

Probability - study of models from
(random) experiments when
the model is fully known.

Statistics - when the model is not
fully known then one tries
to infer about certain aspects of
the model based on observed
values from an experiment.

Step 1:

an
Exam

Assumpt

are

Step 1: Sample $[n]$ observations from a large population and record on a numerical (categorical) fact.

Example -

- Roll dice experiment
- Study on Maternal health & infant mortality

Assumptions: Typically we will get $X_1, X_2, \dots, X_n$ observations that are identically distributed and independent [With Replacement]
 $n \ll \text{population size}$

So far to understand the underlying distribution

- Summary of data
- Histogram, box plots, ...
- Distribution - $Q-Q$ plots.

Key facts :-

- Equal probability for each observed point

Empirical Distribution ; let $X_1, \dots, X_n$ be i.i.d ^{Random variable} observations

Let $S = \{X_1, \dots, X_n\}$ Empirical distribution is a discrete distribution on S with pmf

$$f(t) = \begin{cases} \frac{\#\{i: X_i = t\}}{n} & t \in S \\ 0 & \text{otherwise} \end{cases}$$

Points to keep in mind -

- Random [Every sample produces a different point]
- No assumption made on the underlying population
- Inference on population based on Empirical distribution is called Descriptive statistics

Given $X_1, \dots, X_n$ & empirical distribution f , $S = \{X_1, \dots, X_n\}$

Y - a random variable with p.m.f f

$$P(Y=t) = f(t) \quad t \in \mathbb{R}$$

$$\text{Range}(Y) = S$$

$$E[Y] = \sum_{t \in S} t P(Y=t) = \sum_{t \in S} t \frac{\# \{i: X_i=t\}}{n}$$

[Rearranging] $\quad = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ (Sample mean)

- as n gets large
we are getting
closer to the
underlying distribution

Sample Mean :- let $x_1, \dots, x_n$ be iid observations

Then $\bar{X} = \frac{x_1 + \dots + x_n}{n}$ is the

sample mean.

x_n
(supern)

Theorem 1 :- Suppose $x_1, \dots, x_n$ iid random variables

with $E x_1 = \mu$ $Var(x_1) = \sigma^2$
 $E \bar{X} = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$

Proof:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

Linearity of Expectation
induction

$$E[\bar{X}] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{n\mu}{n} = \mu$$

Even an average
unknown μ

$\bar{X}$

describes the

- In such a case $\bar{X}$ is called an unbiased
estimator for μ

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right)$$

$$= \frac{1}{n^2} \text{Var}(X_1 + \dots + X_n)$$

As n gets large
 $\bar{X}$ is getting closer
to μ

$$\text{Var}(X+Y)$$

$$= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

$\bullet$ X, Y are independent

$$\Rightarrow \text{Cov}(X, Y) = 0$$

$\bullet$ induction

$$= \frac{1}{n^2} [\text{Var}(X_1) + \dots + \text{Var}(X_n)]$$

$$= \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

- n gets large
we are getting
close to the
underlying distribution

$$Y \text{ has } \text{Var}(Y) = 0$$

$$(-)$$

$$\exists c \in \mathbb{R} \text{ s.t. } P(Y=c) = 1$$

$$\text{Var}(\bar{X}) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

- $\bar{X}$ is called a consistent estimator for μ

Sample Variance

$X_1, X_2, \dots, X_n$ are iid r.v's

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

is called the sample variance

Theorem — $X_1, \dots, X_n$ be iid r.v with $E X_i = \mu$
and $\text{Var}(X_i) = \sigma^2$.

$$E[S^2] = \sigma^2$$

Proof

$$E[\bar{X}^2] = \text{Var}(\bar{X}) + (E[\bar{X}])^2$$

(previous theorem) $= \frac{\sigma^2}{n} + \mu^2$

$$E[X_i^2] = \sigma^2 + \mu^2$$

$$E[(n-1)S^2] = E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]$$

$$= \sum_{i=1}^n E[X_i^2] - 2E\left[\sum_{i=1}^n X_i \bar{X}\right] + nE[\bar{X}^2]$$

$$= n(\sigma^2 + \mu^2) - 2nE[\bar{X}^2] + nE[\bar{X}^2] = (n-1)\sigma^2$$

$$E S^2 = \sigma^2 \quad \square$$

S^2 is unbiased estimator for σ^2 .

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} S^2$$

$$E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$$

[Review from Probability]

Let $X_1, \dots, X_n$ be iid random variables with $E[X_i] = \mu$

It $\text{Var}(X_i) = \sigma^2 < \infty$

[Strong law of large numbers]

$$(*) - P\{\bar{X} \rightarrow \mu \text{ as } n \rightarrow \infty\} = 1 \quad \text{a.e. } (E|X_i| < \infty)$$

Proof - M Krane & T-Tao blog $\square$

WLLN: Fix $\varepsilon > 0$. $P(|\bar{X} - \mu| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$

$$P(|\bar{X} - \mu| > \varepsilon)$$

$$\leq \frac{E(\bar{X} - \mu)^2}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty$$

(*)

(*)

#

(*)

(*)

Conclusion:

$$P(\hat{p}_n \rightarrow p \text{ as } n \rightarrow \infty) = 1$$

$\forall \varepsilon > 0$

$$P(|\hat{p}_n - p| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

Central limit Theorem

[De Moivre-Laplace] let $z_1, \dots, z_n$ be iid Bernoulli(p)

$$\hat{p}_n = \frac{\sum_{i=1}^n z_i}{n}$$

$$W_n := \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow[\text{as } n \rightarrow \infty]{\text{distribution}} Z, \text{ where } Z \stackrel{d}{=} N(0,1)$$

$$EW_n = 0$$

$$\text{Var}(W_n) = 1$$

$$\text{ie } \forall x \in \mathbb{R} \quad P(W_n \leq x) \longrightarrow P(Z \leq x) \text{ as } n \rightarrow \infty$$

Proof [Idea]

$$S_n = \sum_{i=1}^n Z_i$$

Note

$$W_n = \frac{S_n - np}{\sqrt{np(1-p)}}$$

$\forall a, b \in \mathbb{R}$

$$\left| P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} < b\right) - \int_a^b \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \right|$$

$$= \left| \sum_{k=\lfloor a\sqrt{np(1-p)} \rfloor}^{\lfloor b\sqrt{np(1-p)} \rfloor} P(S_n = k) - \int_a^b \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \right|$$

$$P(S_n = k)$$

$$\int_a^b \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$$

$$* P(S_n = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$① \phi(y) = \frac{e^{-y^2/2}}{\sqrt{2\pi}}, \quad \left| \sum_{k=\lfloor a\sqrt{n(p-q)} \rfloor}^{\lfloor b\sqrt{n(p-q)} \rfloor} \frac{1}{\sqrt{n(p-q)}} \phi\left(\frac{k-np}{\sqrt{n(p-q)}}\right) - \int_a^b \phi(y) dy \right|$$

$\xrightarrow{n \rightarrow \infty}$ $\bigcirc$
Riemann Sum

$$② n! \underset{n \rightarrow \infty}{\sim} n^n e^{-n} \sqrt{n} \sqrt{2\pi}$$

Use $(*)$, (I) and $(+)$ to conclude $\square$

Central Limit Theorem

$X_1, \dots, X_n$ be iid random variables

$E X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n \rightarrow \infty]{\text{distribution}} Z \quad \text{where } Z \sim N(0, 1)$$

$a < b$, $a, b \in \mathbb{R}$

$$\text{i.e. } \mathbb{P}\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b\right) \longrightarrow \mathbb{P}(a \leq Z < b) \quad \text{as } n \rightarrow \infty$$

(*)

Proof: uses several results $\square$

Consequences

- Very useful - as it assumes very little about the underlying distribution
- Great tool in inferring details about population

Confidence Interval

- Recall 2: 68, 95, 99 rule.

$$P(|Z| < 1.96) = 0.95$$

Central limit Theorem — $X_1, \dots, X_n$ be i.i.d random variables

$E X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n \rightarrow \infty]{\text{distribution}} Z \quad \text{where } Z \sim N(0, 1)$$

$a < b, a, b \in \mathbb{R}$

$$\text{i.e. } P(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b) \longrightarrow P(a \leq Z < b) \quad \text{as } n \rightarrow \infty$$

Proof: uses several results $\square$

Sam $n \rightarrow$ large

to take

$$a = -1.96, b = 1.96$$

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < 1.96\right) \cong 0.95$$

⌋ (**)

$$(=) P\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X} \leq \mu \leq \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right)$$

$$\cong 0.95$$

$$(=) P\left(\mu \in \left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right)\right) \cong 0.95$$

So instead of giving $\bar{X}$ as a point estimator for μ
we can provide a Range / Interval for μ
with 95% Confidence.

Rephrasing $\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X} \right)$ is a 95%

Confidence interval for μ

• On an average, 95% of the time $\mu \in \left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X} \right)$

limitation