = 40, n_{failure} = 60, n = 25)$
$P(X = j)$, $j = 0, 1, \cdots 25$

[3 points] In a forest that has 1000 tigers, 250 are captured, tagged, and released. A few weeks later, a sample of 10 tigers from the forest is captured. If we want to calculate the probability that at least half of those captured tigers are tagged, what type of a random variable X would you use? Clearly state the parameters of the random variable and the probability of interest as $P(X \cdots)$. Do not calculate the probability.

X = no. of tigers that are tagged

Success $\Leftrightarrow$ tagged

$X \sim \text{Hypgeom}(n_{success} = 250, n_{failure} = 750, n = 10)$

$j = 0, 1, 2, \cdots, 10$, $P(X = j) = \dfrac{{}_{250}C_j * {}_{750}C_{10-j}}{{}_{1000}C_{10}}$

$$P(X \geq 5) = \sum_{j=5}^{10} \frac{{}_{250}C_j * {}_{750}C_{10-j}}{{}_{1000}C_{10}}$$

Example: approximately 42% of people have type O blood. On a given day in a blood bank, 120 people arrive to donate blood. What is the probability that 30 of those 120 people have type O blood?

X = no. of people with type O blood

Success $\Leftrightarrow$ type O blood

$X \sim Bin(n = 120, p = 0.42)$ and $P(X = 30)$

✓ $X \sim Hypheom(n_{success} = 42000, n_{failure} = 58000, n = 120)$ and $P(X=30)$

Example: A certain stoplight, when coming from the North, is green approximately 31% of the time. Over the next few days, someone comes to this light 8 times from the North. We are interested in the probability that the person will come across green light 5 times.

X = number of times green light shows up

Success $\Leftrightarrow$ green light shows up

$X \sim Bin(n = 8, p = 0.31)$ and $P(X = 5)$

An assembly line produces products that they put into boxes of 50. The inspector then randomly picks 3 items inside a box to test to see if they are defective. In a box containing 4 defectives, they are interested in the probability that at least one of the three items sampled is defective.

X = no. of defectives

Success $\Leftrightarrow$ defective

$X \sim Hypheom(n_{success} = 4, n_{failure} = 46, n = 3)$

and $P(X \geq 1)$

The negative binomial random variable

- Recall that in the binomial experiment, we are given the two parameters: (1) no. of trials n (2) success probability per trial p. The binomial random variable X
- is equal to the (random) no. of possible successes $j = 0, 1, \ldots, n$. Similarly, in the hypergeometric experiment, we are given the three parameters: (1) no. of successes $n_{success}$ (2) no. of failures $n_{failures}$ (3) sample size (or) no. of trials n. The hypergeometric random variable X is equal to the (random) no. of possible successes $j = 0, 1, \ldots, n$.
- Consider this scenario: we need to complete 3 more survey phone calls successfully to call it a day. The likelihood of someone accepting the call for the survey is 1%. How many calls should we make before we call it a day?

X = counts the no. of $\underbrace{\text{calls}}_{\text{trials}}$

$= 3, 4, 5, 6,$

Given $\begin{cases} r = 3 & (\text{no. of successes we are interested in}) \\ p = 0.01 & (\text{success probability per } \underbrace{\text{call}}_{\text{trial}} \end{cases}$

- Just like the binomial experiment, the trials are independent and success probability is the same across all trials.
- X counts the no. of trials and we call it a negative binomial random variable

- X counts the no. of trials and we call it a negative binomial random variable

$$X \sim \text{NegBin}(r=3,\ p=0.01)$$
$$X=j,\quad j=3,4,5,\dots$$

- In general, $X \sim \text{NegBin}(r, p)$ and

$$X=j,\quad j=r,r+1,r+2,\dots$$

- What is $P(X=j)$, $j=r, r+1, \dots$

- $P(X=j)$ $\rightarrow$ rth success is happening at the jth trial

$$\underbrace{\boxed{r-1\ S,\ j-r\ F}}_{j-1 \text{ trials}} \quad \underbrace{\boxed{S}}_{j\text{th trial}} \ \leftarrow r\text{th success}$$

$\Downarrow$

One way: $\underbrace{SSS..S}_{r-1}\ \underbrace{FF..F}_{j-r}\ \underbrace{S}_{\text{last trial}}$

Another way: $\underbrace{FF..F}_{j-r}\ \underbrace{SS..S}_{r-1}\ \underbrace{S}_{\text{last trial}}$

$$\underbrace{(p*p*\dots*p)}_{r-1 \text{ times}} * \underbrace{(1-p)*\dots*(1-p)}_{j-r \text{ times}} * p$$

$\Rightarrow P(X=j)$ = no. of ways to select $r-1$ slots from $j-1$ slots (without replacement, order does not matter) to put the success S $\boxed{\text{AND}}$ probability of getting $r-1$ successes $\boxed{\text{AND}}$ $j-r$ failures $\boxed{\text{AND}}$ success at the last trial

$$= \underbrace{(j-1)C_{r-1}}_{} * \underbrace{p^{r-1}}_{} * (1-p)^{j-r} * \underbrace{p}_{}$$

$$\Rightarrow \boxed{P(X=j) = (j-1)C_{r-1}\ p^{r}\,(1-p)^{j-r}}$$

Example: at an airport, it is know that approximately 2 out of 10 passengers have a metallic object. If left undetected at the manual security check at the airport entrance, such a metallic object will raise an alarm when the passenger walks through an automated screening machine. It is considered a security breach when the alarm gets raised 20 times a day. What is the probability of a security breach on a particular day when the 100th passenger walks through the automated screening machine?

$p = 2/10$, $r = 20$ $\rightarrow$ no. of successes we are interested

$\Downarrow$

Success $\Leftrightarrow$ passenger raises alarm

$$X \sim \text{NegBin}(r=20,\ p=0.2)$$

$P(X=100)$, note that $j=100$ = probability that the 100th trial will result in the 20th success

An extreme Binomial experiment

$\rightarrow$ given: { no. of trials (or) sample size n; success probability per trial p

$\rightarrow$ Counting: no. of successes $j=0,1,\dots,n$

e.g.
- Sending a bit string over a network
- length of the bit string (how many bits) = 10^5 (n) $\leftarrow$ extreme values
- Probability that a bit is corrupted = 10^{-6} (p)
- Each bit is transmitted independent of the other
- X = no. of corrupt bits at the receiving end
- $X \sim \text{Bin}(n=10^5,\ p=10^{-6})$
- Probability that the message is received intact

$$= P(X=0) = 10^5 C_0\ \underbrace{(10^{-6})^0}_{}\ \underbrace{(1-10^{-6})^{10^5-0}}_{10^5}$$

• Probability that the message is received intact

$$= P(X = \underset{\uparrow j}{0}) = \underbrace{{}^{10^5}C_0}_{1} \underbrace{(10^{-6})^0}_{1} \underbrace{(1-10^{-6})^{10^5-0}}_{(0.99\ldots9)^{10^5}}$$

$$\bullet\ P(X = \underset{\uparrow j}{2}) = \underbrace{{}^{10^5}C_2}_{\Downarrow\ \frac{100000!}{2!(99998)!}} (10^{-6})^2 (1-10^{-6})^{10^5-2}$$

E.g. $n = 10^5$ letters typed in your thesis report

$p = 10^{-3}$ = probability of mistyping a letter

X = no. of mistyped letters in the entire thesis report

$X \sim \text{Bin}(n = 10^5, p = 10^{-3})$ → Extreme case $n >> p$

E.g. $n = 10^6$ people in a city (10 million population)

$P = 10^{-2}$ →
- probability that a person has chickenpox
- probability that a person has natural blue eyes
- probability that a person has 11 fingers
- probability that a person's last name starts with z
- probability that a person is a \$ millionaire
- probability that a person owns a range rover

X = counting the no. of people with the above mentioned characteristic

$X \sim \text{Bin}(n = 10^6, p = 10^{-2})$ → extreme case $n >> p$

E.g. $n = 10^4$ road segments (each road segment is 10km long)

$p = 10^{-3}$ is the probability that an accident can happen in a 10km segment

$X \sim \text{Bin}(n = 10^4, p = 10^{-3})$

What is X counting? X is counting the no. of accidents over the entire highway

E.g. $n = 10^4$ time segments (each time segment is 1 minute long)

$p = 10^{-3}$ is the success probability

success ⟺ stock price goes up 1 Rupee

E.g. $n = 10^2$ time segments (each time segment is 1 millisecond long)

→| 10^{-3}s |← 1st time segment

$p = 10^{-4}$, probability that a photon will strike a pixel over a millisecond

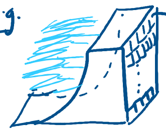
X = no. of success in n trials → $10^2 \times 10^{-3}$s $= 0.1$s

= counting how many photons will strike the pixel over a 0.1s duration

E.g.  → 1cm³ volume

$n = 10^6$ 1cm³ volumes on the front face

E.g.



$\to 1cm^3$ volume

$n = 10^6$ $1cm^3$ volumes on the front face

$p = 10^{-6}$ = probability that a $1cm^3$ volume will have a crack

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

X = counting the total no. of cracks on the front face of the dam

In all the cases above, $\boxed{n >> p}$

no. of trials is much greater than success probability per trial

$$\boxed{\begin{array}{l} X \sim Bin(n, p) \\ P(X=j) = \binom{n}{j} p^j (1-p)^{n-j} \end{array}} \xRightarrow{\text{when } n>>p} \boxed{\begin{array}{l} X \sim Poi(\lambda) \text{ Poisson random variable} \\ P(X=j) = \dfrac{e^{-\lambda}\lambda^j}{j!} \\ \text{where } \lambda = np \end{array}}$$

E.g. The 10km-segment setup $\begin{cases} n = 10^4 \\ p = 10^{-3} \end{cases}$

$$P(X = 10) = \binom{10^4}{10}(10^{-3})^{10}(1-10^{-3})^{10^4-10} \quad \text{(Binomial)}$$

$\uparrow$ j

$$= \begin{cases} \lambda = np = 10^4 * 10^{-3} = 10 \\ e^{-10}\dfrac{10^{10}}{10!} \end{cases} \quad \text{(Poisson)}$$

$\boxed{X \sim Poi(\lambda)}$ λ = parameter of the Poisson random variable

. The expected number of typographical errors on a page of a certain magazine is .2. What is the probability that the next page you read contains (a) 0 and (b) 2 or more typographical errors? Explain your reasoning!

$n = 10^5$ pages, $p = 2 \times 10^{-6} \Rightarrow \begin{cases} X \sim Bin(n = 10^5, p = 2\times10^{-6}) \\ P(X=0) \end{cases}$

$\lambda = np = 0.2 \Rightarrow \begin{cases} X \sim Poi(\lambda = 0.2) \\ P(X=0) \end{cases}$

$$P(X \geq 2) = 1 - (P(X=0) + P(X=1)) = 1 - \left[\frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!}\right]$$

The monthly worldwide average number of airplane crashes of commercial airlines is 3.5. What is the probability that there will be

(a) at least 2 such accidents in the next month;

(b) at most 1 accident in the next month?

Explain your reasoning!

$\lambda = 3.5 \dfrac{\text{accidents}}{\text{month}}$

$P(X \geq 2) = 1 - [P(X=0) + P(X=1)]$

$P(X \leq 1) = P(X=0) + P(X=1)$

$n = 35 \times 10^2$ months

$p = 10^{-3}$ probability of accident per month

$\Rightarrow n * p = 3.5$

$\Downarrow$ months $\Downarrow$ * $\dfrac{\text{probability of accident}}{\cancel{\text{month}}}$

What is the probability that we will have 3 accidents in the next 2 months?

Recall $\lambda = 3.5 \dfrac{\text{accidents}}{\text{months}}$, $\boxed{P(X=3) = \text{probability of 3 accidents in the next month}}$

$\Downarrow$ Change units

Recall $\lambda = 3 \cdot$ ... $\dfrac{\quad}{\text{months}}$ 3 accidents in the next month

$\Downarrow$ Change units

$$\lambda = \frac{3.5 \times 2 \text{ accidents}}{2 \text{ months}} = 7 \; \boxed{\frac{\text{accidents}}{2 \text{ months}}}$$

new value of λ

$$P(X=3) = \frac{e^{-7} \, 7^3}{3!}$$

E.g. On an average, a laptop crashes once every 5000 hours. What is the probability that 10 laptops will crash the next calendar year?

$$\lambda = \frac{1 \text{ crash}}{5000 \text{ hours}} = ? \; \frac{\text{crashes}}{\text{year}}$$

$$P(X=10) = \frac{e^{-\lambda} \lambda^{10}}{10!}$$

$$\lambda = \frac{1 \text{ crash}}{5000 \text{ hours}} = \frac{1}{5000} \; \frac{\text{crashes}}{\text{hour}}$$

$$1 \text{ year} = 24 * 365 \text{ hours}$$

$$\Rightarrow 1 \text{ hour} = \frac{1}{24 * 365} \text{ years}$$

$$\lambda = \frac{24 * 365}{5000} \; \frac{\text{crashes}}{\text{year}}$$