Sampling & Designing

→ How is data produced, drawing samples & designing experiments

Statistical Inference :-

Population: the entire population / group of subjects about which we want information (eg: total Indian population)

Parameter : quantity about population we are interested in (eg: approval % among all Indian voters)

Sample : part of population from which we collect information (eg: 1,000,000 sample size)

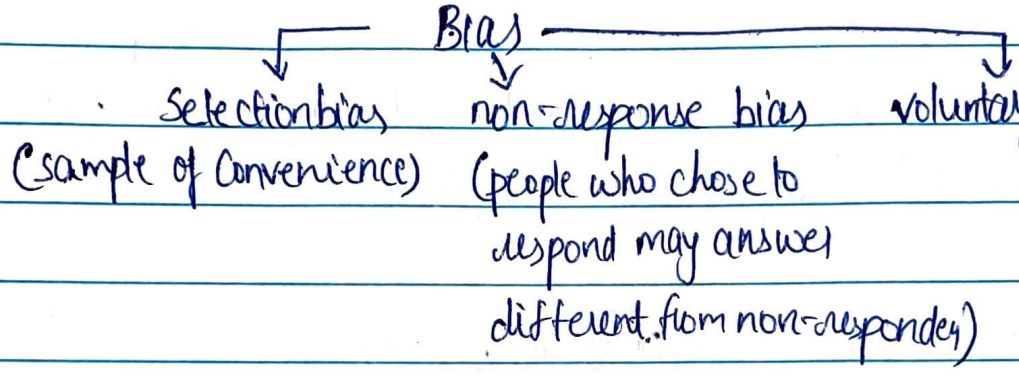
Statistic / Estimate: the quantity we are interested in as measured in sample. (approval % among the sample voters).

⇒ Even a relatively small sample will produce an estimate that is close to the parameter of a very large population.

Simple random sampling & other sampling plans:-

→ Sampling correctly is very important.

→ Selecting 1000 from your neighbours is sample of convenience. This is not a good way as it will introduce bias, sample favouring certain outcome



Ways to avoid these bias is to use chance in a planned way.

→ A simple random sample, selects sample at random without replacement

→ A stratified random sample, divides population into groups of similar subjects called "strata" (eg: rural, urban). Then a simple random sample is chosen. This is resulting in a more precise estimate.

→ Since sample is drawn at random, estimate will be different from the parameter due to "chance error". Another sample gives another "chance error". By taking a bigger sample size, "chance error" can be made small.

estimate = bias + chance error + parameter

But if we increase the sample size, the bias will not get smaller & typically we don't know how large bias is.

Randomized controlled experiments & observation studies:-

Let's say "people who eat red meat has a chance of getting certain cancers than who don't".

→ This is called "association".

→ But this does not mean, eating red meat causes cancer.

→ Association doesn't mean causation.

→ This is an observational study. It measures outcomes of interest & this can be used to establish association

→ Because there maybe "Confounding factors" or "lurking variable"

→ To establish causation, an experiment is required.

A treatment (eating red meat) is assigned to people in treatment group and other is control group. To the end, the confounding factors should be same, then compare outcomes.

① The best way to make sure that 2 groups are similar is to assign a subject at random.

② Control group gets placebo.

③ Experiment is double-blind. Neither subject nor evaluator know which subject is in which group.

⇒ A good experiment requires that the subjects are assigned groups at random so that they operate equally, apart from differences due to chance. Since it is chance effect, we can calculate the size when we evaluate the outcome.

Interpretation of Probability

Probability of an event is defined as proportion of times this event occurs in many repetitions, given it is possible to repeat.

eg:- In 2015, 4 million babies were born & 48.8% were female.

→ P(newborn is female) = 48.8%.

→ Long interpretation of probability can make it difficult to interpret it for single event.

→ Subjective probability is not based on experiment but on subjective probabilities of an event by individuals.

Key to Computing probabilities:

① Probabilities are always between 0 & 1.

② Complement rule ⇒ $P(\bar{A}) = 1 - P(A)$

③ Rule for equally likely outcome (n) ⇒

$$P(A) = \frac{\text{no. of outcomes in } A}{n}$$

④ A & B are mutual exclusive if they cannot occur same time

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

multiplication ⇒ $P(A \& B) = P(A)\,P(B)$ for this two events should be independent, i.e knowing one occurs does not change the probability of others.

⑤ P(atleast One 6 in 3 rolls).

⇒ write 'atleast one 6'

first roll or second roll or third roll

but these are not mutually exclusive, as we can't have 6 on second roll.

∴ P(atleast one 6) = 1 - P(no 6 in 3 rolls)

= 1 - P(no 6 in first) and (second) and (third)

$= 1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$

= 41.4%

Conditional probability :-

→ A spam email contains the word 'money' higher chance than a ham/normal email.

P('money' in email | spam) = 8%. "|" → given that
P('money' | ham) = 1%.

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

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

if A & B are independent, P(A and B) = P(A) P(B) = P(B) P(A)

Q) Let P(spam) = 20%. What is the probability it contains 'money' ? in an email.

A) P(spam) = 20% ⇒ P(ham) = 80%.
P(money | spam) = 8%.
P(money | ham) = 1%.

Event that "money" appears in email can be written as

money and email is spam (or) | money and email is ham

⇒ these 2 events are mutually exclusive, i.e if email is spam, then it can't be ham

∴ P(money appears) = P(money & spam) + P(money & ham)

⇒ (8×20) + (1×80) = P(money | spam) P(spam) + P(money | ham) P(ham)

⇒ 160 + 80 ⇒ $\frac{240\%}{100}$ ⇒ 2.4%

Bayes' Rule :-

From data we know P(money|spam) = 8%. but what we need to build is P(spam|money)

(Bayes' Rule) $P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B \text{ and } A)}{P(A)} = \frac{P(B)\,P(A|B)}{P(A)}$

$\Rightarrow P(\text{spam}|\text{money}) = \frac{P(\text{money}|\text{spam})\,P(\text{spam})}{P(\text{money})} = \frac{8 \times 20}{240} = \frac{2}{3} = 67\%$

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

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

Bayesian Analysis :-

The spam filter classifies email as spam via Bayesian Analysis

→ P(spam) = 20%.

→ After examining keywords like 'money', the posterior probability is improved.

⑥ 1% population has disease, test shows 95% positivity rate if infected is tested. If non-infected is tested, the test shows 2% chance as positive. Given person has tested positive, what is probability they are infected?

A P(disease is present) = 1% ⇒ P(disease not present) = 99%

P(Positive | disease is present) = 95%

P(positive | disease is not present) = 2%

P(disease is present | positive) = ?

$$\therefore P(\text{disease} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{disease})\, P(\text{disease})}{P(\text{positive})}$$

$$\Rightarrow P(\text{positive}) = P(\text{positive} \mid \text{disease})\, P(\text{disease}) + P(\text{positive} \mid \text{no disease})\, P(\text{no ds})$$

$$= (0.95 \times 0.01) + (0.02 \times 0.99)$$

$$= 0.0293$$

$$\therefore P(\text{disease} \mid \text{positive}) = \frac{0.95 \times 0.01}{0.0293} = 0.0324 = 32.4\%$$

Reason is, only 1% of population has disease & test finds most of those. But there's 2% chance of not infected. Since that population is higher, the probability is low.

Warner's Randomized response model :-

Assume surveying "what % students have cheated during online exam". But as students wont answer truthfully, we use Randomization as:

Toss a coin twice (HH, HT, TH, TT), if they get "T", then they answer question-1, else question-2.

q-1 → did you cheat in exam

q-2 → did you get tails on 2nd toss

So the answer will be partly random. A 'yes' could be to any of these questions.

→ While we dont know what individual 'yes' means, we can estimate the proportion of cheaters using all answers collectively.

$$P(yes) = P(yes \& Q1) + P(yes \& Q_2)$$
$$= P(yes|Q_1)\,P(Q_1) + P(yes|Q_2)\,P(Q_2).$$

$$\Rightarrow P(yes|Q_1) = \frac{P(yes) - P(yes|Q_2)\,P(Q_2)}{P(Q_1)}$$

Let 27 answered yes & 30 said no, then

$$P(yes) = \frac{27}{27+30} = 47\%.$$

$$P(yes|Q_1) = \frac{0.47 - (0.5 \times 0.5)}{(0.5)} = 0.44 = 44\%$$

$P(Q_2)$ → just half because it depends on first coin toss

$P(Q_1)$ → half

$P(yes|Q_2)$ → half

Q) A fair coin is tossed 5 times. P(at most 4 tails).

A) P(atmost 4 tails) can be written as

P(1 tail) P(2 tail) P(3 tail) P(4 tail) (not mutually exclusive)

⇒ 1 - P(no tail in 4 tails)

⇒ 1 - P(no tail in 4 tosses) and P(2nd tail) and P(3rd tail) and P(4th tail)

⇒ $1 - \left(\frac{1}{2}\right)^5$

Q) 3 boxes → Box1 → 2 quarters
Box2 → 2 nickels
Box 3 → 1 quarter & 1 nickel.

Pick a random coin from a random box.

If coin picked is quarter, what is chance other coin is also quarter in box.

A) P(selecting Box 1) = P(selecting Box 2) = P(selecting Box 3) = $\frac{1}{3}$

⇒ P(quarter|B1) = 1 P(quarter|B2) = 0 P(quarter|B3) = $\frac{1}{2}$

⇒ $P(B_1|\text{quarter}) = \frac{P(\text{quarter}|B_1)\,P(B_1)}{P(\text{quarter})}$

$P(\text{quarter}) = P(B_1)\,P(\text{quarter}|B_1) + P(B_2)\,P(\text{quarter}|B_2) + P(B_3)\,P(\text{quarter}|B_3)$

$= \left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times 0\right) + \left(\frac{1}{3}\times\frac{1}{2}\right) = \frac{1}{3} + \frac{1}{6}$

⇒ $P(B1|\text{quarter}) = \frac{\left(1\times\frac{1}{3}\right)}{\left(\frac{1}{3}\times 1\right) + \left(\frac{1}{3}\times\frac{1}{2}\right)}$

Q Roll pair of dice. what is the chance of double when you roll a pair of dice? (1,1) (2,2) (3,3) (4,4) (5,5) (6,6)

A mutually exclusive and independent.

P(showing double)]

⇒ 1 − P(not showing double)

↳ (1,2)(1,3) ⋯ (1,6) → 5
(2,1)
(3,1)
(4,1)
(5,1)
(6,1)
} ⇒ 30

⇒ $1 - \frac{1}{36} \Rightarrow \frac{1}{6}$ (out of 36 possible outcomes, 6 are double)

Q Monopoly, if you land in jail, must roll a double on any of next 3 turns, or pay fine. - what is P(get out of jail / no fine)?

A P(double in next 3 turns)

P(in first turn) or P(in 2nd turn) or P(3rd turn) (not mutually exclusive)

⇒ P 1/6 or 1/6 or 1/6

⇒ $1 - (5/6)^3$

Probability of rolling atleast 1 double, is same as not rolling 3 non-doubles ↳ in 3 turns

Q 3% of applicant are admitted. 70% of all applicants have GPA 3.6 ≤ of those admitted, 95% have GPA 3.6 ≤. Chance of applicant being admitted with GPA ≥ 3.6?

$$P(\text{admit} \mid 3.6) = \frac{P(\text{admit} \cap 3.6)}{P(3.6)} = \frac{P(3.6 \mid \text{admit})\, P(\text{admit})}{P(3.6)}$$

A P(admitted) = 3% ⇒ P(not admitted) = 97%

P(3.6 | ~~applicants admitted~~) = 70% = P(applied | 3.6)

~~P(3.6 | admitted)~~ P(admitted | 3.6) = ~~95%~~, then P(3.6 | admitted) = 95%

Q MCQ has 10 questions. Each question has 3 possible answers, one is correct. A student knows correct answer to 4 & guesses other. If then first question was correct, what is chance of it being guess.

A $P(\text{answer is correct}) = \frac{1}{3}$

$P(\text{answer is correct} \mid \text{know}) = \frac{4}{10}$

$P(\text{answer is correct} \mid \text{guess}) = \frac{6}{10}$

$$\Rightarrow P(\text{guess} \mid \text{answer}) = \frac{P(\text{guess} \cap \text{answer})}{P(\text{answer})}$$

$$= \frac{P(\text{guess}) P(\text{answer} \mid \text{guess})}{P(\text{guess}) P(\text{answer} \mid \text{guess}) + P(\text{know}) P(\text{answer} \mid \text{know})}$$

$$= \frac{\frac{1}{3} \times \frac{6}{10}}{\left(\frac{1}{3} \times \frac{6}{10}\right) + \left(1 \times \frac{4}{10}\right)}$$