The entire body of classical statistical inference technique is based on fairly specific assumptions regarding the nature of the underlying population distribution, usually its form and some parameters must be started. Given the right set of assumptions, certain test statistics can be developed using mathematics which is frequently elegant and beautiful. The derived distribution theory is qualified by certain pre-requisite conditions and therefore all conclusions reached using these techniques are exactly valid only so long as the assumptions themselves can be substantiated. However, in a read world problem, everything does not package with level of population of origin. A decision must be made as to what population properties made judiciously be assumed for the model. If the reasonable assumptions are not such that the traditional techniques are applicable, the classical methods may be used and inference conclusions stated only with the appropriate qualifiers. For example, "If the population is normal then....." Frequently in practice, these assumptions which are thought to be reasonable by empirical evidence or past experience are not the desired ones, i.e. those for which a set of standard statistical techniques has been developed. Alternatively, the sample may be too small or the previous experience may be too limited to determine what is a reasonable assumption. Alternatively the researchers may not understand or even beware of the pre-conditions implicit in the derivation of the statistical technique. In each of these situations of blind faith for the scientific method either because of ignorance or with the rationalization that an approximate accurate inference based on recognized and accepted scientific techniques is better than no answer at all or a conclusion based on common sense or intuition. An alternative set of techniques is available which may be classified as distribution free and non-parametric procedures. In a distribution free inference whether for testing or estimation, the methods are based on functions of the sample observation whose corresponding random variables has a distribution which does not depend on the specific distribution functions of the population from which the sample was drawn. Therefore, assumptions regarding the underlying population are not necessary. On the other hand, the term "non-parametric test" implies a test of a hypothesis which is not a statement about the parameter values. The type of statement depends on the distribution accepted for the term parameter. If parameter is interpreted only in the broader sense, the hypothesis can be concerned only with the form of the population. Therefore, distribution-free test and non-parametric test are not synonimous since the first one relates to the distribution of the test statistic and the 2nd one relates to the type of hypothesis to be tested. A distribution-free test may be for a hypothesis concerning the median which is certainly a population parameter within our broad definition.

In spite of the inconsistency in the nomenclature, we shall follow the customery practice and

Consider both types of test procedures as non-parametric
inferences making no distinction between these two
classifications. For the purpose differentiation, the
statistical techniques whose justification in probability
is based on specific assumptions about the population
may be called parametric methods. This implies a
definition of non-parametric statistic then as a
treatment of either non-parametric types of inference
or analogies to standard statistical problem when
specific distribution assumption are replaced by
very general assumptions, and the analysis is
based on some $f^{\underline{n}}$ of sample observations whose
sampling distribution can be determined without
the knowledge of the specific distribution of the
underlying population. The assumption most frequently
require is that the population should be continuous.
More restrictive assumptions are sometimes made, for
example: that population is symmetrical but not to
the extent that the dist$^{\underline{n}}$ is specifically postulated.

The information used in making non-parametric
inference generally relates to some $f^{\underline{n}}$ of the
actual magnitudes of the random variable in the
sample. For example, if the actual observations are
replaced by there relative ranking within the
sample and the prob. dist$^{\underline{n}}$ of some $f^{\underline{n}}$ of thee
sample ranks can be determined by postualating
only very general assumptions about the basic
population, this $f^{\underline{n}}$ will provide a distribution-
free technique for estimation or hypothesis testing.
Inferences based on descriptions. of these
derived samples relate to whatever parameters
are relevant and adoptable, such as the median

for a location parameter. The non-parametric and parametric hypothesis are analogous, both relating to location and identical in the case of a continuous or symmetric population.

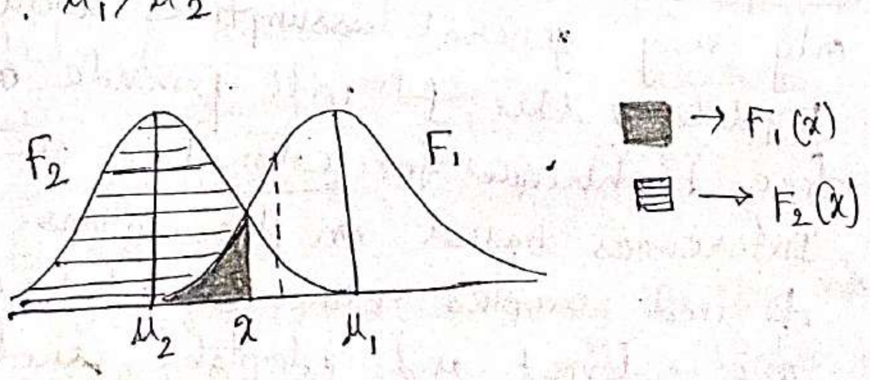
$|t| > t_{\alpha/2, n-1}$ → Rejection Rule

```
n = 0
miu0 = 10
miu1 = 5
sigma = 2
x <- rnorm (n, miu1, sigma)
xbar <- mean(x)
s <- sd(x)
t <- sqrt(n) * (xbar - miu0)/s
cutoff <- qt (0.975, n-1)
if (abs(t) > cutoff)
{
  count = 1
}
else
{ count = 0
}
```

qt ↓ t -এর সেই point যার উপরের area α

$H_0 : \mu_1 = \mu_2$ interpretations are "on an avg"

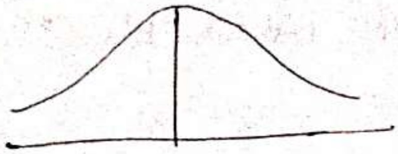
$H_1 : \mu_1 > \mu_2$



$H_0 : F_1(x) = F_2(x) \ \forall x$ $\qquad F(x) = P(X \leq x)$

$H_1 : F_1(x) < F_2(x) \ \forall x$

দুটোই আলাদা $dist^n$ আলাদার graph



One-sample Location Problem

$\boxed{\text{Sign test}}$

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population $F(\cdot)$ with unknown location parameter (median) $\theta_{1/2}$.

The one-sample location problem is defined as

$H_0 : \theta_{1/2} = \theta_0$

against

$H_1 : \theta_{1/2} > \theta_0$

$\theta_{1/2} < \theta_0$

$\theta_{1/2} \neq \theta_0$

Let us define $\psi(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$

Consider the test statistic,

(sign test statistic) $\quad S = \sum_{i=1}^{n} \psi(X_i - \theta_0)$

$=$ No. of X_i's such that $X_i > \theta_0$

$=$ No. of +ve signs among the differences $(X_i - \theta_0)$ [thats why "sign test"]

Non-Parametric justification

$S \sim Bin(n, \pi)$, regardless of the $popu^n$ $F(\cdot)$ where

$\pi = P(X_i > \theta_0)$

Under $H_0 : \pi = \frac{1}{2}$ as θ_0 = median under H_0

Thus the distribution of S under Ho is independe of F(.) and Hence the test provided by S is exactly distribution free under Ho or non-paramet

• Test Procedure :

Now, $\theta_{1/2} > \theta_0$ implies $X_i > \theta_0$ for most of the X_i's. Thus, the value of S under $H_1 : \theta_{1/2} > \theta_0$ is expected to be large than the value of S under Ho.

Hence a right tailed test is appropriate for test

$H_0 : \theta_{1/2} = \theta_0$

against

$H_1 : \theta_{1/2} > \theta_0$

Similarly, a left tailed test is appropriate for testing Ho against $H_1 : \theta_{1/2} < \theta_0$

An equal test is appropriate for testing Ho against $H_1 : \theta_{1/2} \neq \theta_0$

Exact size-α- test for testing Ho $\theta_{1/2} = \theta_0$ against $H_1 : \theta_{1/2} > \theta_0$ is reject Ho at level α if observed $S \geq K_\alpha$ where K_α is the smallest integer satisfying

$$\Rightarrow \sum_{S=K_\alpha}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^n \leq \alpha$$

$n = 10$

$\alpha = 0.05$

$$\sum_{S=8}^{10} \binom{10}{s} \left(\frac{1}{2}\right)^n$$

Similarly exact size α test

$H_0 : \theta_{1/2} = \theta_0$

against

$H_1 : \theta_{1/2} < \theta_0$ is

reject H_0 at level α if observed $S \le K_\alpha$ where K_α is the largest integer satisfying

$$\sum_{s=0}^{K_\alpha} \binom{n}{s} \left(\frac{1}{2}\right)^n \le \alpha$$

Finally exact size-α-test for testing H_0 against $H_1 : \theta_{1/2} \ne \theta_0$ is rejected at level α if either observed $S \geq K_{\alpha/2}$ or $S \le K'_{\alpha/2}$ where $K_{\alpha/2}$ and $K'_{\alpha/2}$ are respectively the smallest and the largest interger statisfying.

$$\sum_{s=K_{\alpha/2}}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}$$

$$\sum_{s=0}^{K'_{\alpha/2}} \binom{n}{s} \left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}$$

Asymtotic distribution of S

Under H_0, $E(S) = \frac{n}{2}$, $V(S) = \frac{n}{4}$

By De Moivre - Laplace CLT

$$\tau = \frac{S - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \xrightarrow{\alpha} N(0,1) \text{ as } n \to \infty$$

Asympotic size-α. test for testing

$H_0 : \theta_{1/2} = \theta_0$

against

$H_1 : \theta_{1/2} > \theta_0$ is

Reject H_0 if $\tau \geq \tau_\alpha$ where τ_α is the upper α point of $N(0,1)$ distribution.

$H_1: \theta_{1/2} < \theta_0 \rightarrow \tau \leq -\tau_{\alpha} = \tau_{1-\alpha}$

$H_1: \theta_{1/2} \neq \theta_0 \rightarrow |\tau| \geq \tau_{\alpha/2}$

Confidence interval for median using sign test

Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics of X.

Lemma:> Let θ_p be the p-th quantile of X, $P(X \leq \theta_p) = p$, $p \in (0,1)$

i.e. $F(\theta_p) = p$

Then

$$P(X_{(r)} < \theta_p < X_{(s)})$$

$$= P(r \leq Y \leq s-1),$$

where Y denotes the no. of ~~X_i's lying between $X_{(r)}$ and $X_{(s)}$~~. X_i's $< \theta_p$.

Proof of the lemma:

$X_{(r)}$ $X_{(s)}$

$$P(X_{(r)} < \theta_p < X_{(s)})$$

$$= P(X_{(r)} < \theta_p) - P(X_{(s)} < \theta_p)$$

The event $X_{(r)} < \theta_p$ implies that there are at least r observations below θ_p. $X_{(r)} < \theta_p \Rightarrow Y \geq r$

Similarly, $X_{(s)} < \theta_p \Rightarrow Y \geq s$

Hence,

$$P(X_{(r)} < \theta_p < X_{(s)}) = P(Y \geq r) - P(Y \geq s)$$

$\cdot\ P(r \le Y \le s-1)$

<u>To find the confidence interval.</u>

$$Y = \sum_{i=1}^{n} Z_i \qquad X_i \overset{iid}{\sim} \text{Cont}(F), \quad i = 1(1)n$$

$$Z_i = \begin{cases} 1, & \text{if } X_i \le \theta_p \\ 0, & \text{ow} \end{cases} \qquad \therefore Y \sim Bin(n, p)$$

$$\text{or, } n - Y \sim Bin(n, 1-p)$$

<u>Sign test statistic</u>

Define, $S(\theta_{M_0})$ = no. of X_i's s.t. $X_i > \theta_{M_0}$

$S(X_0)$ = Sign test statistic = No. of X_i's $> \theta_0$

$$S(\theta_{M_0}) \overset{d}{=} S(0)\Big|_{H_0} \sim Bin\left(n, \tfrac{1}{2}\right)$$

Now,

$$P_{H_0}(S \ge K_{\alpha/2}) \le \frac{\alpha}{2}$$

$$P_{H_0}(S \le K'_{\alpha/2}) \le \frac{\alpha}{2}$$

$$P_{H_0}(S \ge K_{\alpha/2}) + P_{H_0}(S \le K'_{\alpha/2}) \le \alpha$$

$$\text{or, } 1 - P_{H_0}(S < K_{\alpha/2}) + P_{H_0}(S \le K'_{\alpha/2}) \le \alpha$$

$$\text{or, } P_{H_0}(S < K_{\alpha/2}) - P_{H_0}(S \le K'_{\alpha/2}) \ge 1-\alpha$$

$$\text{or, } P_{H_0}(K'_{\alpha/2} + 1 \le S < K_{\alpha/2}) \ge 1-\alpha$$

or, $P_{H_0}\left(K'_{\frac{\alpha}{2}} + 1 \leq S \leq K_{\frac{\alpha}{2}} - 1 \right) \geq 1-\alpha$

From the above lemma,

$$P_{H_0}\left(X_{\left(K'_{\frac{\alpha}{2}}+1\right)} < \theta_{\frac{1}{2}} < X_{\left(K_{\frac{\alpha}{2}}\right)}\right) \geq 1-\alpha$$

$\therefore \left(X_{\left(K'_{\frac{\alpha}{2}}+1\right)}, X_{K_{\frac{\alpha}{2}}} \right)$ is a confidence interval for $\theta_{\frac{1}{2}}$ at the prob. level $(1-\alpha)$ at least. This is dist^n free because this depends upon the order statistics. Because of the symmetry of the dist^n when $p = \frac{1}{2}$,

$$K_{\frac{\alpha}{2}} + K'_{\frac{\alpha}{2}} = n.$$

H.W • Confidence interval for quantile. C.I for θ_p

Note

1. The problem of zero difference does not exist because the population was assumed to be continuous at the median. In reality the inferences are conditional upon the observed no. of non-zero differences.

Alternatively, half of the zero's may be treated as plus and half as minus or plus-minus signs are assigned at random to the zero's

2. Since the hypothesis here states that θ_0 is that value of x which divides the area into two equal parts, H_0 can be equivalently expressed as

$$H_0 : \pi = P(X > \theta_0) = P(X < \theta_0)$$

against

$$H_1 : \pi = P(X > \theta_0) > P(X < \theta_0)$$ (in case of right-sided alternative)

● Paired Sample sign-test (Non-parametric competetor of paired-t test)

The single sample sign-test produces for hypothesis testing and confidence interval of estimation of $\theta_{\frac{1}{2}}$ are equally applicable to paired sample.

Consider a random sample of n pairs.

$$\{(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)\}.$$

Let us define $D_i = X_i - Y_i$, $i = 1(1)n$.

Let $\theta_{\frac{1}{2}}$ denote the median of the distribution of the differences D_i. We assume that the population of differences is continuous at its median $\theta_{\frac{1}{2}}$. so that $P(D = \theta_{\frac{1}{2}}) = 0$.

Note that, this is a test for the median differences which is not necessarily the same as the differences between the two medians say M_X and M_Y. In general it is not true the $M_D = M_X - M_Y$. If X and Y populations are both symmetric and $M_X = M_Y$ and if the difference population

is also symmetric then $M_D = M_X - M_Y$
is a necessary and sufficient condition for M_D

- **Wilcoxon's signed rank test**

$$D_i = x_i - \theta_0 \qquad z_i = \begin{cases} 1 & \text{if } x_i > \theta_0 \\ 0 & \text{if } x_i < \theta_0 \end{cases}$$

$$W^+ = \sum_{i=1}^{n} z_i \, \text{Rank}(|D_i|)$$

$$W^- = \sum_{i=1}^{n} (1 - z_i) \, \text{Rank}(|D_i|)$$

$$W^+ + W^- = \sum_{i=1}^{n} \text{Rank}(|D_i|) = \frac{n(n+1)}{2}$$

Example

$n = 10 \qquad \theta_0 = 7.5$

x_i	D_i	z_i	$\lvert D_i\rvert$	Rank($\lvert D_i\rvert$)
7.75	0.25	1	0.25	4.5
7.91	0.41	1	0.41	6
7.25	−0.25	0	0.25	4.5
6.18	−1.32	0	1.32	10
7.44	−0.06	0	0.06	1
7.29	−0.21	0	0.21	3
6.87	−0.63	0	0.63	8
7.95	0.45	1	0.45	7
6.61	−0.89	0	0.89	9
7.32	−0.18	0	0.18	2

Therefore, $W^+ = 17.5$ $\qquad W^- = 55 - 17.5$

$\frac{n(n+1)}{2} = 55$ $\qquad = 37.5$

Since the ~~or~~ ordinary single sample sign test utilizes only the signs of the differences between each observation and the hypothesised median θ_0, the magnitudes of these observations relative to θ_0 are ignored. Assuming that such information are available, a test statistic which takes into account the individual relative magnitudes might be expected to give better performance. If we assumed that the parent population is symmetric the wilcoxon's signed rank test statistic provides us an alternative test of location which is affected by both the magnitudes and signs of the differences.

- **Test Procedure :**

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a continuous population $F(x)$ with unknown location parameter (median) $\theta_{1/2}$. We assume that F is symmetric about $\theta_{\frac{1}{2}}$. Consider the problem of testing

$$H_0 : \theta_{\frac{1}{2}} = \theta_0$$

against

$$H_1 : \theta_{1/2} > \theta_0$$

$$\theta_{1/2} < \theta_0$$

$$\theta_{\frac{1}{2}} \neq \theta_0$$

Let us define $D_i = X_i - \theta_0 \quad , i = 1(1)n$

under H_0, $\theta_{\frac{1}{2}} = \theta_0$

the differences D_i are symmetrically distributed about 0 then the wilcoxon's signed test statistic is defined as $\text{Min}(W^+, W^-)$

$$\text{and } Z_i = \begin{cases} 1 & \text{if } X_i > \theta_0 \\ 0 & \text{if } X < \theta_0 \end{cases}$$

W^+ can be interpreted as sum of the ranks of the absolute values of those differences which are positive. ~~Eg~~ Similarly W^- can be interpred as sum of the ranks of the absolute values of those differences which are originally negetive. In contrast to the ~~od~~ ordinary one sample sign test the value of W^+ is influenced not only by the no. of positive differences but also by their relative magnitude.

- **Description of the critical region**: If θ_0 is the true median of the symmetrical population then $E(W^+) = E(W^-)$

If the true median exceeds θ_0 then most of the larger ranks will correspond to positive differences. Hence W^+ is expected to be larg under $H_0 : \theta_{1/2} = \theta_0$ then under $H_1 : \theta_{1/2} > \theta_0$

Similarly, W^+ is expected to be smaller under $H_0 : \theta_{1/2} = \theta_0$. then under $H_1 : \theta_{1/2} < \theta_0$

It is more convenient to work with smaller ~~ones~~ sums, tables of the left tailed critical values are generally set up for the random variable which may denote either W^+ or W^-. Large W^+ values correspond to small values of W^-

- Result

W^+ and W^- are identically distributed,

i.e. $P(W^+ \geq c) = P(W^- \geq c)$

Proof >> $P(W^+ \geq c) = P\left(W^+ - \frac{n(n+1)}{4} \geq c - \frac{n(n+1)}{4}\right)$

$= P\left(\frac{n(n+1)}{4} - W^+ \geq c - \frac{n(n+1)}{4}\right)$

Since $W^+ = W^- - \frac{n(n+1)}{2}$ and W^+ is symmetric about $\frac{n(n+1)}{4}$

$$= P\left(\frac{n(n+1)}{2} - W^+ \geq c\right) = P(W^- \geq c)$$

$$\sum_{x=k_{\alpha/2}}^{n} \binom{n}{x}\left(\frac{1}{2}\right)^{n} \leq 0.025 \qquad k_{\alpha/2} = 16$$

$$\sum_{x=0}^{k_{\alpha/2}} \binom{n}{x}\left(\frac{1}{2}\right)^{n} \leq 0.025 \qquad k_{\alpha/2} = 4$$

● <u>Mean and variance of W^+ under H_0</u>

Without loss of generality, let me assume that

$|D_1| < |D_2| < |D_3| < \cdots < |D_n|$

such that rank $(|D_i|) = i$, $i = 1(1)n$

Define $Z_{(i)} = \begin{cases} 1 & \text{if the } i\text{th difference whose absolute value has rank } i \text{ is positive} \\ 0 & \text{otherwise} \end{cases}$

$$W^+ = \sum_{i=1}^{n} Z_i \,\text{Rank}(|D_i|) \quad \text{where } Z_i = \begin{cases} 0 & \text{if } D_i > 0 \\ 0 & \text{if } D_i < 0 \end{cases}$$

It can also be written as

$W^+ = \sum_{i=1}^{n} i Z_{(i)}$, where $Z_{(i)}$'s are independent but not identically distributed

Bernoulli random variables.

But under $H_0 : \mu_{1/2} = \mu_0$.

$Z_{(i)} \overset{iid}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$

$E(Z_{(i)}) = \frac{1}{2}$

$V(Z_{(i)}) = \frac{1}{4}$

$\therefore E_{H_0}(W^+) = \sum_{i=1}^{n} \frac{i}{2} = \frac{n(n+1)}{4}$

under H_0

$$V_{H_0}(W^+) = \sum_{i=1}^{n} \frac{i^2}{4}$$

$$= \frac{n(n+1)(2n+1)}{24}$$

~~Tot~~ To show that,

$$P_{H_0}(Z_{(i)} = 1) = \frac{1}{2} \quad \forall\ i = 1(1)n$$

$$P(Z_{(i)} = 1) = P\left(\text{Rank}(|D_j|) = i,\ D_j > 0\right)$$

$= P$(The i^{th} order statistic among $|D_1|, |D_2|, \ldots |D_n|$ corresponds to a +ve difference)

$$= \int_0^{\infty} \frac{n!}{(i-1)!\,(n-i)!} \left\{F_{|D|}(u)\right\}^{i-1} \left\{1 - F_{|D|}(u)\right\}^{n-i} f_{D}(u)\, du$$

Now,

$$F_{|D|}(u) = P(|D| \le u)$$

$$= P(-u \le D \le u) = F_D(u) - F_D(-u)$$

Under $H_0: \theta_{1/2} = \theta_0$, x_i's are symmetric about θ_0.

So, $D_i = x_i - \theta_0$'s are symmetric about 0

$\therefore F_D(-u) = 1 - F_D(u)$ $\qquad f_{|D|}(u) = 2 f_D(u)$

$\therefore F_{|D|}(u) = 2F_D(u) - 1$ (derivation of $F_{|D|}$)

$1 - F_{|D|}(u) = 2 - 2F_D(u)$

So, $$P_{H_0}(Z_{(i)} = 1) = \int_0^\infty \frac{n!}{(i-1)!\,(n-i)!} \{2F_D(u) - 1\}^{i-1} \{2 - 2F_D(u)\}^{n-i} \, f_D(u)$$

Put, $2F_D(u) - 1 = v$

$\therefore 2 f_D(u) \cdot du = dv$

u	0	∞
v	0	1

~~Symm~~ symmetric about 0 so, $F_D(0) = \frac{1}{2}$

$$P_{H_0}(Z_{(i)} = 1) = \int_0^1 \frac{n!}{(i-1)!\,(n-i)!} v^{i-1} (1-v)^{n-i} \frac{dv}{2}$$

$\cancel{\frac{1}{2}}$

~~$P(Z_{(i)} = 1) = P(\text{Rank}(|D_j|) = i, D_j > 0)$~~

$$= \frac{1}{2} \int_0^1 \frac{\Gamma(n+1)}{\Gamma(i) \cdot \Gamma(n-i+1)} v^{i-1} (1-v)^{n-i} \cdot dv \cdot 1$$

$$= \frac{1}{2} \int_0^1 \frac{1}{B(i, n-i+1)} v^{i-1} (1-v)^{n-i} \, dv$$

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

Nonparametric Justification : Under H_0,

$Z_{(i)} \overset{iid}{\sim} Ber\left(\frac{1}{2}\right)$ ~~is~~ which is independent of the parent population. W^+, being a linear function of $Z_{(i)}$'s, has its distribution independent of F. Hence the test provided by W^+ is exactly distribution free under H_0 and hence non-parametric.

Large Sample Dist^n :

Result : Suppose $Y_1, Y_2, \ldots, Y_n$ are iid with $E(Y_i) = \mu$ and $var(Y_i) = \sigma^2 < \infty$

Then $\sum_{i=1}^{n} a_i Y_i \overset{a}{\sim} N\left(\sum_{i=1}^{n} a_i \mu_i , \sigma^2 \sum_{i=1}^{n} a_i^2\right)$ if

$$\frac{\underset{i \in \{1,2,\ldots n\}}{Max}(a_i^2)}{\sum_{i=1}^{n} a_i^2} \longrightarrow 0 \text{ as } n \to \infty$$

Here $W^+ = \sum_{i=1}^{n} a_i Z_{(i)}$

with $a_i = i \quad (i = 1(1)n)$

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$Max\ a_i^2 = n^2$$

$$\therefore \frac{Max\ a_i^2}{\sum_{i=1}^{n} a_i^2} = \frac{6n^2}{n(n+1)(2n+1)} = \frac{6}{\left(1+\frac{1}{n}\right)(2n+1)}$$

$$\to 0 \text{ as } n \to \infty$$

so, by the above result

$$\frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathcal{D}} N(0,1) \text{ as } n \to \infty$$

$$W^+ = \sum\sum_{1 \le i \le j \le N} W_{ij}$$

where $W_{ij} = \begin{cases} 1 & \text{if } D_i + D_j > 0 \\ 0 & \text{o.w} \end{cases}$

	D_i
7.75	0.25
7.91	0.41
7.25	−0.25
6.18	−1.32
7.44	−0.06
7.29	−0.21
6.87	−0.63
7.95	0.45
6.61	−0.89
7.32	−0.18

i \ j	1	2	3	4	5	6	7	8	9	10
1	1	–	–	–	–	–	–	–	–	–
2	1	1	–	–	–	–	–	–	–	–
3	*	1	0	–	–	–	–	–	–	–
4	0	0	0	0	–	–	–	–	–	–
5	1	1	0	0	0	–	–	–	–	
6	1	1	0	0	0	0	–	–	–	
7	0	0	0	0	0	0	0	–	–	–
8	1	1	1	0	1	1	0	1	–	–
9	0	0	0	0	0	0	0	0	0	–
10	1	1	0	0	0	0	0	0	1	0 0

	D_i	$\mu_0 = 72$ $\lvert D_i \rvert$	Z_i	Rank ($\lvert D_i \rvert$)
67	-5	5	0	3.
73	1	1	1	1
82	10	10	1	4
70	-2	2	0	2
61	-11	11	0	5

$$W^+ = 5$$

i \ j	1	2	3	4	5
1	0	-	-	-	-
2	0	1	-	-	-
3	1	1	1	-	-
4	0	0	1	0	-
5	0	0	0	0	0

$W^+_i = 5$

Under H_0, D_i s are identically distributed.

$$E(W^+) = nE(W_{ii}) + \frac{n(n+1)}{2} E(W_{ij})$$

$$E(W_{ii}) = P(D_i > 0) = P_1$$

$$E(W_{ij}) = P(D_i + D_j > 0) = P_2$$

$P_1 = P(D_i > 0) = \frac{1}{2}$ under H_0

$P_2 = P(D_i + D_j > 0)$ $\quad D_i \to u$, $D_j \to v$

$u+v > 0$

$v > -u$

$$= \int_{-\infty}^{\infty} \int_{-u}^{\infty} f_D(u) f_D(v)\, dv\, du$$

[$\because D_i$ and D_j are independent]

Total terms in $\sum\sum_{1 \le i \le j \le n} W_{ij} = \frac{n(n+1)}{2}$

n terms form W_{ii}

$\frac{n(n-1)}{2}$ terms W_{ij} forms ($i<j$)

$V(W_{ii})$

$V(W_{ij})$

$Cov(W_{ii}, W_{ij})$

$Cov(W_{ii}, W_{hk})$ $i \ne h \le k$

$Cov(W_{ij}, W_{hk})$ $i \ne h, i<j, j \ne k$,

$$= \int_{-\infty}^{\infty} \left[\int_{-u}^{\infty} f_D(v)\, dv \right] f_D(u)\, du$$

$$= \int_{-\infty}^{\infty} \left[1 - F_D(-u) \right] f_D(u)\, du$$

Under H_0, D_i's are symmetric about 0,

$$F_D(-u) = 1 - F_D(u)$$

$$\therefore P_2 = \int_{-\infty}^{\infty} F_D(u) f_D(u)\, du$$

Put $t = F_D(u)$

$dt = f_D(u)$

u	$-\infty$	∞
t	0	1

$$\therefore P_2 = \int_0^1 t\, dt = \frac{1}{2} \text{ under } H_0$$

Thus, $E_{H_0}(W^+) = \frac{n}{2} + \frac{n(n-1)}{4} = \frac{n(n+1)}{4}$

$Var(W^+) = n Var(W_{ii}) +$

$\frac{n(n-1)}{2}\ var(W_{ij}) +$

$2n(n-1)\ Cov(W_{ii}, W_{ik}) +$

$2n\binom{n-1}{2}\ Cov(W_{ij}, W_{ik})$

$+ \binom{n}{4}\ Cov(W_{ij}, W_{hk})$,

$\sum\sum_{i \le j \le 4} W_{ij}$

$= W_{11} + W_{12} + W_{13} + W_{14}$

$+ W_{22} + W_{23} + W_{24}$

$+ W_{33} + W_{34} + W_{44}$

W_{11}	W_{12}	W_{12}	W_{13}
W_{11}	W_{13}	W_{12}	W_{14}
W_{11}	W_{14}	W_{13}	W_{14}
W_{22}	W_{23}	W_{23}	W_{24}
W_{22}	W_{24}	(	
W_{33}	W_{34}		

$$Var(W_{ii}) = E(W_{ii}^2) - [E(W_{ii})]^2$$

$$= E(W_{ii}) - [E(W_{ii})]^2$$

$$E(W_{ii}) = P(D_i > 0) = \frac{1}{2}, \text{ under } H_0$$

$$Var(W_{ii}) = p_1 - p_1^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$Var(W_{ij}) = E(W_{ij}) - [E(W_{ij})]^2$$

$$= p_2 - p_2^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \text{ under } H_0$$

$$Cov(W_{ii}, W_{ik}) = E(W_{ii}, W_{ik}) - E(W_{ii})E(W_{ik})$$

$$E(W_{ik}) = p_2 \qquad \cancel{E(W_{ii}, W_{ik})} \qquad (W_{ii}, W_{ik}) \begin{cases} 1, & \text{if } W_{ii} \\ & W_{ik} \\ 0 & \text{o.w} \end{cases}$$

$$E(W_{ii}) = p_1$$

$$E(W_{ii} W_{ik}) = p_3 \qquad = \begin{cases} 1, & \text{if } D_i > 0 \text{ \& } D_i + D_k \\ 0, & \text{o.w} \end{cases}$$

$$= P(D_i > 0 \cap D_i + D_k > 0)$$

$$= \int_0^{\infty} \int_{-u}^{\infty} f_D(u) f_D(v)\, dv\, du$$

$$= \int_0^{\infty} \left[\int_{-u}^{\infty} f_D(v)\, dv \right] f_D(u)\, du$$

$$= \int_0^{\infty} [1 - F_D(-u)] f_D(u)\, du$$

$$= \int_0^{\infty} F_D(u) f_D(u)\, du \quad [\because \text{symmetric}]$$

Put, $F_D(u) = t$

$f_D(u) = dt$

u	0	∞
t	$1/2$	1

$$\therefore E(W_{ii}\, W_{ik}) = \int_{\frac{1}{2}}^{1} t\,dt = \frac{t^2}{2}\Big|_{\frac{1}{2}}^{1} = \frac{1}{2} - \frac{1}{8} = \frac{3}{8}$$

under H_0

$$\therefore Cov(W_{ii}, W_{ik}) = P_3 - P_1 P_2$$
$$= \frac{3}{8} - \frac{1}{4}$$

$$\boxed{Cov(W_{ii}, W_{ik}) = \frac{1}{8}}$$

For i, j, h, k all distinct,

$$\boxed{Cov(W_{ij}, W_{hk}) = 0}$$

To find. $Cov(W_{ij}, W_{ik})$

$$W_{ij} W_{ik} = \begin{cases} 1 & \text{if } W_{ij} = 1 \ \& \ W_{ik} = 1 \\ 0 & \text{o.w} \end{cases}$$

$$= \begin{cases} 1 & \text{if } D_i + D_j > 0 \ \& \ D_i + D_k > 0 \\ 0 & \text{o.w} \end{cases}$$

$$E(W_{ij} W_{ik}) = P_4 = P(D_i + D_j > 0,\ D_i + D_k > 0)$$
$$= P(D_i > -D_j \cap D_i > -D_k)$$
$$= P(D_i > Max(-D_j, -D_k))$$

$$= P(0 < D_i + D_j < D_i + D_K) + P(0 < D_i + D_K < D_i + D_j)$$

$$= P(-D_i < D_j < D_K) + P(-D_i < D_K < D_j)$$

$$\left[\begin{array}{l} D_i \to u \\ D_j \to v \\ D_K \to w \end{array}\right.$$

$$= \int_{-\infty}^{\infty}\int_{-u}^{\infty}\int_{v}^{\infty} f_D(w) f_D(v) f_D(u)\, dw\, dv\, du$$

$$+ \int_{-\infty}^{\infty}\int_{-u}^{\infty}\int_{w}^{\infty} f_D(v) f_D(w) f_D(u)\, dv\, dw\, du$$

$$= 2\int_{-\infty}^{\infty}\int_{-u}^{\infty}\int_{v}^{\infty} f_D(w) f_D(v) f_D(u)\, dw\, dv\, du$$

$$= 2\int_{-\infty}^{\infty}\int_{-u}^{\infty}\left[\int_{v}^{\infty} f_D(w)\cdot dw\right] f_D(v) f_D(u)\, dv\, du$$

$$= 2\int_{-\infty}^{\infty}\int_{-u}^{\infty}\left[1 - F_D(v)\right]\cdot f_D(v) f_D(u)\, dv\, du$$

$$= 2\left[\int_{-\infty}^{\infty}\int_{-u}^{\infty} f_D(v) f_D(u)\, dv\, du - \int_{-\infty}^{\infty}\int_{-u}^{\infty} F_D(v) f_D(v) f_D(u)\, dv\, du\right]$$

$$= 2(I_1 - I_2) \quad (\text{say})$$

$$I_1 = \int_{-\infty}^{\infty}\left[\int_{-u}^{\infty} f_D(v)\,dv\right] f_D(u)\,du$$

$$= \int_{-\infty}^{\infty}\left[1 - F_D(-u)\right] f_D(u)\,du$$

$$= \int_{-\infty}^{\infty} F_D(u)\, f_D(u)\,du \quad [\text{under } H_0]$$

$$= \frac{1}{2} \quad \text{(Already found)}.$$

$$I_2 = \int_{-\infty}^{\infty}\left[\int_{-u}^{\infty} F_D(v)\, f_D(v)\, \cancel{f_D(u)}\, dv\right] \cancel{du}\, f_D(u)\, dv$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\left[1 - \{F_D(-u)\}^2\right] f_D(u)\,du$$

$$= \frac{1}{2}\left[\int_{-\infty}^{\infty} f_D(u)\,du - \int_{-\infty}^{\infty}\{F_D(-u)\}^2 f_D(u)\,du\right]$$

$$= \frac{1}{2}\left[1 - \int_{-\infty}^{\infty}(1 - F_D(u))^2 f_D(u)\,du\right] \quad \text{under } H_0$$

put $F_D(u) = t$

$f_D(u)\,du = dt$

$$\begin{array}{c|c|c} u & -\infty & \infty \\ \hline t & 0 & 1 \end{array}$$

$$= \frac{1}{2}\left[1 - \int_0^1 (1-t)^2\,dt\right]$$

$$= \frac{1}{2}\left[1 - \left(t + \frac{t^3}{3} - t^2\right)_0^1\right]$$

$$= \frac{1}{2}\left[1 - \left(1 + \frac{1}{3} - 1\right)\right] = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

$\therefore P_4 = 2(I_1 - I_2) = 2\left(\frac{1}{2} - \frac{1}{3}\right) = 2 \times \frac{1}{6} = \frac{1}{3}$, un

$$\text{Cov}(W_{ij}, W_{ik}) = P_4 - P_2^2$$

$$\boxed{\text{Cov}(W_{ij}, W_{ik}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}}$$, under

$$\therefore \text{Var}(W^+) = \frac{n}{4} + \frac{n(n-1)}{2} \cdot \frac{1}{4} + 2n(n-1)\frac{1}{8}$$

$$+ 2n\binom{n-1}{2} \cdot \frac{1}{12} + \binom{n}{4} \cdot 0$$

$$= \frac{n}{4} + \frac{n(n-1)}{8} + \frac{n(n-1)}{4} + 2n \cdot \frac{(n-1)!}{2!\,(n-3)!} \cdot \frac{1}{6}$$

$$= \frac{n}{4} + \frac{n(n-1)}{8} + \frac{n(n-1)}{4} + \frac{n(n-1)(n-2)}{12}$$

$$= \frac{6n + 3n(n-1) + 6n(n-1) + 2n(n-1)(n-2)}{24}$$

$$= \frac{6n + 9n^2 - 9n + 2n(n^2 - 3n + 2)}{24}$$

$$= \frac{6n + 9n^2 - 9n + 2n^3 - 6n^2 + 4n}{24}$$

$$= \cancel{6}\cdot\frac{\cancel{2n^3 + 2n^2 + 2n}}{\cancel{24}}$$

$$= \frac{3n^2 + 2n^3 + n}{24}$$

$$= \frac{n(3n + 2n^2 + 1)}{24}$$

$$= \frac{n\{2n^2 + 2n + n + 1\}}{24} = \frac{n\{2n(n+1) + 1(n+1)\}}{24}$$

$$= \frac{n(n+1)(2n+1)}{4}$$

Problem 1 : A manufacturer of electric bulbs claims that he has developed a new production process which will increase the mean efficiency (in suitable units) from the present value 9.03. The results obtained from an experiment with 15 bulbs from the new process are given below –
9.29, 9.76, 8.93, 10.15, 12.05, 9.02, 8.69, 12.38, 10.87, 11.25, 9.08, 10.00, 11.47, 10.25, 11.56

→ $H_0 : \theta_0 = 9.03$

against

$H_1 : \theta_0 > 9.03$

$S > K_\alpha \rightarrow$ ~~R~~

↳ Rejection rule

$$\sum_{S=K_\alpha+1}^{15} \binom{15}{S}\left(\frac{1}{2}\right)^{15} \leq 0.05 \rightarrow \text{Condition}$$

$K_\alpha = 12$

$S > k_\alpha$ (Reject / critical condition)

$\sum \binom{15}{s}\left(\frac{1}{2}\right)^{15} \leq 0.05$

k_α is the min value that.

$\therefore k_\alpha = 12$

Since,

for $\sum_{s=11}^{15} \binom{15}{s}\left(\frac{1}{2}\right)^{15} = 0.059$

for $\sum_{s=12}^{15} \binom{15}{s}\left(\frac{1}{2}\right)^{15} = 0.017$

Therefore from - trial - error method we find that $\underline{k_\alpha = 12}$

We reject if $\underline{s > 12}$

<u>Interpretaton</u> : Since,

$\sum_{s=12}^{15} \binom{15}{s}\left(\frac{1}{2}\right)^{15} < 0.05$ and

$\sum_{s=11}^{15} \binom{15}{s}\left(\frac{1}{2}\right)^{15} > 0.05$, we take $k_\alpha + 1 = 12$ and $k_\alpha = 11$

$\therefore$ we reject if $s > 11$

2) Below are given the marks obtained by a group of 20 Students in a subject in a college test and in the Subsequent public examination. Test at 1 % level whether the group has improved its performance from the college test to the public exam by using

(i) Signed Test

(ii) Signed rank test (Critical value = 43)

Serial No	Marks in college Test (X_i)	Marks in public exam (Y_i)
1	183	133
2	175	192
3	134	170
4	170	164
5	183	199
6	167	160
7	120	168
8	175	158
9	126	162
10	187	176
11	123	126
12	121	141
13	175	163
14	133	126
15	144	146
16	109	155
17	165	162
18	144	161
19	164	182
20	125	119

i	$Y_i - X_i = D_i$	$\lvert D_i \rvert$	Rank $(\lvert D_i \rvert)$	Z_i
1	-50	50	19	0
2	18	18	12.5	1
3	36	36	15.5	1
4	-6	6	4.5	0
5	16	16	9	1
6	-7	7	6.5	0
7	48	48	18	1
8	-17	17	10.5	0
9	36	36	15.5	1
10	-11	11	8	0
11	3	3	1.5	1
12	20	20	14	1
13	-72	72	20	0
14	-7	7	6.5	0
15	2	2	1	1
16	46	46	17	1
17	-3	3	1.5	0
18	17	17	10.5	1
19	18	18	12.5	1
20	-6	6	4.5	0

Sign-test statistics :-

S = no. of +ve terms among differences is 11.

Signed rank test :

$W^+ = 127$, $W^- = \frac{n(n+1)}{2} - W^+ = 210 - 127 = 83$

test statistics = $\min(W^+, W^-)$

$= 83$

Critical value : 43

Two Sample Location Problem

$H_0 : F_X(x) = F_Y(x) \ \forall x$ (The population $F_X(\cdot)$ and $F_Y(\cdot)$ are identical)

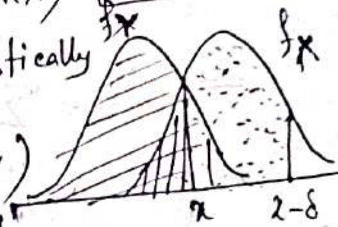
against

the alternatives

$H_1 : F_X(x) > F_Y(x)$ (Y is stochastically larger than X) → $\boxed{\delta > 0}$

$F_X(x) < F_Y(x)$ (Y is stochastically smaller than X) $\boxed{\delta < 0}$

$F_X(x) \neq F_Y(x)$



□, $F_Y(x) = 0.9$

□ → $F_X(x) < 0.9$

$F_Y(x) = F_X(x - \delta)$

Mann-Whitney's U-test :

Ex : X 15, 17, 19, 18, 22, 23

Y 14, 16, 20, 21, 27, 24, 20

Combined sample ascending order

14	Y	No. of Y preceding of X
15	X	
16	Y	$1+2+2+5+5=17$
17	X	
18	X	
19	X	
20	Y	
20	Y	
21	Y	
22	X	
23	X	
24	Y	
27	Y	

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

$$\phi(x, y) = \begin{cases} 1, & \text{if } y < x \\ 0, & \text{if } y > x. \end{cases}$$

$$\phi(X_i, Y_j) = \begin{cases} 1 & \text{if } Y_j < X_i \\ 0 & \text{o.w.} \end{cases}$$

$$\Rightarrow F_X(Y_j) \leq F_X(X_i)$$

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(F_X(X_i); F_X(Y_j))$$

It is used to test the identity of two continuous population F and G. In particular if we take $G(x) = F(x-\delta)$ then the identity of F and G is equivalent to $\delta = 0$. W. consider the following testing problem

$H_0 : \delta = 0$ against

$H_1 : \delta > 0$ (Equivalent to $F(x) > G(x)$ $\forall x$ or Y is stochastically larger than X)

$\delta < 0$ (Equivalent to $F(x) \leq G(x)$. $\forall x$ or
Y is stochastically smaller than X)

$\delta \neq 0$ (Equivalent to $F(x) \neq G(x)$)

Let $X_1, X_2, \ldots, X_{n_1}$ be iid observations from F and $Y_1, Y_2, \ldots, Y_{n_2}$ be iid observations from G (F continuous & $G(x) = F(x-\delta)$). The Mann-Whitney U-statistic is based on the position of Y in the combined sample and is defined as the number of times and Y observations preceeds an X observation in the combined sample of size $n_1 + n_2 = n$. Now, consider the following function

$$\phi(x,y) = \begin{cases} 1 & \text{if } y < x \\ 0 & \text{if } y > x \end{cases}$$

then the statistics is define as

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

- $\delta > 0$
  XXXYXXY..Y
  $U < C$
  (Y sto. larger)
- $\delta < 0$
  YY...YXYXX..X
  $U > C$
  (Y sto smaller)

Non parametric justification

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(X_i, Y_j)$$, under H_0

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \phi(F(X_i), F(Y_j)),$$

Now $F(X_i) \sim U(0,1)$

$F(Y_j) \sim U(0,1)$

Therefore U depends on the observations from $U(0,1)$ distribution which is independent of F. Thus the test provided by U is exactly distribution free and hence non-parametric.

Test Procedure

Note that

$$\phi(X, Y) = \phi(X, X_0 + \delta), \text{ where } X, X_0 \sim F(\cdot) \text{ and } Y = X_0 + \delta$$

$$\lesseqgtr \phi(X, X_0) \text{ according as } \delta > 0 \text{ or } \delta < 0$$

$$\therefore U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

$$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, X_{j0} + \delta), \quad X_i, X_{j0} \sim F(\cdot), \quad Y_j = X_{j0} + \delta$$

$$\lesseqgtr \phi(X_i, X_{j0}). \quad \text{according as } \delta \lesseqgtr 0$$

Therefore $\delta > 0$ implies that U is expected to be smaller under δ than under H_0, so two small values of U will form the critical region i.e. a left tail test is appropriate. Similarly $\delta < 0$ implies that U is expected to be larger under δ than that under H_0. So two large values will form the critical region for H_0 vs H_1:δ Both two large or two small values of U will form the critical region for H_0 vs $H_1 : \delta \neq 0$.

We can show that the distribution of U under H_0 is symmetric about $E(U)$.

Let us define

$$U' = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left[1 - \Phi(X_i, Y_j)\right] =$$ # of times an X observation preceeds an Y observation.

Then the critical region may be given as follows –

Alt	Critical region
$\delta > 0$	$U < C_\alpha$
$\delta < 0$	$U' < C_\alpha$
$\delta \neq 0$	$U < C_{\alpha/2}$ or $U' < C_{\alpha/2}$

Null distribution, mean and variance

In order to determine the size α critical region of the Mann-Whitney U-test, we must now find the null probability distⁿ of U. Under H_0, each of the $\binom{n_1+n_2}{n_1}$ arrangements of the random variables into a combined sequence occurs with equal probability so that

$$f_U(u) = P(U = u) = \frac{r_{n_1,n_2}(u)}{\binom{n_1+n_2}{n_1}},$$ where $r_{n_1,n_2}(u)$ is

the number of distinguishable arrangements of the n_1 X and n_2 Y variables such that in each sequence an Y preceeds an X is exactly u

The values of u for which $f_U(u)$ is the ... between zero and $n_1 n_2$, for the two most extreme orderings.

<u>Result</u> >>

(a) The prob. dist^n of U under H_0 is symmetric about $\frac{n_1 n_2}{2}$.

<u>Proof</u> >> For every particular arrangement Z of the n_1 X and n_2 Y letters, define the conjugate arrangement Z' as the sequence Z written backward. For example if $Z = XXYYXY$, then $Z' = YXYYXX$, Every Y that preceeds an X in the arrangement Z then follows that X in Z', so that if u is the value of the Mann-Whitney U-statistic in Z, then $n_1 n_2 - u$ is the value for Z'

So under H_0, we have

$$r_{n_1, n_2}(u) = r_{n_1, n_2}(n_1 n_2 - u)$$

$$\therefore P\left(U - \frac{n_1 n_2}{2} = u\right) = P\left(U = \frac{n_1 n_2}{2} + u\right)$$

$$= P\left(U = n_1 n_2 - \left(\frac{n_1 n_2}{2} + u\right)\right)$$

$$= P\left(U = \frac{n_1 n_2}{2} - u\right) = P\left(U - \frac{n_1 n_2}{2} = -u\right)$$

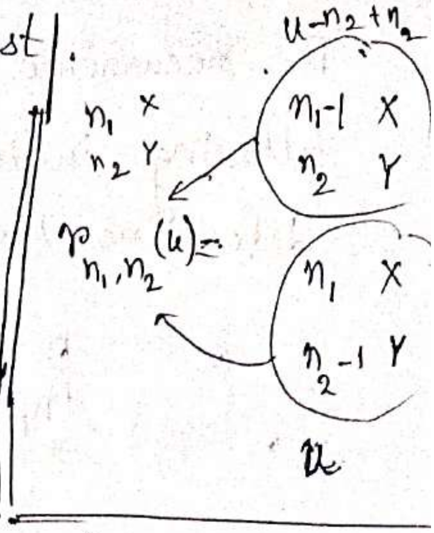
(Proved)

$P\left(U = \frac{n_1 n_2}{2} + u\right)$

$= \frac{r_{n_1 n_2}\left(\frac{n_1 n_2}{2} + u\right)}{\binom{n_1 + n_2}{n_1}}$

Let $\frac{n_1 n_2}{2} + u = v$

Because of this symmetry, only the lower tail critical values need to found for either a one or a two-tailed test].



Recurrence Relation:

Consider a sequence of n_1+n_2 letters being built up by adding a letter to the right of the sequence of n_1+n_2-1 letters.

If the n_1+n_2-1 letters consist of n_1 X and
(n_2-1) Y's, the extra letter must be an Y
but if an Y is added to the right, the numbers
of times an Y preceeds an X is unchanged.
On the other hand if the additional letter is an
X, which would be the case for (n_1-1) X and n_2
Y's in the original sequence, all the Y's preceeds
the new X and there are n_2 of them, so that
U is increased by n_2 units. Therefore we have

$$r_{n_1,n_2}(u) = r_{n_1,n_2-1}(u) + r_{n_1-1,n_2}(u-n_2)$$

$$P_{H_0}(U=u) = \frac{r_{n_1,n_2}(u)}{\binom{n_1+n_2}{n_1}} = \frac{r_{n_1,n_2-1}(u)}{\binom{n_1+n_2}{n_1}} + \frac{r_{n_1-1,n_2}(u-n_2)}{\binom{n_1+n_2}{n_1}}$$

$$= \frac{n_2}{n_1+n_2} \frac{r_{n_1,n_2-1}(u)}{\binom{n_1+n_2-1}{n_2-1}} + \frac{n_1}{n_1+n_2} \cdot \frac{r_{n_1-1,n_2}(u-n_2)}{\binom{n_1+n_2-1}{n_1-1}}$$

$$= \frac{n_2}{n_1+n_2} P_{n_1,n_2-1}(u) + \frac{n_1}{n_1+n_2} P_{n_1-1,n_2}(u-n_2)$$

$$, \; u = 0,1,2 \ldots, n_1 n_2$$

The recurrence relation ~~togetherly~~ together with the following initial and boundary conditions will completely determine the prob. distⁿ of U.

$$P_{n_1, n_2}(0) = \frac{1}{\binom{n_1+n_2}{n_1}}$$

$$P_{n_1, 0}(0) = 1$$

$$P_{n_1, 0}(u) = 0 \,, \; u>0$$

$$P_{0, n_2}(0) = 1$$

$$P_{0, n_2}(u) = 0 \,, \; u>0$$

Mean and Variance of U under H_0:

Define $\pi = P(Y<X)$

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

$$= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{x} f_Y(y)\, dy\right] f_X(x)\, dx$$

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

$$= \int_{-\infty}^{\infty} F_X(x) f_X(x)\, dx \,,$$ under H_0: $F_Y(x) = F_X(x) \; \forall x$

$$= \int_0^1 z\, dz = \frac{1}{2}$$

Under H_1 : $\delta > 0$ i.e. Y is stochastically larger

$$\pi < \frac{1}{2}$$

Similarly under H_1 : $\delta < 0$, i.e. Y is stochastically smaller, $\pi > \frac{1}{2}$

So that the null and the alternative hypothesis may also be represented as

$$H_0 : \pi = \frac{1}{2} \text{ against } H_1 : \pi < \frac{1}{2}$$
$$\pi > \frac{1}{2}$$
$$\pi \neq \frac{1}{2}$$

Thus the $n_1 n_2$ random variables $\phi(X_i, Y_j)$'s can be looked upon as bernoulli with

$$E(\phi(X_i, Y_j)) = \pi \quad \forall\, i, j$$
$$V(\phi(X_i, Y_j)) = \pi(1-\pi) \quad \forall\, i \neq j.$$

New notation

$$\phi_{ij} = \phi(X_i, Y_j)$$

For the joint moments note that this random variables are not independent whenever the X subscripts or Y subscripts are common, so that

$$\text{Cov}(\phi_{ij}, \phi_{hk}) = 0 \quad \forall\, i \neq h,\ j \neq k$$

$$\text{Cov}(\phi_{ij}, \phi_{ik}) = E(\phi_{ij}\phi_{ik}) - E(\phi_{ij})E(\phi_{ik})$$
$$= \pi_1 - \pi^2 \quad \forall\, j \neq k,\ i = 1(1)n_1$$

and $Cov(\phi_{ij}, \phi_{hj}) = E(\phi_{ij}\phi_{hj}) - E(\phi_{ij})E(\phi_{hj})$

$$= \pi_2 - \pi^2, \quad \forall\, i \neq h, \; j = 1(1)n_2$$

Here, $\pi_1 = E(\phi_{ij}, \phi_{ik})$

$$\phi_{ij}\phi_{ik} = \begin{cases} 1, & \text{if } X_i > Y_j \text{ and } X_i > Y_k \\ 0 & \text{o.w} \end{cases}$$

$$E(\phi_{ij}, \phi_{ik}) = P((X_i > Y_j) \cap (X_i > Y_k))$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} f_Y(y_j)\, dy_j\right]\left[\int_{-\infty}^{x} f_Y(y_k)\, dy_k\right] f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} [F_Y(x)]^2 f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} [F_X(x)]^2 f_X(x)\, dx, \quad \text{under } H_0$$

$$= \int_0^1 z^2\, dz = \frac{1}{3}, \quad \text{under } H_0 \; \forall\, i, \; \forall\, j \neq k$$

Here $\pi_2 = E(\phi_{ij}\phi_{hj})$

$$\phi_{ij}\phi_{hj} = \begin{cases} 1, & \text{if } X_i > Y_j \text{ and } X_h > Y_j \\ 0, & \text{o.w} \end{cases}$$

$$E(\phi_{ij}\,\phi_{hj}) = P\big((X_i > Y_j) \cap (X_h > Y_j)\big)$$

$$= \int_{-\infty}^{\infty}\left[\int_{y}^{\infty} f_X(x_i)\,dx_i\right]\left[\int_{y}^{\infty} f_X(x_h)\,dx_h\right] f_Y(y)\,dy$$

$$= \int_{-\infty}^{\infty} [1 - F_X(y)]^2 f_Y(y)\,dy$$

$$= \int_{-\infty}^{\infty} [1 - F_Y(y)]^2 f_Y(y)\,dy, \quad \text{under } H_0$$

$$= \int_0^1 (1-z)^2\,dz = \frac{1}{3}, \quad \text{under } H_0 \quad \forall j,\ \forall i \neq h$$

Thus, $E_{H_0}(U) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{2} = \frac{n_1 n_2}{2}$

$$V(U) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} V(\phi_{ij}) + \sum_{i=1}^{n_1}\sum_{j\neq K=1}^{n_2}\sum Cov(\phi_{ij}, \phi_{iK})$$

$$+ \sum_{j=1}^{n_2}\sum_{i\neq h=1}^{n_1}\sum Cov(\phi_{ij}, \phi_{hj}) + \sum_{i\neq h}^{n_1}\sum\sum_{j\neq K}^{n_2}\sum Cov(\phi_{ij}, \phi_{hK})$$

$$= n_1 n_2 \pi(1-\pi) + n_1 n_2 (n_2 - 1)(\pi_1 - \pi^2) + n_2 n_1 (n_1 - 1)(\pi_2 - \pi^2) + 0$$

$$\therefore V_{H_0}(U) = n_1 n_2 \left(\frac{1}{2}\cdot\frac{1}{2}\right) + n_1 n_2 (n_2 - 1)\left(\frac{1}{3} - \frac{1}{4}\right)$$

$$+ n_2 n_1 (n_1 - 1)\left(\frac{1}{3} - \frac{1}{4}\right)$$

$$= \frac{n_1 n_2}{4}\left[1 + \frac{n_2 - 1}{3} + \frac{n_1 - 1}{3}\right] = \frac{n_1 n_2}{12}(n_1 + n_2 + 1)$$

Since $E_{H_0}\left(\frac{U}{n_1 n_2} = \frac{1}{2}\right)$

& $V_{H_0}\left(\frac{U}{n_1 n_2}\right) \to 0$ as $n_1, n_2 \to \infty$

$\frac{U}{n_1 n_2}$ is a consistent estimator and the test based on $\frac{U}{n_1 n_2}$ is consistent for H_0 against H_1.

For large n_1, n_2

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \overset{a}{\sim} N(0,1)$$

C.R for large sample test

H_1	C.R
$\delta > 0$	$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} > \tau_\alpha$
$\delta < 0$	$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} < -\tau_\alpha$
$\delta \neq 0$	$\left\lvert \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \right\rvert > \tau_{\alpha/2}$

14 Y] 1
15 X] 1
16 Y] 1
17 Y
18 X
X] 1
19
20 Y
20 Y
21 Y] 1
22 X
23 X] 1 Ru
24 Y
27 Y] 1

Definition of
Similar
followed b

Example :

Null hypo
sample
with
$Y_1, Y_2,$
size

Wald-Wolfowitz's Run test :

14	Y] 1
15	X] 1
16	Y] 1
17	X
18	X] 1
19	X
20	Y
20	Y] 1
21	Y
22	X
23	X] 1
24	Y
27	Y] 1

Run = 7

X	Y
15	14
17	16
19	20
18	21
22	27
23	24
	20

$X_1 \cdot X_2 \cdot \cdot X_{n_1}$

$Y_1 \; Y_2 \cdot \cdot \cdot Y_{n_2}$

$X_1 \; X_2 \cdots X_{n_1}, Y_1 \; Y_2 \cdots Y_{n_2}$

Ascending order

Definition of Run : A run is a sequence of similar objects or symbols preceeded and followed by dissimilar ones.

Example : $\underbrace{Y}_{1} \underbrace{X}_{1} \underbrace{Y}_{1} \underbrace{X X X}_{1} \underbrace{Y Y Y}_{1} \underbrace{X X}_{1} \underbrace{Y Y}_{1}$

Here total no. of runs = 7

Null hypothesis : Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample of size n_1 drawn from a population with continuous distribution function $F(\cdot)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be another random sample of size n_2 drawn from another population with continuous distribution function $G(\cdot)$ such that

$G(x) = F(x-\delta)$; $\delta \in \mathbb{R}$.

F and G are both univariate and the samples are drawn independently of each other. We are going to test $H_0 : \delta = 0$

against

$$H_1 : \begin{array}{l} \delta < 0 \\ \delta > 0 \\ \delta \neq 0 \end{array}$$

<u>Test statistics and the testing procedure</u>

1. Let $\underset{\sim}{Z} = (X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2})$ be the combined sample. Arrange all $N = n_1 + n_2$ observations in increasing order as

$$Z_1 \le Z_2 \le \cdots \le Z_N,$$ where each ~~capital~~ Z_i is either an X observation or an Y observation.

2. Replace each observation by X or Y according as the population its comes from.

3. Count the total no. of runs. This is our test statistics, denoted by r.

<u>Critical Region</u> : If H_0 is not true then the no. of runs in each case will be dereased. So the critical region will be defined by the left tail i.e. H_0 is rejected at level α if $r \le r_\alpha$ where r_α is the largest integer satisfying $P_{H_0}(r \le r_\alpha) \le \alpha$

This is one ... of the
well known non-parametric tests for the two sample location problem.

$n_1 = 7$ $\quad$ $n_2 = 4$

X	Y
Y	X
X	Y
Y	X
X	Y
Y	X
X	Y
Y	X
X	X
X	X
X	X

9 $\quad$ 8

X	Y
Y	X
X	Y
Y	X
X	Y
Y	Y
X	Y
Y	X

$n_1 = 4$

$n_2 = 4$

run = 8.

Runs $= 2 \cdot \min(n_1, n_2) + 1$

min run

X Y

$= 2$

$\boxed{r \text{ lies between } 2 \text{ and } 2 \cdot \min(n_1, n_2) + 1}$

$P(r = 2d)$ $\quad$ $\dfrac{2\binom{n_1 - 1}{d-1}\binom{n_2 - 1}{d-1}}{\binom{n_1 + n_2}{n_1}}$

d $\quad$ d

(partition) $X \,|\, \cdots \,|\, X$ $\quad$ n_1 $\quad$ $d-1$

Null distribution : If H_0 is true then all the $\binom{n_1 + n_2}{n_1}$ distinguisable arrangements of n_1 X and n_2 Y's in a line are equally likely. We find the no. of arrangement of this $(n_1 + n_2)$ objects which gives a total of

r runs. We consider two cases separately - $r \to$ even and $r \to$ odd

Case I : $r = 2d$ (even)

This will happen if we have d runs of X and d runs of Y. The first run may be either an X or an Y. In order to get d runs of X, we have to partition the n_1 X into d groups, none of which is to be empty ($d \leq n_1$). This can be done by placing the X's in a line and putting each of $(d-1)$ bars between any two X in the line. There are overall $(n_1 - 1)$ positions in which each bar can be placed, thus giving a total of $\binom{n_1 - 1}{d-1}$ distinguishable arrangements of the n_1 X's in d groups. In a similar way there are $\binom{n_2 - 1}{d-1}$ distinguisable arrangements of the n_2 Y's in d groups. Hence the total no. of distinguishable arrangements giving d runs of X and d runs of Y beginning with a run of X is $\binom{n_1 - 1}{d-1} \cdot \binom{n_2 - 1}{d-1}$. Since under H_0 the $\binom{n_1 + n_2}{n_1}$ arrangements are equally likely, it follows that the total no. of distinguisable arrangements giving d runs of X and d runs of Y beginning with a run of Y is also ~~Hence~~

$$\binom{n_1 - 1}{d-1} \cdot \binom{n_2 - 1}{d-1}.$$

Hence

$$P_{H_0}(r = 2d) = \frac{2\binom{n_1-1}{d-1}\binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}} \quad \text{—(1)}$$

Case 2 : r odd $(r = 2d+1)$

To have $2d+1$ runs, either we have $d+1$ runs of X and d runs of Y, starting with a X run or d runs of X and $d+1$ runs of Y, starting with a Y run. The total number of distinguishable arrangements for the former is $\binom{n_1-1}{d-}\binom{n_2-1}{d-1}$ and that for the later is $\binom{n_1-1}{d-1}\binom{n_2-1}{d}$

under Ho, these arrangements are all equally likely.

$$P_{H_0}(r = 2d+1) = \frac{\binom{n_1-1}{d}\binom{n_2-1}{d-1} + \binom{n_1-1}{d-1}\binom{n_2-1}{d}}{\binom{n_1+n_2}{n_1}} \quad \text{—(2)}$$

The critical values of r based on (1) and (2) are given by Swed and Eisenhart.

For $n_1, n_2 > 10$, one can use the statistic

$\dfrac{r - E(r)}{\sqrt{V(r)}}$ which is asymptotically Normal (0,1) under H_0.

Treatment of ties : Break ties in all possible ways and calculate r for each such way of breaking ties. No problem if r values are identical or ~~not~~ as each such value of r leads to the same conclusion. Otherwise accept the largest value of r as a conservative test procedure. If there are too many ties accross the samples it is better to discard the samples and take the new ones.

$$\boxed{\binom{4}{2} = \frac{4!}{2!\,2!} = 6}$$

Example $\underset{\sim}{X} = (13, 18, 19, 19)$

$\underset{\sim}{Y} = (15, 15, 17, 19, 19, 21)$

Seq.	13	15	15	17	18	19	19	19	19	21	No. of runs
1	X	Y	Y	Y	X	X	X	Y	Y	Y	4
2	X	Y	Y	Y	X	X	Y	X	Y	Y	6
3	X	Y	Y	Y	X	X	Y	Y	X	Y	6
4	X	Y	Y	Y	X	Y	Y	X	X	Y	6
5	X	Y	Y	Y	X	Y	X	Y	X	Y	8
6	X	Y	Y	Y	X	Y	X	X	Y	Y	6

We take $r = 8$ as it maximizes the probability.

Expectation and Variance of r, under H_0

n_1 X 's Total no. of observations,

n_2 Y 's $N = n_1 + n_2$

Define, $$I_j = \begin{cases} 1 & \text{if } j^{th} \text{ symbol} \neq (j-1)^{th} \text{ symbol} \\ 0 & \text{otherwise} \end{cases}$$

$$j = 2, 3, \ldots N$$

Then
$$r = 1 + I_2 + I_3 + I_4 + \ldots + I_N$$

$$P(I_j = 1) = \frac{2n_1 n_2}{N(N-1)} \quad \forall\, j = 2(1)N$$

$$E(I_j) = \frac{2n_1 n_2}{N(N-1)}$$

$$E(r) = 1 + \sum_{j=2}^{N} \frac{2n_1 n_2}{N(N-1)} = 1 + \frac{2n_1 n_2}{N}$$

$$V(I_j) = E(I_j^2) - \{E(I_j)\}^2$$

$$= E(I_j)\left[1 - E(I_j)\right]$$

$$= \frac{2n_1 n_2}{N(N-1)}\left[1 - \frac{2n_1 n_2}{N(N-1)}\right]$$

$$Var(r) = var\left(1 + \sum_{j=2}^{N} I_j\right)$$

$$= \sum_{j=2}^{N} V(I_j) + \sum_{j \neq k}^{N}\sum cov(I_j, I_k)$$

To find $\sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} Cov(I_j, I_k)$

$$\sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} Cov(I_j, I_k) = \sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} \left[E(I_j I_k) - E(I_j)E(I_k) \right]$$

2nd term

$$\sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} E(I_j) E(I_k) = \sum_{j=2}^{N} \sum_{k=2}^{N} E(I_j) E(I_k) - (N-1)\{E(I_j)\}^2$$

$$= (N-1)^2 \left[\frac{2n_1 n_2}{N(N-1)}\right]^2 - (N-1)\left[\frac{2n_1 n_2}{N(N-1)}\right]^2$$

$$= \left(\frac{2n_1 n_2}{N}\right)^2 \left(1 - \frac{1}{N-1}\right)$$

$$= \frac{4n_1^2 n_2^2 (N-2)}{N^2(N-1)}$$

1st term

$$\sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} E(I_j I_k)$$

Total no. of terms $= (N-1)(N-2)$

~~The no.~~ Total no. of terms in which $|j-k| = 1$ is $2(N-2)$

[Here $I_{j-1} I_j = 1$ or $I_j I_{j+1} = 1$]

$|j-k| = 1$

$I_2\ I_3 \quad I_3\ I_4$

$I_5\ I_6 \quad I_7\ I_8$

$\vdots$

$I_{N-1}\ I_N$

- - - - -

$I_3\ I_2 \quad I_4\ I_3$

$\vdots$

(Type

No.

For

Prob. of

$XYX \to n_1$

$YXY \to n_2$

Possibilities

$j-1$	j	$k-1$	k
X	Y	X	Y
X	Y	Y	X
Y	X	X	Y
Y	X	Y	X

(Type $XYX \to I_2 = 1, I_3 = 1$).

No. of remaining terms in which $|j-k|>1$ is

$$(N-1)(N-2) - 2(N-2) = (N-2)(N-3)$$

[Ex: ~~XYXY~~

$$\begin{matrix} XYYX \\ YXXY \end{matrix} > \begin{matrix} I_2 = 1 & I_3 = 0 \\ I_4 = 1 & \end{matrix}]$$

For the terms with $|j-k| = 1$

$$E(I_j I_k) = P(I_j = 1, I_k = 1) \quad (|j-k| = 1)$$

Prob. of
$XYX \to n_1 n_2 (n_1 - 1)$
$YXY \to n_2 n_1 (n_2 - 1)$

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

$$= \frac{n_1 n_2 (n_1 + n_2 - 2)}{N(N-1)(N-2)} = \frac{n_1 n_2 \cancel{(N-2)}}{N(N-1)\cancel{(N-2)}}$$

$$= \frac{n_1 n_2}{N(N-1)}$$

For the remaining terms

$$E(I_j I_k) = \underbrace{P(I_j = 1, I_k = 1)}_{} \quad (|j-k|>1)$$

$$E(I_j I_k) = \frac{4 n_1 n_2 (n_1 - 1)(n_2 - 1)}{N(N-1)(N-2)(N-3)}$$

Position
$j-1$
j
$k-1$
k

Possibilities

	$j-1$	j	$k-1$	k
I_3 I_4	X	Y	X	Y
I_7 I_8	X	Y	Y	X
I_N	Y	X	X	Y
I_4 I_3	Y	X	Y	X

jth symbol $\neq (j-1)$th symbol

k^{th} symbol $\neq (k-1)^{th}$ symbol

$$\therefore \text{1st term} = 2(N-2)\frac{n_1 n_2}{N(N-1)} + \cancel{(N-2)}(N-3) \cdot \frac{4 n_1 n_2 (n_1 - 1)(n_2 - 1)}{N(N-1)\cancel{(N-2)}\cancel{(N-3)}}$$

$E(I_k)]$

$) - (N-1)\{E(I_j)\}^2$

$\left[\frac{2n_1 n_2}{N(N-1)}\right]^2$

$$= \frac{2n_1n_2}{N(N-1)}\left[(N-2) + 2(n_1-1)(n_2-1)\right]$$

$$= \frac{2n_1n_2}{N(N-1)}\left[n_1+n_2-\cancel{2}+2n_1n_2-2n_1-2n_2+\cancel{2}\right]$$

$$= \frac{2n_1n_2}{N(N-1)}\left[2n_1n_2-n_1-n_2\right]$$

$$= \cancel{\frac{2n_1n_2}{N(}}$$

$$\therefore \sum_{\substack{j=2 \\ j\neq k}}^{N}\sum_{k=2}^{N} Cov(I_j\, I_k) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{N(N-1)} - \frac{4n_1^2n_2^2(N-2)}{N^2(N-1)}$$

$$= \frac{2n_1n_2}{N(N-1)}\left[2n_1n_2-n_1-n_2-\frac{2n_1n_2(N-2)}{N}\right]$$

$$= \frac{2n_1n_2}{N^2(N-1)}\left[2n_1n_2(n_1+n_2)-(n_1+n_2)^2 - 2n_1n_2(N-2)\right]$$

$$= \frac{2n_1n_2}{N^2(N-1)}\left[\cancel{(n_1+n_2)n_1}\; 2n_1^2n_2 + 2n_1n_2^2 - \right]$$

$$= \frac{2n_1n_2}{N(N-1)}\left[2n_1n_2-n_1-n_2-\frac{2n_1n_2(n_1+n_2-2)}{n_1+n_2}\right]$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \Big[\cancel{2n_1^2 n_2} + \cancel{2n_1 n_2^2} - n_1^2 - n_1 n_2 - n_2^2 - n_1 n_2 - \cancel{2n_1^2 n_2} - \cancel{2n_1 n_2^2} + 4n_1 n_2 \Big]$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \left[-n_1^2 - n_2^2 + 2n_1 n_2 \right]$$

$$= -\frac{2n_1 n_2}{N^2(N-1)} \cdot (n_1 - n_2)^2$$

$$Var(r) = \sum_{j=2}^{N} V(I_j) + \sum_{\substack{j=2 \\ j \neq k}}^{N} \sum_{k=2}^{N} Cov(I_j, I_k)$$

$$= (N-1) \left[\frac{2n_1 n_2}{N(N-1)} \left\{ 1 - \frac{2n_1 n_2}{N(N-1)} \right\} \right] - \frac{(n_1 - n_2)^2 \, 2n_1 n_2}{N^2(N-1)}$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \left(N(N-1) - 2n_1 n_2 \right) - \frac{(n_1 - n_2)^2 \, 2n_1 n_2}{N^2(N-1)}$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \left[N(N-1) - 2n_1 n_2 - (n_1 - n_2)^2 \right]$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \left[\cancel{n_1 + n_2} \; N(N-1) - 2n_1 n_2 - (n_1^2 + n_2^2) + 4n_1 n_2 \right]$$

$$= \frac{2n_1 n_2}{N^2(N-1)} \left[N(N-1) - N^2 + 2n_1 n_2 \right]$$

$$= \frac{2n_1 n_2}{N^2(N-1)}\left[2n_1 n_2 - N\right] = \frac{2n_1 n_2}{N^2(N-1)}\left(2n_1 n_2 - n_1 - n_2\right).$$

Properties and uses of run test :

The run test is consistent against all types of differences in population. The very generality of the test ~~we can~~ weakens splits performance against specific alternatives — say difference in location only or difference in scale only. Its primary usefulness is in the preliminary analysis of the data when no particular form of the alternative is yet formulated. Then if the null hypothesis of equality is rejected by the run test, further studies can be initiated with some other test designed for specific alternatives in an attempt to identify the types of differences between the parameters.

Run test for testing randomness of sample

Run test is one of the best known and easiest to apply for testing randomness in a sequence of observations. The data may be dichopomous to start with (e.g. FMMFMFM). If actual measurements are collected, the data may be classified as a dichopotomous sequence according as each observation is above or below

some fixed numbers, often the calculated sample median or mean.
In the latter case, any observation equal to this fixed
number are ignored in the analysis and the sample
size is adjusted accordingly. The run test can be
used for one sided or two sided alternatives.
If the alternative hypothesis is simply non randomness
a two sided test should be used since the presence
of a trend ~~wh~~ would usually be indicated by a
clustering of similar objects which is reflected by an
unusually small no. of runs. A one sided test is
more appropriate for trend alternatives.