U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 106/ MASTAT – 106 Non Parametrics

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur.

Dr. Shruti **Member-Secretary**
Sr. Assistant Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Prof. Shashi Bhushan **Writer**
Department of Statistics
University of Lucknow, Lucknow

Dr. Shruti **Writer**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. S. K. Pandey **Editor**
Department of Statistics
University of Lucknow, Lucknow

Dr. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 106/ MASTAT – 106 **NON PARAMETRICS**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Blocks & Units Introduction

The present SLM on ***Non Parametrics*** consists of eleven units with three blocks.

The ***Block - 1 – Order Statistics,*** is the first block, which is divided into four units.

The ***Unit - 1 – Basic Distribution Theory,*** is the first unit of present self learning material, which describes Order statistics, Distribution of maximum, minimum and r-th order statistic, Joint distribution of r-th and s-th order statistic.

In ***Unit – 2 – Asymptotic Distribution Theory,*** the main emphasis on the Moments of order statistics, asymptotic distributions of an order statistic, asymptotic relative efficiency, non parametric estimation of distribution function, Glivenko-Cantelli fundamental theorem

In ***Unit – 3 – Distribution Free Intervals,*** we have focussed mainly on Distribution of range function of order statistics, distribution free confidence intervals for quintiles, distribution free tolerance interval, distribution free bounds for moments, Fooleries limits

In ***Unit – 4 – Rank order Statistics,*** is being discussed the Rank order statistics, Dwass' technique, Ballot theorem its generalization, extension and application to fluctuations of sums of random variables.

The ***Block - 2 – Sequential Analysis*** is the second block with two units.

In ***Unit – 5 – Sequential Tests*** is discussed with SPRT and its properties, Wald's Fundamental identity, OC and ASN functions, Wald's equation, Wolfowitz generalization of FRC bound, Stein's two stage procedure.

In ***Unit – 6 – Sequential Estimation*** has been discussed, Asymptotic theory of sequential estimation, sequential estimation of normal mean.

The ***Block - 3 – Nonparametric Tests and Inference*** has three units.

Unit – 7 – One- sample Location Tests dealt with One and two sample location tests, Sign test. Wilcoxon test, Median test.

Unit – 8 – Other non- parametric tests dealt with Mann- Whitney U- Test, Application of U-statistic to rank tests. One sample and two sample Kolmagorov-Smirnov tests. Run tests.

Unit – 9 – Nonparametric Inference, The Kruskal-Wallis one way ANOVA Test, Friedman's two-way analysis of variance by ranks, efficiency criteria and theoretical basis for calculating ARE, Pitman ARE.

At the end of every block/unit the summary, self assessment questions and further readings are given.

Block 1- Order Statistics

Unit 1: Basic Distribution Theory

Structure

Unit 2: Distribution Free Intervals

Structure

UNIT 1: BASIC DISTRIBUTION THEORY

1.1 ORDER STATISTICS

Definition: The observation occupying r^{th} place in ascending order of the sample values is known as the r^{th} order statistic. We denote it by Y_r or $X_{(r)}$ so that Y_1= $X_{(1)}$ represents the minimum of the sample observations while $Y_n = X_{(n)}$ is the maximum of sample observations.

The definition of order statistics does not require that the *X's* to be identically distributed, nor do we need them to be independent. Also, it was not presumed that the parent distributions be continuous, nor that their densities exist. Although, most of the classical results dealing with order statistics were originally derived in more restrictive settings. Generally, it is assumed that the X's were independent and identically distributed (i.i.d.) with common continuous (cumulative) distribution function *F(x)*, and having a density function *f(x)* and, henceforth, we will assume the *X's* to be so.

The following list, though, not exhaustive, but may serve help to convince the reader that this text will not be focusing on some abstract concepts of little *practical utility*:

1. ***Robust Location Estimates.*** Suppose that *n* independent measurements are available, and we wish to estimate their assumed common mean. It has long been recognized that the sample mean, though attractive from many viewpoints, suffers from an extreme sensitivity to outliers and model violations.

Estimates based on the median or the average of central order statistics are less sensitive to model assumptions.

2. Detection of Outliers. If one is confronted with a set of measurements and is concerned with determining whether some have been incorrectly made or reported, attention naturally focuses on certain order statistics of the sample. Usually the largest one or two and/or the smallest one or two are deemed most likely to be outliers.

3. *Censored Sampling.* Fifty expensive machines are started up in an experiment to determine the expected life of a machine. If, as is to be hoped, they are fairly reliable, it would take an enormously long time to wait for all machines to fail. Instead, great savings in time and machines can be effected if we base our estimates on the first few failure times (i.e., the first few order statistics from the conceptual sample of i.i.d. failure times).

4. *Waiting for the Big One.* Disastrous floods and destructive earthquakes recur throughout history. Dam construction has long focused on so called 100-year floods. Presumably the dams are built big enough and strong enough to handle any water flow to be encountered except for a level expected to occur only once every 100 years. Whether one agrees or not with the 100-year disaster philosophy, such inferences are concerned with the distribution of large order statistics from a possibly dependent, possibly not identically distributed sequence.

5. *Strength of Materials.* The adage that a chain is no stronger than its weakest link underlies much of the theory of strength of materials, whether they be threads, sheets, or blocks. By considering failure potential in infinitesimally small sections of the material, one quickly is led to strength distributions associated with limits of distributions of sample minima, which is again an order statistic.

6. *Reliability.* The example of a cord composed of *n* threads can be extended to lead us to reliability applications of order statistics. It may be that failure of one thread will cause the cord to break (the weakest link), but more likely the cord will function as long as *k* (a number less than *n)* of the threads remain unbroken.

7. *Quality Control.* Each candy bar should weigh 2.1 ounces; just a smidgen over the weight stated on the wrapper. No matter how well the candy pouring machine was adjusted at the beginning of the shift, minor fluctuations will occur, and potentially major aberrations might be encountered (if a peanut gets stuck in the control valve). We must be alert for correctable malfunctions causing unreasonable variation in the candy bar weight. Enter the quality control man with his *X and R charts* or his *median and R charts.* If the median (or perhaps the mean) is far from the target value, we must shut down the line.

8. *Selecting the Best.* Field trials of corn varieties involved carefully balanced experiments to determine which of several varieties is most productive. Obviously we are concerned with the maximum of a set of probably not identically distributed variables in such a setting.

9. *Inequality Measurement.* The income distribution in India (where a few individuals earn most of the money) is clearly more unequal than that of

United Kingdom (where progressive taxation has a leveling effect). How does one make such statements precise? The usual approach involves order statistics of the corresponding income distributions. The particular device used is called a *Lorenz curve.* It summarizes the percent of total income accruing to the poorest *p* percent of the population for various values of *p.* Mathematically this is just the scaled integral of the empirical quantile function. A high degree of convexity in the Lorenz curve signals a high degree of inequality in the income distribution.

10. *Olympic Records.* Bob Beamon's 1968 long jump remains on the *Olympic* record book. Few other records last that long. If the best performances in each Olympic Games were modeled as independent identically distributed random variables, then records would become more and more scarce as time went by. Such is not the case. The simplest explanation involves improving and increasing populations. Thus the 1964 high jumping champion was the best of, say, *Nx* active international-caliber jumpers. In 1968 there were more high-caliber jumpers of probably higher caliber. So we are looking, most likely, at a sequence of not identically distributed random variables. But in any case we are focusing on maxima, that is, on certain order statistics.

11. *Allocation of Prize Money.* At the end of the annual Bob Hope golf tournament the player with the lowest score gets first prize. The second lowest score gets second prize, etc. In 1991 the first five prizes were: $198,000, $118,800, $74,800, $52,800, and $44,000. Obviously we are dealing with order statistics here. Presumably the player with the highest ability level will most likely post the lowest score.

12. *Characterizations and Goodness of Fit.* The exponential distribution is famous for its so-called lack of memory. The usual model involves a light bulb or other electronic device. For example, if $X_1,\dots,X_n$ are i.i.d. exponential, then their spacings $(X_{(i)} - X_{(j)})$ are again exponential and, remarkably, are independent. It is only in the case of exponential random variables that such spacings properties are encountered. A vast literature of exponential characterizations and related goodness-of-fit tests has consequently developed. It is interesting to note that most tests of goodness of fit for any parent distribution implicitly involve order statistics, since they often focus on deviations between the empirical quantile function and the hypothesized quantile function.

Let $X_1,\ X_2,\dots,\ X_n$ be a random sample of size n taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a<x<b$. Let Y_1 be the minimum of $X_1,\ X_2,\dots.,\ X_n$ is called the first order statistics, Y_2 be the next minimum is called the second order statistics and so on so that Y_n be the maximum of $X_1,\ X_2,\dots.,\ X_n$ is called the n^{th} order statistics. Then $Y_1<Y_2<\dots.<Y_n$ is known as order statistics of the random sample $X_1,\ X_2,\dots,\ X_n$.

1.2 DISTRIBUTION OF MAXIMUM

Let X_1, X_2,..., X_n be a random sample of size *n* taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a<x<b$. Let $F_{Y_n}(x)$ be the c .d. f. of maximum or the *n*-th order statistics Y_n at the point x is given by

$$F_{Y_n}(x)=P[Y_n \le x]$$

$$=\mathrm{P}[Max\ of\ X_1, X_2,..., X_n \le x]$$

$$=\mathrm{P}[X_1 \le x, X_2 \le x,..., X_n \le x]$$

$$=\mathrm{P}[X_1 \le x]P[X_2 \le x]....P[X_n \le x] \qquad (X_1, X_2,..., X_n \text{ are mutually independent})$$

$$=F_{X_1}(x)F_{X_2}(x)...F_{X_n}(x) \tag{1}$$

where $F_{X_i}(x)$ is c. d. f. of X_i for *i*= 1, 2,..., *n* and since $X_1, X_2,..., X_n$ are identically distributed with c. d. f. $F(x)$ such that

$$F_{X_1}(x)=F_{X_2}(x)=...=F_{X_n}(x)=F(x) \tag{2}$$

Using (2) in (1), we get

$$F_{Y_n}(x)=\{F(x)\}^n \tag{3}$$

Since in case continuous population, the density function is given by

p. d. f. of Y_n = $f_{Y_n}(x)$

$$=\frac{dF_{Y_n}(x)}{dx}$$

$$=n\{F(x)\}^{n-1}\frac{dF(x)}{dx} \qquad \left[\because f(x)=\frac{dF(x)}{dx}\right]$$

$$= \begin{cases} n\left[F(x)\right]^{n-1} f(x) & ; a < x < b \\ 0 & ; \text{otherwise} \end{cases} \tag{4}$$

Example: Find out the p.d.f. of the maximum of the sample values from a sample of size n drawn from $U(0,\theta)$ parent.

Solution: Consider the p.d.f. of $U(0,\theta)$ given by

$$f(x) = \begin{cases} \dfrac{1}{\theta} & ; 0 < x < \theta \\ 0 & ; otherwise \end{cases}$$

so that its d.f. is given by

$$F(x) = \begin{cases} 0 & ; x \leq 0 \\ \dfrac{x}{\theta} & ; 0 < x \leq \theta \\ 1 & ; x \geq \theta \end{cases}$$

Let $X_{(n)}$ be the n^{th} order statistics or maximum of sample values in a sample of size *n*. Then

$$f_{X(n)}(x) = n\left\{F(x)\right\}^{n-1} f(x) \quad ; 0 < x < \theta$$

$$= n\left(\frac{x}{\theta}\right)^{n-1} \frac{1}{\theta} \qquad ; 0 < x < \theta$$

$$= n\,\frac{x^{n-1}}{\theta^{n}} \qquad ; 0 < x < \theta$$

Hence, the p.d.f. of $X_{(n)}$ is given by

$$f_{X_{(n)}}(x) = \begin{cases} \dfrac{nx^{n-1}}{\theta^{n}} & ; 0 < x < \theta \\ 0 & ; otherwise \end{cases}$$

PROBABILITY INTEGRAL TRANSFORM

If X is a random variable of a continuous type having p.d.f. $f(x)$ *and distribution function* $F(x)$*, then* $Z = F(x)$ *has a uniform distribution* $U(0,1)$.

Proof: Given X is a continuous random variable with p. d. f. $f(x)$ and c. d. f. $F(x)$ then we wish to prove that $Z = F(x)$ follows $U(0,1)$. Consider the p.d.f. of *z* given by

$$f(z) = (\operatorname{mod} J)(\text{put } x \text{ in terms of } z \text{ in } f(x))$$

where J stands for the jacobian of transformation. For particular (or specific) values of *z* and *x*, we may write

$$z = F(x)$$

$$J = \left|\frac{dx}{dz}\right| = \left|\frac{1}{dz/dx}\right| = \left|\frac{1}{dF(x)/dx}\right| = \frac{1}{f(x)} \qquad (\because f(x) \geq 0)$$

so that

$$f(z) = \begin{cases} 1 & 0 < z < 1 \\ 0 & otherwise \end{cases}$$

$$\Rightarrow Z = F(x) \sim U(0,1) \qquad \text{Q.E.D.}$$

Remark: The importance of probability integral transform is that the order statistics $X_{(1)},\ldots,X_{(n)}$ in a sample from any continuous distribution with c.d.f. $F(x)$ are transformed by order preserving probability integral transform $u = F(x)$ into $U_{(1)},\ldots,U_{(n)}$.

Example: If X is a uniform random variable with distribution function F(x), prove that

$$E\left[X_{(r)}\right]=\frac{n!}{(r-1)!(n-r)!}\int_0^1 Y^{r-1}(1-Y)^{n-r}\,h(Y)\,dY$$

where $h(Y)=F^{-1}(Y)$

Solution: Let $X\sim U(a,b)$ and consider

$$E\left[X_{(r)}\right]=\int_a^b x\,f_{X_{(r)}}(x)\,dx \qquad ;\ a<x<b$$

$$=\int_a^b x\frac{n!}{(r-1)!(n-r)!}\{F(x)\}^{r-1}\{1-F(x)\}^{n-r}\,f(x)\,dx$$

Let $F(x)=y$

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

so that

$$E\left[X_{(r)}\right]=\int_0^1 F^{-1}(y)\frac{n!}{(r-1)!(n-r)!}y^{r-1}(1-y)^{n-r}\,dy$$

$$=\frac{n!}{(r-1)!(n-r)!}\int_0^1 y^{r-1}(1-y)^{n-r}\,h(y)\,dy \qquad \left[\because F^{-1}(y)=h(y)\right] \qquad \text{Q.E.D.}$$

Example: Let $x_{1,}x_2,x_3$ *be independent random variable with p. d. f.*

$$f(x)=\exp\left[-(x-\theta)\right]I_{(\theta,\infty)}(x)$$

Determine the constant $c=c(\theta)$ *for which* $P\left(\theta<x_{(3)}<c\right)=0.96$

Solution: Since

$$f(x) = \mathrm{e}^{[-(x-\theta)]} I_{(\theta,\infty)}(x)$$

where $I_{(\theta,\infty)}(x) = \begin{cases} 1 & if\ \theta \le x \le \infty \\ 0 & otherwise \end{cases}$

In other words

$$f(x) = \begin{cases} e^{-(x-\theta)} & ;\theta \le x \le \infty \\ 0 & ;otherwise \end{cases}$$

$$F(x) = \Pr[X \le x] = \int_{-\infty}^{x} f(x)dx$$

$$= \int_{-\infty}^{\theta} f(x)dx + \int_{\theta}^{x} f(x)dx$$

$$= 0 + \int_{\theta}^{x} e^{-(x-\theta)}\, dx$$

$$= \left[-e^{-(x-\theta)}\right]_{\theta}^{x}$$

$$= \left[1 - e^{-(x-\theta)}\right]$$

Therefore, $F(x) = \begin{cases} 1 - e^{-(x-\theta)} & ;\theta \le x \le \infty \\ 0 & ;otherwise \end{cases}$

We may write $P\left(\theta < x_{(3)} < c\right) = 0.96$ as

$$\int_{\theta}^{c} f_{x_{(3)}}(x)dx = 0.96$$

where $f_{x_{(3)}}(x)$ stands for the p. d. f. of third order statistic where *n*=3, so that

$$\int_{\theta}^{c} 3\{F(x)\}^2 f(x)dx = 0.96$$

Let $F(x)=t \Rightarrow f(x)dx=dt$ so that

$$\int_{F(\theta)}^{F(c)} 3t^2dt=0.96$$

$$\Rightarrow \int_{0}^{F(c)} 3t^2dt=0.96$$

where $F(\theta)=0$ and $F(c)=1-e^{-(c-\theta)}$ so we get

$$t^3\Big|_0^{F(c)}=0.96$$

$$\left[F(c)\right]^3=0.96$$

$$1-e^{-(c-\theta)}=(0.96)^{1/3}$$

$$1-(0.96)^{1/3}=e^{-(c-\theta)}$$

$$c=\theta-\ln\left[1-(0.96)^{1/3}\right]$$

is the required value of c, such that $\Pr\left[\theta<x_{(3)}<c\right]=0.96$.

1.3 DISTRIBUTION OF MINIMUM

Let X_1, X_2,..., X_n be a random sample of size n taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a<x<b$. Let $F_{Y_1}(x)$ be the c .d. f. of minimum or the first order statistics Y_1 at the point x is given by

$$F_{Y_1}(x)=\mathrm{P}[Y_1\le x]$$

$$=1-\mathrm{P}[Y_1>x]$$

$$=1-\mathrm{P}\left[\min(X_1,X_2,...,X_n)>x\right]$$

$$= 1 - \mathrm{P}\left[X_1 > x, X_2 > x, \ldots, X_n > x\right]$$

$$= 1 - P\left[X_1 > x\right]P\left[X_2 > x\right]\ldots P\left[X_n > x\right] \quad (X_1, X_2, \ldots., X_n \text{ are mutually ind.})$$

$$= 1 - \left\{1 - P\left[X_1 \le x\right]\right\}\left\{1 - P\left[X_2 \le x\right]\right\}\ldots\left\{1 - P\left[X_n \le x\right]\right\}$$

$$= 1 - \left\{1 - F_{X_1}(x)\right\}\left\{1 - F_{X_2}(x)\right\}\ldots\left\{1 - F_{X_n}(x)\right\}$$

where $F_{X_i}(x)$ is c. d. f. of X_i for i= 1, 2,..., n and since $X_1, X_2, \ldots, X_n$ are identically distributed with c. d. f. $F(x)$ such that

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F(x) \tag{2}$$

Using (2) in (1), we get

$$F_{Y_1}(x) = 1 - \left\{1 - F(x)\right\}^n \qquad ; a < x < b$$

Since in case continuous population, the density function is given by

p. d. f. of Y_1 = $f_{Y_1}(x)$

$$= \frac{dF_{Y_1}(x)}{dx}$$

$$= n\left\{1 - F(x)\right\}^{n-1}\frac{dF(x)}{dx} \qquad \left[\because f(x) = \frac{dF(x)}{dx}\right]$$

$$= \begin{cases} n\left\{1 - F(x)\right\}^{n-1} f(x) & ; a < x < b \\ 0 & ; \text{otherwise} \end{cases} \tag{4}$$

Example: Let $X_j\left(j = 1, 2, \ldots, n\right)$ be i.i.d. negative exponential random variable with parameter λ then show that the distribution of $X_{(1)}$ is a negative exponential distribution with parameter $n\lambda$. Conversely, show that if $X_j\left(j = 1, 2, \ldots, n\right)$ are i.i.d.

random variables and $X_{(1)}$ *follows a negative exponential distribution with parameter* $n\lambda$ *then the common distribution of X's is negative exponential with parameter* λ.

Solution:

It is given as $f(x)=\begin{cases}\lambda e^{-\lambda x} & ;0<x<\infty\\ 0 & ;elsewhere\end{cases}$

Therefore $F(x)=\begin{cases}1-e^{-\lambda x} & ;0<x\le\infty\\ 0 & ;otherwise\end{cases}$

Let $F_{X_{(1)}}(x)$ be the c. d. f. of $X_{(1)}$ and since, $X_1, X_2, ..., X_n$ are identically distributed with c. d. f. $F(x)$. Therefore,

$$F_{X_{(1)}}(x)=F_{X_{(2)}}(x)=....=F_{X_{(n)}}(x)=F(x)$$

Hence,

$$F_{X_{(1)}}(x)=1-\{1-F(x)\}^n$$

$$=1-\{1-1+e^{-\lambda x}\}^n$$

$$=1-e^{-\lambda xn}$$

therefore,

$$f_{X_{(1)}}(x)=\frac{dF_{X_{(1)}}(x)}{dx}$$

$$=\frac{d}{dx}\left(1-e^{-\lambda xn}\right)$$

$$=n\lambda e^{-\lambda xn}$$

$X_{(1)}$ follows negative exponential with parameter $n\lambda$.

Conversely suppose it is given that $X_{(1)}$ follows negative exponential with parameter $n\lambda$ so that

$$F_{X_{(1)}}(x) = 1 - e^{-\lambda xn}$$

Also, $F_{X_{(1)}}(x) = 1 - \{1 - F(x)\}^n$

Equating both we have,

$$1 - \{1 - F(x)\}^n = 1 - e^{-n\lambda x}$$

$$1 - F(x) = e^{-x\lambda}$$

$$F(x) = 1 - e^{-\lambda x}$$

$$\Rightarrow f(x) = \frac{d}{dx}F(x)$$

$$= \frac{d}{dx}\left(1 - e^{-\lambda x}\right)$$

$$= \lambda e^{-\lambda x}$$

which is negative exponential with parameter λ . Hence, the common distribution of X's is negative exponential with parameter λ.

1.4 DISTRIBUTION OF r-th ORDER STATISTIC

Let X_1, X_2,..., X_n be a random sample of size n taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a < x < b$. Let $F_{Y_r}(x)$ be the c. d. f. and $f_{Y_r}(x)$ be the p. d. f. of the r^{th} order statistic Y_r at the point x is given by

$$f_{Y_r}(x) = \frac{dF_{Y_r}(x)}{dx}$$

$$=\lim_{h\to 0}\frac{F\left(x+h\right)-F\left(x\right)}{h}$$

$$=\lim_{h\to 0}\frac{\Pr\left(Y_r\le x+h\right)-\Pr\left(Y_r\le x\right)}{h}$$

$$=\lim_{h\to 0}\frac{1}{h}\mathrm{P}\left(x\le Y_r\le x+h\right)$$

$$\overset{\downarrow}{a}\underbrace{\overline{\qquad\qquad\qquad\qquad\qquad\qquad}}_{r-1\;X\text{'s}<x}\overset{\downarrow}{x}\underbrace{\overline{\qquad\quad}}_{\text{one }X<x+h}\overset{\downarrow}{x+h}\underbrace{\overline{\qquad\qquad\qquad\qquad\qquad}}_{n-r\;\;X\text{'s}>x+h}\overset{\downarrow}{b}$$

$$=\lim_{h\to 0}\frac{1}{h}\mathrm{P}\left[r-1\text{of the }X\text{'}s\le x,\ \text{one}\,X\text{ in}\left(x,x+h\right],\ \ n-r\text{ of the }X\text{'}s>x+h\right]$$

Using multinomial law, we have

$$f_{Y_r}\left(x\right)=\lim_{h\to 0}\frac{1}{h}\frac{n!}{\left(r-1\right)!\left(n-r\right)!}\left\{\mathrm{P}\left(X\le x\right)\right\}^{r-1}\mathrm{P}\left(x\le X\le x+h\right)\left\{\mathrm{P}\left(X>x+h\right)\right\}^{n-r}$$

$$=\lim_{h\to 0}\frac{1}{h}\frac{n!}{\left(r-1\right)!\left(n-r\right)!}\left\{F\left(x\right)\right\}^{r-1}\left\{F\left(x+h\right)-F\left(x\right)\right\}\left\{1-F\left(x+h\right)\right\}^{n-r}$$

$$=\lim_{h\to 0}\frac{n!}{\left(r-1\right)!\left(n-r\right)!}\left\{F\left(x\right)\right\}^{r-1}\left\{\frac{F\left(x+h\right)-F\left(x\right)}{h}\right\}\left\{1-F\left(x\right)\right\}^{n-r}$$

$$=\begin{cases}\dfrac{n!}{\left(r-1\right)!\left(n-r\right)!}\left\{F\left(x\right)\right\}^{r-1}\left\{1-F\left(x\right)\right\}^{n-r}\ f\left(x\right) & ;a<x<b\\ 0 & ;\text{otherwise}\end{cases}$$

Remark: It is interesting to note that the $F_{Y_1}\left(x\right)=1-\left\{1-F\left(x\right)\right\}^n$ and $F_{Y_n}\left(x\right)=\left\{F\left(x\right)\right\}^n$ are special cases of the general result of $F_{Y_r}\left(x\right)$ given by

$$F_{Y_r}\left(x\right)=P\left(Y_r\le x\right)$$

$$=P\left\{\text{at least }r\text{ of the }X_i\text{ are less than or equal to }x\right\}$$

$$=\sum_{i=r}^{n}\binom{n}{i}F^i\left(x\right)\left[1-F\left(x\right)\right]^{n-i}$$

since the summand is the binomial probability of getting exactly i of the $X_1,\ldots,X_n$ less than or equal to *x*. Also, one more useful relationship that exists between the binomial sums and incomplete beta functions is

$$F_{Y_r}(x) = I_{F(x)}(r, n-r+1)$$

where $I_p(a,b)$ is an incomplete beta function defined as

$$I_p(a,b) = \int_0^p y^{a-1}(1-y)^{b-1}\,dy \quad ; a>0, b>0$$

Therefore, $F_{Y_r}(x)$ can be calculated from the tables of $I_p(a,b)$ and the percentage points of Y_r can be obtained by inverse interpolation of these tables.

Example: Obtain the upper 5% point of Y_4 *in sample of 5 from standard normal distribution.*

Solution: We need to find *x* such that

$$\begin{aligned} F_{Y_4}(x) &= 0.95 \\ &= I_{F(x)}(4, 5-4+1) = I_{F(x)}(4,2) \end{aligned}$$

so that

$$0.05 = I_{1-F(x)}(2,4)$$

thereby giving

$$0.0764 = 1 - F(x)$$

Hence, from normal tables, we have *x* = 1.43.

Example: Let $Y_1 < Y_2 < Y_3 < Y_4$ *denote the order statistics of the random sample of size 4 from the population with p. d. f.*

$$f(x)=\begin{cases}2x & ;0\le x\le 1\\ 0 & ;otherwise\end{cases}$$

Obtain the p. d. f. of Y_3 *and* $P(1/2<Y_3)$

Solution: It is given that n=4 and

$$f(x)=\begin{cases}2x & ;0\le x\le 1\\ 0 & ;otherwise\end{cases}$$

Hence, for $x<0$

$$F(x)=\Pr[X\le x]=\int_{-\infty}^{x}0\,dx=0$$

For $x\le 1$

$$F(x)=\Pr[X\le x]=\int_{-\infty}^{0}f(x)\,dx+\int_{0}^{x}f(x)\,dx=x^2$$

For $x>1$

$$F(x)=\Pr[X\le x]=\int_{-\infty}^{0}f(x)\,dx+\int_{0}^{1}f(x)\,dx+\int_{1}^{x}f(x)\,dx=1$$

Therefore, $F(x)=\begin{cases}0 & ;x<0\\ x^2 & ;0\le x\le 1\\ 1 & ;x\ge 1\end{cases}$

Putting r=3 and n=4 in (1) we have,

p. d. f. of Y_3= $f(Y_3)$

$$=\frac{4!}{2!\,1!}\{F(y_3)\}^2\{1-F(y_3)\}\,f(y_3) \qquad ;\ 0\le y_3\le 1$$

$$=12\left[y_3^2\right]^2\left[1-y_3^2\right]2y_3 \qquad ;0\le y_3\le 1$$

$$=24\left[y_3^5\right]\left[1-y_3^2\right] \qquad ;0\le y_3\le 1$$

so that

$$f(Y_3)=\begin{cases} 24y_3^5\left(1-y_3^2\right) & ;0\le y_3\le 1 \\ 0 & ;otherwise \end{cases}$$

Now, $\Pr\left(Y_3>1/2\right)=1-\Pr\left(Y_3\le 1/2\right)$

$$=1-\int_0^{1/2} f\left(y_3\right)\,dy_3$$

$$=1-24\int_0^{1/2}\left(1-y_3^2\right)y_3^5\,dy_3$$

Let $y_3^2=t\ \ \therefore 2y_3\,dy_3=dt$

$$\Pr\left(Y_3>1/2\right)=1-12\int_0^{1/4}t^2\left(1-t^2\right)dt$$

$$=1-\left[4t^3-3t^4\right]_0^{1/4}=1-\frac{1}{64}\left[4-\frac{3}{4}\right]=1-\frac{13}{256}$$

$$=\frac{243}{256}$$

Example: Let $f\left(x,\theta\right)=1/\theta\ ,0<x<\theta$ *and* x_1,x_2,x_3 *be a random sample of size 3 from this parent distribution and let* Y_1,Y_2,Y_3 *be order statistic of this sample so that* $Y_1=\min\left(x_1,x_2,x_3\right)$ *and* $Y_3=\max\left(x_1,x_2,x_3\right)$. *Obtain* $\Pr\left(Y_2\ge\theta/2\right)$.

Solution: Given

$$f\left(x\right)=f\left(x,\theta\right)=1/\theta\ ,0<x<\theta$$

Now for $x<\theta$

$$F\left(x\right)=P\left[X\le x\right]=\int_{-\infty}^{x}f\left(x\right)dx=0$$

$$= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{x} f(x)\,dx = 0 + \int_{0}^{x} \frac{1}{\theta}\,dx$$

$$= \frac{x}{\theta}$$

For $x > \theta$

$$F(x) = P[X \le x] = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\theta} f(x)\,dx = \int_{\theta}^{x} f(x)\,dx = 1$$

Therefore, $F(x) = \begin{cases} 0 & ; x \le 0 \\ x/\theta & ; 0 < x \le \theta \\ 1 & ; x \ge \theta \end{cases}$

Thus, $\Pr(Y_2 \ge \theta/2) = \int_{\theta/2}^{\theta} f(w)\,dw$

where $f(w)$ is the p. d. f. of second order statistic for n=3. So

$$f(w) = \frac{3!}{(2-1)!(3-2)!}\{F(w)\}^{2-1}\{1-F(w)\}^{3-2} f(w)$$

$$= 6\{F(w)\}\{1-F(w)\} f(w) \qquad ; 0 < w < \theta$$

So, $\Pr(Y_2 \ge \theta/2) = \int_{\theta/2}^{\theta} 6F(w)\{1-F(w)\} f(w)\,dw$

Let $F(w) = t \Rightarrow f(w)dw = dt$ and $F(\theta/2) = 1/2, F(\theta) = 1$

so that

$$\Pr(Y_2 \ge \theta/2) = \int_{1/2}^{1} 6t(1-t)dt$$

$$= 6\left[\frac{t^2}{2} - \frac{t^3}{3}\right]_{1/2}^{1} = \left[3t^2 - 2t^3\right]_{1/2}^{1} = 3 - 2 - \frac{3}{4} + \frac{1}{4}$$

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

Example: Let $Y_1 < Y_2 < Y_3 < Y_4$ *be the order statistics of the random sample of size 4 from the distribution having probability density function*

$$f(x)=\begin{cases} e^{-x} & ;0<x<\infty \\ 0 & ;otherwise \end{cases}$$

Find $\Pr(3 \le Y_4)$.

Solution: Given that

$$f(x)=\begin{cases} e^{-x} & ;0<x<\infty \\ 0 & ;otherwise \end{cases}$$

Now for $x \le 0$

$$F(x)=\Pr[X \le x]=\int_{\infty}^{0} f(x)\,dx+\int_{0}^{x} f(x)\,dx$$

$$=0+\int_{0}^{x} e^{-x}dx$$

$$=1-e^{-x}$$

Therefore,

$$\Pr[Y_4 \ge 3]=\Pr[3 \le 4]$$

$$=\int_{3}^{\infty} f(w)\,dw$$

where $f(w)$ is the p. d. f. of 4th order statistic Y_4 for n=4

$$\because f(w)=4\left[F(w)\right]^{4-1} f(w)\,dw \qquad ;0<w\le\infty$$

so that

$$\Pr\left[Y_4 \geq 3\right] = \int_3^{\infty} 4\left[F(w)\right]^3 f(w)\,dw$$

Putting $F(w) = t \ \therefore f(w)\,dw = dt$, we have

$$\Pr\left[Y_4 \geq 3\right] = \int_{F(3)}^{F(\infty)} 4[t]^3 dt = \left[t^4\right]_{F(3)}^{F(\infty)} \qquad \text{since } \left[F(3)\right] = 1 - e^{-3} \text{ and } \left[F(\infty)\right] = 1$$

$$= 1 - \left(1 - e^{-3}\right)^4$$

1.5 JOINT DISTRIBUTION OF r-th AND s-th ORDER STATISTIC

Let X_1, X_2,..., X_n be a random sample of size n taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a < x < b$. Let $F_{rs}(x,y)$ be the joint c. d. f. and $f_{rs}(x,y)$ be the joint p. d. f. of the r^{th} and s^{th} $(r < s)$ order statistic Y_r and Y_s at the point (x, y), *x<y,* is given by

$$f_{rs}(x,y) = \frac{d^2 F_{rs}(x,y)}{dxdy}$$

$$= \lim_{\substack{h \to 0 \\ k \to 0}} \frac{\mathrm{P}\left(x < Y_r \leq x+h,\ y < Y_s \leq y+k\right)}{hk}$$

$$= \lim_{\substack{h \to 0 \\ k \to 0}} \frac{1}{hk}\mathrm{P}\begin{bmatrix} r-1 \text{of the } X\text{'s} \leq x,\ \text{one } X \text{ in}\left(x, x+h\right], s-r-1 \text{ of the } X\text{'s in}\left(x+h, y\right], \\ \text{one } X \text{ in}\left(y, y+k\right],\ n-s \text{ of the } X\text{'s} > y+k \end{bmatrix}$$

$$\overset{\downarrow}{a}\underbrace{\qquad\quad}_{r-1\ X\text{'s} <x}\overset{\downarrow}{x}\underbrace{\qquad\quad}_{\text{one } X<x+h}\overset{\downarrow}{x+h}\underbrace{\qquad\quad}_{s-r-1\ X\text{'s} <y}\overset{\downarrow}{y}\underbrace{\qquad\quad}_{\text{one } X<y+k}\overset{\downarrow}{y+k}\underbrace{\qquad\quad}_{n-s\ X\text{'s} >y+k}\overset{\downarrow}{b}$$

Using multinomial law, we have

$$= \lim_{\substack{h \to 0 \\ k \to 0}} \frac{1}{hk} \frac{n!}{(r-1)!1!(s-r-1)!1!(n-s)!} \begin{bmatrix} \left\{\mathrm{P}\left(X \leq x\right)\right\}^{r-1} \mathrm{P}\left(x \leq X \leq x+h\right) \left\{\mathrm{P}\left(x+h \leq X \leq y\right)\right\}^{s-r-1} \\ \mathrm{P}\left(y \leq X \leq y+k\right) \left\{\mathrm{P}\left(X > y+k\right)\right\}^{n-s} \end{bmatrix}$$

$$=\lim_{\substack{h\to 0\\ k\to 0}}\frac{1}{hk}\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\begin{bmatrix}\{F(x)\}^{r-1}\{F(x+h)-F(x)\}\{F(y)-F(x+h)\}^{s-r-1}\\ \{F(y+k)-F(y)\}\{1-F(y+k)\}^{n-s}\end{bmatrix}$$

$$=\lim_{\substack{h\to 0\\ k\to 0}}\frac{1}{hk}\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\{F(x)\}^{r-1}\lim_{h\to 0}\{F(y)-F(x+h)\}^{s-r-1}$$

$$\lim_{k\to 0}\{1-F(y+k)\}^{n-s}\lim_{k\to 0}\frac{1}{k}\{F(y+k)-F(y)\}\lim_{h\to 0}\frac{1}{h}\{F(x+h)-F(x)\}$$

$$=\begin{cases}\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\{F(x)\}^{r-1}\{F(y)-F(x)\}^{s-r-1}\{1-F(y)\}^{n-s}\ f(x)f(y);a<x<y<b\\ 0 \qquad ;\text{otherwise}\end{cases}$$

Example: Let $Y_1<Y_2<Y_3<Y_4$ be the order statistics of a random sample of size 4 from the probability distribution function

$$f(x)=\begin{cases}e^{-x} & ;0<x<\infty\\ 0 & ;otherwise\end{cases}$$

Show that Y_2 and Y_4-Y_2 are stochastically independent.

Solution: For $x\le\infty$

$$F(x)=\Pr(X\le x)=\int_{-\infty}^{x}f(x)dx$$

$$=\int_{-\infty}^{0}f(x)dx+\int_{0}^{x}f(x)dx$$

$$=0+\left[-e^{-x}\right]_0^x$$

$$=1-e^{-x}$$

Hence, $F(x)=\begin{cases}1-e^{-x} & ;0<x\le\infty\\ 0 & ;otherwise\end{cases}$

Let $Z_1=Y_2$ and $Z_2=Y_4-Y_2$

Then, the joint p. d. f. of Y_2 and Y_4 is given by

$$g_{24}\left(y_2,y_4\right)=\frac{5!}{(5-9)!(2-1)!(9-2-1)!}\left\{F\left(y_2\right)\right\}^{2-1}\left\{F\left(y_4\right)-F\left(y_2\right)\right\}^1\left\{1-F\left(y_4\right)\right\}^1 f\left(y_2\right)f\left(y_4\right)$$

$$;0<y_2<y_4<\infty$$

$$=120\left\{1-e^{-y_2}\right\}\left\{1-e^{-y_4}-1+e^{-y_2}\right\}\left\{1-1+e^{-y_4}\right\}\left\{e^{-y_2}\right\}\left\{e^{-y_4}\right\}$$

$$=120\left\{e^{-y_2}\right\}\left\{1-e^{-y_2}\right\}\left\{e^{-y_2}-e^{-y_4}\right\}\left\{e^{-2y_4}\right\} \qquad ;0<y_2<y_4<\infty \qquad (1)$$

For specific values we may write

$z_1=y_2$ and $z_2=y_4-y_2$

i.e. $y_2=z_1$ and $y_4=z_2+z_1$

The jacobian of transformation is

$$J=\frac{\partial\left(y_2,y_4\right)}{\partial\left(z_1,z_2\right)}=\begin{vmatrix}1 & 0\\ 1 & 1\end{vmatrix}=1$$

Hence the joint p. d. f. of z_1 and z_2 is

$f\left(z_1,z_2\right)=$ (mod J)(put y_2 and y_4 in terms of z_1 and z_2 in $g_{24}\left(y_2,y_4\right)$)

$$=120e^{-z_1}\left(1-e^{-z_1}\right)\left(e^{-z_1}-e^{-z_1-z_2}\right)e^{-2\left(z_1+z_2\right)} \qquad 0<z_1<\infty,0<z_2<\infty$$

$$=120e^{-4z_1}\left(1-e^{-z_1}\right)\left(1-e^{-z_2}\right)e^{-2z_2}$$

Now,

$$f\left(z_1\right)=\int_0^\infty f\left(z_1,z_2\right)dz_2$$

$$=120e^{-4z_1}\left(1-e^{-z_1}\right)\int_0^\infty\left(1-e^{-z_2}\right)e^{-2z_2}\,dz_2 \qquad ;\ 0<z_1<\infty$$

Let $e^{-z_2} = t$ and $e^{-z_2} dz_2 = dt$ so that

$$f\left(z_1\right) = 120\, e^{-4z_1}\left(1 - e^{-z_1}\right)\int_1^0 t\left(1-t\right)\left(-dt\right) \qquad ;\ 0 < z_1 < \infty$$

$$= 120\, e^{-4z_1}\left(1 - e^{-z_1}\right)\int_0^1 \left(t - t^2\right) dt \qquad ; 0 < z_1 < \infty$$

$$= 120\, e^{-4z_1}\left(1 - e^{-z_1}\right)\left(\frac{t^2}{2} - \frac{t^3}{3}\right)_0^1 \qquad ; 0 < z_1 < \infty$$

$$= 120\, e^{-4z_1}\left(1 - e^{-z_1}\right)\left(\frac{1}{2} - \frac{1}{3}\right) \qquad ;\ 0 < z_1 < \infty$$

$$= 20\, e^{-4z_1}\left(1 - e^{-z_1}\right) \qquad ; 0 < z_1 < \infty \qquad (3)$$

Similarly,

$$f\left(z_2\right) = \int_0^\infty f\left(z_1, z_2\right) dz_1$$

$$= 120\, e^{-2z_2}\left(1 - e^{-z_2}\right)\int_0^\infty \left(1 - e^{-z_1}\right) e^{-4z_1}\, dz_1 \qquad ; 0 < z_2 < \infty$$

Let $e^{-z_1} = t \ \Rightarrow\ -e^{-z_1} dz_1 = dt$

$$f\left(z_2\right) = 120\, e^{-z_2}\left(1 - e^{-z_2}\right)\int_0^1 t^3\left(1-t\right) dt \qquad ; 0 < z_2 < \infty$$

$$= 120\, e^{-z_2}\left(1 - e^{-z_2}\right)\left(\frac{t^4}{4} - \frac{t^5}{5}\right)_0^1 \qquad ; 0 < z_2 < \infty$$

$$= 6\, e^{-2z_2}\left(1 - e^{-z_2}\right) \qquad ;\ 0 < z_2 < \infty \qquad (4)$$

From (2), (3) and (4), we have,

$$f\left(z_1, z_2\right) = f\left(z_1\right) f\left(z_2\right)$$

showing $Z_1 = Y_2$ and $Z_2 = Y_2 - Y_4$ are stochastically independent.

JOINT DISTRIBUTION OF ORDER STATISTICS

Let X_1, $X_2, \dots,$ X_n be a random sample of size n taken from a continuous population whose p. d. f. is $f(x)$ and c .d. f. is $F(x)$ for $a < x < b$. Let $F(y_1, \dots, y_n)$ be the joint c. d. f. and $f(y_1 ,\dots ,y_n$ be the joint p. d. f. of all the order statistics $Y_1, \dots, Y_n$ at the point $(y_1, \dots, y_n)$ is given by

$$f(y_1, \dots, y_n) = \frac{\partial^n F(y_1, \dots, y_n)}{\partial y_1 \dots \partial y_n}$$

$$= \lim_{\substack{\Delta y_1 \to 0 \\ \dots \\ \Delta y_n \to 0}} \frac{\mathrm{P}(y_1 < Y_1 \le y_1 + \Delta y_1, \dots, y_n < Y_n \le y_n + \Delta y_n)}{\Delta y_1 \dots \Delta y_n}$$

$$= \lim_{\substack{\Delta y_1 \to 0 \\ \dots \\ \Delta y_n \to 0}} \frac{\mathrm{P}\{\text{one } X \text{ in } (y_1, y_1 + \Delta y_1], \dots, \text{one } X \text{ in } (y_n, y_n + \Delta y_n]\}}{\Delta y_1 \dots \Delta y_n}$$

$$\overset{\downarrow}{a}\underbrace{\qquad\qquad}\overset{\downarrow}{y_1}\underbrace{\qquad\quad}_{\text{one } X}\overset{\downarrow}{y_1 + \Delta y_1}\underbrace{\qquad\qquad}_{\dots}\overset{\downarrow}{y_n}\underbrace{\qquad\quad}_{\text{one } X}\overset{\downarrow}{y_n + \Delta y_n}\underbrace{\qquad\qquad}\overset{\downarrow}{b}$$

Using multinomial law, we have

$$= \lim_{\substack{\Delta y_1 \to 0 \\ \dots \\ \Delta y_n \to 0}} \frac{n!}{1! \dots 1!} \frac{\mathrm{P}(y_1 < Y_1 \le y_1 + \Delta y_1)}{\Delta y_1} \dots \frac{\mathrm{P}(y_n < Y_n \le y_n + \Delta y_n)}{\Delta y_n}$$

$$= \frac{n!}{1! \dots 1!} \lim_{\Delta y_1 \to 0} \left\{ \frac{F(y_1 + \Delta y_1) - F(y_1)}{\Delta y_1} \right\} \dots \lim_{\Delta y_n \to 0} \left\{ \frac{F(y_n + \Delta y_n) - F(y_n)}{\Delta y_n} \right\}$$

$$= \begin{cases} n!\, f(y_1) \dots f(y_n) & ;\ \text{a}<y_1 < \dots < y_n < b \\ 0 & ;\text{otherwise} \end{cases}$$

UNIT 2: DISTRIBUTION FREE INTERVALS

2.1 DISTRIBUTION OF RANGE FUNCTION OF ORDER STATISTICS

Let $Y_i\left(i=1,2,...,n\right)$ be an ith order statistic of the random sample $X_1, X_2, ..., X_n$ drawn from a continuous population whose c. d. f. $F\left(x\right)$ and p. d. f. is $f\left(x\right)$ for $a<x<b$. We define the sample range as

$R=Y_n-Y_1$.

In order to find the p. d. f. of R we first need to find the joint p. d. f. of Y_1 and Y_n is given by

$$g\left(y_1, y_n\right)=\frac{n!}{\left(n-2\right)!}\left\{F\left(y_n\right)-F\left(y_1\right)\right\}^{n-2} f\left(y_n\right) f\left(y_1\right) \qquad ;a<y_1<y_n<b$$

$$=n\left(n-1\right)\left\{F\left(y_n\right)-F\left(y_1\right)\right\}^{n-2} f\left(y_n\right) f\left(y_1\right) \qquad ;a<y_1<y_n<b$$

Let us now consider the transformation

$R=Y_n-Y_1$

$U=Y_n$

For specific values, we write

$r=y_n-y_1$

$u=y_n$

Then,

$y_1=u-r$ and $y_n=u$

Thus, the transformation $y_1 = u - r$ and $y_n = u$ maps $\{(y_1, y_n); a < y_1 < y_n < b\}$ onto $\{(r,u); a < r < u < b\}$ so that the joint p. d. f. of *R* and *U* is given by

$f_{RU}(r,u)$= (mod *J*){putting y_1 and y_n in terms of u and r in $g(y_1, y_n)$}

where *J* stands for the jacobian of transformation given by

$$J = \frac{\partial(y_1, y_n)}{\partial(r,u)} = \begin{vmatrix} \frac{\partial y_1}{\partial r} & \frac{\partial y_1}{\partial u} \\ \frac{\partial y_2}{\partial r} & \frac{\partial y_2}{\partial u} \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

so the joint p. d. f. of R and *U* takes form

$$f_{RU}(r,u) = n(n-1)\{F(u) - F(u-r)\}^{n-2} f(u-r) f(u) \qquad ; a < r < u < b$$

In order to obtain the p. d. f. of *R* we integrate out *U* from the joint p. d. f. of R and *U* and get

$$f_R(r) = \int_a^b f_{RU}(r,u)\,du$$

$$= \begin{cases} n(n-1)\int_r^b \{F(u) - F(u-r)\}^{n-2} f(u-r) f(u)\,du & ; a < r < b \\ 0 & \text{; otherwise} \end{cases}$$

Example: If $f(x) = \lambda e^{-\lambda x} \qquad ; 0 < x < \infty$

and $F(x) = 1 - e^{-\lambda x} \qquad ; 0 < x < \infty$

Find the distribution of sample range.

Solution: We know distribution of sample range

$$f(r) = n(n-1)\int_r^\infty [F(u) - F(u-r)]^{n-2} f(u-r) f(u)\,du$$

$$=n\left(n-1\right)\int_{r}^{\infty}\left[1-e^{-\lambda x}-1+e^{-\lambda u+\lambda r}\right]^{n-2}\lambda e^{-\lambda u+\lambda r}\lambda e^{-\lambda u}du$$

$$=n\left(n-1\right)\lambda^{2}\int_{r}^{\infty}\ e^{-\lambda u\left(n-2\right)}\left(-1+e^{\lambda r}\right)^{n-2}e^{-2\lambda u}e^{\lambda r}du$$

$$=n\left(n-1\right)\lambda^{2}\left(-1+e^{\lambda r}\right)^{n-2}e^{\lambda r}\int_{r}^{\infty}e^{-\lambda un}du$$

$$=n\left(n-1\right)\lambda^{2}\left(-1+e^{\lambda r}\right)^{n-2}e^{\lambda r}\left[\frac{e^{-\lambda un}}{-n\lambda}\right]_{r}^{\infty}$$

$$=\left(n-1\right)\lambda\left(-1+e^{\lambda r}\right)^{n-2}e^{\lambda r}e^{-\lambda nr}$$

$$=\left(n-1\right)\lambda\left(-1+e^{\lambda r}\right)^{n-2}e^{\lambda r}e^{-\lambda nr}\qquad\text{Q.E.D.}$$

Also

$$\left(n-1\right)\lambda\int_{0}^{\infty}\ \left(-1+e^{\lambda r}\right)^{n-2}e^{-r\lambda n}e^{\lambda r}dr$$

$$\left(n-1\right)\int_{1}^{\infty}\ e^{-\lambda u\left(n-2\right)}\left(t-1\right)^{n-2}\left(\frac{1}{t}\right)^{n}dt\qquad\because e^{\lambda r}=t,\lambda e^{\lambda r}\,dr=dt$$

$$=\left(n-1\right)\int_{1}^{\infty}\left(t\right)^{n-2}\left(1-\frac{1}{t}\right)^{n-2}\frac{1}{t^{n}}dt$$

Let $\left(1-\frac{1}{t}\right)=\vartheta$ so that $\left(\frac{1}{t^{2}}\right)dt=dv$

$$I=\left(n-1\right)\int_{0}^{1}\left(\vartheta\right)^{n-2}dv$$

$$=\left(n-1\right)\left[\frac{u^{n-1}}{n-1}\right]_{0}^{1}$$

$$=1$$

Hence,

$$f(r)=(n-1)\lambda\left(e^{-r\lambda}-1\right)^{n-2}e^{-nr\lambda}e^{\lambda r} \qquad 0<r<\infty$$

2.2 DISTRIBUTION FREE CONFIDENCE INTERVALS FOR QUANTILES

QUANTILES OF A DISTRIBUTION

Let X be a continuous random variable with p. d. f. $f(x)$ and c. d. f. $F(x)$. Let p be a positive proper fraction and the equation $F(x)=p$ as a unique solution for x, this unique root is denoted by the symbol ξ_p and is called the quantiles of order p.

Thus,

$$\Pr\left[X\le\xi_p\right]=F\left(\xi_p\right)=p \qquad 0<p<1$$

If $F(x)$ is not strictly increasing, $F(x)=p$ may hold in some interval, in this case any point in the interval would serve as a quantile of order p.

Example. The quantile of order $1/2$ is the median of the distribution and

$$\Pr\left[X\le\xi_{0.5}\right]=F\left(\xi_{0.5}\right)=1/2$$

Example: Let $x_i\,(i=1,2,\ldots,n)$ be i. i. d. random variable with p. d. f. $f(x)$ of the continuous type. If m is the median of the distribution, find the probability that

i) *All exceed's m*

ii) *The maximum never exceeds m*

Solution: Since m is the median of the continuous distribution. Therefore $F(m)=\Pr(X\le m)=1/2$ and $\Pr(X\ge m)=\Pr(X\le m)=1/2$

Now,

i)Pr(all exceed's m)= $\Pr\left(X_{(1)} \geq m\right)$

$$= \int_{m}^{\infty} n\left\{1-F\left(x\right)\right\}^{n-1} f\left(x\right)dx$$

Let $1-F\left(x\right)=t \quad \Rightarrow -f\left(x\right)dx = dt$

so that

Pr (all exceed's $m) = \int_{1/2}^{0} -n\,t^{n-1}\,dt$

$$= \int_{0}^{1/2} n\,t^{n-1}\,dt = \left[t^{n}\right]_{0}^{1/2}$$

$$= \left(\frac{1}{2}\right)^{n}$$

(ii) Pr (none of the X's exceeds median)= Pr(the maximum never exceeds m)

$$= \Pr\left[X_{(n)} \leq m\right]$$

$$= \int_{0}^{m} n\left[F\left(x\right)\right]^{n-1} f\left(x\right)dx$$

Let $F\left(x\right)=t \quad \Rightarrow f\left(x\right)dx = dt$

so that

Pr (none of the X's exceeds $m) = \int_{0}^{1/2} n\,t^{n-1}\,dt = \left[t^{n}\right]_{0}^{1/2}$

$$= \left(\frac{1}{2}\right)^{n}$$

Q.E.D.

CONFIDENCE INTERVAL FOR DISTRIBUTION QUANTILES

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from a continuous distribution with distribution function $F(x)$. Let $Y_1 < Y_2 < \ldots < Y_n$ be the order statistics of the sample. Let $Y_i < Y_j$, we consider the event $Y_i < \xi_p < Y_j$. For the ith order statistic Y_i to be less than ξ_p it must be true that at least *i* of the x values are less than ξ_p. Moreover, for the *j*th order statistic to be greater than ξ_p fewer than *j* of the *x* values are less than ξ_p. That is, if we say that we have a "success" when an individual *x* value is less than ξ_p, then, in the n independent trials, there must be at least *i* success but fewer than *j* success for the event $Y_i < \xi_p < Y_j$ to occur. But since the probability of success on each trial is

$$\Pr\left[X \le \xi_p\right] = F\left(\xi_p\right) = p,$$

the probability of this event is

$$\Pr\left[Y_i < \xi_p < Y_j\right] = \sum_{w=i}^{j-1} \frac{n!}{w!(n-w)!} p^w (1-p)^{n-w}$$

The probability of having at least *i*, but less the *j* success. When particular values of *n*, *i*, and *j* are specified, this probability can be computed. Let this probability be γ

i.e. $\Pr\left[Y_i < \xi_p < Y_j\right] = \gamma$

then we say that the probability is γ that the random interval (Y_i, Y_j) includes the quantile of order p. if the experimental values of Y_i and Y_j are respectively,

y_i and y_j, the interval $\left(y_i, y_j\right)$ serves as $100\gamma\%$ confidence interval for ξ_p, the quantile of order *p*.

Example: Find the smallest value of n for which $\Pr\left[Y_1 < \xi_{0.5} < Y_n\right] \geq 0.99$*, where* $Y_1 < Y_2 < \ldots < Y_n$*, are order statistics of random sample of size n from a distribution of continuous type and* ξ_p *is a quantile of order p.*

Solution: Consider

$$\Pr\left[Y_1 < \xi_{0.5} < Y_n\right] \geq 0.99$$

$$= \sum_{w=1}^{n-1} {}^nC_w \left(0.5\right)^w \left(1-0.5\right)^{n-w} \geq 0.99$$

so that

$$\sum_{w=1}^{n-1} {}^nC_w \left(1/2\right)^n \geq 0.99 \qquad (1)$$

Also, we know that

$$\sum_{w=0}^{n} {}^nC_w \left(1/2\right)^n = 1 = \left(1/2+1/2\right)^n \qquad \therefore \left(q+p\right)^n = \sum_{r=0}^{n} {}^nC_r p^r \left(q\right)^{n-s}$$

$$\left(1/2\right)^n + \sum_{w=1}^{n-1} {}^nC_w \left(1/2\right)^n + \left(1/2\right)^n = 1$$

This gives

$$\sum_{w=1}^{n-1} {}^nC_w \left(1/2\right)^n = 1 - \left(1/2\right)^n \qquad (2)$$

From (1) and (2) we get

$$1 - 2\left(1/2\right)^n \geq 0.99$$

$$1 - 0.99 \geq 2\left(1/2\right)^n$$

$2(1/2)^n \leq 0.01$ (3)

(3) holds for *n*=8, 9, …, hence smallest *n* is 8.

TOLERANCE INTERVAL

Let $X_1, X_2, \ldots, X_n$ denotes a random sample of size n taken from a distribution having a positive and continuous p. d. f. $f(x)$ if and only if a<x<b. let $F(x)$ be its distribution function. Consider the random variables $F(X_1)$, $F(X_2)$, … $F(X_n)$. These random variables are mutually stochastically independent and each follows $U(0,1)$.

Let $Z_1, < Z_2 < \ldots < Z_n$ be the order statistics of the random sample $F(X_1)$, $F(X_2)$,…, $F(X_n)$. If $Y_1 < Y_2 < \ldots < Y_n$ are the order statistics of the original sample $X_1, X_2, \ldots, X_n$ then

$Z_1 = F(Y_1), Z_2 = F(Y_2), \ldots Z_n = F(Y_n)$

Let us consider the difference $Z_j - Z_i = F(Y_j) - F(Y_i)$ for every i<j

Now $F(Y_j) = \Pr(X \leq Y_j)$

And $F(Y_i) = \Pr(X \leq Y_i)$

But $\Pr(X = Y_j) = \Pr(X = Y_i) = 0$ (as distribution is continuous)

Thus $Z_j - Z_i = \Pr(Y_i < X < Y_j)$

Let p be a positive fraction if

$F\left(Y_j\right)-F\left(Y_i\right)\geq p$

Then at least 100p% of the probability for the distribution of X is between y_i and y_j

Let $\gamma=\Pr\left[F\left(Y_j\right)-F\left(Y_i\right)\geq p\right]$

$$=\Pr\left[Z_j-Z_i\geq p\right]$$

$$=\int_0^{1-p}\int_{p+z_i}^{1}h_{ij}\left(Z_{i,}Z_j\right)dZ_jdZ_i$$

where $h_{ij}\left(Z_i,Z_j\right)$ is joint p. d. f. of Z_i and Z_j.

Then, the random interval $\left(Y_i,Y_j\right)$ has probability γ of containing at least 100*p*% of the probability for the distribution of *x* is the tolerance interval of 100*p*% of the probability distribution of *x*. If now y_i and y_j denote respectively, experimental values of Y_i and Y_j, the interval $\left(y_i,y_j\right)$ either does or does not contain at least 100*p*% of the probability for the distribution of *x* and y_i and y_j are known as the tolerance limits for 100*p*% of the probability distribution of *x*.

Example: Let Y_1 and Y_n be the smallest (i.e. the first and the n^{th} order statistics of a random sample of size n from the continuous distribution $F\left(x\right)$. Find the smallest n such that $P\left[\left\{F\left(Y_n\right)-F\left(Y_1\right)\right\}\geq 0.5\right]$ is at least 0.95.

Solution: Consider

$P\left[\left\{F\left(Y_n\right)-F\left(Y_1\right)\right\}\geq 0.5\right]\geq 0.95$

$$1-P\left[\left\{F\left(Y_n\right)-F\left(Y_1\right)\right\}\leq 0.5\right]\geq 0.95$$

$$P\left[\left\{F\left(Y_n\right)-F\left(Y_1\right)\right\}\leq 0.5\right]\leq 0.5$$

$$P\left[0<Z_n-Z_1\leq 0.5\right]\leq 0.5$$

$$\int_0^{0.5} r_{n-1}\left(\vartheta\right)d\vartheta\leq 0.5$$

but $$h_{j-i}\left(\vartheta\right)=\begin{cases}\dfrac{\overline{|n+1}}{\overline{|\left(j-i\right)}\,\overline{|\left(n-j+i+1\right)}}\vartheta^{j-i-1}\left(1-\vartheta\right)^{n-(j-i)} & ;0<\vartheta<1\\ 0 & ;\text{elsewhere}\end{cases}$$

therefore, $$h_{n-i}\left(\vartheta\right)=\frac{\overline{|n+1}}{\overline{|\left(n-1\right)}\,\overline{|\left(n-n+1+1\right)}}\vartheta^{n-2}\left(1-\vartheta\right)^{n-n+1}$$

$$=\frac{n!}{\left(n-2\right)!}\vartheta^{n-2}\left(1-\vartheta\right)^{1}$$

$$=n\left(n-1\right)\vartheta^{n-2}\left(1-\vartheta\right)$$

Thus,

$$n\left(n-1\right)\int_0^{0.5}\vartheta^{n-2}\left(1-\vartheta\right)d\vartheta\leq 0.05$$

$$n\left(n-1\right)\int_0^{0.5}\left(\vartheta^{n-2}-\vartheta^{n-1}\right)d\vartheta\leq 0.05$$

$$n\left(n-1\right)\left(\frac{\vartheta^{n-1}}{n-1}-\frac{\vartheta^{n}}{n}\right)_0^{0.5}\leq 0.05$$

$$n\left(n-1\right)\left(\frac{\left(0.5\right)^{n-1}n-\left(0.5\right)^{n}\left(n-1\right)}{n\left(n-1\right)}\right)\leq 0.05$$

$$\left(\frac{1}{2}\right)^{n-1}\left(\frac{2n-n+1}{2}\right)\leq 0.05$$

$$\left(n+1\right)\left(\frac{1}{2}\right)^{n}\leq 0.05$$

The smallest *n* satisfying the above equation in *n*=8.

2.4 COVERAGES

Let $X_1, X_2, ..., X_n$ denotes a random sample of size n taken from a distribution having a positive and continuous p. d. f. $f\left(x\right)$ if and only if a<x<b. let $F\left(x\right)$ be its distribution function. Consider the random variables $F\left(X_1\right)$, $F\left(X_2\right)$, ... $F\left(X_n\right)$. These random variables are mutually stochastically independent and each follows $U\left(0,1\right)$. Thus $F\left(X_1\right)$, $F\left(X_2\right)$, ... $F\left(X_n\right)$ is a random sample of size *n* from U(0,1).

Let $Z_1, < Z_2 < ... < Z_n$ be the order statistics of the random sample $F\left(X_1\right)$, $F\left(X_2\right)$,..., $F\left(X_n\right)$. If $Y_1 < Y_2 < ... < Y_n$ are the order statistics of the original sample $X_1, X_2, ..., X_n$ then

$$Z_1 = F\left(Y_1\right), Z_2 = F\left(Y_2\right), ... Z_n = F\left(Y_n\right)$$

Now, consider the random variables

$$C_1 = W_1 = F\left(Y_1\right) = Z_1$$

$$C_2 = W_2 = F\left(Y_2\right) - F\left(Y_1\right) = Z_2 - Z_1$$

$$C_3 = W_3 = F\left(Y_3\right) - F\left(Y_2\right) = Z_3 - Z_2 .$$

$$\vdots$$

$C_n = W_n = F(Y_n) - F(Y_{n-1}) = Z_n - Z_{n-1}$

Then random variable W_1 or C_1 is called the coverage of the random interval $\{x; -\infty < x < Y_1\}$ and the random variable W_i or C_i, i=1, 2,..., n is called a coverage of random interval $\{x; Y_{i-1} < x < Y_i\}$

Joint p. d. f. of $W_1, W_2, ..., W_n$ or $C_1, C_2, ..., C_n$

We have

$C_1 = W_1 = Z_1$

$C_2 = W_2 = Z_2 - Z_1$

$C_3 = W_3 = Z_3 - Z_2$.

$\vdots$

$C_n = W_n = Z_n - Z_{n-1}$

For specific values

$c_1 = w_1 = z_1$

$c_2 = w_2 = z_2 - z_1$

$c_3 = w_3 = z_3 - z_2$.

$\vdots$

$c_n = w_n = z_n - z_{n-1}$

The inverse function of this associated transformation are given by

$z_i = w_1 + w_2 + ... + w_i$

$= c_1 + c_2 + ... + c_i$ for every i=1, 2, ..., n

Now jacobian of transformation

$$J = \frac{\partial(z_1, z_2, ..., z_n)}{\partial(w_1, w_2, ..., w_n)}$$

$$= \begin{vmatrix} \frac{\partial z_1}{\partial w_1} & \frac{\partial z_1}{\partial w_2} \cdots & \frac{\partial z_1}{\partial w_n} \\ \frac{\partial z_2}{\partial w_1} & \frac{\partial z_2}{\partial w_2} \cdots & \frac{\partial z_2}{\partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_n}{\partial w_1} & \frac{\partial z_n}{\partial w_2} & \frac{\partial z_n}{\partial w_n} \end{vmatrix}$$

$$= 1$$

Therefore, mod(J) = 1

Now

$h(w_1, w_2, ..., w_n) = r(c_1, c_2, ..., c_n)$

= (mod J){put $z_1, z_2, ..., z_n$ in terms of $w_1, w_2, ..., w_n$ in the joint p. d. f. of $Z_1, Z_2, ..., Z_n$}

But $h(z_1, z_2, ..., z_n) = n!$

Thus, $h(w_1, w_2, ..., w_n) = n!$ $;0 < w_i, i = 1, 2, ..., n;$ $w_1 + w_2 + ... + w_n < 1$

=0 ;elsewhere

Example: Show that each of the coverages has the beta p. d. f.

$$k(w) = \begin{cases} n(1-w)^{n-1} & ;0 < w < 1 \\ 0 & ;elsewhere \end{cases}$$

Solution: Since the joint p. d. f. of the coverages $k(w_1, w_2, ..., w_n)$is symmetric in $w_1, w_2, ..., w_n$ axis given by

$$k\left(w_1, w_2, ..., w_n\right) = \begin{cases} n! & ;0 < w_i \\ & i = 1, 2, ..., n \\ & w_1 + w_2 + ... + w_n < 1 \\ 0 & ;elsewhere \end{cases}$$

it is evident that the distribution of every sum r , $r < n$ of these coverages $w_1, w_2, ..., w_n$ is exactly the same for fixed value of r .

Consider if $i < j$ and $r = j - i$, the distribution of any sum of j - i coverages

$$Z_j - Z_i = F\left(Y_j\right) - F\left(Y_i\right) = w_{i+1} + w_{i+2} + ... + w_j$$

$$= \left(w_1 + w_2 + ... + w_j\right) - \left(w_1 + w_2 + ... + w_i\right)$$

$$= w_{i+1} + w_{i+2} + ... + w_j \qquad \left(w_i < w_j\right)$$

is exactly the same as that of

$$z_{j-1} = F\left(Y_{j-i}\right) = w_1 + w_2 + ... + w_{j-i}$$

but we know the p. d. f. of

$$h_{j-i}\left(v\right) = \frac{n!}{\left\{\left(j-i\right)-1\right\}!\left(n-\left(j-i\right)\right)!} v^{\left\{\left(j-i\right)-1\right\}} \left(1-v\right)^{n-\left(j-i\right)}$$

$$= \begin{cases} \dfrac{\lceil\overline{n+1}}{\lceil\overline{\left(j-i\right)}\lceil\overline{\left(n-j+i+1\right)}} v^{j-i-1}\left(1-v\right)^{n-\left(j-i\right)} & ;0 < v < 1 \\ 0 & ;elsewhere \end{cases}$$

Consequently, $Z_j - Z_i = F\left(Y_j\right) - F\left(Y_i\right)$ has above mentioned p. d. f. Putting r=1 such that j=2 and i=1, we have p. d. f. of w_1 given by

$$k\left(w_1\right) = \frac{\lceil\overline{n+1}}{\lceil\overline{1}\lceil\overline{n}}\left(1-w_1\right)^{n-1} \qquad 0 < w_1 < 1$$

$$= n\left(1-w_1\right)^{n-1}$$

But similarly if *j*=3 and *i*=2

$$k\left(w_2\right)=\frac{n!}{1!\left(n-1\right)!}\left(1-w_2\right)^{n-1} \qquad 0<w_2<1$$
$$=n\left(1-w_2\right)^{n-1}$$

Therefore in general, we can say that each of the coverages has the beta p. d. f.

$$k\left(w\right)=\begin{cases} n\left(1-w\right)^{n-1} & ;0<w<1 \\ 0 & ;elsewhere \end{cases}$$

Example: Let c_i *denote the* i^{th} *coverage, find expectation of* c_i.

Solution: Since each of the coverage c_i, *i*=1,2,...,*n* has the beta p. d. f.

$$k\left(c\right)=\begin{cases} n\left(1-c\right)^{n-1} & ;0<c<1 \\ 0 & ;elsewhere \end{cases}$$

and $c_1=Z_1=F\left(Y_1\right)$ follows $U\left(0,1\right)$. The expectation of each c_i is given by

$$E\left(c_i\right)=\int_0^1 nc\left(1-c\right)^{n-1}dc$$

$$=n\int_0^1 c\left(1-c\right)^{n-1}dc$$

$$=n\left[\left[\left\{\frac{c\left\{-\left(1-c\right)^n\right\}}{n}\right\}\right]_0^1-\int_0^1\frac{\left\{-\left(1-c\right)^n\right\}}{n}dc\right]$$

$$=n\left[0+\frac{1}{n}\int_0^1\left\{\left(1-c\right)^n\right\}dc\right]=\left[\frac{\left\{-\left(1-c\right)^{n+1}\right\}}{n+1}\right]_0^1$$

$$=\frac{1}{n+1}$$

PGSTAT-11 / MASTAT-11 (Old)

PGSTAT—15 / MASTAT- 15 (New)

NON-PARAMETRICS

BLOCK 2- NON- PARAMETRIC TESTS

PGSTAT-11 / MASTAT-11 (Old)

PGSTAT—15 / MASTAT- 15 (New)

NON-PARAMETRICS

Block 2- Non- Parametric Tests

Unit 3: One- sample and Two Sample Location Non- Parametric Tests

Structure

Unit 4: Other non- parametric tests

Structure

3.1 Introduction of Non- Parametric Methods

The parametric inferential methods are based on stringent assumptions about the probability distribution of the parent population like the form of probability distribution is apriori available, availability of observations either on ratio scale or atleast on interval scale etc. However, these assumptions may not be satisfied in many practical situations. For instance, the measurement on the units under study is often made on nominal or ordinal scale owing to practical difficulties. Usually, we do not know the distribution characterizing the phenomena of the experiment. However, we can often choose a sufficiently large class of distributions $\left\{F_\theta(x)\right\}$ invariably indexed by an unknown parameter θ. The range of θ is Ω which is called the parameter space. The statistician has to decide upon the particular probability distribution which explains most the phenomena of the experiment. That is, the statistician has to make a decision about the value of the parameter, by means of the observable random variable X. However, in many situations the outcome X is a complicated set of numbers. If at all feasible, he would like to condense his data and come out with a magic number which contains all the relevant information about the parameter θ. In such situations where the stringent assumptions of the parametric inferential methods are not satisfied, we resort to non-parametric methods. The non-parametric methods rely on relatively mild assumptions about the probability distribution of the parent population.

The statistical methods which are not concerned with estimation of testing for parameter(s) of probability distribution functions are known as NON-PARAMETRIC METHODS. Nonparametric statistical procedures are widely used due to their simplicity, applicability under fairly general assumptions and robustness to outliers in the data. Hence they are popular statistical tools in industry, government and various other disciplines. Also, there an extensive amount of literature is available on nonparametric statistics ranging from theory to applications.

The term non-parametric is sometimes synonymously used with distribution free methods as if both have the same meaning. There is a slight difference between the two methods. The statistical inferential procedures whose validity does not depend on the form of probability distribution of the population from which the sample has been drawn are known as DISTRIBUTION FREE METHODS. The distribution free procedures are primarily devised for non-parametric problems; hence the two terms are used interchangeably. Also, the non-parametric methods are devised for no parameter problems.

Non-parametric Inferences

The classical statistical inference techniques are based on the assumptions regarding the nature of the population distribution from which the samples are drawn. i.e. form of the population distribution and the parameters of the population distribution. The exact sample tests are based on the assumption that the parent population is normal. Most of the standard statistical techniques are based on the assumptions of normality, independence and homoscedasticity.

Remark

The statistical methods which remain valid under violation of assumptions of normality, independence & homoscedasticity are called 'robust'.

Parametric Test

The parametric tests are those tests in which certain conditions are imposed about the parameters of the population from which the samples are drawn.

Ex- t-test, F-test.

General assumption of parametric tests

The parent population from which the samples are drawn is assumed to be normal.

The form of the basic distribution is always known.

Non-parametric Test

The non-parametric test are those tests in which no assumption, regarding the test of the population from which the samples are drawn is made.

The non-parametric tests are the tests for a hypothesis which is not a statement about the parameter values. Here, the hypothesis is concerned with either form of the population (e.g- goodness of fit) or with some characteristic of the probability distribution of the sample data (e.g- test of randomness).

General assumption of non-parametric tests

1. The parent population is continuous.
2. The sample observations are independent.
3. The distribution of the parent population is symmetrical.
4. The lower order moments exist.

3.2 Objectives of this Unit

The objective of this unit is to understand the basic concepts of non-parametric methods and to apply these methods in practice.

3.3 Advantages and Dis-advantages of Non- Parametric Methods

Advantages of Non-parametric tests

1. Non-parametric tests are quick and easy to apply and do not require complicated sample theory.
2. No assumption is required about the form of the distribution of the parent population from which the samples are drawn.
3. Non-parametric tests can be used even in the situations where actual measurements are unavailable and the data are obtained only as ranks. i.e. if measurements scale is nominal or ordinal, non-parametric methods can be used.

4. The probability statements obtained from most of the non-parametric tests are exact probabilities.
5. Non-parametric tests are used in the situation where sample data are taken from several different populations.
6. With non-accurate and dirty data (e.g: contaminated observations, outliers etc.), many non-parametric methods are appropriate.
7. Non-parametric tests requires no. of minimum sample size for valid and reliable results.
8. Non-parametric tests requires minimal calculation.

Disadvantages of Non-parametric tests

1. If all the assumptions of a statistical model are satisfied by the data and if the observations are of required strength, then non-parametric tests are wasteful of time and data.
2. Non-parametric tests are designed to test the statistical hypothesis only and not for estimating the parameters.
3. Power efficiency on non-parametric tests are always less than parametric tests.
4. No non-parametric test exists for testing interactions in the analysis of variance model unless specific assumptions about the additivity of the model are made.
5. It is not possible to determine the actual power of non-parametric test due to want of actual situation or actual probability distribution.

NON-PARAMETRIC TEST FOR LOCATION

The following are the non-parametric test for location parameter of a population or the non-parametric tests for the location parameters of two populations.

In non-parametric theory, the most frequently used measure of location is "population median" M or $K_{0.5}$, which is the unique real solution of the equation.

$$F(M)=\frac{1}{2} \text{ or } F(K_{0.5})=\frac{1}{2}$$

3.4 Sign Test

The sign test is a non-parametric test for the location parameter median M of a population.

3.4.1 Hypothesis and Assumptions

In this test, we make the assumption of independence and homoscedasticity but the assumption of normality for the parent population is not required.

We wish to test the null hypothesis

$H_0 : M = M_0$ (a given value)

against

(i) a one sided alternative

$H_1 : M < M_0$ (left tailed test)

$H_1 : M > M_0$ (right tailed test)

(ii) a two sided alternative

$H_1 : M \neq M_0$

1.4.2 Test Procedure

Let $X_{(1)},\dots,X_{(n)}$ be the order statistics corresponding to a random sample $X_1,\dots\dots X_n$ of size n drawn from the population having distribution function F with unknown median M , where F is assumed to be continuous in the neighborhood of M so that $P(X = M) = 0$.

By definition of median, we have

$$P(X > M) = P(X < M) = \frac{1}{2}$$

If the sample data are consistent with the hypothetical value of median M_0, then on the average half of the sample observations will be greater than M_0. We replace each observation greater than M_0 by a plus sign (+) and each observation smaller than M_0 by a minus sign (-). Further, we count the numbers of plus signs and the minus signs and denoted it by r and s respectively, with $r + s \leq n$. The number of plus signs (r) may be used to test H_0.

When the population is dichotomized, the sampling distribution of r given $(r+s)$ is binomial with parameter $p = P(X > M_0) = \frac{1}{2}$. Thus the testing of H_0 becomes an equivalent testing for the hypothesis that the binomial parameter p has the value $\frac{1}{2}$ .i.e. $H_0 : p = \frac{1}{2}$

The critical region for

$H_0 : M = M_0$ or $H_0 : p = \frac{1}{2}$

against $H_1 : M \neq M_0$ or $H_1 : p \neq \frac{1}{2}$

for α level of significance is given by

$r \geq r_{\alpha/2}$ and $r \leq r_{\alpha/2}$.

where $r_{\alpha/2}$ is the smallest integer such that

$$\sum_{k=r_{\alpha/2}}^{r+s} {}^{r+s}C_k . p^{r+s} \leq \frac{\alpha}{2}$$

i.e. $$\sum_{k=r_{\alpha/2}}^{r+s} {}^{r+s}C_k . \left(\frac{1}{2}\right)^{r+s} \leq \frac{\alpha}{2}$$

and $r'_{\alpha/2}$ is the largest integer such that

$$\sum_{k=0}^{r'_{\alpha/2}} {}^{r+s}C_k . p^{r+s} \leq \frac{\alpha}{2}$$

i.e. $$\sum_{k=0}^{r'_{\alpha/2}} {}^{r+s}C_k . \left(\frac{1}{2}\right)^{r+s} \leq \frac{\alpha}{2}$$

For testing $H_0 : M = M_0$ or $H_0 : p = \frac{1}{2}$

against $H_1 : M > M_0$ or $H_1 : p > \frac{1}{2}$

the critical region for α level of significance is given by

$$r \geq r_\alpha$$

where r_α is the smallest integer such that

$$\sum_{k=r_\alpha}^{r+s} {}^{r+s}C_k . p^{r+s} \le \alpha$$

i.e. $\sum_{k=r_\alpha}^{r+s} {}^{r+s}C_k . \left(\frac{1}{2}\right)^{r+s} \le \alpha$

In this alternative hypothesis the sample will have excess of plus signs.

For testing $H_0 : M = M_0$ or $H_0 : p = \frac{1}{2}$

against $H_1 : M < M_0$ or $H_1 : p < \frac{1}{2}$

the critical region for α level of significance is given by

$$r \ge r_\alpha^{'}$$

where $r_\alpha^{'}$ is the largest integer such that

$$\sum_{k=0}^{r_\alpha^{'}} {}^{r+s}C_k . p^{r+s} \le \alpha$$

i.e. $\sum_{k=0}^{r_\alpha^{'}} {}^{r+s}C_k . \left(\frac{1}{2}\right)^{r+s} \le \alpha$

In this alternative hypothesis the sample will have less plus signs.

Large sample test

If $(r+s) > 25$, then we use the normal approximation to the binomial to perform the test.

In this case, under H_0

$$Z = \frac{r - E\left(r\right)}{\sqrt{V\left(r\right)}} \to N(0,1)$$

since $E\left(r\right) = \frac{\left(r+s\right)}{2}$ and $V\left(r\right) = \frac{\left(r+s\right)}{4}$

Hence, under H_0

$$Z = \frac{r - \left(\frac{(r+s)}{2}\right)}{\sqrt{\frac{(r+s)}{4}}} = \frac{r-s}{\sqrt{(r+s)}} \to N(0,1)$$

3.4.3 Example

Example1. Test the null hypothesis that the median length θ of ear-head of a variety of wheat is $\theta_0 = 9.9$ cm. against the alternative that $\theta_0 \neq 9.9$ cm., with $\alpha = 0.05$, on the basis of the following 25 ear-head measurements:

9.5	**8.9**	**10.5**	**11.5**	**8.5**	**9.4**	**10.6**	**8.8**	**11.7**	**10.5**
11.2	9.2	9.8	9.5	9.9	10.9	10.2	9.1	10.8	9.4
11.6	8.7	8.3	11.3	8.1					

Solution: First, we determine the signs of all measurement and replace each measurement greater than θ_0 by + sign and each measurement less than θ_0 by − sign. Measurement which is equal to θ_0 is ignored.

9.5	**8.9**	**10.5**	**11.5**	**8.5**	**9.4**	**10.6**	**8.8**	**11.7**	**10.5**
(−)	(−)	(+)	(+)	(−)	(−)	(+)	(−)	(+)	(+)
11.2	9.2	9.8	9.5	9.9	10.9	10.2	9.1	10.8	9.4
(+)	(−)	(−)	(−)	ignored	(+)	(+)	(−)	(+)	(−)
11.6	8.7	8.3	11.3	8.1					
(+)	(−)	(−)	(+)	(−)					

From the above table, we observe that no. of plus signs $r=11$ and the no. of minus signs$=s=13$ and one observation is ignored.

So we have to test whether $r=11$ support the hypothesis $H_0:\theta_0=9.9$, or equivalently to judge how likely are 11 successes (the number of plus signs) to occur in 24 trials from a binomial distribution with $p=0.5$. The critical region for the level α two-sided test is given by

$$r \geq r_{\alpha/2} \text{ and } r \leq r'_{\alpha/2},$$

where $r_{\alpha/2}$ is the smallest and $r'_{\alpha/2}$ is the largest integer such that

$$\sum_{r_{\alpha/2}}^{n} {}^{n}C_x \left(\frac{1}{2}\right)^n \leq \alpha/2$$

and
$$\sum_{0}^{r'_{\alpha/2}} {}^{n}C_x \left(\frac{1}{2}\right)^n \leq \alpha/2$$

From binomial tables, we find that $r_{0.025}=18$ and $r'_{0.025}=6$ for $n=24$ and $p=0.5$. Thus, for $r=11$ null hypothesis is to be accepted.

Note: The critical region for one-sided test alternative

$$\sum_{r_\alpha}^{24} {}^{24}C_x \left(\frac{1}{2}\right)^{24} \leq 0.05$$

since under the alternative hypothesis the sample will have an excess of plus signs. In the case of the other one-sided alternative, viz.

$$H : \theta < 9.9\,\text{cm. or } H : p < 0.5$$

The critical region for the level α will be $r \leq r_\alpha^{'}$, where $r_\alpha^{'}$ is the largest integer such that

$$\sum_{0}^{r_\alpha} {}^{24}C_x \left(\frac{1}{2}\right)^{24} \leq \alpha$$

Example. The weights of 12 persons before they are subjected to a change of diet and after a lapse of six months are recorded below:

S. No.	Weight (in kg.)	
	Before	After
1	57	62
2	48	55
3	55	62
4	45	53
5	62	59
6	42	45
7	49	45

8	60	55
9	65	64
10	51	55
11	46	50
12	58	66

Test whether there has been any significant gain in weight as a result of the change of diet.

Solution: Let y and x be the weight of a person before and after the change of diet, then the hypothesis to be tested is $H_0 : \theta = 0$ and the alternative is $H : \theta > 0$, where θ is the median of the distribution of differences d_i. The gain in weight $\left(d_i\right)$ for 12 persons are:

$+5, +7, +7, +8, -3, +3, -4, -5, -1, +4, +6, -8$

Here, the no. of plus signs $= 7$ and the no. of minus signs $= 5$. Under the null hypothesis, the expected number of plus signs among the differences in a sample of 12 pairs is 6. The sampling distribution of the number of plus signs is the binomial distribution with probability of plus signs 0.5. From table, we find that the probability of 7 or more plus signs is 0.387. So the null hypothesis is accepted at the 5% level.

3.4.4 Merits and Demerits

Merits

It is very simple to calculate.

It requires minimum effort for calculation.

Demerits:

The disadvantage of the sign test is that, although it takes account of signs of the deviations, it makes no allowance for their magnitudes.

3.5 Wilcoxon Test

One sample Wilcoxon signed-rank test

It is a non-parametric test for the location parameter (median) of a population.

3.5.1 Hypothesis and Assumptions

In this test, we make the assumption of independence and homoscedasticity but do not assume normality for the parent population. Also, if we assume that the parent population is continuous and symmetric, the Wilcoxon signed rank test is more efficient than the sign test for testing median of the population, since it takes into account both the magnitudes and signs of the deviations.

We wish to test the null hypothesis

$H_0 : M = M_0$ (a given value)

against

a: one sided alternative

$H_1 : M < M_0$ (left tailed test)

$H_1 : M > M_0$ (right tailed test)

or **b:** two sided alternative

$H_1 : M \neq M_0$

3.5.2 Test Procedure

Let $x_1, x_2, ...,x_n$ be a random sample of size n drawn from a population which is continuous and symmetric about median *M*. Then, under H_0, the differences $D_i = X_i - M_0,\ \forall i = 1,2,....,n$ are symmetrically distributed about zero, so that the positive and negative differences of the equal absolute value have the same probability of occurrence. Thus,

$P(D_i \geq C) = P(D_i \geq -C)$

or $P(D_i \geq C) = 1 - P(D_i \leq C)$

Suppose we order these absolute differences $|D_1|, |D_2|, ..., |D_n|$ from smallest to largest and assign them ranks $i = 1,2,....,n$. Let T^+ be the sum of ranks of the positive D_i and T^- be the sum of the ranks of the negative D_i.

If H_0 is true (i.e. M_0 is the true median of the symmetrical population), then expectation of T^+ equals the expectation of T^-. Since the sum of all the ranks is a constant given by

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

The tests based on T^+, T^- and T^-, T^+ will be equivalent (since they are linearly related). In practice, the minimum of T^+ and T^- is used as the test statistic.

Let us define a new random variable:

$$D_{(i)} = \begin{cases} 1 & \text{if } \mathrm{D_i} > 0 \, \text{for}\, \mathrm{i}^{\text{th}} \text{ smallest} |\mathrm{D_i}| \\ 0 & \text{if } \mathrm{D_i} < 0 \, \text{for}\, \mathrm{i}^{\text{th}} \text{ smallest} |\mathrm{D_i}| \end{cases}$$

$D_{(i)}$ are independent Bernoulli random variables but are not identically distribute such that

$$E\left[D_{(i)}\right]=p_i$$

$$V\left[D_{(i)}\right]=p_i\left(1-p_i\right)$$

and $\operatorname{cov}\left[D_{(i)},D_{(j)}\right]=0,\ i\neq j$

We can write

$T^+=\sum_{i=1}^{n} iD_{(i)}$ and $T^-=\sum_{i=1}^{n} i\left[1-D_{(i)}\right]$

Thus $E\left[T^+\right]=\sum_{i=1}^{n} iE\left[D_{(i)}\right]=\sum_{i=1}^{n} ip_i$

and $V\left[T^+\right]=\sum_{i=1}^{n} i^2V\left[D_{(i)}\right]=\sum_{i=1}^{n} i^2p_i\left(1-p_i\right)$

under H_0, i.e. when $p_i=\frac{1}{2}$

$$E\left[T^+\right]=\sum_{i=1}^{n} i\left(\frac{1}{2}\right)=\frac{1}{2}\sum_{i=1}^{n} i=\frac{n(n+1)}{4}$$

and $V\left[T^+\right]=\sum_{i=1}^{n} i^2\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right)=\frac{1}{4}\sum_{i=1}^{n} i^2=\frac{n(n+1)(2n+1)}{24}$

Similarly for T^-,

Let $T=\min\left[T^+,T^-\right]$ and T_α be such that $P\left[T\leq T_\alpha\right]=\alpha$.

Then the critical regions for α level of significance for testing $H_0 : M = M_0$ against different types of alternative are given as

Alternative Hypothesis	Critical Region
$H_1 : M > M_0$	$T^- \leq T_\alpha$
$H_1 : M < M_0$	$T^+ \leq T_\alpha$
$H_1 : M \neq M_0$	$T^+ \leq T_{\alpha/2}$ or $T^- \leq T_{\alpha/2}$

If $n > 25$, then distribution of T is taken to be approximation normal i.e. under H_0 we have

$$Z = \frac{T - E[T]}{\sqrt{V[T]}} \rightarrow N(0,1)$$

where $T = \min\left[T^+, T^-\right]$ and

$$E[T] = \frac{n(n+1)}{4}$$

$$V[T] = \frac{n(n+1)(2n+1)}{24}$$

Also, the sample size n is adjusted to include only non-zero differences.

3.5.3 Example. Test the null hypothesis that the median length θ of ear-head of a variety of wheat is $\theta_0 = 9.9$ cm. against the alternative that $\theta_0 \neq 9.9$ cm., with $\alpha = 0.05$, on the basis of the following 25 ear-head measurements:

9.5	8.9	10.5	11.5	8.5	9.4	10.6	8.8	11.7	10.5

11.2	9.2	9.8	9.5	9.9	10.9	10.2	9.1	10.8	9.4
11.6	8.7	8.3	11.3	8.1					

Solution: First, we determine

S.no.	x_i	$d_i = x_i - \theta_0$	Rank of $\lvert d_i \rvert$
1	9.5	-0.4	3.5
2	8.9	-1	13.5
3	10.5	0.6	7.5
4	11.5	1.6	20.5
5	8.5	-1.4	18.5
6	9.4	-0.5	5.5
7	10.6	0.7	9.5
8	8.8	-1.1	15
9	11.7	1.8	23.5
10	10.5	0.6	7.5
11	11.2	1.3	17
12	9.2	-0.7	9.5
13	9.8	-0.1	1
14	9.5	-0.4	3.5
15	9.9	0	
16	10.9	1	13.5

17	10.2	0.3	2
18	9.1	-0.8	11
19	10.8	0.9	12
20	9.4	-0.5	5.5
21	11.6	1.7	22
22	8.7	-1.2	16
23	8.3	-1.6	20.5
24	11.3	1.4	18.5
25	8.1	-1.8	23.5

Here $T^{+} = 153.5$, $T^{-} = 146.5$, so that $T = 150$From table, for $n = 24$ and $\alpha = 0.05$, we have $T_{\alpha} = 81$. Since T^{+} and $T^{-}$are both greater than T_{α}, there is not sufficient evidence to reject H_{0}.

In the case of the one-sided alternative $H : \theta < 9.9\text{cm}$. ($H : \theta > 9.9\text{cm}$.), we shall compare $T^{+} = 153.5$ $\left(T^{-} = 146.5\right)$with the critical value $T_{\alpha} = 81$, at $\alpha = 0.025$, and arrive at same conclusion that there is no ground for rejecting H_{0} (in favour if the appropriate on-sided alternative) since $T > T_{\alpha}$.

Example4. The weights of 12 persons before they are subjected to a change of diet and after a lapse of six months are recorded below:

S.no.	Weight (in kg.)	
	Before	After

1	57	62
2	48	55
3	55	62
4	45	53
5	62	59
6	42	45
7	49	45
8	60	55
9	65	64
10	51	55
11	46	50
12	58	66

Test whether there has been any significant gain in weight as a result of the change of diet.

Solution: Let y and x be the weight of a person before and after the change of diet, then the hypothesis to be tested is $H_0 : \theta = 0$ and the alternative is $H : \theta > 0$, where θ is the median of the distribution of differences d_i. The gain in weight $\left(d_i\right)$ and the absolute rank for 12 persons are:

S.no.	Weight (in kg.)		$d_i = x_i - y_i - \theta_0$	Rank of $\lvert d_i \rvert$
	y_i	x_i		

1	57	62	+5	6.5
2	48	55	+7	9.5
3	55	62	+7	9.5
4	45	53	+8	11.5
5	62	59	-3	2.5
6	42	45	+3	2.5
7	49	45	-4	4.5
8	60	55	-5	6.5
9	65	64	-1	1
10	51	55	+4	4.5
11	46	50	+6	8
12	58	66	-8	11.5

Here, $T^{+} = 52$ and $T^{-} = 26$; here T^{-} will be used. From table, we have, for $n = 12$ and $\alpha = 0.01$(one-sided), $T_{\alpha} = 10$. Since $T^{-} > T_{\alpha}$, therefore we conclude that there is no sufficient evidence to reject the null hypothesis that there is no effect of diet in favour of the alternative hypothesis at the 1% level.

3.5.4 Merits and Demerits

The Wilcoxon Signed rank test takes into account the magnitude of the deviations.

As one of the assumption made here is that intendance of observations continuity everywhere and symmetry which is not practically possible all the time.

Comparision of Sign test and Wilcoxon signed rank test

1. In sign test, the assumptions required are independence of observations and the population is continuous at media. In Wilcoxon signed rank test, the assumptions required are the population is continuous everywhere and it is symmetric about median.
2. In sign test, we consider only the directions of the deviations while in Wilcoxon signed rank test, we consider directions of the deviations as well as the magnitudes of the directions. Thus Wilcoxon signed rank test is more efficient than the sign test.
3. Both the tests are useful generally for the same type of problem. But only Wilcoxon signed test is suitable for a test of symmetry as well.

3.6 Median Test

If $N = m + n$ is even then

Median = any number between $\frac{N}{2}$th and $\left(\frac{N+2}{2}\right)$th order statistic

Let U be the number of X sample observations that are less than the sample median for the combined sample.

The test based on U, the number of observations from X sample median which are less than the combined sample median, is called the sample median. Then

$$t=\begin{cases}\dfrac{N-1}{2}, & if \text{ N is odd}\\ \dfrac{N}{2}, & if \text{ N is even}\end{cases}$$

The probability distribution of U for fixed t is

$$f(u)=\frac{{}^{m}C_{u}\,{}^{n}C_{t-u}}{{}^{m+n}C_{t}};\ u=0,1,2,........,t$$

where $t=\dfrac{N}{2}$.

If H_0 is true, then $P(X<M)=P(X>M);\ \forall M$ and here M is combined sample median. i.e. the two populations have a common median which is estimated by M.

3.6.1 Hypothesis and Assumptions

The general location alternative is

$$H_L : F_X(x)=F_Y(x-\theta);\ \text{ for some } x\ \&\ \theta\neq 0$$

if U is too large, then

$H_L : F_X(x)\geq F_Y(x);$ if $\theta>0$ and $\forall x$

i.e. $H_L : F_X(x)>F_Y(x);$ if $\theta>0$ and for some x

i.e. the median of the X population is smaller than the median of Y population.

If U is too small, then

$H_L : F_X(x)\leq F_Y(x);$ if $\theta<0$ and $\forall x$

i.e. $H_L : F_X\left(x\right) < F_Y\left(x\right);$ if $\theta < 0$ and for some x

i.e. the median of the X population is greater than the median of Y population.

The critical region for α level of significance is given as

Alternative Hypothesis	Critical Region
$\theta > 0$ or $M_X < M_Y$	$u \geq c_{\alpha}^{'}$
$\theta < 0$ or $M_X > M_Y$	$u \leq c_{\alpha}$
$\theta \neq 0$ or $M_X \neq M_Y$	$u \leq c_{\frac{\alpha}{2}}$ or $u \geq c_{\frac{\alpha}{2}}^{'}$

3.6.2 Test Procedure

1. Consider the observations in the order in which they are obtained.
2. Determine the median of those observations i.e. determine the sample median M.
3. For each observation note that whether it is above or below the sample median. Denote the observation below the sample median M by B or (-) sign and observations above the sample median M by A or (+) sign . The zero values will be ignored.
4. Denote the number of minus signs or the numbers by B's by n_1 and the number of plus signs or the number of A's by n_2 .
5. Count the number of runs and denote this number by R.
6. Reject the null hypothesis H_0 the sample is random if

   $R \geq R_1$ or $R \leq R_u$.

where R_1 and R_u are critical of R to be determined from the distribution of R n_1 and n_2. The critical values of R required for significance have been have been tabulated.

3.6.3 Example

Suppose in a random sample of size 30, there 12 runs above and below the sample median where n_1=number of minus (-) sings=10

n_2= number of minus (+) sings= 20

Test the hypothesis the sample is random.

Solution

R= Number of runs above and below the sample median =12

n_1=number of minus (-) sings=10

n_2= number of minus (+) sings= 20

from table the lower critical value of R, R_1=9

the upper critical values of R, R_u=20

since $9 < R < 20$

the hypothesis of randomness is accepted at 5% level of signifance. i.e. sample is random.

3.6.4 Merits and Demerits of Median test

Median test when the sample observations are divided into two types on the basis of deviations from sample median.

UNIT 4: OTHER NON- PARAMETRIC TESTS

ONE SAMPLE AND TWO SAMPLE LOCATION TEST

4.4 Mann-Whitney U Test

4.4.1 Introduction and Assumptions

Mann Whitney U test is a non-parametric test for testing that the two populations differ in their location. It is useful to the t-test, if the assumption of t-test are violated, we use Mann Whitney U test. We assume that the two samples are drawn from continuous populations.

Let we have two populations X and Y with cumulative distribution functions F_X and F_Y respectively. A random sample of size m is drawn from the X population and another random sample of size n is drawn from the Y population, denoted as

$$X_1, X_2, ...X_m \text{ and } Y_1, Y_2, ...Y_n$$

These $N = m + n$ observations drawn from two populations are arranged in order of magnitude from smallest to largest.

Like run test, this test is based on the idea that the particular pattern is exhibited when m observations of X random variable and n observations of Y random variables are arranged together in increasing order of magnitude.

The test criterion is based on the positions of Y's in the combined ordered sequence. A sample pattern where most of the Y's are greater than the most of the X 's or vice-versa can be used as statistical criteria for rejection of null hypothesis of identical distribution.

Since, in this case, we see that there is no random missing in the sample observation. The Mann Whitney U statistic is defined as the number of times Y proceeds X in the combined ordered arrangement of two independent random samples.

4.4.2 Test Procedure

If mn random variable are defined as

$$D_{ij} = \begin{cases} 1 \quad if\ Y_j < X_i; \quad \forall i = 1, 2, \ldots.., m \\ 0 \quad if\ Y_j > X_i; \quad \forall j = 1, 2, \ldots.., n \end{cases}$$

Thus Mann-Whitney U statistic is defined as

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

We wish to test the null hypothesis

$$H_0 : F_X(x) = F_Y(x) \ ; \ \forall x$$

i.e. two samples are drawn from the identical populations.

The general location alternative is

$H_L : F_X(x) = F_Y(x - \theta);$ for some x & $\theta \neq 0$

If U is too large, then

$F_X(x) \geq F_Y(x);$ $\forall x$ and if $\theta > 0$

i.e. $F_X(x) > F_Y(x);$ for some x if $\theta > 0$

If U is too large, then

$F_X(x) \leq F_Y(x);\ \forall x \text{and if} \theta < 0$

i.e. $F_X(x) < F_Y(x); \text{for some} x \text{if } \theta > 0$

We define,

$$\pi = P(D_{ij} = 1) = P(Y < X)$$

$$= P[-\infty < X < \infty, -\infty < Y < X]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(y)\, f(x)\, dy\, dx$$

$$\pi = \int_{-\infty}^{\infty} F_Y(x)\, f(x)\ dx$$

Under H_0, i.e. $H_0 : F_X(x) = F_Y(x)$

Then

$$\pi = \int_{-\infty}^{\infty} F_X(x)\, f(x)\ dx$$

For solving above integration, let $F_X(x) = v$ and differentiate this equation w.r.t x, we get $f(x) = dv$. Also, the limits changes as $x = -\infty \Rightarrow F(-\infty) = 0 = v$ and $x = \infty \Rightarrow F(\infty) = 1 = v$

Therefore, integral reduces to

$$\pi = \int_{0}^{1} v\ dv = \left[\frac{v^2}{2}\right]_0^1 = \frac{1}{2}$$

Hence $H_0 : F_X(x) = F_Y(x)$ or $H_0 : \pi = \frac{1}{2}$

Also $H_L : F_Y(x) \geq F_X(x)$; $\forall x$

Is equivalent to $H_L : \pi \geq \frac{1}{2}$; $\forall x$

i.e. $H_L : F_Y(x) > F_X(x)$; for some x

is equivalent to $H_L : \pi > \frac{1}{2}$; for some x

and $H_L : F_Y(x) \leq F_X(x)$; $\forall x$

is equivalent to $H_L : \pi \leq \frac{1}{2}$; $\forall x$

i.e. $H_L : F_Y(x) < F_X(x)$; for some x

is equivalent to $H_L : \pi < \frac{1}{2}$; for some x

The mn random variables D_{ij} are Bernoulli variables, with parameter π. i.e.

$$E\left[D_{ij}\right] = E\left[D_{ij}^2\right] = \pi$$

$$V\left[D_{ij}\right] = \pi(1-\pi)$$

We define the parameters π_1 and π_2 as,

$$\pi_1 = P\left(Y_j < X_i \cap Y_k < X_i\right) = \int_{-\infty}^{\infty} \left[F_Y(x)\right]^2 f(x)dx$$

and

$$\pi_2 = P\left(X_i > Y_j \cap X_h < X_j\right) = \int_{-\infty}^{\infty} \left[1 - F_Y(x)\right]^2 f(y)dy$$

Since $U = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}$

Therefore mean and variance of U are defined as

$$E[U] = \sum_{i=1}^{m}\sum_{j=1}^{n} E\left[D_{ij}\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} \pi$$

$$E[U] = mn\pi$$

and

$$V[U] = V\left(\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}\right)$$

$$= mn\pi(1-\pi) + mn(n-1)\left(\pi_1 - \pi^2\right) + mn(m-1)\left(\pi_2 - \pi^2\right)$$

$$= mn\left[\pi - \pi^2 + (n-1)\left(\pi_1 - \pi^2\right) + (m-1)\left(\pi_2 - \pi^2\right)\right]$$

$$= mn\left[\pi - \pi^2 + (m+n-1) + (n-1)\pi_1 + (m-1)\pi_2\right]$$

$$V[U] = mn\left[\pi - \pi^2 + (N-1) + (n-1)\pi_1 + (m-1)\pi_2\right]$$

as $m,n \to \infty$

$E[U/mn] = \pi$ and $V[U/mn] \to 0$

Hence U/mn is a consistent estimator of π.

If we define the another random variable

$$U^{'} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(1 - D_{ij}\right)$$

The critical region for α level of significance is given as

Alternative hypothesis	Critical region
$\pi \le \frac{1}{2}$ or $F_Y(x) \le F_X(x)$	$U \le C_\alpha$
$\pi \ge \frac{1}{2}$ or $F_Y(x) \ge F_X(x)$	$U' \le C_\alpha$
$\pi \ne \frac{1}{2}$ or $F_Y(x) \ne F_X(x)$	$U \le C_{\frac{\alpha}{2}}$ or $U' \le C_{\frac{\alpha}{2}}$

Under H_0, i.e. $H_0 : F_X(x) = F_Y(x)$

Then $\pi = \frac{1}{2}$

and $\pi_1 = \pi_2 = \frac{1}{3}$

Thus $E[U] = \frac{mn}{2}$

$$V[U] = mn\left[\frac{1}{2} - \frac{1}{4}(N-1) + \frac{(n-1)}{3} + \frac{(m-1)}{3}\right]$$

$$V[U] = mn\left[\frac{1}{12} - \frac{N}{4} + \frac{1}{3}(n+m)\right]$$

$$= mn\left[\frac{1}{12} - \frac{N}{4} + \frac{N}{3}\right]$$

$$V[U] = \frac{mn(N+1)}{12}$$

If N is large, then under H_0

$$Z = \frac{U - E[U]}{\sqrt{V[U]}} = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \to N(0,1)$$

4.4.3 Example. The following are the marks secured by two batches of salesmen in the final test taken after completion of training. Use the U-test with $\alpha = 0.02$ for the null hypothesis that the samples are drawn from identical distributions against the alternative that the distributions differ in location only.

Batch A: 28, 25, 27, 29, 25, 19, 23, 26, 30, 22, 21, 28

Batch B: 20, 24, 25, 26, 18, 28, 23

Solution: Here $n_1 = 7$, $n_2 = 12$ and $N = n_1 + n_2 = 12$

$U = 51,\ U^{'} = 26$

where U is the number of times x_i precedes y_j among all (x_i, y_j)pairs and $U^{'}$ is the number of times y_j precedes x_i among all (x_i, y_j)pairs assuming no $x = y$ ties. From table, we find that for two-tail test $n_1 = 7$ and $n_2 = 12$ at the level 0.02, the critical value is 14. Since 20(the smaller of U and $U^{'}$) is greater than 14, so we have no reason to believe that the samples are not drawn from identical distribution.

4.4.4 Merits and Demerits

It is a good substitute for t-test when the conditions imposed on parent populations are not met .

4.4.5 Application of U-statistic to rank tests

TEST OF GOODNESS OF FIT

This type of test are designed for a null hypothesis which is a statement about the form of the cumulative distribution function or probability function of the parent population from which the sample is drawn.

Let a random sample of size n is drawn from a population with unknown cumulative distribution function say F. We want to test the null hypothesis

$$H_0 : F(x) = F_0(x)\,;\ \forall x$$

against the alternative hypothesis

$$H_1 : F(x) \neq F_0(x)\,;\ \text{for some } x$$

If F_0 is specified with all its parameters, then H_0 is a simple hypothesis. If F_0 is not completely specified, then H_0 is a composite hypothesis and the unknown parameters are to be estimated from the sample data in order to perform any test. The alternative hypothesis in both the cases will be composite therefore rejection of H_0 does not provide any result.

4.5 One sample Kolmogorov-Smirnov (K-S) Test

Goodness of fit tests are used when only the form of the population is in question, with the hope that the null hypothesis will be found accepted. The two types of goodness of fit tests are:

1. Chi Square goodness of fit test
2. Kolmogrov Siminirov test

Chi Square goodness of fit test

4.5.1 Hypothesis and Assumptions

If a random sample of size n is drawn from a population with unknown cumulative distribution function F.

We wish to test the null hypothesis

$H_0 : F(x) = F_0(x)\,;\ \forall x$

against the alternative hypothesis

$H_1 : F(x) \neq F_0(x)\,;$ for some x

In order to apply the chi-square test in continuous distribution, the sample data must be grouped according to some scheme in order to form a frequency distribution.

Assuming that the population distribution F_0 is completely specified by the null hypothesis H_0 , we can obtain the probability p_i that a random observation will be classified in the i^{th} category $(i = 1, 2, ..., k)$.

These probabilities multiplied by n, the sample size, give the expected frequencies under H_0. i.e.

$$E_i = np_i,\ (i = 1, 2, ..., k)$$

Let the n observations have been grouped into k mutually exclusive categories, O_i and E_i are the observed and expected frequencies respectively, for the i^{th} group $(i = 1, 2, ..., k)$.

We use the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \tag{1}$$

with $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$

The exact sampling distribution of this test statistic is complicated. But for large samples, it has χ^2 distribution with $(k-1)$ degree of freedom. This approximation is good for every $E_i \geq 5$.

For $E_i < 5$, we combine the adjacent categories till the expected frequency in the combined category is at least 5.

If

$$cal\, \chi^2 > tab\, \chi^2_{\alpha,(k-1)}$$

then H_0 is rejected at α level of significance.

If F_0 is completely specified, then H_0 is a composite hypothesis and the unknown parameters are to be estimated from the sample data in order to perform the test.

In this case, the test statistic described by (1) has χ^2 distribution with $(k-r-1)$ degree of freedom, where r is the number of independent parameters of F_0 estimated from the sample data.

Thus $mS_m(x)$ is the number of X sample observations that are less than or equal to x. And $nT_n(x)$ is the number of Y sample observations that are less than or equal to x.

For large m and n, the deviations between two empirical distribution functions, $\left|S_m(x)-T_n(x)\right|$ should be small for all values of x.

Thus the test statistic

$$D_{m,n} = \max_x \left|S_m(x)-T_n(x)\right|$$

is called Kolmogrov-Smirnov two sample test statistic.

The probability distribution of $D_{m,n}$ does not depend upon F_X and F_Y as long as F_X and F_Y are continuous.

Therefore, $D_{m,n}$ may be called a distribution free statistic.

The directional deviations are defined as

$$D_{m,n}^{+} = \max_x \left|S_m(x)-T_n(x)\right|$$

$$D_{m,n}^{-} = \max_x \left|T_n(x)-S_m(x)\right|$$

$D_{m,n}^{+}$ and $D_{m,n}^{-}$ are called one-sided kolmogrov-smirnov statistic. These are also distribution free.

We wish to test the null hypothesis

$$H_0 : F_X\left(x\right) = F_Y\left(x\right) \;\; ; \;\; \forall x$$

i.e. under H_0, the population distributions are identical and we have two samples from the sample population.

against

(i) One sided alternative

$H_1 : F_X\left(x\right) > F_Y\left(x\right) \;\; ; \;\; \forall x$ (right tailed test)

The appropriate test statistic is

$$D^{+}_{m,n} = \max_x \left| S_m\left(x\right) - T_n\left(x\right) \right|$$

or

$H_1 : F_X\left(x\right) < F_Y\left(x\right) \;\; ; \;\; \forall x$ (left tailed test)

The appropriate test statistic is

$$D^{-}_{m,n} = \max_x \left| T_n\left(x\right) - S_m\left(x\right) \right|$$

Comparison of Chi-square test with Kolmogrov-Siminirov test for goodness of fit

1. Both types of tests are distribution free because the sampling distribution of the test statistic does not depend on the cumulative distribution function.
2. The chi-square tests are specially designed for use with categorical data, while K-S tests are for random samples from the continuous populations.
3. The chi-square test is sensitive to vertical deviations between the observed and expected histograms, whereas the K-S test is based on vertical deviations between the observed and expected cumulative distribution functions.
4. K-S test can be applied for any sample size, while chi-square test can be applied for large sample size when each expected cell frequency is not too small.
5. The advantage of K-S test is that the exact sampling distribution of K-S test statistic is known and tabulated, whereas the sampling distribution of chi-square test statistic is approximately chi-square for finite sample size.
6. When H_0 is composite, the chi-square test is easily modified by reducing the number of degrees of freedom (as some parameters are estimated) while K-S test can't be modified in the situation.
7. The K-S test is more powerful and more flexible than the chi-square test.
8. The chi-square test also comes in the category of parametric tests whereas K-S test is only a non-parametric.

9. In K-S test, we can use one side test also which is not possible chi-square test.

4.6 Two sample Kolmogorov-Smirnov Test

4.6.1 Hypothesis and Assumptions

Suppose a random variable is continuously distributed in each of two populations, the distribution functions being denoted by F and G. Further, suppose that independent random samples, say $x_1, x_2, x_3, \dots\dots\dots, x_m$ and $y_1, y_2, y_3, \dots\dots\dots, y_n$ have been drawn from the two continuous distribution F_m and G_n respectively.

Here our problem is to test the hypothesis that the to distribution are identical i.e.

$$H_0 : F(\theta) = G(\theta)$$

against $$H_1 : F(\theta) \neq G(\theta); \qquad \forall t$$

Then an appropriate test criterion for testing hypothesis is K-S statistic which is as follows

$$D_{mn} = \max_{-\infty<t<\infty} \left| F_m(\theta) - G_n(\theta) \right|$$

If the hypothesis is true , one expects the value of D_{mn} to be small, while a large value of D_{mn} may be taken as an indication that the parent distributions are not identical.

4.7 Run test

Runs

If we are given an ordered sequence of two or more types of symbols, a run is defined to be a succession of one or more identical symbol which are followed and proceed by a different symbol or no symbol at all.

In any situation, if the sample observations may not behave random, it is necessary to test the randomness of the sequence before the usual statistical methods based on randomness are applied.

Too few runs, too many runs, a run of excessive length or too many runs of excessive length etc. can be used as statistical criteria for rejection of the null hypothesis of randomness, since these situations should occur rarely in a truly random sequence.

A null hypothesis of randomness would be rejected if the total number of runs is either too small or too large.

Advantages

1. Test of randomness are an important addition to the statistical theory, because almost all the classical statistical techniques are based on the assumption of a random sample.
2. The run tests are applicable to either qualitative or quantitative data.

Distribution of Runs

Let us suppose an ordered sequence of n elements of two types, n_1 of the first type i.e. values of x and n_2 of the second type i.e. the values of y such that $n_1 + n_2 = n$.

If $r_1 = \text{number of runs of type 1st elements i.e. X's}$

$r_2 = \text{number of runs of type 2nd elements i.e. Y's}$

The total number of runs in this sequence is

$r_1 + r_2 = r \quad ; \; r \leq n$

The probability distribution of the random variable 'R' is obtained as follows:

We can select n_1 positions for the n_1 values of X from $(n_1 + n_2)$ positions in ${}^{n_1+n_2}C_{n_1}$ ways.

The probability of each arrangement $= \dfrac{1}{{}^{n_1+n_2}C_{n_1}}$

Now, we have to determine how many of these arrangements yield $R = r$. Here, two cases arise:

Case (i):- When r is odd i.e. $r = 2k+1$; $k \in I^+$ i.e. there are $(k+1)$ runs of ordered values of X and k runs of ordered values of Y or vice-versa.

First we consider the number of ways of obtaining $(k+1)$ runs of n_1 values of X. This can be done in ${}^{n_1-1}C_k$ ways.

Similarly, we consider the number of ways of obtaining k runs of n_2 values of Y. This can be done in ${}^{n_2-1}C_{k-1}$ ways.

The joint operation can be performed in $\left({}^{n_1-1}C_k\right)\left({}^{n_2-1}C_{k-1}\right)$ ways.

Secondly, considering the number of ways of obtaining $(k+1)$ runs of n_2 values of Y. This can be done in ${}^{n_2-1}C_k$ ways.

Similarly, we consider the number of ways of obtaining k runs of n_1 values of X . This can be done in ${}^{n_1-1}C_{k-1}$ ways.

The joint operation can be performed in $\left({}^{n_1-1}C_{k-1}\right)\left({}^{n_2-1}C_k\right)$ ways.

Thus,

$$P\left(r=2k+1\right)=\frac{\left({}^{n_1-1}C_k\right)\left({}^{n_2-1}C_{k-1}\right)+\left({}^{n_1-1}C_k\right)\left({}^{n_2-1}C_{k-1}\right)}{\left({}^{n_1+n_2}C_{n_1}\right)}$$

Case (ii):- When r is even i.e. $r=2k\ ;k\in I^+$ i.e. there are k runs of ordered values of X and k runs of ordered values of Y or vice-versa.

First we consider the number of ways of obtaining k runs of n_1 values of X . This can be done in ${}^{n_1-1}C_{k-1}$ ways.

Similarly, we consider the number of ways of obtaining k runs of n_2 values of Y . This can be done in ${}^{n_2-1}C_{k-1}$ ways.

The joint operation can be performed in $\left({}^{n_1-1}C_{k-1}\right)\left({}^{n_2-1}C_{k-1}\right)$ ways.

Secondly, considering the number of ways of obtaining k runs of n_2 values of Y . This can be done in ${}^{n_2-1}C_{k-1}$ ways.

Similarly, we consider the number of ways of obtaining k runs of n_1 values of X . This can be done in ${}^{n_1-1}C_{k-1}$ ways.

The joint operation can be performed in $\left({}^{n_1-1}C_{k-1}\right)\left({}^{n_2-1}C_{k-1}\right)$ ways.

Thus, $P(r=2k)=\dfrac{2\left({}^{n_1-1}C_{k-1}\right)\left({}^{n_2-1}C_{k-1}\right)}{\left({}^{n_1+n_2}C_{n_1}\right)}$

Thus the probability distribution of R, the total number of runs of $n_1+n_2=n$ objects, n_1 of type 1st and n_2 of type 2nd, is given as:

$$f(x)=\begin{cases}\dfrac{2\left({}^{n_1-1}C_{r/2-1}\right)\left({}^{n_2-1}C_{r/2-1}\right)}{\left({}^{n_1+n_2}C_{n_1}\right)} & ;\ \text{if } r \text{ is even}\\ \dfrac{\left({}^{n_1-1}C_{r-1/2}\right)\left({}^{n_2-1}C_{r-3/2}\right)+\left({}^{n_1-1}C_{r-3/2}\right)\left({}^{n_2-1}C_{r-1/2}\right)}{\left({}^{n_1+n_2}C_{n_1}\right)} & ;\ \text{if } r \text{ is odd}\end{cases}$$

where $r=2,3,\ldots,n_1+n_2$

Test of Randomness

Sometimes, it is desirable to test whether the sample observations can be regarded as random or not. To test the randomness of the sample observations, we use run test.

Let $X_1, X_2, \ldots., X_n$ be a random sample of size n taken from continuous distribution. In the given sequence $X_1, X_2, \ldots., X_n$ for each observation we note whether it is above or below the sample median.

4.7.1 Hypothesis and Assumptions

Run test is used for examining whether or not a set of observations constitutes a random sample from an infinite population. Test for randomness is of major importance because the assumption of randomness underlies statistical inference. In addition, tests for randomness are important for time series analysis. Departure from randomness can take many forms.

H_0: Sample values come from a random sequence

H_1: Sample values come from a non-random sequence.

4.7.2 Test Procedure

Let r be the number of runs (a run is a sequence of signs of same kind bounded by signs of other kind). For finding the number of runs, the observations are listed in their order of occurrence. Each observation is denoted by a '+' sign if it is more than the previous observation and by a '-' sign if it is less than the previous observation. Total number of runs up (+) and down (-) is counted. Too few runs indicate that the sequence is not random (has persistency) and too many runs also indicate that the sequence is not random (is zigzag).

Critical Value: Critical value for the test is obtained from the table for a given value of n and at desired level of significance (α). Let this value be r_c.

Decision Rule: If r_c (lower) $\leq r \leq$ r_c (upper), accept H_0. Otherwise reject H_0.

Tied Values: If an observation is equal to its preceding observation denote it by zero. While counting the number of runs ignore it and reduce the value of n accordingly.

Large Sample Sizes: When sample size is greater than 25 the critical value r_c can be obtained using a normal distribution approximation.

The critical values for two-sided test at 5% level of significance are

r_c (lower) $= \mu - 1.96\sigma$

r_c (upper) $= \mu + 1.96\sigma$

For one-sided tests, these are

r_c (left tailed) $= \mu - 1.65\sigma$, if $r \leq r_c$, reject H_0

r_c (right tailed) $= \mu + 1.65\sigma$, if $r \geq r_c$, reject H_0,

where

$$\mu = \left(\frac{2n-1}{3}\right) \text{ and } \sigma = \sqrt{\left(\frac{16n-29}{90}\right)}$$

4.7.3 Example

Data on value of imports of selected agricultural production inputs from U.K. by a county (in million dollars) during recent 12 years is given below: Is the sequence random?

5.2 5.5 3.8 2.5 8.3 2.1 1.7 10.0 10.0 6.9 7.5 10.6

Solution:

H_0: the sequence is random.

H_1: the sequence is not random.

5.2 5.5 3.8 2.5 8.3 2.1 1.7 10.0 10.0 6.9 7.5 10.6

+ - - + - - + 0 - + +

Here n = 11, the number of runs r = 7. Critical n values for α = 5% (two sided test) from the table are r_c (lower) = 4 and r_c (upper) = 10.

Since r_c (lower) $\leq r \leq$ r_c (upper), i.e., observed r lies between 4 and 10, H_0 is accepted. The sequence is random.

4.7.4 Merits and Demerits

The number of runs a sequence indicative of randomness.

any set patterns of symbols in a sequence shows lack of randomness .

Too many or too less runs show lack of randomness.

4.8 Pitman ARE. (Asymptotic Relative Efficiency)

The Relative efficiency depends on the choice of α, the choice of β and the particular being considered if H_1 is composite. In order to provide an overall comparison of one test with another it is clear that relative efficiency leaves much to be desired. We would prefer a comparison test that does not depend on our choice of α, β or a particular alternative possible under H_1 if H_1 is composite, which it usually is. One way this sometimes may be accomplished is described briefly as follows,

Consider a sequence of tests , all with the same fixed α. If the sequence of tests is consistent , β will become smaller as the sample size n_1 gets larger. Instead of allowing β to become smaller, we would consider a different alternative each time (under the composite alternative hypothesis) for each different value of n_1 where, each time, the alternative considered is one that allows β to remain constant from test to test. Thus, as n_1 becomes larger, α and β remain fixed and the alternative being considered varies.

For each value of n_1 a value of n_2 is calculated so the second test has the same α and β under the alternative considered. Then there is a sequence of values of relative efficiency n_2/n_1, one for each test in the original sequence of tests. If n_1/n_2 approaches a constant as n_1 becomes large, and if that constant is the same no matter which values of α and β are being used, then that constant is called the Asymptotic Relative Efficiency of the first test to be second test or more correctly the first sequence of tests to the second sequence of tests. Sometimes the name Pitman's efficiency is used for this definition of asymptotic relative efficiency to distinguish it from other definitions of asymptotic relative efficiency.

Definition: Let n_1 and n_2 be the sample sizes required for two tests T_1 and T_2 to have the same power under the same level of significance. If α and β remain fixed, the limit of n_2/n_1 as n_1 approaches infinity is called the asymptotic relative efficiency (A.R.E.) of the first test to the second test, if that limit is independent of α and β.

A book by Noether (1976 a) contains many of the more important results of studies of A.R.E. See also Stuart (1954) and Ruist (1955) for further discussions.

Thus A.R.E. often provides a compact summary of the relative summary of the relative efficiency between two tests.

Two Sample Problem

In two sample problem, we are concerned with the data which consists of two independent random samples; i.e. random samples are drawn independently from each of two populations. Not only the elements within each sample are independent, but also every element in the first sample is independent of every element in the second sample.

We have two populations called as X and Y populations, with cumulative distribution functions F_X and F_Y respectively.

A random sample $X_1, X_2, ...X_m$ of size m is drawn from the population X and another random sample $Y_1, Y_2, ...Y_n$ of size n is drawn from the population Y.

Generally the hypothesis of interest in two sample problem is that the two samples are drawn from the identical populations. i.e.

$$H_0 : F_X(x) = F_Y(x) \ ; \ \forall x$$

We shall discuss three types of alternatives:-

(a) In the first type of alternative, we consider the alternative hypothesis that the two populations differ in any manner i.e. the two populations may differ in location or in dispersion or in skewness or in kurtosis etc.

(i) The two sided alternative is

$$H_1 : F_X(x) \neq F_Y(x) \ ; \ \text{for some}\, x$$

(ii) A one sided alternative is

$$H_1 : F_X(x) \leq F_Y(x) \ ; \ \forall x$$

i.e. $H_1 : F_X(x) < F_Y(x)$; for some x

i.e. the variable is stochastically larger than the variable Y.

or

$H_1 : F_X(x) \geq F_Y(x)$; $\forall x$

$H_1 : F_X(x) > F_Y(x)$; for some x

i.e. the variable is stochastically smaller than the variable Y.

For this type of problem, we shall discuss the following tests:-

1. Wald-Wolfowitz Run Test
2. Kolmogrov-simirnov two sample Test

(b) In the second type of alternatives, we consider the alternative hypothesis that the two populations differ in location only, this type of alternative is called the location alternative.

$$H_L : F_X(x) = F_Y(x-\theta); \text{ for some } x \;\&\; \theta \neq 0$$

i.e. the cumulative distribution function of Y is shifted to left if $\theta < 0$

i.e. $F_X(x) \leq F_Y(x)$; $\forall x$ or $F_X(x) < F_Y(x)$; for some x

and

the cumulative distribution function of Y is shifted to right if $\theta > 0$

i.e. $F_X(x) \geq F_Y(x)$; $\forall x$ or $F_X(x) > F_Y(x)$; for some x

For this type of problem, we shall discuss the following tests:-

1. Median Test
2. Mann-Whitney U Test

3. Wilcoxon Test

(c) In the third type of alternative hypothesis, we consider the alternative hypothesis that the two populations differ in scale parameter only, this type of alternative is called the scale alternative.

$$H_S : F_X\left(\theta x\right) = F_Y\left(x\right); \text{ for some } x \;\&\; \theta \neq 1$$

i.e. the cumulative distribution function of Y is with compressed scale if $\theta > 1$ and the cumulative distribution function of Y is with enlarged scale if $\theta < 1$.

For this type of problem, we shall discuss the following tests:-

1. Mood Test
2. Sukhatme Test

Wald-Wolfowitz Run Test

This two sample test is based on the assumption that the populations under consideration are continuous.

We wish to test the hypothesis that the two independent samples have been drawn from the identical populations against the alternative that the two populations differ in any manner i.e. in location, in dispersion, in skewness or in kurtosis etc.

Let $X_1, X_2, ...X_m$ and $Y_1, Y_2, ...Y_n$ be two random samples of sizes m and n respectively drawn from two populations. These $N = m + n$ observations drawn from two populations are arranged in order of magnitude from smallest to largest, keeping in view which of the observations correspond to the X sample and which to Y sample.

For example, with $m=4$ & $n=5$, the arrangements might be

$XYYXXYXYY$, $m+n=9$.

We have 6 runs, 3 runs of $X's$ and $Y's$.

The total number of runs in the ordered pooled sample is indicative of the degree of random mixing. We wish to test the null hypothesis

$H_0 : F_X(x) = F_Y(x) \;\; ; \;\; \forall x$

against $H_1 : F_X(x) \neq F_Y(x) \;\; ; \;\; \text{for some}\, x$

where F_X & F_Y are the cumulative distribution functions of the populations.

Let r be the total number of runs in the group of N observations.

A run is defined to be a succession of one or more identical symbols which are followed and proceed by a different symbol or no symbol at all.

Under H_0, the two samples are drawn from the same population. i.e. Under H_0, the two samples are expected to be well mixed and r is expected to be large.

But r is small, if the two samples come from the different populations. i. e. if H_0 is fase.

If all the values of Y are greater than all the values of X (or vice-versa), then there will be only two runs.

Since too few runs will provide the critical region (or rejection region for null hypothesis H_0).

The Wald-Wolfowitz run test for α level of significance has the critical region

$$r \leq r_\alpha$$

where r_α is the largest integer such that

$$P\left[r \le r_\alpha / H_0\right] \le \alpha$$

If H_0 is true, then all the ${}^{m+n}C_n = {}^{m+n}C_m$ different possible arrangements of m $X's$ and n $Y's$ in a line are equally likely.

When r is odd i.e. $r = 2k+1$; $k \in I^+$.i.e. there are $(k+1)$ runs of ordered values of X and k runs of ordered values of Y or vice-versa. Then,

$$P\left[r = 2k+1 / H_0\right] = \frac{\left({}^{m-1}C_k\right)\left({}^{n-1}C_{k-1}\right) + \left({}^{m-1}C_{k-1}\right)\left({}^{n-1}C_k\right)}{{}^{m+n}C_m}$$

When r is even i.e. $r = 2k$; $k \in I^+$.

i.e. there are k runs of ordered values of X and k runs of ordered values of Y or vice-versa. Then,

$$P\left[r = 2k / H_0\right] = \frac{2\left({}^{m-1}C_k\right)\left({}^{n-1}C_{k-1}\right)}{{}^{m+n}C_n}$$

Under H_0, the mean and variance of r are given as

$$E\left[r\right] = \frac{2mn}{m+n} + 1$$

$$V\left[r\right] = \frac{2mn\left(2mn - m - n\right)}{\left(m+n\right)^2\left(m+n-1\right)}$$

For large m,n under H_0

$$Z = \frac{r - E\left[r\right]}{\sqrt{V\left[r\right]}} \square N\left(0,1\right)$$

Note: It is the test for equality of distributions based on runs.

Rank Order Statistics

If the rank order statistics of a random sample $X_1, X_2, \ldots\ldots, X_n$ are denoted by $r(x_1), r(x_2), \ldots\ldots, r(x_n)$.

The i^{th} rank order statistic $r(x_i)$ is called the rank of the i^{th} observation in the unordered sample.

Ex: $r(x_i) = i$

The functional definition of the rank of any x_i in a set of n observations is given as,

$$r(x_i) = \sum_{j=1}^{n} S(x_i - x_j)$$

where $S(u) = \begin{cases} 1 & ;\text{if } u \geq 0 \\ 0 & ;\text{if } u \geq 0 \end{cases}$

Linear Rank Statistics

If the two independent random samples $X_1, X_2, \ldots\ldots, X_m$ and $Y_1, Y_2, \ldots\ldots, Y_n$ are drawn from the two populations with cumulative distribution functions F_X and F_Y respectively.

We consider the null hypothesis

$H_0 : F_X(x) = F_Y(x)$; $\forall x, F$ unknown

The set of $m + n = N$ observations are assigned ranks $1, 2, \ldots\ldots, N$.

The functional definition of the rank of observations in the combined sample (with no ties) is given as,

$$r(x_i)=\sum_{j=1}^{m}S(x_i-x_j)+\sum_{j=1}^{n}S(x_i-y_j)$$

$$r(y_i)=\sum_{j=1}^{n}S(y_i-y_j)+\sum_{j=1}^{n}S(y_i-x_j)$$

where $S(u)=\begin{cases}1 & ;\text{if } u\geq 0\\ 0 & ;\text{if } u\geq 0\end{cases}$

we denote the combined ordered sample by a vector of indicator random variables as follows-

Let $Z=(z_1,z_2,.....,z_N)$ be the combined ordered sample. Then we describe

$$z_i=\begin{cases}1 & ;\text{if i}^{\text{th}}\text{random variable in the combined ordered sample is X}\\ 0 & ;\text{if i}^{\text{th}}\text{ random variable in the combined ordered sample is } Y\end{cases}\ ;\ \forall i=1,2,.....,N$$

The vector Z indicates the rank order statistics of the combined samples. The linear rank order statistics is defined as

$$T_N=\sum_{i=1}^{N}a_iz_i$$

Where a_i are given numbers or weights.

Note: under H_0

$$E(z_i)=\frac{m}{N}$$

$$V(z_i)=\frac{mn}{N^2}$$

$$\operatorname{cov}(z_i, z_j) = \frac{-mn}{N^2(N-1)} \quad , \forall i, j = 1, 2, \ldots\ldots, N$$

Mood Test for Dispersion

If we have two populations called as X and Y with cumulative distribution functions F_X and F_Y respectively. A random sample of size m is drawn from X population and another random sample of size n is drawn from Y population denoted as:

$X_1, X_2, \ldots\ldots, X_m$ and $Y_1, Y_2, \ldots\ldots, Y_n$

These $m+n=N$ observations drawn from the two populations are arranged in order of magnitude from smallest to largest.

In this combined ordered sample of N observations (with no ties), the average rank is the mean of first N integer. i.e. $\left(\frac{N+1}{2}\right)$.

The deviation of the i^{th} ordered variable about its mean rank is $\left[1-\left(\frac{N+1}{2}\right)\right]$.

The amount of deviation is an indication of the relative spread.

In linear rank statistic, we may take weights either the absolute value of the deviations or the squared values of the deviations to measure the relative spread.

In Mood test, we take weights as the squared values of the deviations. We define the Mood Test Statistic as

$$M_N = \sum_{i=1}^{N} \left[i - \frac{N+1}{2} \right]^2 z_i$$

It gives the sum of squares of the deviations of the X ranks from the average combined rank.

We wish to test the null hypothesis that the two samples are drawn from the identical populations.

$$H_0 : F_Y(x) = F_X(x) \ ; \ \forall x$$

The general scale alternative is

$$H_s : F_Y(x) = F_X(\theta x) \ ; \ \forall x \text{ and } \theta \neq 1$$

If M_N is too small, then

$$H_s : F_Y(x) \geq F_X(\theta x) \ ; \ \forall x \text{ and } \theta > 1$$

i.e. $H_s : F_Y(x) > F_X(\theta x) \ ; \ \forall x \text{ and } \theta > 1$

If M_N is too large, then

$$H_s : F_Y(x) \leq F_X(\theta x) \ ; \ \forall x \text{ and } \theta < 1$$

$$H_s : F_Y(x) < F_X(\theta x) \ ; \ \forall x \text{ and } \theta < 1$$

Since,

$$M_N = \sum_{i=1}^{N} \left[i - \frac{N+1}{2} \right]^2 z_i$$

Then mean and variance of Mood's test statistic is

$$E[M_N] = \frac{m(N^2 - 1)}{12}$$

Also variance is obtained as

$$V[M_N] = E\left[M_N - E[M_N]\right]^2$$

By solving it, we get

$$V\left[M_N\right] = \frac{mn\left(N+1\right)\left(N^2-4\right)}{180}$$

When m,n are large, then under H_0

$$Z = \frac{M_N - E\left[M_N\right]}{\sqrt{V\left[M_N\right]}} \sim N(0,1)$$

Sukhatme Test for Dispersion

If we have two populations called as X and Y with cumulative distribution functions F_X and F_Y respectively. A random sample of size m is drawn from X population and another random sample of size n is drawn from Y population denoted as:

$X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$

These $m+n=N$ observations drawn from the two populations are arranged in order of magnitude from smallest to largest.

Here the X and Y populations have or can be adjusted to have equal medians, without loss of generality we assume that this common median is zero.

In this case, we arrange the observations such that most of the negative Y 's should proceed negative X 's, and most of the positive Y 's should follow positive X 's, if Y 's have a larger spread than X 's.

If mn indicator random variables are defined as

$$D_{ij} = \begin{cases} 1 & \text{if } Y_j < X_i < 0 \ or \ 0 < X_i < Y_j; \quad \forall i = 1, 2, \ldots.., m \\ 0 & \text{otherwise}; \quad \forall j = 1, 2, \ldots.., n \end{cases}$$

Thus Sukhatme test statistic is defined as

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

i.e.

$$\pi = \int_{-\infty}^{0} \left[F_Y(x) - F_X(x) \right] f(x)dx + \int_{0}^{\infty} \left[F_X(x) - F_Y(x) \right] f(x)dx$$

$$+ \int_{-\infty}^{0} F_X(x) \, f(x)dx - \int_{0}^{\infty} F_X(x) \, f(x)dx + \int_{0}^{\infty} f(x)dx$$

$$\pi = \int_{-\infty}^{0} \left[F_Y(x) - F_X(x) \right] f(x)dx + \int_{0}^{\infty} \left[F_X(x) - F_Y(x) \right] f(x)dx + \frac{1}{4}$$

Under H_0, $\pi = \frac{1}{4}$

Hence $H_0 : F_Y(x) = F_X(x)$ or $H_0 : \pi = \frac{1}{4}$

The mn random variable D_{ij} are Bernoulli variables, with parameter π .i.e.

$$E\left[D_{ij} \right] = E\left[D_{ij}^2 \right] = \pi$$

$$V\left[D_{ij} \right] = \pi(1 - \pi)$$

We define the parameters π_1 and π_2 as

$$\pi_1 = P\left[\left(Y_j < X_i < 0 \ or \ 0 < X_i < Y_j \right) \cap \left(Y_k < X_i < 0 \ or \ 0 < X_i < Y_k \right) \right]$$

$$= P\left[\left(Y_j < X_i < 0\right)\cap\left(Y_k < X_i < 0\right) + \left(0 < X_i < Y_j\right)\cap\left(0 < X_i < Y_k\right)\right]$$

$$\pi_1 = \int_{-\infty}^{0}\left[F_Y\left(x\right)\right]^2 f(x)dx + \int_{0}^{\infty}\left[1 - F_Y\left(x\right)\right]^2 f(x)dx$$

and

$$\pi_2 = P\left[\left(Y_j < X_i < 0 \; or \; 0 < X_i < Y_j\right)\cap\left(Y_j < X_h < 0 \; or \; 0 < X_h < Y_j\right)\right]$$

$$= P\left[\left(Y_j < X_i < 0\right)\cap\left(Y_j < X_{hi} < 0\right) + \left(0 < X_i < Y_j\right)\cap\left(0 < X_h < Y_j\right)\right]$$

$$\pi_2 = \int_{-\infty}^{0}\left[\frac{1}{2} - F_X\left(y\right)\right]^2 f(y)dy + \int_{0}^{\infty}\left[F_X\left(y\right) - \frac{1}{2}\right]^2 f(y)dy$$

Since $T = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}$

Then mean and variance of T is defined as

$$E\left[T\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} E\left(D_{ij}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\pi = mn\pi$$

and $V\left[T\right] = V\left(\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}\right)$

$$V\left[T\right] = mn\left[\pi - \pi^2\left(N-1\right) + \left(n-1\right)\pi_1 + \left(m-1\right)\pi_2\right]$$

As $m, n \to \infty$

$$E\left[T / mn\right] = \pi$$

$$V\left[T / mn\right] \to 0$$

Hence T / mn is an unbiased ad consistent estimator of π .

If we define $T^{'}$ as

$$T^{'} = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}^{'}$$

where

$$D_{ij}^{'} = \begin{cases} 1 & \text{if } X_i < Y_j < 0 \ or \ 0 < Y_j < X_i \\ 0 & \text{otherwise} \end{cases}$$

The critical region for α level of significance is given as

Alternative Hypothesis	Critical Region
$\pi < \frac{1}{4} \ (\theta > 1)$	$T \le C_{\alpha}$
$\pi > \frac{1}{4} \ (\theta < 1)$	$T^{'} \le C_{\alpha}^{'}$
$\pi \ne \frac{1}{4} \ (\theta \ne 1)$	$T \le C_{\frac{\alpha}{2}} \ or \ T^{'} \le C_{\frac{\alpha}{2}}^{'}$

Under H_0, i.e. $H_0 : F_Y(x) = F_X(x)$

Then $\pi = \frac{1}{4}$

and $\pi_1 = \pi_2 = \frac{1}{12}$

Thus $E[T] = \frac{mn}{4}$

$$V[U] = mn\left[\frac{1}{2} - \frac{1}{4}(N-1) + \frac{(n-1)}{3} + \frac{(m-1)}{3}\right]$$

$$V[U] = \frac{mn(N+7)}{48}$$

If N is large, then under H_0

$$Z = \frac{T - E[T]}{\sqrt{V[T]}} \to N(0,1)$$

i.e. $Z = \dfrac{U - \dfrac{mn}{4}}{\sqrt{\dfrac{mn(N+7)}{48}}} \to N(0,1)$

4.9 CONTINGENCY TABLE

A contingency table is an array of natural numbers in matrix from where those natural numbers represent counts, or frequencies. For example, an entomologist observing insects may say he observed 37 insects, or he may say he observed

Moths	**Grasshoppers**	**others**	**Total**
12	22	3	37

using 1×3 (one by three) contingency table. This is one way contingency table because it has only one row.

The entomologist may wish to be more specific and use a 2×3contingency table, as follows.

	Moths	**Grasshoppers**	**others**	**Total**
Alive	3	21	3	27
Dead	9	1	0	10
Total	12	22	3	37

The totals, consisting of two row totals, three column totals, and grand total. It is a two way contingency table and may be extended to include several rows (r) and several columns (c) as an $r\times s$ contingency table.

4.9.1 THE 2×2CONTINGENCY TABLE

In general $r\times c$contingency table is an array of natural numbers arranged in to r rows and c columns and thus has rc *cells* or places for the numbers. This section is concerned only with the case where r = 2 and c = 2, the 2×2

contingency table, because there are four cells, 2×2 contingency table is also called the *fourfold* contingency table.

One application of the 2×2 contingency table arise when N objects (or persons), possible selected at random from some population, are classified in to one of two categories before a treatment is applied or an event takes place. After the treatment is applied the same N object are again examined and classified in to two categories. The question to be answered is, “Does the treatment significantly alter the proportion of object in each of two categories?” The appropriate statistical procedure was seen to be a variation of the sign test known as the McNemar test. The McNemar test is often able to detect subtle differences, primarily because the same sample is used in the two situations (such as “before” and “after”). Another way of testing the same hypothesis tested with the McNemar test is by drawing a random sample from the population before the treatment and then comparing it with another random sample drawn from the population after the treatment. The additional variability introduced by using to different random sample is undesirable because it tends to obscure the changes in the population caused by the treatment. However, there are times when it is not practical, or even possible, to use the same sample twice. Then the procedures to be described in the section may be used.

In the first procedure, two random samples are drown, one from each of two populations, two test the null hypothesis that the probability of event A (some specified event)is the same for both populations. The null hypothesis may also be stated as “the proportion of the population with characteristic A is same for both populations.”

4.9.2 The chi-squared test for differences in probabilities, 2×2

Data: A random sample of n_1 observations is drawn from one population (or before a treatment is applied) and each observation is classified in to either class 1 or class 2, the total numbers in the two classes being o_{11} and o_{12} respectively,

Where $o_{11}+o_{12}=n_1$. A second random sample of n_2 observations is drawn from a second population (or the first population after some treatment is applied) and the number of population in class 1 or class 2 is o_{21} or o_{22} respectively, where $o_{21}+o_{22}=n_2$. The data are arranged in to the following 2×2 contingency table.

Assumptions

1. Each sample is a random sample.
2. The two sample are mutually independent.
3. Each observation may be categorized in to class 1 or class 2.

Test Statistic: If any column total is zero, the test statistic is define as $T_1=0$. Otherwise,

$$T_1=\frac{\sqrt{N}(O_{11}O_{22}-O_{12}O_{21})}{\sqrt{n_1n_2C_1C_2}} \qquad (1)$$

Null distribution the exact distribution of T_1 is difficult to tabulate because of all the different combination of values possible for o_{11}, o_{12}, o_{21} and o_{22}. Therefore the large sample approximation is used, which is the standard normal distribution whose quintiles are given in Table.

Hypothesis: Let the probability that a randomly selected element will be in class 1 be denoted by p_1 in population 1 and p_2 in population 2. Note that it is not necessary for p_1 and p_2 to be known. The hypotheses merely specify a relationship between them.

A. (Two-Tailed Test)

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Reject H_0 at the approximate level α if T_1 is less than the $\alpha / 2$ quintile of a standard normal random variable Z, or if T_1 is grater then the 1-$\alpha / 2$ quintile of Z, where the quintiles of Z are given in table.

The p-value is twice the smaller of the probabilities that Z is less then the observed value of T_1 or grater then the observed value of T_1, from table.

Note that for the above hypotheses, T_1^2 is often use instead of T_1 as the test statistic. Then the rejection region is the upper tail of the chi-squared distribution with 1 degree of freedom given in table.

B. (Lower-Tailed Test)

$$H_0 : p_1 \geq p_2$$

$$H_1 : p_1 < p_2$$

Reject H_0 at the approximate level α if T_1 is less than the α quintile of a standard normal random variable Z, where the quintiles of Z are given in table.

The p-value is the probability that Z is less than the observed value of T_1, obtained from table.

C. **(Upper-Tailed Test)**

$$H_0 : p_1 \leq p_2$$

$$H_1 : p_1 < p_2$$

Reject H_0 at the approximate level α if T_1 is greater than the $1-\alpha$ quintile of a standard normal random variable Z, where the quintiles of Z are given in table.

The p-value is the probability that Z is greater than the observed value of T_1, obtained from table.

EXAMPLE 1

Two Carloads of manufactured items are sampled randomly to determine if the proportion of defective items is different for the two carloads. From the first carload 13 of the 86 items were defective. From the second carload 17 of the 74 items were considered defective.

	Defective	**Non defective**	**Totals**
Carload 1	13	73	86
Carload 2	17	57	74
Totals	30	130	160

The assumptions are met, and so the two-tailed test is use to test

H_0: The proportion of defective is equal in two carloads using the test statistic

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

$$= \frac{\sqrt{160}((13)(57) - (73)(17))}{\sqrt{(86)(74)(30)(130)}}$$

$$= -1.2695$$

The 0.975 quintile of a standard normal random variable is found from Table A1 to be 1.9600. therefore the rejection region of approximate size 0.05 consist of all value of T_1 grater then 1.9600, or less then -1.9600. The observed value is -1.2695, so the null hypothesis is accepted at the $\alpha = 0.05$ level of significance.

The p-value is twice the probability of Z being less then the observed value -1.2695, which is found from the table as 0.102, so the p-value is approximately 0.204.Therefore the decision to accept H_0 seems to be a fairly safe one.

The following example illustrates the use of one-tailed test.

EXAMPLE 2

At the U.S. Naval Academy a new lighting system was installed throughout the midshipmen's living quarters. It was claimed that the new lighting system resulted in poor eyesight due to continual strain on the eyes of the midshipmen. Consider a (fictitious) study to test the null hypothesis,

H_0: The probability of good vision is less now then it was

Let p_1 be the probability that a randomly selected graduating midshipman had good vision under the old lighting system and let p_2 be the corresponding probability with the new light. Then the preceding hypotheses may be restated as

$$H_0 : p_1 \leq p_2$$

$$H_1 : p_1 > p_2$$

Which matches the set C of hypotheses. The random sample are taken to be the entire graduation class just prior to the installation class to spend 4 years using the new light system for population 2. it is hoped that these sample will behave the same as would random samples from the entire population of graduating seniors, real and potential.

Suppose the results were as fallows.

	Good vision	**Poor vision**	
Old lights	O_{11}=714	O_{12}=111	n_1=825
New Lights	O_{21}=662	O_{22}=154	n_2=816
Totals	1376	265	1641

Decision rule C defines the critical region $\alpha = 0.05$ to be all values of T_1 greater than 1.6449 from table. Computation of T_1 gives

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

$$= \frac{\sqrt{1641}((714)(154) - (111)(662))}{\sqrt{(825)(816)(1376)(265)}}$$

$$= 2.982$$

So the null hypotheses is clearly rejected. From Table we see that the null hypotheses could have been rejected at a level of significance as small as about 0.002, so that p-value is 0.002.

We may there for conclude that the population represented by the two graduation classes do differ with respect to the proportions having poor eyesight, and the direction predicted. That is, population 2 (with the new light) has poor eyesight then population 1 (with the old light). Whether the poorer eyesight is result of the new lights has not been shown. However, an association of poor eyesight with the new lights has been shown in this hypothetical example.

4.9.3 Fisher's Exact Test

Data: The N observations in the data are summarized in a 2×2 contingency table as previously both of the row totals, r and N-r and both of the column totals, c and N-c, are determined beforehand and are therefore fixed not random.

	Col 1	**Col 2**	
Row 1	x	$r-x$	r
Row 2	$c-x$	$N-r-c+x$	$N-r$
Total	c	$N-c$	N

Assumptions :

1. Each observation is classified into exactly one cell.
2. The row and column totals are fixed, not random.(However see the comment at the end for random totals in rows, columns, or both.)

Test Statistic:

The test statistic T_2 is the number of observations in the cell in row 1, column 1.

Null Distribution:

The exact distribution of T_2 when H_0 is true is given by the hyper geometric distribution

$$P(T_2 = x) = \frac{\binom{r}{x}\binom{N-r}{c-x}}{\binom{N}{c}} \qquad x = 0,1..........,\min(r,c)$$

$$= 0 \qquad \textit{for all other values of } x \qquad (1)$$

For a large approximation use

$$T_3 = \frac{x - \frac{rc}{N}}{\sqrt{\frac{rc(N-r)(N-c)}{N^2(N-1)}}}$$

which has the standard normal distribution given in table as an approximation. If row totals or column totals, or both, are random it is more accurate to use T_1 given by $T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$

in the large sample approximation.

Hypotheses :

Let p_1 be the probability of an observation in row 1 being classified into column 1. Let p_2 be the probability of an observation in row 2 being classified in column 1.Let t_{obs} be the observed value of T_2.

A. (Two-tailed test)

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

First find the p- value using equation (1) . The p-value is twice the smaller of $P(T_2 \leq t_{obs})$ or $P(T_2 \geq t_{obs})$. Reject H_0 at the level of significance α if the p-value is less than or equal to α.

B. (Lower-tailed test)

$$H_0 : p_1 \geq p_2$$
$$H_1 : p_1 < p_2$$

Find the p- value $P(T_2 \leq t_{obs})$ using equation (1). Reject H_0 at the level of significance α if $P(T_2 \leq t_{obs})$ is less than or equal to α.

C. (Upper-tailed test)

$$H_0 : p_1 \leq p_2$$
$$H_1 : p_1 > p_2$$

Find the p- value $P(T_2 \geq t_{obs})$ using equation (1). Reject H_0 at the level of significance α if $P(T_2 \geq t_{obs})$ is less than or equal to α.

Example

Fourteen newly hired business majors, 10 males and 4 females, all equally qualified, are being assigned by the bank president to their new jobs. Ten of the new jobs are as tellers , and four are as account representatives. The null hypothesis is that males and females have equal chances at getting the more desirable account representative jobs. The one-sided alternative of interest is that females are more likely than males to get the account representative jobs.

Only one female is assigned a teller position. Can the null hypothesis be rejected? The information given is sufficient to fill in the following 2×2 contingency table, because the row totals and column totals are already known.

	Account representative	**Teller**	
Males	1	9	10
Females	3	1	4
Total	4	10	N =14

$$H_0 : p_1 \geq p_2$$
$$H_1 : p_1 < p_2$$

The exact lower-tailed p-value is given by Equation (1) as

$$P(T_2 \le 1) = P(T_2 = 0) + P(T_2 = 1)$$

$$P(T_2 = x) = \frac{\binom{10}{0}\binom{4}{4}}{\binom{14}{4}} + \frac{\binom{10}{1}\binom{4}{3}}{\binom{14}{4}}$$

$$= \frac{1}{1001} + \frac{40}{1001} = 0.041$$

The null hypothesis is rejected at $\alpha = 0.05$.