# PRIOR AND POSTERIOR DISTRIBUTION

Multiplication law of probability

Let, A and B be two dependent events then, the probability of getting events A and B is given by

$$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B/A)$$

INITIALLY OCCURS

$$\text{or, } P(A \cap B) = P(B) \cdot P(A/B)$$

where, $P(A/B)$ = Conditional probability of A given B
$P(B/A)$ = Conditional probability of B given A.

If events A and B are independent, then,

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

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

$P(B/A) = \frac{P(A \cap B)}{P(A)}$ ; $P(A) \neq 0$

# Conditional Probability

Conditional probability refers to the happening of an event when the information about happening of other event is given.

Let, A and B be two dependent events, then the conditional probability of A given B,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \; ; \; P(B) \neq 0$$

$$\text{Similarly, } P(B/A) = \frac{P(A \cap B)}{P(A)} \; ; \; P(A) \neq 0$$

If A and B are independent then,

$$P(A/B) = P(A) \text{ and } P(B/A) = P(B) \quad [\because P(A \cap B) = P(A) \cdot P(B)]$$

- QUESTIONS -

Q1.

Let, A = event that accidents are alcohol related

B = event that accidents occured at night

$P(B) = 0.6$

$P(A) = 0.52$

$P(A \cap B) = 0.37$

a. Probability that an accident was alcohol related given that it occured at night,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.37}{0.6} = \underline{0.616}$$

i.e. if accidents occured at night then there is 61% chances that the accidents are alcohol related.

b.

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.37}{0.52} = \underline{0.711}$$

Q2.

Let, A: event in which pilots strike

B: event in which drivers strike

$P(A) = 0.75$

$P(B) = 0.65$

$P(A/B) = 0.90$

a. $P(A \cap B) = P(B) \cdot P(A/B)$

$= 0.65 * 0.90$

$= \underline{0.585}$

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

$= \frac{0.585}{0.75}$

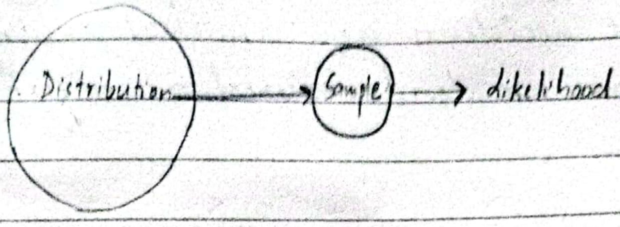
$= \underline{0.78}$

classmate

Date

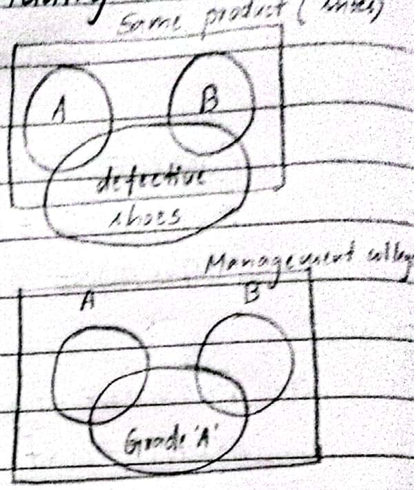
Page

Q3.




Common event

Mutually exclusive events




Let, $E_1$ = screws manufactured by plant I

$E_2$ = screws manufactured by plant II.

$S$ = Standard quality screws manufactured by plant

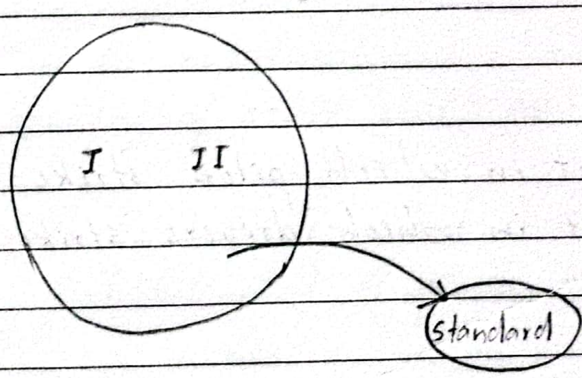
$S \subset E_1 \cup E_2$

Given, $P(E_1) = 0.80$

$P(E_2) = 0.20$

$P(S/E_1) = 0.85$

$P(S/E_2) = 0.65$

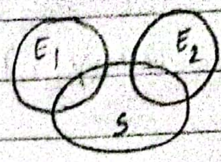



$P(E_1) = 0.8$ — $P(S/E_1) = 0.85$ — $P(E_1 \cap S) = P(E_1) \cdot P(S/E_1) = 0.8 * 0.85 = 0.68$

$P(E_2) = 0.2$ — $P(S/E_2) = 0.65$ — $P(E_2 \cap S) = P(E_2) \cdot P(S/E_2) = 0.2 * 0.65 = 0.13$

Probability of getting standard quality screw is given by

$$P(S) = P(E_1 \cap S) + P(E_2 \cap S)$$
$$= 0.68 + 0.13$$
$$= 0.81$$




(a) Probability that a randomly selected standard quality screw is manufactured by plant I,

$$P(E_1/S) = \frac{P(E_1 \cap S)}{P(S)}$$
$$= \frac{0.68}{0.81}$$
$$= \underline{0.839}$$

(b) ... manufactured by plant II

$$P(E_2/S) = \frac{P(E_2 \cap S)}{P(S)}$$
$$= \frac{0.13}{0.81}$$
$$= \underline{0.1604}$$

or $1 - P(E_1/S)$

$$= 1 - 0.839$$
$$= 0.16$$

$$\boxed{P(A/B) + P(A'/B) = 1}$$

Q11. H = Husband is watching TV
H' = Husband is not watching TV
W = Wife is watching TV

$P(H) = 0.60$ , $P(H') = 0.40$

[illegible], 2024

Q8.

Let, G = Event that football team will play on grass
A = " " " " " " " artificial turf
K = Event that the player incurs a knee injury

Given, $P(A) = 0.3 \Rightarrow P(G) = 0.70$

$P(K/G) = 0.40$

$P(K/A) = 0.40 + 60\%$ of $0.40$

$= 0.40 + 0.24$

$= 0.64$

a. The probability that a randomly selected player incurs a knee injury,

$P(K) = P(G \cap K) + P(A \cap K)$

$= P(G) \cdot P(K/G) + P(A) \cdot P(K/A)$

$= 0.70 * 0.40 + 0.30 * 0.64$

$= \underline{0.472}$

b. The probability that a randomly selected football player with a knee injury incurred the injury playing on grass,

$$P(G/K) = \frac{P(G \cap K)}{P(K)}$$

$$= \frac{P(G) \cdot P(K/G)}{P(K)} = \frac{0.7 * 0.4}{0.472}$$

$$= 0.593$$

Q13. Let, W = Women with breast cancer
W' = Women without breast cancer
T = Test is positive

Common event $\longrightarrow$ Mutually exclusive events

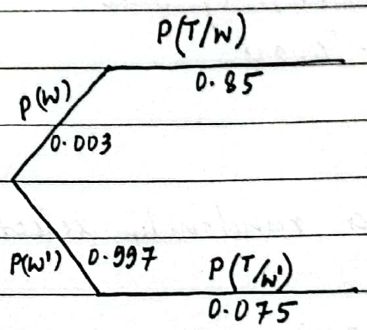
Given,

$P(T/W) = 0.85$ , $P(W) = 0.3\% = 0.003$

$P(W') = 1 - 0.003 = 0.997$

$P(T'/W') = 0.925 \Rightarrow P(T/W') = 1 - 0.925$ $\left[ \because P(A/B) + P(A'/B) = 1 \right.$

$= 0.075$




$P(W \cap T) = 0.003 * 0.85 = 0.00255$

$P(W' \cap T) = 0.997 * 0.075 = 0.074$

Probability that the test is positive $P(T) = P(W \cap T) + P(W' \cap T)$

$= 0.00255 + 0.074$

$= 0.0765$

If the test is positive, then the probability that the woman has cancer is given by

$$P(W/T) = \frac{P(W \cap T)}{P(T)}$$

$$= \frac{P(W) \cdot P(T/W)}{P(T)}$$

$$= \frac{0.003 * 0.85}{0.0765}$$

$$= \frac{0.0025}{0.0765}$$

$$= \underline{0.032}$$

# Statement of Baye's theorem

Let $E_1, E_2, \ldots, E_n$ be mutually exclusive events defined on the sample space 'S' such that $P(E_i) > 0 \; \forall i = 1, 2, \ldots, n$. Let $A$ be any arbitrary event which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$ then Baye's theorem states that,

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

$$= \frac{P(E_i) \cdot P(A/E_i)}{P(A)}$$

where,

$$P(A) = P(A \cap E_1) + \cdots + P(A \cap E_n)$$

$$= P(E_1) \cdot P(A/E_1) + \cdots + P(E_n) \cdot P(A/E_n)$$

$$= \sum_{i=1}^{n} P(E_i) \cdot P(A/E_i)$$

# Likelihood Function

The joint probability mass function or joint probability density function is called likelihood function.

Let $X = x_1, x_2, \ldots, x_n$ be a sample drawn from a population with probability density function $f(x; \theta)$, then the likelihood function of sample observations $x_1, x_2, \ldots, x_n$ is given by

$$L = f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta)$$

$$= \prod_{i=1}^{n} f(x_i; \theta)$$

[Here, $x_i$'s are identically, independently distributed (iid) ].

Example:

\# Let $x_i \sim P(\lambda)$ $\forall$ $i = 1, 2, \ldots, n$, then, the likelihood function of $x_1, x_2, \ldots, x_n$ is given by

$$L = \prod_{i=1}^{n} P(x_i ; \lambda)$$

$$= P(x_1 ; \lambda) \ldots P(x_n ; \lambda)$$

$$= \frac{e^{-\lambda} \lambda^{x_1}}{x_1 !} \cdots \frac{e^{-\lambda} \lambda^{x_n}}{x_n !}$$

$$= \frac{e^{-n\lambda} \cdot \lambda^{\Sigma x_i}}{\prod_{i=1}^{n} (x_i !)}$$

Normal distribution with parameter $\mu$ and $\sigma^2$.

\# Let $x_i \sim N(\mu, \sigma^2)$
then, the likelihood function of $x_1, x_2, \ldots, x_n$ is given by:

$$L = \prod_{i=1}^{n} f(x_i ; \mu, \sigma^2)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

If $X \sim N(\mu, \sigma^2)$ then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} ; \quad -\infty < x < \infty$$

$-\infty < \mu < \infty$

$\sigma > 0$

- PRIOR AND POSTERIOR DISTRIBUTION -

In Bayesian Inference, we treat the parameter '$\theta$' as a random variable having probability density function or probability mass function $f(\theta)$; $\theta \in \Theta$. The probability distribution of

Statistics → Descriptive, Inference

Inference ⇓ Decision Making

'θ' defines the prior probability density / mass function f(θ) is called prior distribution

Let us consider a random sample $X = x_1, x_2, \ldots, x_n$ be drawn from a population with pdf / pmf $f_\theta(x)$ where θ

\# is considered as a parameter θ.

Aug 3, 2024

NOTE: If $X \sim B(m, n)$ then

$$f(x) = \frac{x^{m-1}(1-x)^{n-1}}{\beta(m,n)}$$

If $X \sim G(\alpha, \beta)$, then

$$f(x) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma\alpha} \; ; \quad x > 0$$

If $X \sim \exp(\theta)$

$$f(x) = \theta e^{-\theta x} \; ; \quad x > 0$$

$$\beta(m, n) = \frac{\Gamma m \, \Gamma n}{\Gamma m+n}$$

$$\Gamma m = (m-1)!$$

If $X \sim Gpo(\theta)$

$$f(x) = \theta e^{-\theta x}$$

$$E(X) = \frac{1}{\theta}$$

If $X \sim Gpo(\theta)$

$$f(x) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}$$

$$E(X) = \theta$$

\# (Continue)

Here, we denote the pdf/pmf of X, $f_\theta(x)$ by $f(x/\theta)$, the conditional pdf/ ~~pmf~~ of X given θ, called likelihood function of $x/\theta$

Now, the pdf/pmf of $x$ and $\theta$ is given by

$$f(x,\theta) = f(\theta)\cdot f(x/\theta)$$

∴ The marginal pdf/pmf of $x$ given by

$$f(x) = \int_{\theta} f(x,\theta)\, d\theta$$

$$= \int_{\theta} f(\theta)\cdot f(x/\theta)\, d\theta$$

[For continuous]

| wt. → X \ Y ↓ hight | $y_1$ | $y_2$ | ... | $y_m$ |
|---|---|---|---|---|
| $x_1$ | $P_{11}$ | $P_{12}$ | ... | $P_{1m}$ |
| ⋮ | | | | |
| $x_n$ | | | | |

$$P(X = x_1) = \sum_{j=1}^{m} P_{1j}$$

$$f(x) = \sum_{\theta} f(\theta)\, f(x/\theta) \quad \text{[For discrete]}$$

Now, applying Baye's theorem, we get

$$f(\theta/x) = \frac{f(x,\theta)}{f(x)}$$

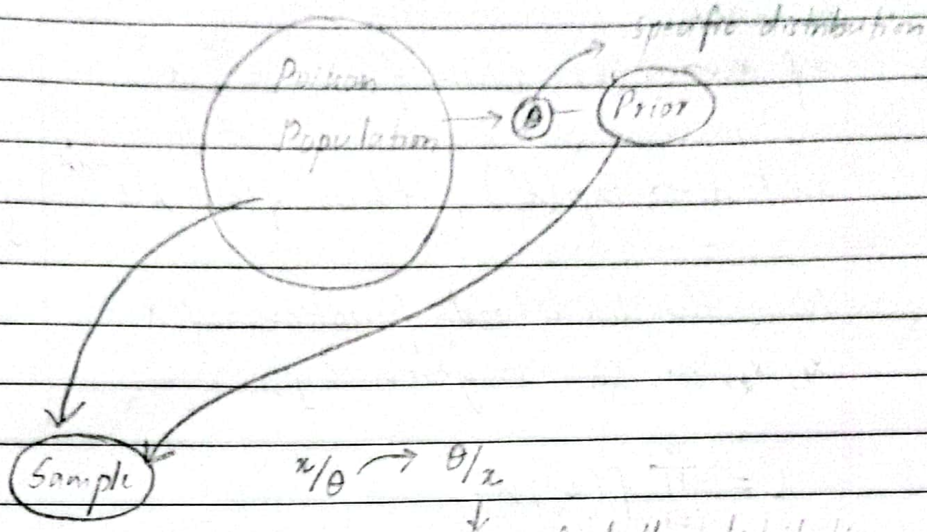
$f(x/\theta)$

$$\boxed{\therefore f(\theta/x) = \frac{f(\theta)\cdot f(x/\theta)}{f(x)}}$$

The conditional probability density function $f(\theta/x)$ is called posterior probability density function and the distribution defined by posterior probability function $f(\theta/x)$ is called posterior distribution.

The prior distribution gives the distribution of $\theta$ before the sample is drawn while the posterior distribution of $\theta$ gives the conditional distribution of $\theta$ after the sample is drawn, so the posterior distribⁿ reflects both the prior information and the sample information.

[ 11 ]




$x/\theta \rightarrow \theta/x$

we find this distribution

$$f(\theta/x) \propto \underset{\text{Prior}}{f(\theta)} \cdot \underset{\text{Likelihood}}{f(x/\theta)}$$

NOTE: A usual way of expressing the posterior density function is to use proportionality.

We have,

$$\underset{\text{Posterior of } \theta}{f(\theta/x)} = \frac{\overset{\text{Prior}}{f(\theta)} \cdot \overset{\text{Likelihood}}{f(x/\theta)}}{f(x)} \rightarrow \text{Normalizing constant}$$

$$\Rightarrow f(\theta/x) \propto f(\theta) \cdot f(x/\theta)$$

$$\Rightarrow \underset{\text{Posterior}}{f(\theta/x)} \propto \underset{\text{Prior}}{f(\theta)} * \text{Likelihood}$$

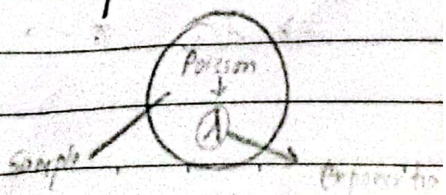
$$\Rightarrow \text{Posterior} \propto \text{Prior} * \text{Likelihood}.$$

# EXAMPLE

Let $X/\lambda \sim P(\lambda)$ and prior distribution of '$\lambda$' is exponential with parameter $\lambda'$.

Find the posterior distribution of $\lambda$.

$[f(\lambda/x) = ?]$



Solution

If $X \sim P(\lambda)$, then $f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

and $\lambda \sim Gp(\lambda')$, then $f(\lambda) = \lambda' e^{-\lambda'\lambda}$

If $X \sim P(\lambda)$,
$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

If $X \sim Gp(\theta)$
~~$P(X=x)$~~ =
$f(x) = \theta e^{-\theta x}$

Now, the likelihood function of $X = x_1, x_2, \dots, x_n$ is given by,

$$L = \prod_{i=1}^{n} f(x_i; \lambda)$$

$$= f(x_1; \lambda) \dots f(x_n; \lambda)$$

$$\Rightarrow L = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \dots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i)!} \rightarrow \text{Constant}$$

$$L = e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} \cdot \text{constant}$$

Now, the prior pdf of $\lambda$ is

$$f(\lambda) = \lambda' e^{-\lambda'\lambda}$$
$$= e^{-\lambda'\lambda} \cdot \text{constant}$$

Now, the posterior distribution of $\lambda$ is calculated as

Posterior $\propto$ Prior * Likelihood

$$\Rightarrow f(\lambda/x) \propto \{e^{-\lambda'\lambda}\} * \{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}\}$$

$$\Rightarrow f(\lambda/x) \propto e^{-\lambda(n+\lambda')}\lambda^{\left(\sum_{i=1}^{n} x_i + 1\right) - 1}$$

$$\Rightarrow \lambda/x \sim G\left(\sum_{i=1}^{n} x_i + 1, n + \lambda'\right)$$

If $X \sim G(\alpha, \beta)$ then,
$f(x) = \frac{e^{-\beta x} x^{\alpha-1} \beta^{\alpha}}{\Gamma\alpha}$

Q. The annual number of claims arising from a particular group of policies follow a Poisson distribution with mean $\mu$. The prior distribution of $\mu$ is exponential with mean 30.
In the previous two years, the number of claims arising from the group were 28 and 26 respectively.
Determine the posterior distribution of $\mu$.

Solution
$X$ = annual number of claims

Here, $X$ follows Poisson distribution i.e. $X \sim P(\mu)$.
so, $P(X = x) = \dfrac{e^{-\mu} \mu^{x}}{x!}$

Now, the likelihood function,
$$L = \prod_{i=1}^{2} f(x_i, \mu)$$
$$= f(x_1, \mu) * f(x_2, \mu)$$
$$= \frac{e^{-\mu} \mu^{x_1}}{x_1!} * \frac{e^{-\mu} \mu^{x_2}}{x_2!}$$

$$= e^{-2\mu} \cdot \mu^{x_1 + x_2} \cdot \text{constant} \quad [\because x_1 = 28, x_2 = 26]$$

$$\Rightarrow L = e^{-2\mu} \mu^{54} \cdot \text{constant}$$

~~$\mu \sim$ Exp with mean 30~~
~~$f(\mu) = \frac{1}{30} e^{-\mu/30}$~~

Further, $\mu$ follows exponential distribution with mean 30.
$$\therefore f(\mu) = \frac{1}{30} e^{-\mu/30}$$

$= e^{-\mu/30} \cdot \text{constant}$

Now, posterior distribution of $\mu$ is calculated as
Posterior $\propto$ Prior $*$ Likelihood

$$\Rightarrow f(\mu/x) \propto \{e^{-\mu/30}\} * \{e^{-24\mu} \mu^{54}\}$$

$$\Rightarrow f(\mu/x) \propto e^{-61/30\,\mu} \mu^{55-1}$$

$\therefore$ $\mu/x$ follows gamma distribution with $\alpha = 55$, $\beta = \frac{61}{30}$

$$\therefore \mu/x \sim G(55, 61/30)$$

August 6, 2024

Q. If $X/\lambda \sim Exp(\lambda)$ and prior of $\lambda \sim Exp(\lambda')$ find the posterior distribution of $\lambda$.

Solution

Since, $X \sim Exp(\lambda)$ so the likelihood function $L$ of sample observations $x_1, x_2, \ldots, x_n$ is

If $X \sim Exp(\theta)$
$f(x) = \theta e^{-\theta x}$

(Posterior $\propto$ Prior x Likelihood

$$L = \prod_{i=1}^{n} f(x_i; \lambda) = \{\lambda e^{-\lambda x_1}\} \ldots \{\lambda e^{-\lambda x_n}\}$$

$$= \lambda^n e^{-\lambda \sum x_i}$$

Further,

$$f(\lambda) = \lambda' e^{-\lambda' \lambda}$$

$$= e^{-\lambda' \lambda} \cdot \text{constant}$$

Now, the posterior distribution of $\lambda$ is obtained as
Posterior $\propto$ Prior x Likelihood

$$\Rightarrow f(\lambda/x) \propto \{e^{-\lambda'\lambda}\}\{\lambda^{n} e^{-\lambda\sum_{i=1}^{n} x_i}\}$$

$$\Rightarrow f(\lambda/x) \propto e^{-\lambda(\lambda'+\sum x_i)} \lambda^{(n+1)-1}$$

$$\Rightarrow \lambda/x \sim G(n+1, \lambda'+\sum x_i)$$

# Assignment

If $X/\lambda \sim G(\alpha, \lambda)$ and prior $\lambda \sim G(\alpha', \lambda')$ then find the posterior distribution of $\lambda$.

Q. If $X/p \sim Bin(m, p)$ and prior $p \sim \beta(\alpha, \lambda)$, then find the posterior distribution of 'p'.

Solution

Since $X/p \sim B(m, p)$ so

$$f(X/p) = {}^{m}C_{x}\, p^{x} (1-p)^{m-x}$$

If $X \sim B(n, p)$ then $f(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$

Now, the likelihood function L of sample observations $x_1, x_2, \dots, x_n$

$X \sim \beta(\alpha, \lambda)$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\lambda-1}}{\beta(\alpha, \lambda)}$$

$$\beta(m, n) = \frac{\Gamma m \,\Gamma n}{\Gamma m+n}, \quad \Gamma m = (m-1)!$$

$$L = \prod_{i=1}^{n} f(x_i; m, p)$$

$$= \left\{ {}^{m}C_{x_1}\, p^{x_1} (1-p)^{m-x_1} \right\} \cdots \left\{ {}^{m}C_{x_n}\, p^{x_n} (1-p)^{m-x_n} \right\}$$

$$= \left\{ p^{\sum_{i=1}^{n} x_i} \right\} * \left\{ (1-p)^{mn - \sum_{i=1}^{n} x_i} \right\} . \text{ constant}$$

Further, $p \sim \beta(\alpha, \lambda)$

$$\therefore f(p) = \frac{p^{\alpha-1}(1-p)^{\lambda-1}}{\beta(\alpha, \lambda)} = p^{\alpha-1} . (1-p)^{\lambda-1} . \text{constant}$$

Now, the posterior distribution of p is calculated as

Posterior $\propto$ Prior * Likelihood

$$\Rightarrow f(P/x) \propto \left\{ p^{\alpha-1} . (1-p)^{\lambda-1} \right\} \left\{ p^{\sum_{i=1}^{n} x_i} (1-p)^{mn - \sum_{i=1}^{n} x_i} \right\}$$

$$\Rightarrow f(P/x) \propto p^{\left(\sum_{i=1}^{n} x_i + \alpha\right) - 1} (1-p)^{\left(mn - \sum_{i=1}^{n} x_i + \lambda\right) - 1}$$

$$\Rightarrow P/x \sim \beta\left(\alpha + \sum_{i=1}^{n} x_i,\; mn - \sum_{i=1}^{n} x_i + \lambda\right)$$

Hence, posterior of 'p' follows Beta distribution with parameters $\left(\alpha + \sum_{i=1}^{n} x_i\right)$ & $\left(mn - \sum_{i=1}^{n} x_i + \lambda\right)$.

# CONJUGATE PRIOR

For a given likelihood, if the posterior distribution belongs to the same family as the prior distribution then prior is called conjugate prior.

Q. The no. of claims per month has a negative binomial distribution with parameters $k$ and $p$. The number of claims observed over 'n' months are $x_1, x_2, \ldots, x_n$.

(i) Show that the beta distribution is the conjugate prior for 'p'.

Solution

Given, $X \sim NB(k, p)$ so,

$$f(x) = {}^{x+k-1}C_x \, p^k (1-p)^x$$

Now, the likelihood function

$$L = \prod_{i=1}^{n} f(x_i; k, p)$$

$$= \left\{ {}^{x_1+k-1}C_{x_1} \, p^k (1-p)^{x_1} \right\} \cdots \left\{ {}^{x_n+k-1}C_{x_n} \, p^k (1-p)^{x_n} \right\}$$

$$= \left\{ p^{nk} \right\} \left\{ (1-p)^{\sum_{i=1}^{n} x_i} \right\} \cdot \text{constant}$$

If $X \sim NB(k, p)$

$f(x) = {}^{x+k-1}C_x \, p^k (1-p)^x$

If $X \sim \beta(m, n)$, then

$f(x) = \dfrac{x^{m-1} (1-x)^{n-1}}{\beta(m, n)}$

Further, $p \sim \beta(\alpha, \lambda)$

$$\therefore f(p) = \frac{p^{\alpha-1} (1-p)^{\lambda-1}}{\beta(\alpha, \lambda)}$$

$$= p^{\alpha-1} (1-p)^{\lambda-1} \cdot \text{constant}$$

Now, the posterior of $p$ is calculated as

Posterior $\propto$ Prior $*$ Likelihood

$$\Rightarrow f(P/x) \propto \left\{ p^{\alpha-1} \cdot (1-p)^{\lambda-1} \right\} \left\{ p^{nk} (1-p)^{\sum_{i=1}^{n} x_i} \right\}$$

$$\Rightarrow f(P/x) \propto p^{(\alpha + nk)-1} (1-p)^{(\sum_{i=1}^{n} x_i + \lambda)-1}$$

$$\Rightarrow P/x \sim \beta\left(\alpha + nk, \sum_{i=1}^{n} x_i + \lambda\right)$$

Hence, posterior of $p$ follows beta distribution. Thus, beta distribution is the conjugate prior for $p$.

# Assignment

Q. 10 iid observations from a $P(\lambda)$ distribution gave 3, 4, 3, 1, 5, 5, 2, 3, 3, 2. Assuming an $Exp(0.2)$ as prior distribution of $\lambda$.
Find the posterior distribution of $\lambda$.

Q. A random sample of size 10 from a Poisson distribution with mean $\lambda$ yields the following data values:
3, 4, 3, 1, 5, 5, 2, 3, 3, 2.
The prior distribution of $\lambda$ is Gamma $(5, 2)$.
Find the posterior distribution of $\lambda$ and also find the mean of distribution.

$X \sim G(\alpha, \beta)$

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

August 8, 2024

Q. The number of claims in a week arising from a certain group insurance policies has a Poisson $dist^n$ with mean $\lambda$. Seven claims were incurred in the last week. The prior $dist^n$ of $\lambda$ is uniform on the integers 8, 10, and 12.

(i) Determine the posterior $dist^n$ of $\lambda$.

(ii) Find the mean of posterior $dist^n$.

Solution

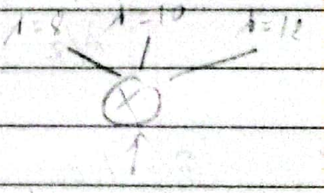
Let $X$ = number of claims in a week

Here $X \sim P(\lambda)$ so $P(X = x) = \dfrac{e^{-\lambda} \lambda^x}{x!}$

Further, $\lambda$ is uniform on the integers 8, 10, & 12.

| | | | |
|---|---|---|---|
| $\therefore \lambda$: | 8 | 10 | 12 |
| $P(\lambda)$ | $1/3$ | $1/3$ | $1/3$ |

To obtain the posterior distribution, we must calculate $P(\lambda = 8/X = 7)$, $P(\lambda = 10/X = 7)$ and $P(\lambda = 12/X = 7)$




Now, probability of

$P(\lambda = 8/X = 7) \propto P(\lambda = 8) \cdot P(X = 7/\lambda = 8)$

$= \dfrac{1}{3} \cdot \left( \dfrac{e^{-8}\, 8^7}{7!} \right)$

$= \underline{0.0465}$

$P(\lambda = 10/X = 7) \propto P(\lambda = 10) \cdot P(X = 7/\lambda = 10)$

$= \dfrac{1}{3} \cdot \left( \dfrac{e^{-10}\, 10^7}{7!} \right) = \underline{0.0300}$

$$P(\lambda=12/x=7) \propto P(\lambda=12) \cdot P(x=7/\lambda=12)$$

$$= \frac{1}{3} \cdot \left( \frac{e^{-12}\, 12^{7}}{7!} \right)$$

$$= \underline{0.014}$$

$$\therefore P(x=7) = P(x=7, \lambda=8) + P(x=7, \lambda=10) + P(x=7, \lambda=12)$$

$$= 0.465 + 0.03 + 0.014$$

$$= \cancel{0.509} \quad \underline{0.091}$$

Now, the posterior distⁿ

$$P(\lambda=8/x=7) = \frac{0.0465}{0.091} = 0.510$$

$$P(\lambda=10/x=7) = \frac{0.030}{0.091} = 0.329$$

$$P(\lambda=12/x=7) = \frac{0.014}{0.091} = 0.153$$

(i)

| $\lambda$ | 8 | 10 | 12 |
|---|---|---|---|
| $\lambda/x$ | 0.510 | 0.329 | 0.153 |

(ii) Now, expected value of posterior distⁿ

$$E(\lambda/x) = \Sigma \lambda P(\lambda/x)$$

$$= 8(0.510) + 10(0.329) + 12(0.153)$$

$$= \underline{9.206}$$

**Assignment**

The no. of claims received per day from a certain portfolio has a Poisson distⁿ with mean $\lambda$. The prior distⁿ of $\lambda$ is as follows

| | | | |
|---|---|---|---|
| $\lambda$ : | 1 | 2 | 3 |
| $P(\lambda)$ : | 0.3 | 0.5 | 0.2 |

Given that 3 claims were received last day, determine the posterior distⁿ of $\lambda$ and the mean.

Hypothesis
↓
Assumption for population parameter

Avg wt $\boxed{55\ kg}$

Sample

statistic

| | Population | Sample |
|---|---|---|
| Mean | $\mu$ | $\bar{x}$ |
| Variance | $\sigma^2$ | $s^2$ |
| Proportion | $P$ | $p$ |
| Correlation | $\rho$ | $r$ |

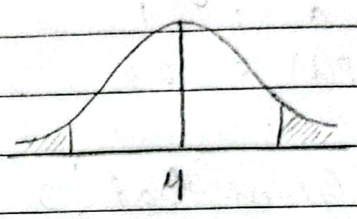
## Types of Hypothesis

i) Null Hypothesis

$H_0 : \mu = \mu_0$

ii) Alternative Hypothesis

$H_1 : \mu \neq \mu_0$ (Two tailed test)



$H_1 : \mu > \mu_0$ (Right tailed test) → One tailed test

$H_1 : \mu < \mu_0$ (Left tailed test)

eg: 55 kg or not

eg: ~~more~~ Less than 55 kg

eg: more than 55

# Types of Error

| Situation \ Decision making | $H_0$ is true | $H_0$ is false |
|---|---|---|
| Reject $H_0$ | Wrong decision (Type I error) | Correct decision |
| Do not reject $H_0$ | Correct decision | Wrong decision (Type II error) |

* Type first error

$\alpha$ = Prob. (type first error) = Level of significance

also called producer's risk.

* Type second error

$\beta$ = Prob ( type second error)

Consumer's risk.

$\beta$ = Prob ( accept $H_0$ ~~when~~ / $H_0$ is false)

$1-\beta$ = Prob ( reject $H_0$ / $H_0$ is false)

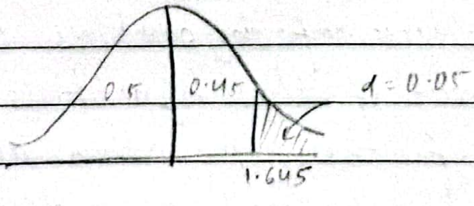
= Prob ( correct decision)

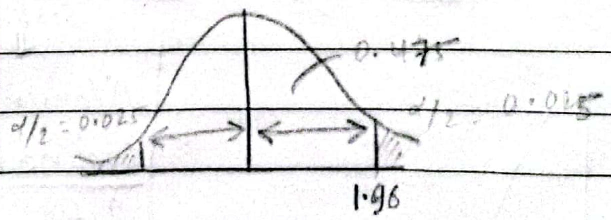
= Power of the test.

# Critical value ( significant value)

The value which separates the acception & rejection ~~of~~ region is called the critical value.

Depends on
* Size of $\alpha$ ( level of significance)
* Whether the test is one tailed or two tailed.



# Degree of freedom

Number of values in the sample that can be choosen freely. For 'n' sample observations, the degree of freedom is $(n-1)$.

↓

For small sample sized test

# Identification of one tailed and two tailed test.

Aug 10, 2024

UNIT 6 : Non-parametric test: (NP-Test)

Parametric test

t- test
Z - test
F- test

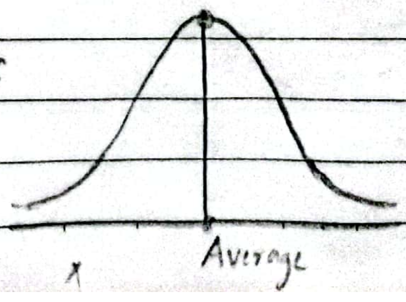
} Follows normal distribution

* Introduction

Non-parametric tests are defined as those statistical test of hypothesis in which parametric values are not involved.

In other words, those tests of hypothesis are said to be non-parametric test if the hypothesis does not involve any parameter of the population and if measurements are on the nominal or ordinal scale.

As $n \to \infty$, f → Z




Scaling
⇓

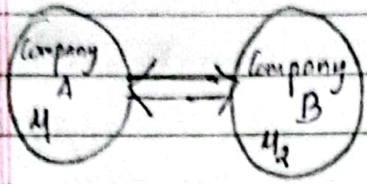
| | |
|---|---|
| Nominal, Ordinal | Qualitative data |
| Ratio, Interval | Quantitative data |

(VVI)

# Difference between parametric and non-parametric test

| Parametric | Non-parametric |
| --- | --- |
| i) Parametric tests are those tests whose models specify certain conditions about the parameter of the population from which the sample has been drawn. | i) N-P tests are those tests whose models do not specify any conditions about the parameters of the population from which the sample has been drawn. |
| ii) Parametric tests are mostly applied only to the data which are measured in interval or ratio scale. | ii) N-P tests are applied only to data which are measured in nominal and ordinal scale. |
| iii) Parametric tests are the most powerful. | iii) N-P tests are less powerful or weaker than the parametric tests. |
| iv) Parametric tests are designed to test the statistical hypothesis of one or more parameters of the population. | iv) N-P tests are designed only to test statistical hypothesis which does not involve any parameter of the population. |

# Mann–Whitney U test




Alternative

| Small Sample size | Large sample size |
|---|---|
| $n_1$ and $n_2 \le 10$ | $n_1 > 10$, $n_2 > 10$ |
| $n_1 + n_2 \le 20$ | $n_1 + n_2 > 20$<br>z-test |

↳ $Md_1 = Md_2$ ✓

or, $E(X) = E(Y)$

This test is the most powerful non-parametric test for testing hypothesis of difference between two independent location of two independent random samples. This is non-parametric alternative test of t-test for difference of means.

average

Procedure.

Null hypothesis ($H_0$): $Md_1 = Md_2$ i.e. there is no significant difference between two sample medians or two popⁿ medians.

Alternative hypothesis ($H_1$): $Md_1 \neq Md_2$ i.e. there is a significant difference between two sample medians or two population medians. (Two tailed test)

or $H_1: Md_1 > Md_2$ (Right) } one tailed test
or, $H_1: Md_1 < Md_2$ (left) }

Sample size ($n_1 \le 10$ & $n_2 \le 10$)

Test-statistic: Under $H_0$,

$U_0$ = minimum of $\{U_1$ and $U_2\}$

where $U_1 + U_2 = n_1 n_2$

p-value $\Rightarrow U_0, n_1, n_2$



| $x_1$ | $x_2$ |
|---|---|
| 10 (4) | 15 (8.5) |
| 18 (11) | 12 (6.5) |
| 12 (6.5) | 19 (12) |
| 15 (8.5) | 25 (13) |
| 17 (10) | 6 (2) |
| 5 (1) | 9 (3) |
| | 11 (5) |
| $n_1 = 6$ | $n_2 = 7$ |

for 12, $\frac{6+7}{2} = 6.5$

Common rank

(Take average)

Increasing or ascending order

Rank of $x_1$    Rank of $x_2$

Sum of ranks, $R_1 =$    $R_2 =$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

→ where, $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$R_1$ = sum of ranks given to first sample data.

$R_2$ = " " " " " second " ".

P-value > $\alpha$, then we do not reject $H_0$ or insignificant

P-value < $\alpha$ then we reject $H_0$ or significant

Critical value: The critical value of the test-statistic is obtained from Mann-Whitney U-test table at $\alpha$ level of significance and ($n_1$ & $n_2$).

Decision: If $U_0 > U_{\alpha,(n_1, n_2)}$ then we do not reject $H_0$ otherwise reject $H_0$.

$\alpha = 0.05$
Two tailed
$n_1 = 7$, $n_2 = 8$

| $n_2 \backslash n_1$ | 7 |
|---|---|
| 8 | 10 |

| $n_2 \backslash n_2$ | 8 |
|---|---|
| 7 | 10 |

p-value

$n_1 = 5$ $\quad$ $U_0 = 6$
$n_2 = 10$

$\alpha = 0.05$
p-value = 0.0097
Multiply by 2 for two-tailed test.
$2P = \alpha$

The p-value is obtained from Mann-Whitney probability table for $(n_1, n_2)$ and $U_0$.

For two tailed, P-value = $2P_0$
For one tailed, P-value = $P_0$

If p-value > $\alpha$, then we do not reject $H_0$ otherwise reject $H_0$.

August 13, 2024 [SMALL SAMPLE SIZE]

Q14.

$n_1 < 10$ & $n_2 < 10$

$n_1 = 5$
$n_2 = 4$
$\alpha = 0.05$ = level of significance

If level of significance '$\alpha$' is not given, then consider 5%.

$H_0: Md_1 = Md_2$ i.e. There is no difference between two average numbers of correction of trained and untrained managers.
$H_1: Md_1 \neq Md_2$ i.e. there is difference between two average numbers of correction of trained and untrained managers (Two tailed test)

Test statistic: Under $H_0$,
$U_0$ = minimum of {$U_1$ and $U_2$}

where, $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$ (Note that, $n_1 n_2 = U_1 + U_2$)

Calculation:

| Trained | Untrained | Ranks of trained | Ranks of untrained |
|---|---|---|---|
| 78 | 110 | 7 | 9 |
| 64 | 70 | 4 | 5 |
| 75 | 53 | 6 | 3 |
| 45 | 57 | 1 | 2 |
| 92 | | 8 | |
| | | $R_1 = 26$ | $R_2 = 19$ |

Now,

$$U_1 = 5\times4 + \frac{5(5+1)}{2} - 26$$

$$= 20 + 15 - 26$$

$$= 9$$

$$U_2 = 5\times4 + \frac{4\times5}{2} - 19$$

$$= 20 + 10 - 19$$

$$= 11$$

$\therefore U_0 = 9$

Critical value : We have, $\alpha = 0.05$ and $(n_1, n_2) = (5,4)$

$\therefore U_{(5,4),0.05}$ (two-tailed test) $= 1$

Decision : Since $U_0 > U_{(5,4),0.05}$, so we do not reject $H_0$. Hence, there is no difference between the average number of correction of trained and untrained managers.

p-value = 0.4524
∴ p-value = 2(0.4524)
= 0.9048

$\alpha = 0.05$

∵ p-value > $\alpha$, so we do not reject $H_0$.

Assignment
Q15 and 17

Q. The heart beating rate of 5 vegetarians and 5 non vegetarians are recorded below:

| | | | | | |
|---|---|---|---|---|---|
| Vegetarians: | 56 | 67 | 82 | 60 | 75 |
| Non-vegetarians: | 53 | 42 | 75 | 58 | 65 |

Is the mean heart beating rate of non vegetarians significantly high. Use Mann Whitney U test at $\alpha = 5\%$.

Solution. $n_1 = 5$, $n_2 = 5$
$\alpha = 0.05$

$H_0$: $Md_1 = Md_2$ i.e. the mean heart beating rate of non-vegetarian is not significantly high than that of vegetarian.
$H_1$: $Md_1 < Md_2$ i.e. the mean heart beating rate of non-vegetarian is significantly higher than vegetarian (one tailed test)

Test statistic: Under $H_0$
$U_0$ = minimum of $\{U_1, U_2\}$

where, $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

Combined

Calculation

| Vegetarian | Non-vegetarian | Ranks of vegetarian | Ranks of n |
|---|---|---|---|
| 56 | 53 | 3 | |
| 67 | 42 | 7 | |
| 82 | 75 | 10 | |
| 60 | 58 | 5 | |
| 75 | 65 | 8.5 | |
| | | $R_1 = 33.5$ | |

Now,

$$U_1 = 5\times5 + \frac{5\times6}{2} - 33.5$$

$$= 25 + 15 - 33.5$$

$$= 6.5$$

$$U_2 = 5\times5 + \frac{5\times6}{2} - 21.5$$

$$= 25 + 15 - 21.5$$

$$= 18.5$$

$\therefore U_0 = 6.5$

Critical value; We have $\alpha = 0.05$ and $(n_1, n_2) = (5,5)$

$\therefore U_{(5,5),\,0.05}$ (one-tailed test) $= 4$

Decision: Since $U_0 > U_{(5,5),\,0.05}$. Hence, ~~there~~ the mean heart beating rate of non-vegetarian is not significantly high than that of vegetarian.

.1111
.1548
.13385 > 0.0

vegetarian

2

1

8.5

4

6

$R_2 = 21.5$

# [LARGE SAMPLE SIZE]

For large sample size
$U_0 \sim N(\mu_{U_0}, \sigma^2_{U_0})$

large sample size i.e. $n_1 > 10$ and $n_2 > 10$

For large sample size, the distribution of $U_0$ is approximately by normal distribution with mean $= \frac{n_1 n_2}{2}$

and variance $= \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$

Test statistic: Under $H_0$

$$Z = \frac{U_0 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

In case of tied observation, the corrected value of s.d. is calculated as

$$\sigma_U = \sqrt{\frac{n_1 n_2}{n(n-1)} \left[ \frac{n^3 - n}{12} - \frac{\Sigma (t_i^3 - t_i)}{12} \right]}$$

where, $n = n_1 + n_2$

$t_i =$ number of times $i^{th}$ rank is repeated

$X \sim N(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma^2}$$

$U_0 \sim N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right)$

$$Z = \frac{U_0 - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

P123

Q16.

Solution

$n_1 = 11$

$n_2 = 9$

$H_0: Md_1 = Md_2$ i.e. yields of wheats per unit area is not different

$H_1: Md_1 \neq Md_2$ i.e. yields of wheats per unit area is different (Two-tailed test)

$H_0: Md_1 = Md_2$ i.e. there is no significant difference in the average yields between two varieties of wheat I and wheat II.

$H_1: Md_1 \neq Md_2$ i.e. there is significant difference in the average yields between two varieties of wheat I and wheat II (two-tailed test).

Calculation:

| Wheat I | Wheat II | Combined Ranks of wheat I | Ranks of |
|---|---|---|---|
| ~~15.9~~ | ~~16.4~~ | 9 | 12.5 |
| ~~15.3~~ | ~~16.8~~ | 5 | 15 |
| ~~16.4~~ | ~~17.1~~ | 12.5 | 17 |
| ~~14.9~~ | ~~16.9~~ | 3 | 16 |
| ~~15.3~~ | ~~18.0~~ | 5 | 19 |
| ~~16.0~~ | ~~15.6~~ | 10.5 | 8 |
| 14.6 | ~~18.1~~ | 2 | 20 |
| ~~15.3~~ | ~~17.2~~ | 5 | 18 |
| ~~14.5~~ | 15.4 | 1 | 7 |
| ~~16.6~~ | | 14 | |
| 16.0 | | 10.5 | |
| | | $R_1 = 77.5$ | $R_2 =$ |

$t_1 = 3$

$t_2 = 2$

Now,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 11 \times 9 + \frac{11 \times 12}{2} - 77.5$$

$$= 87.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$= 11 \times 9 + \frac{9 \times 10}{2} - 132.5$$

$$= 99 + 45 - 132.5$$

$$= 11.5$$

$\therefore U_0 =$ minimum of $\{U_1 \; \& \; U_2\}$

$= 11.5$

Test-statistic: Since, the problem is large sample size so

at II

$$Z = \frac{U_0 - E(U_0)}{\sqrt{Var(U_0)}}$$

Now, $E(U_0) = \frac{n_1 n_2}{2} = \frac{11(9)}{2} = 49.5$

and $V(U_0) = \frac{n_1 n_2}{n(n-1)} \left[ \frac{n^3 - n}{12} - \frac{(3^3 - 3) + (2^3 - 2) + (2^3 - 2)}{12} \right]$

$= 172.46$

132.5

$\therefore Z = \frac{11.5 - 49.5}{\sqrt{172.46}} = -2.89 \Rightarrow |Z\text{-cal}| = 2.89$

Critical value: We have, $\alpha = 0.01$

$$Z_{tab} = Z_{0.01} \text{ (two-tailed test)} = 2.575$$

Decision: Since $|Z_{cal}| > Z_{tab}$, so we reject $H_0$.

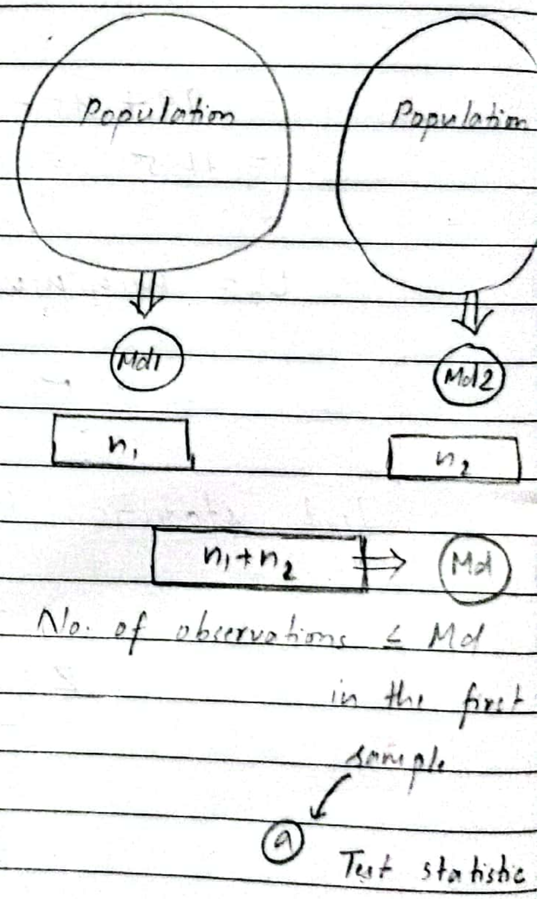
Assignment Q18

August 15, 2024

Median Test

Median Test:

Median test is non-parametric test used to test the difference in medians of two independent distributions. In other words, this test is applied to test whether two independent random samples have been drawn from two populations with same median or not. Also, it is used to test whether the two treatments applied in an experiment are equally effective or not.




Case I:

Small sample size i.e. $n_1 \le 10$ and $n_2 \le 10$.

Null Hypothesis ($H_0$): $Md_1 = Md_2$
i.e. there is no significant difference between two samples medians.

Alternative Hypothesis ($H_1$): $Md_1 \neq Md_2$ i.e. there is a significant difference between two sample medians. (Two Tailed test)

or $H_1: Md_1 > Md_2$ }
or $H_1: Md_1 < Md_2$ } One Tailed test.

Test statistic: Under $H_0$, the test-statistic is obtained as follows:

(i) Combined ~~with~~ both the samples and arrange them in ascending order of magnitude such that $n = n_1 + n_2$.

(ii) Calculate the median of the combined sample and count the number of observations less or equal to median in the first sample which is called the test statistic and is denoted by 'a'.

Now,

$$P_0 = P(A = a)$$

$$= \frac{{}^{n_1}C_a \; {}^{n_2}C_{K-a}}{{}^{n_1+n_2}C_K}$$

where, $a = 0, 1, 2, \ldots,$ minimum of $(n_1, K)$

$K = \frac{n_1 + n_2}{2}$ ← round up when in decimal

Critical value:

We can obtain P-value ($P_0$), the probability associated with the ~~extreme~~ value as extreme as observed 'a' for $n_1$ and $n_2$.

$$\therefore P_0 = P(A \geq a)$$

Decision

For one-tailed test, P-value $= P_0 > \alpha$ then we do not reject $H_0$ otherwise reject $H_0$. For two-tailed test, P-value $= 2P_0 > \alpha$, then we do not reject $H_0$ otherwise reject $H_0$.

Q7.

$n_1 = 10$ and $n_2 = 10$

$H_0$ : $Md_1 = Md_2$ i.e. the two training programs are equally effective.

$H_1$ : $Md_1 \neq Md_2$ i.e. the two training programs are not equally effective. (Two-tailed test)

Test-statistic : Under $H_0$, the test statistic is 'a'.

Calculation:

Arranging observations of both samples in ascending order of magnitude.

30, 32, 34, 39, 40, 42, 43, 44, 45, 46, 47, 47, 48, 49, 50, 51, 55, 55, 59, 71

| (10) stem | (1) leaf |
|---|---|
| ~~3~~ | ~~2 9 0 4~~ |
| 4 | ~~6 8 2 8 9~~ 4 |
| 5 | ~~1 5 8 9 5~~ |
| 6 | |
| 7 | ~~1~~ |

$Md =$ Value of $\left(\frac{n+1}{2}\right)^{th}$ term ; $n = n_1 + n_2$

$= \frac{20+1}{2}$

$= 10.5^{th}$ term

$\therefore Md = \frac{46+47}{2} = 46.5$

Now, $a =$ no. of observations less or equal to Md (46.5) in the first sample.

$= 7$

Calculation of P-value

$P_0 = P(A \geq a)$ ; $K = \frac{n_1 + n_2}{2} = \frac{10+10}{2} = 10$

$$= \sum_{a=7}^{10} \frac{{}^{n_1}C_a \; {}^{n_2}C_{K-a}}{{}^{n_1+n_2}C_K}$$

$$= \frac{1}{{}^{20}C_{10}} \left[ {}^{10}C_7 \; {}^{10}C_{10-7} + {}^{10}C_8 \; {}^{10}C_{10-8} + {}^{10}C_9 \; {}^{10}C_{10-9} + {}^{10}C_{10} \; {}^{10}C_{10-10} \right]$$

$$= \frac{16526}{184756}$$

$$= 0.0894$$

Decision : P-value = $2P_0 = 2 \times 0.089$
$= 0.178$ [$\because$ The problem is two-tailed]

and $\alpha = 0.05$

Since, P-value > $\alpha$, so we do not reject $H_0$.
Hence, both the training programs are equally effective.

***

* Assignment 10

Large Sample size ($n_1 > 10$ and $n_2 > 10$)

| | No. of obs. $\leq$ Md | No. of obs. > Md | Total |
|---|---|---|---|
| I[st] Sample | a | b | a+b |
| II[nd] Sample | c | d | c+d |
| Total | a+c | b+d | a+b+c+d = n |

In this case, the median test is equivalent to $\chi^2$-test.

Therefore, the test-statistic is

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2_{(1)}$$

<u>Critical value</u>: The critical value of the test statistic is obtained from $\chi^2$-table at $\alpha$ level of significance and 1 dof.

$$\chi^2_{tab} = \chi^2_{\alpha, 1}$$

<u>Decision</u>: If $\chi^2_{cal} \geq \chi^2_{tab}$, then we reject $H_0$

otherwise we do not reject $H_0$.

Q9.

$n_1 = 12, \; n_2 = 12$

$n = n_1 + n_2 = 12 + 12 = 24$

$H_0$: $Md_1 = Md_2$ i.e. the marks distbn of two teachers do not differ significantly.

$H_1$: $Md_1 \neq Md_2$ i.e. the marks distbn of two teachers differ significantly (Two-tailed test).

<u>Test-statistic</u>: Under $H_0$,

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

<u>Calculation</u>:

Arranging observations of both samples in ascending order of magnitude:

31, 36, 43, 45, 48, 60, 62, 66,

73, 73, 74, 76, 77, 77, 77, 78, 78, 80, 82,

85, 88, 89, 91, 92

| S | L |
|---|---|
| 3 | 6 1 |
| 4 | 3 5 8 |
| 5 | |
| 6 | 6 0 2 |
| 7 | 3 3 7 8 4 7 8 6 7 |
| 8 | 9 2 0 8 5 |

Md = value of $\left(\frac{n+1}{2}\right)^{th}$ term ; $n = n_1 + n_2$
$= 12 + 12 = 24$

$= \left(\frac{24+1}{2}\right)^{th}$ term

$= 12.5^{th}$ term

$\therefore Md = \frac{76+77}{2}$

$= 76.5$

2x2 Contingency table

| | $x_i \le Md$ | $x_i > Md$ | Total |
|---|---|---|---|
| Sample I | a = 7 | b = 5 | 12 |
| Sample II | c = 5 | d = 7 | 12 |
| Total | (a+c) = 12 | (b+d) = 12 | 24 |

$\therefore \chi^2 = \frac{24(7\times7 - 5\times5)^2}{12\times12\times12\times12}$

$= 0.666$

Critical value:
We have, $\alpha = 0.05$ and degree of freedom (df) = 1

$\therefore \chi^2_{tab} = \chi^2_{1, 0.05} = 3.841$

Decision: Since $\chi^2_{cal} < \chi^2_{tab}$, so we do not reject $H_0$. Hence, the marks distbⁿ of two teachers do not differ significantly.

August 17, 20..

# Kolmogorov Smirnov Test (K-S Test)
[ One sample, two samples ]

Kolmogorov-Smirnov (K-S) test is a test of goodness-of-fit. It is alternative to Chi-square test for goodness-of-fit when sample size is small.

$\chi^2$ - test for the goodness-of-fit.

Expected = $x \cdot P(x)$

$f(x) =$

$P(x) :$

$F(x)$ = Cumulative probability distn ~~density~~ function.

$-\infty$ ——— $a$ ——— $+\infty$

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

$f(x) = \frac{dF(x)}{dx}$

## One-Sample

Procedure:

$H_0$: $F(x) = F_0(x)$ i.e. there is no significant difference between observed and expected frequency. or frequencies are uniformly distributed.

$H_1$: $F(x) \neq F_0(x)$ i.e. there is significant difference between observed and expected frequency or frequencies are not uniformly distributed (Two Tailed test)

or $H_1$: $F(x) > F_0(x)$
or $H_1$: $F(x) < F_0(x)$ } (One Tailed Test).

Test-statistic : Under $H_0$,

$D_0$ = Maximum $|F_e(x) - F_0(x)|$

where, $F_e(x) = \frac{cf_e}{n}$ ; $cf_e$ = expected cumulative frequency

$n = \Sigma f$ = Total frequency / sample size.

$f_e$ = expected frequency = $n p_i$ ; $p_i$ = probability of each category

$= \frac{\Sigma f}{\text{no. of categories}}$

$n \le 40$ Small
$n > 40$ large

$$F_0(x) = \frac{cfo}{n} \quad ; \quad cfo = \text{observed cumulative frequency}$$

Critical value: The critical value of the test-statistic is obtained from K-S table at $\alpha$ level of significance and size of sample 'n'.

$$\therefore D_{tab} = D_{n,\alpha}$$

Decision: If $D_0 \ge D_{n,\alpha}$ then we reject $H_0$ otherwise do not reject $H_0$.

Q17.

$H_0$: The computers of five hard disk are uniformly infected.
$H_1$: The computers of five hard disk are not uniformly infected. (Two Tailed test).

Test-statistic: Under $H_0$,

$$D_0 = \text{Maximum} \left| F_e(x) - F_0(x) \right|$$

Calculation

| Capacity of hard disk | No. of virus infected (O) | cfo | $F_0(x)$ | Expected (E) |
|---|---|---|---|---|
| 500 | 11 | 11 | 11/50 | |
| 320 | 15 | 26 | 26/50 | |
| 1000 | 20 | 46 | 46/50 | |
| 2000 | 3 | 49 | 49/50 | |
| 4000 | 1 | 50 | 50/50 | |

$\longrightarrow$

Calculation :

| Capacity of hard disk | (O) No. of virus infected | cfo | $F_0(x)$ | Expected (E) = np | cfe | $F_e(x)$ |
|---|---|---|---|---|---|---|
| 500 | 11 | 11 | 11/50 | 10 | 10 | 10/50 |
| 320 | 15 | 26 | 26/50 | 10 | 20 | 20/50 |
| 1000 | 20 | 46 | 46/50 | 10 | 30 | 30/50 |
| 2000 | 3 | 49 | 49/50 | 10 | 40 | 40/50 |
| 4000 | 1 | 50 | 50/50 | 10 | 50 | 50/50 |

$\therefore D_0 = 16/50 = 0.32$

Critical value: We have, $\alpha = 0.05$ and $n = 50$

$\therefore D_{n,\alpha}$ (Two Tailed Test) $= D_{50, 0.05}$ (hoo)

$= \frac{1.36}{\sqrt{50}}$

$= 0.192$

Decision : Since $D_0 > D_{n,\alpha}$, we reject $H_0$. Hence, the computers of five hard disk are not uniformly infected.

Assignment

Q. A random sample of 20 volume based internet connected have following speed of internet connection in mps:

2.7, 2.9, 3.0, 3.1, 2.8, 3.0, 2.9, 3.0, 2.6, 3.1, 3.2, 3.1, 3.0, 2.9, 3.3, 3.0, 2.8, 2.9, 3.0, 2.9.

Apply Kolmogorov Smirnov test for testing that the internet speed are equally distributed. Use $\alpha =$ ~~0.0~~ 5%.

2.6 1

2.7 1

2.8 2

2.9 5

1/50
6/50
16/50
9/50
0

# Two Samples Test:

[Small Sample Size, $n_1 = n_2 \le 40$ or $n_1 \neq n_2 \le 20$]

[Large Sample size, $n_1 = n_2 > 40$ or $n_1 \neq n_2 > 20$]

<u>Null Hypothesis</u>:

$H_0$: $F(x) = F(y)$ i.e. the two samples are drawn from same population.

$H_1$: $F(x) \neq F(y)$ i.e. the two samples are not drawn from same population. (Two Tailed test)

$$\left.\begin{array}{l} \text{or } H_1: F(x) > F(y) \\ \text{or, } H_1: F(x) < F(y) \end{array}\right\} \text{(One Tailed Test)}$$

<u>Small Sample Size</u>

Test-statistic: Under $H_0$,

$$D_0 = \text{maximum} \left| F(x) - F(y) \right| ; \quad F(x) = \frac{cf_x}{n_1}, \quad F(y) = \frac{cf_y}{n_2}$$

<u>Critical value</u>: The critical values of the test-statistic is obtained from K-S table at $\alpha$ level of significance for sample sizes $n_1$ and $n_2$.

$$D_{tab} = D_{(n_1, n_2), \alpha}$$

Different population

Difference ⇒ No diff ⇒ Same population

Q11. $H_0$: There is no significant difference between two types of motors.

$H_1$: There is significant difference between two types of motors. (Two Tailed Test)

<u>Test - statistic</u> : Under $H_0$,

$$D_0 = \text{Max} \, |F(x) - F(y)|$$

| dif in years | fx | fy | cfx | F(x) | cfy | F(y) | \|F(x) - F(y)\| |
|---|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 0 | 0/9 | 1 | 1/9 | 1/9 |
| 3 | 3 | 2 | 3 | 3/9 | 3 | 3/9 | 0 |
| 4 | 2 | 2 | 5 | 5/9 | 5 | 5/9 | 0 |
| 5 | 2 | 1 | 7 | 7/9 | 6 | 6/9 | 1/9 |
| 6 | 1 | 2 | 8 | 8/9 | 8 | 8/9 | 0 |
| 7 | 1 | 1 | 9 | 9/9 | 9 | 9/9 | 0 |
| | $n_1 = 9$ | $n_2 = 9$ | | | | | |

$$\therefore D_0 = \frac{1}{9}$$

$\therefore$ <u>Critical value</u> : We have, $\alpha = 0.05$, $n_1 = 9$, $n_2 = 9$

$$\therefore D_{(n_1, n_2), \alpha} = D_{(9,9), 0.05} \quad \text{(two-tailed test)}$$

$$= 5/9$$

<u>Decision</u> : Since $D_0 < D_{n_1, n_2, \alpha}$, we don't reject $H_0$. Thus, the two samples are drawn from same population. (Or there is no significant difference between two motors).