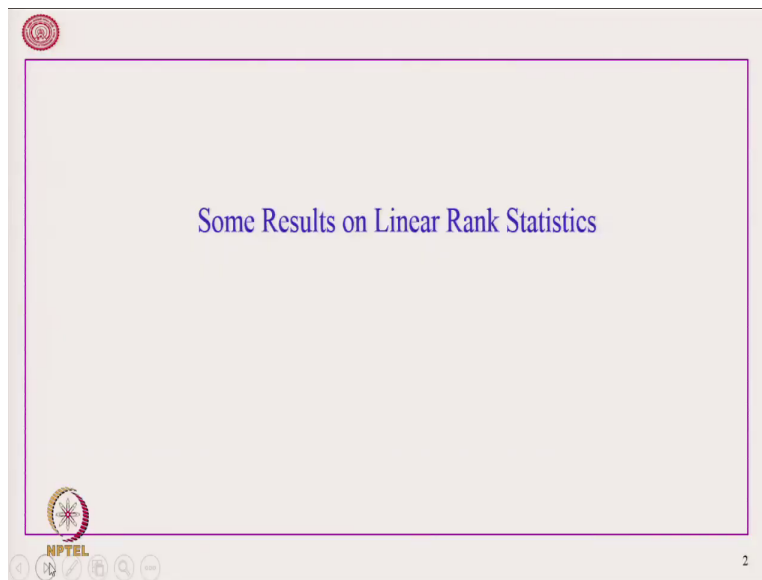


Nonparametric Statistical Inference
Professor Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture 5

Nonparametric Statistical Inference

Welcome students to MOOCs series of lectures on nonparametric statistical inference, this is lecture number 5. As I said at the end of the last class, that in this class, we shall study some mathematical results on linear rank statistics.

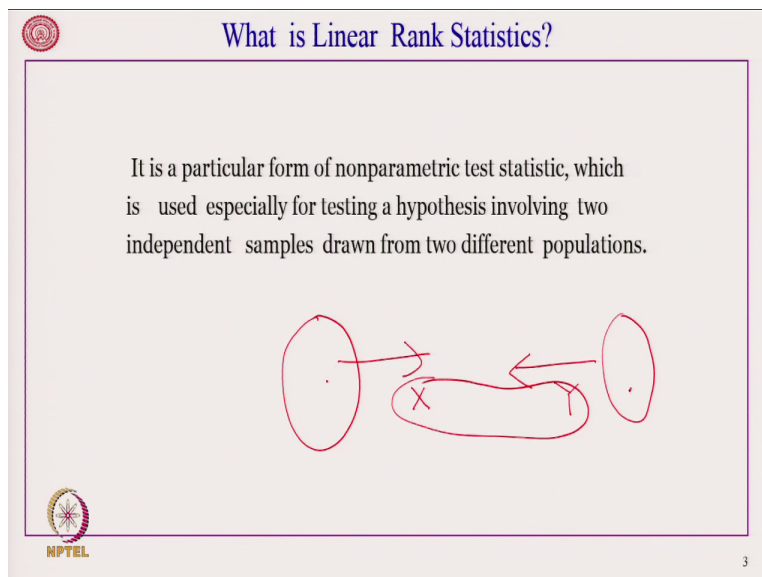
(Refer Slide Time: 00:43)



Some Results on Linear Rank Statistics

NPTEL

2



What is Linear Rank Statistics?

It is a particular form of nonparametric test statistic, which is used especially for testing a hypothesis involving two independent samples drawn from two different populations.



NPTEL

3

So, what is linear rank statistics? It is a particular form of nonparametric test statistic, which is used especially for testing a hypothesis involving 2 independent samples drawn from 2 different populations. So, consider 1 population from which we are drawing a sample X. There is another population from which you are drawing a sample Y.

And we want to compare these 2 samples to understand the behavior or the properties of these 2 populations.

(Refer Slide Time: 01:29)



What is Linear Rank Statistics?

We have already seen application of **Linear rank statistic**, for various tests such as:

- **Wilcoxon Rank Sum Test** : to compare medians of X and Y
- **Moods Test**: to compare dispersion of X and Y.



NPTEL

4

We have already seen some applications of linear rank statistics for various tests. In particular, you remember that we have discussed Wilcoxon Rank Sum Test to compare the medians of X and Y also we have done Moods test, **Freund Ansari test** and David Barton's test, they also have use of linear rank statistic.

(Refer Slide Time: 01:55)



What is Linear Rank Statistics?

It has been defined as a linear function of the indicator variables, that is

$$T_N(Z) = \sum_{i=1}^N a_i Z_i$$

where a_i s are given constants called *weights* or *scores*.

Z_i indicates whether the i^{th} sample in the pooled sorted array is from X population or not.

$T_N(Z) = a^T Z$

7: $\begin{array}{ccccccc} 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ X & Y & X & X & Y & Y & X \\ \hline Z & (1 & 0 & 1 & 1 & 0 & 0 & 1) \end{array}$


5

So, if we remember, the linear rank statistic is of the form, $T_N(Z) = \sum_{i=1}^N a_i Z_i$

where Z_i is an indicator variable. In all our examples, we typically looked at the Z_i value is equal to 1 for X and Z_i value is equal to 0 for Y. Therefore, if we have 1 0 1 1 0 0 1 that means, there are 7 observations of which 4 are from X and 3 are from Y. And not only that, in the sorted order, they look like this.

Therefore, Z vector will look like 1 0 1 1 0 0 1. Now, what is a_i ? a_i is a coefficient depending upon the nature of the test we will use different a_i 's. As we will see later or we have already said therefore, $T_N(Z)$ is equal to the dot product of a vector with Z written as a column.

(Refer Slide Time: 03:30)



What is Linear Rank Statistics?

We shall restrict ourselves to those Linear Rank Statistics which are particularly sensitive to differences only in location and only in scale.

Spread Median



6

For this course, we shall restrict ourselves only to those linear rank statistic, which are particularly sensitive for differences in location that is median and only in scale that is spread.

(Refer Slide Time: 03:47)



Assumption:

Two independent random samples $X_1, X_2, \dots, X_m$ & $Y_1, Y_2, \dots, Y_n$
Are drawn from two populations with continuous distributions F_X and F_Y , respectively.

↓ ↓



7

The assumptions is that we have taken 2 independent random samples, X_1, X_2, X_m from 1 population Y_1, Y_2, Y_n from another population and essentially we want to see that the corresponding cumulative distribution functions that is F_X and F_Y are same or not that is what we

want to test. So, samples are taken like that from the F_X and F_Y distribution only important assumption that they are continuous.

(Refer Slide Time: 04:31)



Assumption:

Two independent random samples $X_1, X_2, \dots, X_m$ & $Y_1, Y_2, \dots, Y_n$
Are drawn from two populations with continuous distributions
 F_X and F_Y , respectively.

Under the Null Hypothesis we want to test whether these two
samples are from the same continuous distribution $F(x)$:

i.e. $H_0: \underline{F_X(x)} = \underline{F_Y(x)} = \underline{F(x)}$ $\forall x \in \mathbb{P}$
unknown



NPTEL

8

And as I said, under the null hypothesis we want to test whether they have the same distribution that is a $F_X(x)$ is equal to $F_Y(x)$ is equal to F_X , which is unknown for all x belonging to $\mathbb{R}$.

(Refer Slide Time: 04:55)



To test this we use

- (1) the ordered arrangement of the combined sample from the two populations
- (2) in particular, their **rank**s in the combined population.



9

And to test this, we use the ordered arrangement of the combined sample from the 2 populations in particular, their ranks in that combined population.

(Refer Slide Time: 05:06)

Mathematical Definition

The intuitive idea of ranks for X observations & Y observations can be defined mathematically as follows.

The rank of x_i in the combined population is:

$$r_{XY}(x_i) = \sum_{k=1}^m S(x_i - x_k) + \sum_{k=1}^n S(x_i - y_k)$$

Where the S indicates a step function:

$$S(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$


10

So, mathematically, we can define rank for X observations and Y observations in the following way. The rank of x_i , the i th x observation in the combined population is

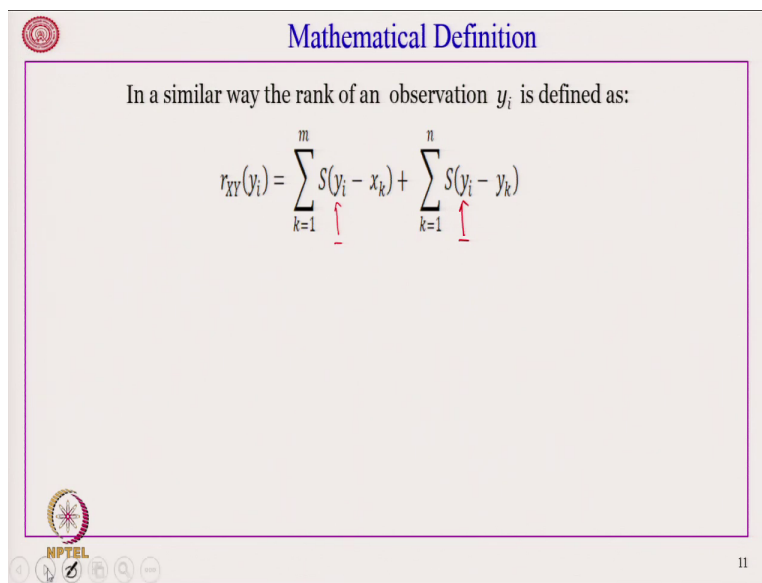
$$r_{XY}(x_i) = \sum_{k=1}^m S(x_i - x_k) + \sum_{k=1}^n S(x_i - y_k)$$

where S indicates a step function,

$$S(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

Therefore, S function looks like this it is 0, before 0, from 0 onwards it is 1. So, this is a type of step function that S(x) denotes.

(Refer Slide Time: 06:22)



Mathematical Definition

In a similar way the rank of an observation y_i is defined as:

$$r_{XY}(y_i) = \sum_{k=1}^m S(y_i - x_k) + \sum_{k=1}^n S(y_i - y_k)$$

The slide includes a NPTEL logo in the bottom left corner and a slide number '11' in the bottom right corner.

In a similar way, the rank of an Y observation y_i can be defined as summation over k is equal to 1 to m $S(y_i - x_k)$ plus k is equal to 1 to n $S(y_i - y_k)$. So, this is the 1 for which you are trying to find the rank and this we are subtracting all X values and Y values and based on that difference, we are trying to calculate the rank.

(Refer Slide Time: 06:58)

Illustration:

Sample	1.1	2.2	3.3	4.4	5.5	6.6	7.7
Rank	1	2	3	4	5	6	7
Diff.	$5.5 - 1.1$	$5.5 - 2.2$	$5.5 - 3.3$	$5.5 - 4.4$	$5.5 - 5.5$	$5.5 - 6.6$	$5.5 - 7.7$
Sign	+	+	+	+	0	-	-
S()	1	1	1	1	1	0	0

$$S(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$
 Therefore, Rank of 5.5 is 5

Let me illustrate with an example. Suppose my combined sample is 1.1, 2.2 3.3 like up to 7.7; what is the rank of 5.5? So, it is an Y observation let us assume, that blacks are Y's and X's are in red. So, its rank as we can easily see is 5, how to do that. So, you are subtracting all the observations from 5.5. Therefore, 5.5 minus 1.1, 5.5 minus 2.2, 5.5 minus 3.3, up to 5.5 minus 7.7. Quite naturally, these are all plus and this is 0 and these are minus. Therefore, S values is 1, 1, 1, 1 and 0 here. So, how many ones are there? There are 5 ones. Therefore, rank of 5.5 is equal to 5 that in any case, we have tabulated here.

(Refer Slide Time: 08:18)

Illustration:

In a similar way the rank of 1.1 is 1.

Sample	1.1	2.2	3.3	4.4	5.5	6.6	7.7
Diff.	$1.1 - 1.1$	$1.1 - 2.2$	$1.1 - 3.3$	$1.1 - 4.4$	$1.1 - 5.5$	$1.1 - 6.6$	$1.1 - 7.7$
Sign	0	-	-	-	-	-	-
S()	1	0	0	0	0	0	0

$\therefore \text{Rank}(1.1)$

NPTEL

13

Another example what is going to be the rank of 1.1? So, what we are doing from 1.1 we are subtracting all the observations and all of them are minus except 0. So, in the S function, we get only 1, 1 and remaining all zeros. Therefore, rank of 1.1 is equal to 1.

(Refer Slide Time: 08:46)



In this lecture we prove some important results on the distribution properties of Linear Rank Statistics of the form

$$T_N(Z) = \sum_{i=1}^N a_i Z_i$$


14



In this lecture we prove some important results on the distribution properties of Linear Rank Statistics of the form

$$T_N(Z) = \sum_{i=1}^N a_i Z_i$$

Where : $N = m + n$

$$Z_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ observation in the combined population is an } x \\ 0 & \text{if } i^{\text{th}} \text{ observation in the combined population is an } y \end{cases}$$


15

In this lecture, we prove some important results on the distribution properties of the linear rank statistic of the form

$$T_N(Z) = \sum_{i=1}^N a_i Z_i$$

As I said earlier that we will be focusing on X observations. Therefore, Z_i is equal to 1 if i^{th} observation in the combined population is an X and it is 0. If i^{th} observation in the combined population is an y.

(Refer Slide Time: 09:18)

Result 1

Under the null hypothesis $H_0: F_X(z) = F_Y(z) = F(z) \forall z$, we have

$\forall i = 1, 2, \dots, N$

1. $E(Z_i) = \frac{m}{N}$ ← $m = \# \text{ X observation}$
2. $Var(Z_i) = \frac{mn}{N^2}$ ← $N = \text{Total no. of observation}$
3. $Cov(Z_i, Z_j) = -\frac{mn}{N^2(N-1)}$ ← $n = \# \text{ Y observation}$

$\forall j = 1, 2, \dots, N; i \neq j$

NPTEL 16

So, result 1 under the null hypothesis that is, both of them are from the same distribution,

$$F_X(z) = F_Y(z) = F(z) \forall z, \forall i = 1, 2, \dots, N$$

Expected value of Z_i is equal to m by N , m is equal to number of X observations and N is total number of observations. Variance of Z_i is equal to mn upon N square, where n is equal to number of y observations and covariance of Z_i and Z_j is equal to minus mn upon N square into N minus 1 when j is different from i .

1. $E(Z_i) = \frac{m}{N}$
2. $Var(Z_i) = \frac{mn}{N^2}$
3. $Cov(Z_i, Z_j) = -\frac{mn}{N^2(N-1)}$

$\forall j = 1, 2, \dots, N; i \neq j$

(Refer Slide Time: 10:35)



Proof

Observe:

Under the null hypothesis H_0 , Z_i is a Bernoulli random variable where success is considered when the i^{th} random variable in the combined ordered sample is from the X sample

$\therefore$ Each Z_i can be thought of as a Bernoulli r.v. with $p = \frac{m}{N}$



$x \ x \ x \ x \ x \ x \ x \ x$
 $P(Z_i = 1) = \frac{m}{N}$ m of them are 1 $\leftarrow X$
 $P(Z_i = 0) = \frac{n}{N}$ n of them are 0 $\leftarrow Y$
 Similar to coin tossing with $p = \frac{m}{N}$.

Under the null hypothesis Z_i is a Bernoulli distribution or Bernoulli random variable where success is considered when the i^{th} random variable in the combined ordered sample is from X . I hope you understand it, but let us assume that these are the n ordered observations, out of which m of them are 1, and n of them are 0.

Because these are coming from X and these are coming from Y . Therefore, when we randomly pick up 1 value, probability Z_i is equal to 1 is equal to m by N and probability Z_i is equal to 0 is equal to n by capital N . Therefore, it is similar to a coin tossing with p is equal to, m by N . Therefore, each Z_i can be thought of, as a Bernoulli random variable with p is equal to, m by N .

(Refer Slide Time: 12:46)

Proof

Observe:

Under the null hypothesis H_0 , Z_i is a Bernoulli random variable where success is considered when the i^{th} random variable in the combined ordered sample is from the X sample.

Since, the event of i^{th} rank corresponding to X and Y are equally likely under H_0 , the probability of success is:

$$P(Z_i = 1) = \frac{m}{N}$$
$$\therefore P(Z_i = 0) = \frac{n}{N}$$

Thus, $E(Z_i) = 1 * P(Z_i = 1) + 0 * P(Z_i = 0) = 1 * \frac{m}{N} + 0 * \frac{n}{N} = \frac{m}{N}$



18

This is what we have written here. That probability Z_i is equal to 1 is equal to m by N , Z_i is equal to 0 is equal to n by N . Therefore, the expected value of Z_i is equal to

$$1 * P(Z_i = 1) + 0 * P(Z_i = 0) = 1 * \frac{m}{N} + 0 * \frac{n}{N} = \frac{m}{N}$$

(Refer Slide Time: 13:14)

Now,

$$E(Z_i^2) = 1^2 * P(Z_i = 1) + 0^2 * P(Z_i = 0) = 1 * \frac{m}{N} + 0 * \frac{n}{N} = \frac{m}{N}$$

$$\therefore \text{Var}(Z_i) = E(Z_i^2) - E(Z_i)^2 = \left(\frac{m}{N}\right) - \left(\frac{m}{N}\right)^2 = \frac{m}{N} - \frac{m^2}{N^2}$$

$$= \frac{mN - m^2}{N^2}$$

$$= \frac{m(N - m)}{N^2}$$

$$= \frac{mn}{N^2}$$

Handwritten notes on the slide include: $\text{Var}(X) = E(X^2) - (E(X))^2$ and a checkmark next to the final result.

Now variance, we know that variance of a random variable X is equal to the expected value of X square minus expected value of X whole square. So, let us first calculate expected value of Zi square, this is equal to

$$E(Z_i^2) = 1^2 * P(Z_i = 1) + 0^2 * P(Z_i = 0) = 1 * \frac{m}{N} + 0 * \frac{n}{N} = \frac{m}{N}$$

Therefore, variance of Zi is equal to

$$E(Z_i^2) - E(Z_i)^2 = \left(\frac{m}{N}\right) - \left(\frac{m}{N}\right)^2 = \frac{m}{N} - \frac{m^2}{N^2}$$

$$= \frac{mN - m^2}{N^2}$$

If we take m common, it is

$$\frac{m(N - m)}{N^2}$$

$$= \frac{mn}{N^2}$$

Therefore, mn upon capital N square is the variance of each Zi.

(Refer Slide Time: 14:26)



Result 1

Under the null hypothesis $H_0: F_X(z) = F_Y(z) = F(z) \forall z$, we have
 $\forall i = 1, 2, \dots, N$

1. $E(Z_i) = \frac{m}{N}$ ✓
2. $Var(Z_i) = \frac{mn}{N^2}$ ✓
3. $Cov(Z_i, Z_j) = -\frac{mn}{N^2(N-1)}$
 $\forall j = 1, 2, \dots, N; i \neq j$



20

So, these 2 results have been proved. Now, we need to prove that covariance Z_i, Z_j is equal to

$$-\frac{mn}{N^2(N-1)}$$

(Refer Slide Time: 14:43)



We know that $Cov(X, Y) = E(X*Y) - E(X)*E(Y)$

With that we compute $Cov(Z_i, Z_j) = E(Z_i*Z_j) - E(Z_i)*E(Z_j)$



21

We know that covariance of $X Y$ is equal to expected value of $X Y$ minus the expected value of X times the expected value of Y . Therefore, when we are trying to compute covariance of Z_i and

Z_j this is going to be expected value of $Z_i Z_j$ minus expected value of Z_i times the expected value of Z_j .

(Refer Slide Time: 15:11)



We know that $\text{Cov}(X, Y) = E(X*Y) - E(X)*E(Y)$
With that we compute $\text{Cov}(Z_i, Z_j) = E(Z_i*Z_j) - E(Z_i)*E(Z_j)$

Now, $E(Z_i Z_j) = 1 * \underbrace{P(Z_i = 1, Z_j = 1)}$

$Z_i * Z_j = \begin{cases} 1 & \text{when both are 1} \\ 0 & \text{if at least one of them is 0} \end{cases}$



NPTEL

22

Now expected value of $Z_i Z_j$ is equal to

$$1 * P(Z_i = 1, Z_j = 1)$$

Why? Because $Z_i Z_j$, the product is going to be 1 when both are 1 and this is going to be 0 if at least 1 of them is 0.

(Refer Slide Time: 15:55)



Now ,

Both Z_i and Z_j will be 1 if out of m observations of X one is at the i^{th} position and one is at the j^{th} position of the combined population.



NPTEL

23

Now, both Z_i and Z_j will be 1 if out of the m of observations X 1 is at the i^{th} position and 1 is at the j^{th} position in the combined population.

(Refer Slide Time: 16:09)



Now ,

Both Z_i and Z_j will be 1 if out of m observations of X one is at the i^{th} position and one is at the j^{th} position of the combined population.

The no. of ways of choosing 2 out of m observation of X is $\binom{m}{2}$

and

The no. of possible choices to be assigned ranks i and j is $\binom{N}{2}$



NPTEL

24

Therefore, the number of ways of choosing 2 out of m observations of X is mC_2 and the total number of possible choices is capital NC_2 .

(Refer Slide Time: 16:24)


$$\begin{aligned}\text{Thus, } P(Z_i = 1, Z_j = 1) &= \frac{\binom{m}{2}}{\binom{N}{2}} = \frac{\frac{m(m-1)}{2}}{\frac{N(N-1)}{2}} = \frac{m(m-1)}{N(N-1)} \\ \Rightarrow E(Z_i Z_j) &= \frac{m(m-1)}{N(N-1)}\end{aligned}$$


NPTEL

25

Hence, probabilities Z_i is equal to 1 and Z_j is equal to 1 is

$$\frac{\binom{m}{2}}{\binom{N}{2}} = \frac{\frac{m(m-1)}{2}}{\frac{N(N-1)}{2}} = \frac{m(m-1)}{N(N-1)}$$

Therefore, the expected value of Z_i is Z_j is equal to

$$\frac{m(m-1)}{N(N-1)}$$

(Refer Slide Time: 16:53)



$$\begin{aligned}
 &\therefore \text{Cov}(Z_i Z_j) \\
 &= E\left(\underbrace{(Z_i - E(Z_i))}_{\checkmark} \underbrace{(Z_j - E(Z_j))}_{\checkmark}\right) \\
 &= E\left(\underbrace{Z_i Z_j}_{\checkmark} - \underbrace{Z_i E(Z_j)}_{\checkmark} - \underbrace{Z_j E(Z_i)}_{\checkmark} + \underbrace{E(Z_i)E(Z_j)}_{\checkmark}\right) \\
 &= E\left(\underbrace{Z_i Z_j}_{\checkmark} - \underbrace{Z_i \frac{m}{N}}_{\checkmark} - \underbrace{Z_j \frac{m}{N}}_{\checkmark} + \underbrace{\frac{m^2}{N^2}}_{\checkmark}\right) \\
 &= \underbrace{E(Z_i Z_j)}_{\checkmark} - \underbrace{\frac{m}{N} E(Z_i)}_{\checkmark} - \underbrace{\frac{m}{N} E(Z_j)}_{\checkmark} + \underbrace{\frac{m^2}{N^2}}_{\checkmark} \\
 &= \frac{m(m-1)}{N(N-1)} - \frac{m}{N} \cdot \frac{m}{N} - \frac{m}{N} \cdot \frac{m}{N} + \frac{m^2}{N^2}
 \end{aligned}$$


26

Therefore, covariance of $Z_i Z_j$ which is defined as

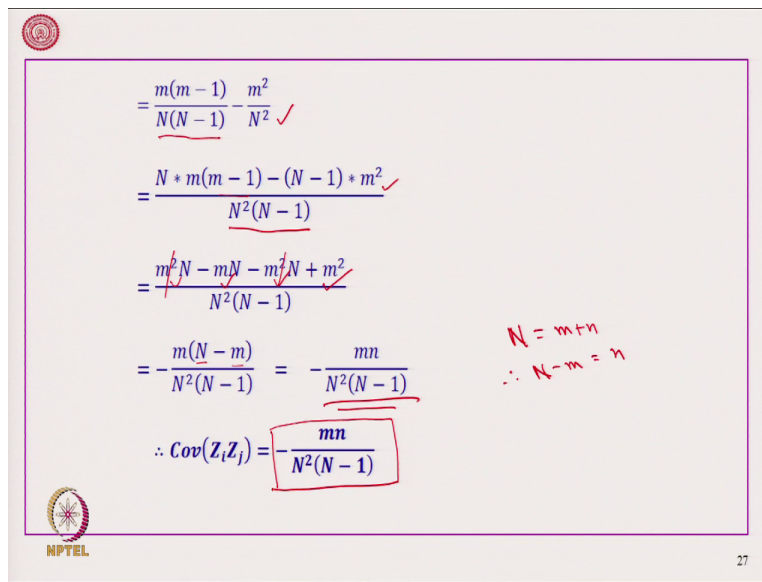
$$\begin{aligned}
 &E\left((Z_i - E(Z_i))(Z_j - E(Z_j))\right) \\
 &E\left(Z_i Z_j - Z_i E(Z_j) - Z_j E(Z_i) + E(Z_i)E(Z_j)\right)
 \end{aligned}$$

Which comes out to be expected value of Z_i is Z_j minus Z_i into m upon capital N , because for each Z the expected value is small m upon capital N . Similarly, expected value of this is also m upon capital N and both of them are giving me small m square upon capital N square. Now, by linearity of expectation with this is equal to

$$E(Z_i Z_j) - \frac{m}{N} E(Z_i) - \frac{m}{N} E(Z_j) + \frac{m^2}{N^2}$$

This we have already calculated. So, this is coming out to be m square upon capital N square with a minus sign plus m square upon capital N square.

(Refer Slide Time: 18:35)



$$\begin{aligned}
 &= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} \checkmark \\
 &= \frac{N * m(m-1) - (N-1) * m^2}{N^2(N-1)} \checkmark \\
 &= \frac{\cancel{m^2}N - m\cancel{N} - \cancel{m^2}N + m^2}{N^2(N-1)} \checkmark \\
 &= -\frac{m(N-m)}{N^2(N-1)} = -\frac{mn}{N^2(N-1)} \\
 &\therefore \text{Cov}(Z_i, Z_j) = \boxed{-\frac{mn}{N^2(N-1)}}
 \end{aligned}$$

$N = m+n$
 $\therefore N-m = n$

Which is equal to

$$\frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2}$$

Let us now expand them. Therefore, we get in the denominator

$$N^2(N-1)$$

numerator is coming out to be

$$N * m(m-1) - (N-1) * m^2$$

So, these 2 cancel out, if I take minus common, we are left with capital N minus small m and capital N is equal to m plus n. Therefore, N minus m is equal to small n and therefore, we get the result. So, this is proved.

(Refer Slide Time: 19:47)

In a similar way the conditional probability

$$P(Z_i = 1 | Z_j = 1) = \frac{P(Z_i = 1 \cap Z_j = 1)}{P(Z_j = 1)}$$

$$= \frac{m(m-1)}{N(N-1)} * \frac{N}{m} = \frac{(m-1)}{(N-1)}$$

Intuitively,
 if $Z_j = 1$ is given $\Rightarrow$ in the j^{th} position there is an X.
 $\therefore Z_i = 1$ can happen in how many ways an X can be chosen out of $(m-1)$ many remaining X's.

$P(Z_i = 1 | Z_j = 1)$
 $\& P(Z_i = 0 | Z_j = 1)$
 $\& P(Z_i = 0 | Z_j = 0)$

NPTEL 28

In a similar way, let us prove another result, which is from conditional probability. What is the conditional probability that is Z_i is equal to 1 given Z_j is equal to 1. We know that from the definition of conditional probability probability Z_i is equal to 1 given Z_j is equal to 1 is equal to

$$\frac{P(Z_i = 1 \cap Z_j = 1)}{P(Z_j = 1)}$$

which is coming out to be,

$$\frac{m(m-1)}{N(N-1)} * \frac{N}{m}$$

which is coming out to be

$$\frac{(m-1)}{(N-1)}$$

Intuitively it is clear. If Z_j is equal to 1 is given implies that in the j^{th} position there is an X. Therefore, Z_i is equal to 1 can happen in how many ways an X can be chosen out of small m minus 1 many remaining X's. Thus, we get a minus 1 in the numerator in the denominator. It is

very natural that out of the remaining capital N minus 1 observation anyone can come there. Therefore, the probability is going to be small m minus 1 upon capital N minus 1.

Try to prove the results, what is the probability Z_i is equal to 1 given Z_j is equal to 0 and probability Z_i is equal to 0 given Z_j is equal to 1 and probability Z_i is equal to 0 given Z_j is equal to 0. I will leave these as an exercise for you. This is simple combinatorial problems, but you will be able to do it.

(Refer Slide Time: 22:55)



Result 2

Let $T_N = \sum_{i=1}^N a_i Z_i$ where $a_1, a_2, \dots, a_N$ are constants. Then

under the Null Hypothesis $H_0: F_X(z) = F_Y(z) = F(z) \forall z$

a) $\underline{E(T_N)} = m \sum_{i=1}^N \left(\frac{a_i}{N} \right)$

b) $\underline{Var(T_N)} = \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right]$



29

Now, let us go to the result 2 . Let T_N is equal to

$$\sum_{i=1}^N a_i Z_i$$

where a_1, a_2 are our constants. Then under the null hypothesis expected value of T_N is equal to

$$m \sum_{i=1}^N \frac{a_i}{N}$$

and variance of T_N is equal to

$$\frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right]$$

(Refer Slide Time: 23:31)



$$a) \quad E(T_N) = m \sum_{i=1}^N \frac{a_i}{N}$$

Proof: $E(T_N) = \sum_{i=1}^N a_i E(Z_i)$

$$= \sum_{i=1}^N a_i \left(\frac{m}{N} \right) \checkmark$$

$$= m \sum_{i=1}^N \frac{a_i}{N}$$

Handwritten notes on the right:

$$T_N = a_1 Z_1 + a_2 Z_2 + \dots + a_N Z_N$$

$$\therefore E(T_N)$$

$$= a_1 (E(Z_1)) + a_2 (E(Z_2)) + \dots + a_N (E(Z_N))$$


30

So, let us prove it the expected value of T_n is equal to

$$m \sum_{i=1}^N \frac{a_i}{N}$$

It is very straightforward, because T_n is equal to a_1 times Z_1 plus a_2 times Z_2 plus a_N times Z_N .

Therefore, expected value of T_n is equal to

$$\sum_{i=1}^N a_i E(Z_i)$$

So, this is what we have written is equal to

$$\sum_{i=1}^N a_i \left(\frac{m}{N} \right)$$

If we take m out of the summation, we get the result. It is a very straightforward thing, but what about variance?

(Refer Slide Time: 24:44)



$$b) \text{Var}(T_N) = \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right] \quad \checkmark$$

Proof:

$$\begin{aligned} \text{Var}(T_N) &= \sum_{i=1}^N a_i^2 \text{Var}(Z_i) + \sum_{i=1}^N \sum_{j \neq i}^N a_i a_j \text{Cov}(Z_i, Z_j) \\ &= \sum_{i=1}^N a_i^2 \left(\frac{mn}{N^2} \right) + \sum_{i=1}^N \sum_{j \neq i}^N a_i a_j \left(\frac{mn}{N^2(N-1)} \right) \end{aligned}$$

$\text{Var}(aX + bY)$
 $= a^2 \text{Var}(X) \checkmark$
 $+ b^2 \text{Var}(Y) \checkmark$
 $+ 2ab \text{Cov}(X, Y)$

$$= \frac{mn \sum_{i=1}^N a_i^2}{N^2} - \frac{mn \sum_{i=1}^N \sum_{j \neq i}^N a_i a_j}{N^2(N-1)}$$


31

How do we show that variance of T_n is equal to this complicated quantity. So, let us start we know that variance of aX plus bY is equal to a square times variance of X plus b square times variance of Y plus $2ab$ covariance of X and Y . When we generalize it, we get variance of T_n is equal to a_i square times variance of Z_i , i is equal to 1 to n plus, we are looking at all pairs and trying to find the covariance of Z_i and Z_j , this is equal to σa_i square multiplied by mn upon N square.

This we have already found that variance of Z_i is equal to this plus the double summation $a_i a_j$ multiplied by minus mn upon N square into N minus 1, this we have already found out. Therefore, this is coming out to be a mn upon N square σa_i square minus, mn comes out of the summation, therefore, we get double summation i and j , i not equal to j , $a_i a_j$ divided by N square into N minus 1.

(Refer Slide Time: 26:36)



$$\begin{aligned}
 &= \frac{mn}{N^2(N-1)} \left[(N-1) * \sum_{i=1}^N a_i^2 - \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N a_i a_j \right] \\
 &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \sum_{i=1}^N a_i^2 - \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N a_i a_j \right] \\
 &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i^2 + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N a_i a_j \right) \right] \\
 &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right] \\
 \therefore \text{Var}(T_N) &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right]
 \end{aligned}$$


Now, if we take a mn upon N square into N minus 1 common. We have capital N minus 1 multiplied by sigma a_i square minus this summation double summation i is equal to 1 to N , j is equal to 1 to N , $a_i a_j$, j not equal to i . Now, we break it as N into sigma a_i square minus sigma a_i square minus this double summation. Now, we put this inside the summation. Therefore, what we are getting N time sigma a_i square minus sigma a_i square and sigma sigma $a_i a_j$.

This is nothing but sigma over a_i whole square. Therefore, the whole thing boils down to mn , N into N square minus 1 multiplied by N into sigma a_i square minus sigma whole square and thus we get the result for variance.

(Refer Slide Time: 27:48)

Let us Revisit Wilcoxon Rank Sum Test

NPTEL

33

The Wilcoxon Ranksum test statistic W is : $\sum_{i=1}^N i Z_i$ *ai*

We have stated that **under H_0** (Please Refer to Lecture 3)

$$E(W) = \frac{m(N+1)}{2}$$

and

$$Var(W) = \frac{mn(N+1)}{12}$$

It is obvious that

$$\underline{E(W)} = m \sum_{i=1}^N \frac{a_i}{N} = m \sum_{i=1}^N \frac{i}{N} = \frac{m(N+1)}{2}$$

stated in Lecture 3

$\rightarrow \frac{m}{N} \cdot \frac{N(N+1)}{2}$

NPTEL

34

Now, let us revisit Wilcoxon Rank Sum Test. We have found out that in the Wilcoxon Rank Sum Test, this statistic is W , which is the sum of ranks of the X observations and therefore, we have written it as $\sum i Z_i$. Therefore, the expected value of W is equal to

$$\frac{m(N+1)}{2}$$

and variance of W is equal to

$$\frac{mn(N+1)}{12}$$

This we have stated in lecture 3. Now, let us prove it the expected value of W is equal to m times sigma over i is equal to 1 to N, ai upon N. This we have just found out now replace ai with i because that is the value of ai in this summation that is ai. Therefore, we can write it as m times sigma over i 1 to N, i upon capital N and this is, is equal to. If I write, 1 step further m upon capital N into N into N plus 1 by 2, N gets cancelled and therefore, we get the result.

(Refer Slide Time: 29:17)

Wilcoxon Rank Sum Test: Revisit

$$\begin{aligned}
 \text{Var}(W) &= \frac{mn}{N^2(N-1)} [N \sum_{i=1}^N a_i^2 - (\sum_{i=1}^N a_i)^2] \\
 &= \frac{mn}{N^2(N-1)} [N \sum_{i=1}^N i^2 - (\sum_{i=1}^N i)^2] \quad \text{put } a_i = i \\
 &= \frac{mn}{N^2(N-1)} \left[N \left(\frac{N(N+1)(2N+1)}{6} \right) - \left(\frac{N(N+1)}{2} \right)^2 \right] \\
 &= \frac{mn}{N^2(N-1)} \left[\frac{N^2(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4} \right] \\
 &= \frac{mn}{N^2(N-1)} * \frac{N^2}{12} [2 * (N+1)(2N+1) - 3 * (N+1)^2]
 \end{aligned}$$

35

In a similar way, what is going to be the variance. Now, variance of W is equal to this term that we have already calculated. Now, let us put the value of ai. Therefore, this boils down to

$$\frac{mn}{N^2(N-1)} [N \sum_{i=1}^N a_i^2 - (\sum_{i=1}^N a_i)^2]$$

Now, summation over first N natural integer square is equal to an into n plus 1 into 2n plus 1 by 6 minus N into N plus 1 by 2 whole square multiplied by this quantity. So, let us take out N square upon 12 from this. Therefore, we are left with 2 into N plus 1 into 2 N plus 1 minus 3 into N plus 1 whole square.

(Refer Slide Time: 30:22)



$$\begin{aligned}
 &= \frac{mn}{N-1} * \frac{N+1}{12} [2 * (2N+1) - 3 * (N+1)] \\
 &= \frac{mn}{N-1} * \frac{N+1}{12} [4N + 2 - 3N - 3] \\
 &= \frac{mn}{N-1} * \frac{N+1}{12} (N-1) \\
 &= \frac{mn(N+1)}{12}
 \end{aligned}$$

Therefore $Var(W) = \frac{mn(N+1)}{12}$



36

Which is equal to

$$\frac{mn}{N-1} * \frac{N+1}{12} [4N + 2 - 3N - 3]$$

is equal to

$$\frac{mn}{N-1} * \frac{N+1}{12} [4N + 2 - 3N - 3]$$

This comes out to be N minus 1 which cancels with this. Therefore, we are left with mn into capital N plus 1 divided by 12. So, that is the variance of the Wilcoxon Rank Sum Test statistic W, we have used the result, but we proved it today.

(Refer Slide Time: 31:09)

Another important aspect of a Linear Rank Statistic is its

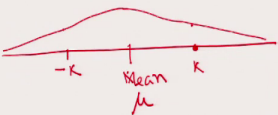
Symmetry

A statistic T_N is considered symmetric about its mean μ if for every $k \neq 0$

$$P[T_N(Z) - \mu = k] = P[T_N(Z) - \mu = -k]$$

i.e. $P[T_N(Z) = \mu + k] = P[T_N(Z) = \mu - k]$

$T_N(Z)$



Note that k need not be an integer

NPTEL

38

Now, another important aspect of linear rank statistic is symmetry that means, whether it is symmetric around this mean value. That means, a statistic T_N is considered symmetric about its mean μ . If for every k not equal to 0,

$$P[T_N(Z) - \mu = k] = P[T_N(Z) - \mu = -k]$$

Or in other words,

$$P[T_N(Z) = \mu + k] = P[T_N(Z) = \mu - k]$$

So, that is what we call it symmetric. For illustration consider $T_N Z$. So, suppose this is the distribution, this is the mean value of $T_N Z$ which we are calling μ .

It is considered to be symmetric. If the value at a distance k from μ is same as the value at a distance minus k . That is what we are saying here. Note that k need not be an integer. It can be any real value.

(Refer Slide Time: 32:50)

Another important aspect of a Linear Rank Statistic is its

Symmetry

A statistic T_N is considered symmetric about its mean μ if for every $k \neq 0$

$$P[T_N(Z) - \mu = k] = P[T_N(Z) - \mu = -k]$$

i.e. $P[T_N(Z) = \mu + k] = P[T_N(Z) = \mu - k]$



If the Null distribution of a Linear Ranks Statistic is symmetric
We need to generate only half of the probabilities.



39

If the null distribution of linear rank statistic is symmetric, the advantage is that we need to generate only half of the probabilities as you can easily understand if we know the probability density for this part, we know what is the density for the other half. So, that is why symmetric distribution is very useful many times.

(Refer Slide Time: 33:20)



The question is: How to check the above?

*we cannot check
for all possible k .*



40



Intuitive Idea

Suppose that corresponding to every vector Z of m 1's & n 0's
There exists another vector Z' also of m 1's & n 0's such that,
whenever ,

$T_N(Z) = \mu + k$ the value of $T_N(Z') = \mu - k$ ✓

Then the frequency distribution of $\mu + k$ is same as that of $\mu - k$, and
the distribution is symmetric.



41

The question is how to check the above? How do you know whether a distribution is symmetric? We cannot check for all possible k therefore; we need to find out device. So, the intuitive idea is following suppose that corresponding to every vectors Z of m 1's and n 0's, there exists another vectors Z prime also of m 1's and n 0's such that $T_N Z$ is equal to μ plus k .

Whenever that happens that for Z , we get the value of $T_N Z$ is equal to μ plus k , then for Z prime, the value of t and Z prime is going to be μ minus k that means, what for each case of μ plus k , there is a corresponding case of μ minus k . Therefore, the number of cases for this is

same as the number of cases for this and therefore, their probabilities are also going to be the same.

(Refer Slide Time: 34:45)

Intuitive Idea

Suppose that corresponding to every vector Z of m 1's & n 0's
There exists another vector Z' also of m 1's & n 0's such that,
whenever ,

$$T_N(Z) = \mu + k \quad \text{the value of} \quad T_N(Z') = \mu - k$$

Then the frequency distribution of $\mu + k$ is same as that of $\mu - k$, and
the distribution is symmetric.

Thus the condition for symmetry can be modified as :

$$\underline{T_N(Z) + T_N(Z')} = \underline{2\mu}$$

$Z \quad Z' \rightarrow T_N(Z) + T_N(Z') = 2\mu$

42

Intuitive Idea

Suppose that corresponding to every vector Z of m 1's & n 0's
There exists another vector Z' also of m 1's & n 0's such that,
whenever ,

$$T_N(Z) = \mu + k \quad \text{the value of} \quad T_N(Z') = \mu - k$$

Then the frequency distribution of $\mu + k$ is same as that of $\mu - k$, and
the distribution is symmetric.

Thus the condition for symmetry can be modified as :

$$T_N(Z) + T_N(Z') = 2\mu$$

The following result gives a simple relation to ensure symmetry,
Where Z and Z' are called conjugates

43

That is what we have written there. Then the frequency distribution of $\mu + k$ is same as that of $\mu - k$ and therefore, the distribution is going to be symmetric. Thus, condition for symmetric can be modified as

$$T_N(Z) + T_N(Z') = 2\mu$$

So, whenever we get, for each Z another Z prime, such that $T_n Z$ plus $T_n Z$ prime is equal to twice μ . And that happens for all Z and therefore, we can say that the distribution of T_n is symmetric. Now, there are several results available for that. So, we are stating some of the results, which are simple relations which ensure symmetry when Z and Z prime are called conjugates.

(Refer Slide Time: 35:50)

Result 3

The null distribution of $T_N(z)$ is symmetric about its mean μ , whenever the weights satisfy the relation :

$$\underline{a_i + a_{N-i+1} = c} \quad \forall \quad \underline{i = 1, 2, \dots, N}$$

where c is a constant, and

$$\mu = m \sum_{i=1}^N \frac{a_i}{N}$$

Handwritten notes on the slide:

- $a_2 + a_{N-2+1} = c$
- $i=1 : a_1 + a_N = c$
- $i=2 : a_2 + a_{N-1} = c$
- $i=3 : a_3 + a_{N-2} = c$
- each such pair will have the same summation

Diagram showing the sequence $a_1, a_2, a_3, \dots, a_{N-3}, a_{N-2}, a_{N-1}, a_N$ with arcs connecting a_i to a_{N-i+1} for $i=1, 2, 3$.

So, let us go to prove the result number 3, the null distribution of $T_N Z$ is symmetric about its mean μ whenever the weight vector a_i satisfies the following that

$$a_i + a_{N-i+1} = c \quad \forall \quad i = 1, 2, \dots, N$$

where c is a constant and μ is equal to as you already know is

$$m \sum_{i=1}^N \frac{a_i}{N}$$

So, let us look at it very carefully, it is saying a_i plus a_{N-i+1} is equal to c . So, put i is equal to 1. So, therefore, it is saying a_1 plus a_N is equal to c put i is equal to 2 it is saying a_2 plus a_{N-1} is equal to c , i is equal to 3 implies a_3 plus a_{N-2} is equal to c . Therefore, what we are saying is that if $a_1, a_2, a_3, a_4, a_{N-3}, a_{N-2}, a_{N-1}$ and a_N if this is the vector then this plus this, this plus this, this plus this, this plus this, each such pair will have the same summation.

(Refer Slide Time: 37:56)

For any vector $Z = (Z_1, Z_2, \dots, Z_N)$ of m 1's and n 0's, define the conjugate vector

$Z' = (Z'_1, Z'_2, \dots, Z'_N)$ of m 1's and n 0's, where $Z'_i = Z_{N-i+1}$

Then $T_N(Z) + T_N(Z')$

$$= \sum_{i=1}^N a_i Z_i + \sum_{i=1}^N a_i Z_{N-i+1}$$

Handwritten notes: Z' is the reverse of Z . Example: $Z = 11010$, $Z' = 01011$.

$$= \sum_{i=1}^N a_i Z_i + \sum_{j=1}^N a_{N-j+1} Z_j \quad (j = N-i+1)$$

Handwritten notes: $a_i = a_{N-j+1}$. Changing j to i .

$$= \sum_{i=1}^N (a_i + a_{N-i+1}) Z_i$$

$$= c \sum_{i=1}^N Z_i$$

Handwritten note: $\therefore T_N(Z) + T_N(Z')$ is const

Handwritten note: $= cm$

Let us first prove the result. So, corresponding to each Z of m 1's and n 0's we define the conjugate vectors Z_1 prime, Z_2 prime, Z_n prime also have m 1's and n 0's, where Z_i prime is equal to Z minus i plus 1. Therefore, if this is the Z vector, Z prime is going to be the reverse of that. That means we look at it from this side. Therefore, if Z is equal to 11010, Z prime is equal to 01011.

So, that is the basic property. In this case $T_N Z$ plus $T_N Z$ prime is equal to

$$\sum_{i=1}^N a_i Z_i + \sum_{i=1}^N a_i Z_{N-i+1}$$

because Z_i Prime is equal to Z N minus i plus 1 is equal to $\sum a_i Z_i$ plus. Now, we are making a change of variable we are calling N minus i plus 1 is equal to j . Therefore, Z N minus i plus 1 becomes Z_j and a_i is equal to a N minus j plus 1 because i is equal to Z N minus j plus 1 is equal to by change of variable j to i is equal to

$$\sum_{i=1}^N (a_i + a_{N-i+1}) Z_i$$

$$c \sum_{i=1}^N Z_i$$

is equal to c times m .

Therefore, $T_n Z$ plus $T_n Z$ prime is a constant.

(Refer Slide Time: 40:19)



So we are through if we can show $cm = 2\mu$

We have: $E(T_N(Z)) = m \sum_{i=1}^N \frac{a_i}{N} = \mu$

Similarly, $E(T_N(Z')) = m \sum_{i=1}^N \frac{a'_i}{N} = \mu$

Therefore together we get $cm = 2\mu$

Hence $T_N(Z) + T_N(Z') = 2\mu$

Therefore the condition for symmetry is met.

Thus T_N is symmetric. ✓ □



47

So, we are through if we can show that cm is equal to twice μ , because if you remember we said that distribution will be symmetric if $T_N Z$ plus $T_N Z$ prime is equal to twice μ . Now, the expected value of $T_N Z$ is equal to m times σa_i upon N is equal to μ . Similarly, expected value of $T_N Z$ prime is equal to m times σa_i prime upon N is equal to μ .

Where a_i prime is the constant we defined for Z prime. Therefore, together we get cm is equal to twice μ . Therefore,

$$T_N(Z) + T_N(Z') = 2\mu$$

Therefore, the condition for symmetry is met with and therefore T_N is symmetric, proved.

(Refer Slide Time: 41:21)

Illustration

Consider $m = 3$ and $n = 2$.
Therefore 10 possible arrangements of three 1's and two 0's ✓ ✓

$$\frac{5!}{3!2!} = \frac{4 \cdot 5}{2} = 10$$

48

Let me now illustrate this with a simple example m is equal to 3 and n is equal to 2. That means you are looking at Z vectors of 3 1's and 2 0's. And the number of possible ways of doing it is 10. Because this is factorial 5, upon factorial 3, factorial 2 is equal to 4 into 5 upon 2 is equal to 10.

(Refer Slide Time: 41:50)



	$a:$ 1 2 3 2 1	$c=6$
11100	6	✓10
11010	5	8✓
11001	4	✓11
10101	5	✓9
10110	6	6✓
10011	4	7✓
01110	7	✓9
01101	6	✓12
01011	5	✓10
00111	6	8✓

$T_n(Z)$

Case 1:

4: 2 ✓
5: 3 ✓
6: 4 ✓
7: 1 ✓

Mean = 5.4

Case 2:

6: 1 ✓
7: 1 ✓
8: 2 ✓
9: 2 ✓
10: 2 ✓
11: 1 ✓
12: 1 ✓

Mean = $\frac{\mu = 9}{c = 6} = \frac{9}{6} = 1.5$

$\therefore 2\mu = 18 = cm$

$\begin{matrix} 6 & 3 \end{matrix}$



49

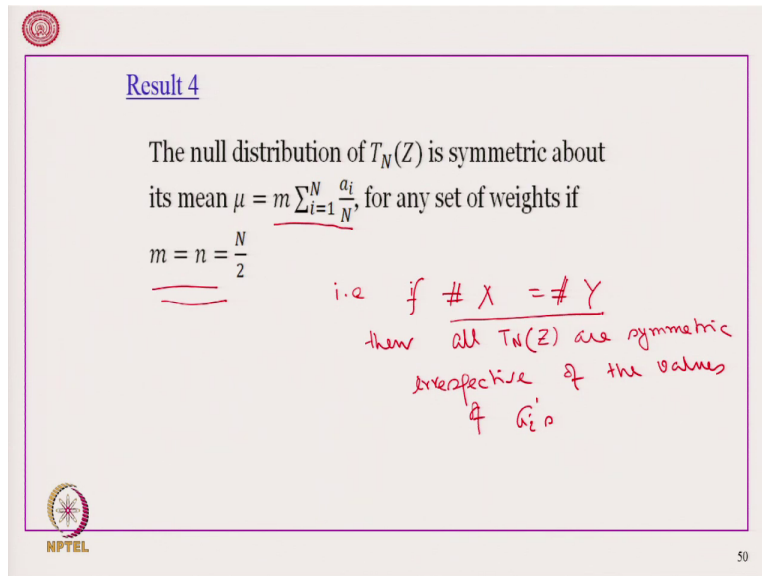
So, therefore, we need to consider 10 permutations, that is what we have done 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Let us consider the a vector to be this and a vector like this. Note that here, the condition is not maintained that a_i plus a N minus i plus 1 is constant, but in this case we can see that this is 2 plus 4, 5 plus 1 and 3 plus 3, all of them are going to be 6.

So, in this case c is equal to 6. Now, if we calculate the value of $T_n Z$; this is coming out to be, in this case 6, 5, 4, 5, 6, 4, 7, 6, 5, 6 this you can easily calculate for illustration, when it is 11100 its dot product with 1, 2, 3, 2, 1 is going to be 1 plus 2 plus 3 is equal to 6 on the other hands. If I consider 0110, then its dot product is going to be 2 plus 3 plus 2, therefore this is going to be 7. Therefore, what is going to be the distribution of $T_n Z$, we can see that it takes the value 4, two times 5, three times 6, four times and 7, one time.

And the mean is coming out to be 5.4 because 6 plus 5, 11, plus 4, 15 plus 5, 20 plus 6, 26, 30, 37, 43, 48 plus 6 is equal to 54. Therefore, mean is equal to 5.4. In a similar way you calculate for this a vector, we will find that our distribution is coming out to be 6 once, 7 once 8 twice, 10 twice, 9 twice, 11 once and 12 once. Therefore, we can see that the mean of the distribution is coming out to be 9, c is equal to 6 we have already calculated, therefore, twice μ is equal to 18 is equal to cm , where c is equal to 6, a m is equal to number of x is equal to 3. And we can see

that this is symmetric around the mean. It is not approved because proof we have already given this is only to visualize the result.

(Refer Slide Time: 44:41)



Result 4

The null distribution of $T_N(Z)$ is symmetric about its mean $\mu = m \sum_{i=1}^N \frac{a_i}{N}$, for any set of weights if $m = n = \frac{N}{2}$

i.e. if $\#X = \#Y$ then all $T_N(Z)$ are symmetric irrespective of the values of a_i 's

NPTEL 50

Now, let me give you another result, the null distribution of $T_N(Z)$ is symmetric about its mean μ Which is, is equal to $m \sum_{i=1}^N a_i / N$ for any set of weights, if m is equal to capital N by 2, that is, if number of X is equal to number of Y , then all $T_N(Z)$'s are symmetric; irrespective of the value of a_i 's. You consider any constant a_i 's, the distribution will remain symmetric, provided number of X is equal to number of Y .

(Refer Slide Time: 45:44)

Proof

Note that here $m = n$.

Let us define for any vector $Z = (Z_1, Z_2, \dots, Z_N)$ of m 1's and n 0's, a conjugate vector also of m 1's and n 0's as $Z' = (Z'_1, Z'_2, \dots, Z'_N)$ where $Z'_i = 1 - Z_i$.

Note that this works only because $m = n$

Then

$$T_N(Z) + T_N(Z') = \sum_{i=1}^N a_i Z_i + \sum_{i=1}^N a_i (1 - Z_i)$$

i.e. if in Z the i th position is 1 then in Z' the i th position is 0 & vice versa

$$= \sum_{i=1}^N a_i Z_i + \sum_{j=1}^N a_j - \sum_{i=1}^N a_i Z_i = \sum_{j=1}^N a_j$$

NPTEL

So, proof note that a m is equal to N , therefore, let us define for on any vector Z is equal to Z_1, Z_2, Z_n of m 1's and n 0's, a conjugate vector also of m 1's and n 0's as Z prime is equal to Z_1 prime, Z_2 prime, Z_n prime where Z_i prime is equal to $1 - Z_i$. That means, if in Z the i th position is 1, then in Z prime, the i th position is 0 and vice versa.

Because number of X is equal to number of Y that is a m is equal n . Therefore, by doing this transformation, we still maintain the number of 1's and 0's. Therefore,

$$\sum_{i=1}^N a_i Z_i + \sum_{i=1}^N a_i (1 - Z_i)$$

because this is coming from Z_i prime is equal to $\sigma a_i Z_i$ plus σa_i minus $\sigma a_i Z_i$ is equal to σa_i , because, these 2 cancels.

(Refer Slide Time: 47:32)



Now,

$$\begin{aligned}\mu &= m \sum_{i=1}^N \frac{a_i}{N} \checkmark \\ &= \frac{N}{2} * \sum_{i=1}^N \frac{a_i}{N} \checkmark \\ &= \frac{1}{2} * \sum_{i=1}^N a_i \\ &\Rightarrow \sum_{i=1}^N a_i = 2\mu \quad \Rightarrow \sum a_i = 2\mu\end{aligned}$$

Hence, we have $\underline{T_N(Z) + T_N(Z')} = \underline{2\mu}$.

Thus, the null distribution of $T_N(Z)$ is symmetric.



52

Now, μ is equal to m by $N \sum a_i$. Now, small m is equal to N by $2 \sum a_i$ by N . N cancels therefore, we have half times $\sum a_i$ implies that $\sum a_i$ is equal to twice μ . Therefore, $T_N(Z) + T_N(Z')$ is equal to twice μ . Therefore, by our condition of symmetry, the null distribution of $T_N(Z)$ is symmetric.

(Refer Slide Time: 48:10)

Illustration

Consider $m = n = 2$ Therefore $N = 4$

$a: 1 \ 3 \ 5 \ 4 \quad 3 \ 3 \ 7 \ 5$ Case 2:

1100	4 ✓	6 ✓
1010	6 ✓	10 ✓
1001	5 ✓	8 ✓
0110	8 ✓	10 ✓
0101	7 ✓	8 ✓
0011	9 ✓	12 ✓

Case 1:

$4:1 \ 5:1 \ 6:1 \ 7:1 \ 8:1 \ 9:1 \quad \mu = 39/6 = 6.5$

$T_N(Z) + T_N(Z') = 10 + 3 = 13$

$T_N(Z) + T_N(Z') = 12 + 6 = 18$

$\mu = 54/6 = 9$

$T_N(Z) + T_N(Z') = 18 = 2 * 9$

NPTEL

53

So, illustration, we have taken simple case, 2 m's and 2 n's, therefore, capital N is equal to 4. We have chosen 2 arbitrary weight vectors 1 3 5 4 and 3 3 7 5. In this case, this is the distribution 1 4, 1 5, 1 6, 1 7, 1 8 and 1 9; μ is equal to 6.5 and therefore, we can see that $T_N Z$ plus $T_N Z$ prime is equal to 2 into 6.5 is equal to 13. And as we can see here, that 9 plus 4 is equal to 13, 7 plus 6 is equal to 13. And 8 plus 5 is equal to 13. Similarly, for case 2, the distribution will come out to be 6 once, 8 twice, 10 twice, and 12 once.

The μ is coming out to be 9, and we can see that T_N plus $T_N Z$ prime is equal to 12 plus 6 is equal to 18, 10 plus 8 is equal to 18 and 8 plus 10 is equal to 18. Therefore, we get that is equal to 2 times 9.

(Refer Slide Time: 49:30)

Result 5:

The null distribution of $T_N(Z)$ is symmetric about its mean $\mu = m \sum_{i=1}^N \frac{a_i}{N}$ if N is even and the weights are:

$$a_i = i \text{ for } i \leq \frac{N}{2} \text{ and } a_i = N - i + 1 \text{ for } i > \frac{N}{2}$$


Result 5:

The null distribution of $T_N(Z)$ is symmetric about its mean $\mu = m \sum_{i=1}^N \frac{a_i}{N}$ if N is even and the weights are:

$$a_i = i \text{ for } i \leq \frac{N}{2} \text{ and } a_i = N - i + 1 \text{ for } i > \frac{N}{2}$$

For a Z satisfying above we define the conjugate Z' as follows:



Now, let me give you another result, which is the result number 5. This is the last result that we are going to do in this class. The null distribution of $T_N(Z)$ is symmetric about its mean μ , which we know is $m \sum_{i=1}^N \frac{a_i}{N}$, if N is even, and weights are a_i is equal to i for i less than equal to $N/2$, and a_i is equal to $N - i + 1$ for i greater than $N/2$.

That is, if the total number of observation is even, and the weights given our a_i is equal to i for the first half of the values, and it is $N - i + 1$ for the other half of the values, then the

distribution of T_n is going to be symmetric about its mean. For Z satisfying the above, we define a conjugate Z prime as follows.

(Refer Slide Time: 50:36)

$$Z'_i = Z_{i + N/2} \text{ for } i \leq \frac{N}{2}$$

$$Z'_i = Z_{i - N/2} \text{ for } i > \frac{N}{2}$$

For illustration: Suppose $N = 8 \Rightarrow$ there are 4 1's and 4 0's

Suppose $Z = \underline{1 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0}$

Therefore $Z' = \underline{0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1}$

Now $a_i = \underline{1 \ 2 \ 3 \ 4 \ 4 \ 3 \ 2 \ 1}$

Since $a_i = i \text{ for } i \leq \frac{N}{2} \text{ and } a_i = N - i + 1 \text{ for } i > \frac{N}{2}$

We need to prove that T_N is symmetric.

Z prime i is equal to Z_i plus N by 2 for i less than equal to N by 2 and Z_i prime is equal to Z_i minus N by 2 for i is greater than N by 2. So, let me illustrate with an example. Suppose N is equal to 8 and there are 4 1's and 4 0's. So, consider a particular Z which is 1 0 1 1 0 1 0 0. Therefore, how do you construct a conjugate Z prime, for i is equal to 1 to N by 2, Z_i prime is Z_i plus N by 2. Therefore, Z_1 prime is equal to 0, coming from here. Z_2 prime is equal to 1 coming from here. Similarly, this is 3 from here, and this is 4 from here.

For i greater than N by 2 that means for 5, 6, 7, 8. The relationship is Z_i prime is getting the value of Z_i minus N by 2. Therefore, Z_5 prime is getting the value of Z_1 , Z_6 prime is getting the value of Z_2 , Z_7 prime is getting the value of Z_3 and Z_8 prime is getting the value of Z_4 . So, like that, corresponding to each Z , we can construct an appropriate Z prime. And notice that this is also belonging to the same vector with 4 1's and 4 0's.

Now, therefore, the weights are going to be it is a_i is equal to i for i less than equal to N by 2. Therefore, for the first 4 the weights are 1, 2, 3, 4 and a_i is equal to N minus i plus 1 for i greater than N by 2. Therefore, when i is equal to 5, it is 8 minus 5 plus 1 therefore 4 when i is equal to 6, it is 8 minus 6 plus 1 that is 3. Similarly, 2 and 1. Therefore, we can see that the weights are

decreasing in a symmetric manner as well. We need to prove that T_n is symmetric around this mean.

(Refer Slide Time: 53:06)

$$\begin{aligned}
 & T_N(Z) + T_N(Z') \\
 &= \sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{i=\frac{N}{2}+1}^N (N-i+1) Z_i + \sum_{i=1}^{\frac{N}{2}} i Z_{N+1-i} + \sum_{i=\frac{N}{2}+1}^N (N-i+1) Z_{i-\frac{N}{2}} \\
 &= \sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{i=\frac{N}{2}+1}^N (N-i+1) Z_i + \sum_{j=\frac{N}{2}+1}^N \left(j - \frac{N}{2}\right) Z_j + \sum_{j=1}^{\frac{N}{2}} \left(\frac{N}{2} - j + 1\right) Z_j
 \end{aligned}$$

So, let us prove it mathematically, $T_n Z$ plus $T_n Z$ prime is equal to $\sum_{i=1}^{N/2} i Z_i$ and $\sum_{i=N/2+1}^N (N-i+1) Z_i$. So, that is coming from $T_n Z$. In a similar way $T_n Z$ prime is equal to $\sum_{i=1}^{N/2} i Z_{N/2+1-i}$, because that is how the Z prime has been allocated values and from $\sum_{i=N/2+1}^N (N-i+1) Z_{i-N/2}$, is equal to for $\sum_{i=1}^{N/2} i Z_i$ plus $\sum_{i=N/2+1}^N (N-i+1) Z_i$.

Now, here what we are doing, we are making a change of variable. So, on this side we make a change of variable $N/2 + i = j$. Therefore, $i = j - N/2$. Therefore, what we are doing? We are putting instead of i , $j - N/2$ and in place of $N/2 + i$ we are putting j and since i is going from 1 to $N/2$, j is going from $N/2 + 1$ to N .

In a similar way, we are again making a change of variable and what we are getting is that j is equal to 1 to $N/2$, it is $N/2 - j + 1$ multiplied by Z_j ,

(Refer Slide Time: 55:22)



$$\begin{aligned}
 &= \left(\sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{j=1}^{\frac{N}{2}} \left(\frac{N}{2} - j + 1 \right) Z_j \right) \\
 &\quad + \left(\sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{j=\frac{N}{2}+1}^N \left(j - \frac{N}{2} \right) Z_j \right) \\
 &= \sum_{i=1}^{\frac{N}{2}} \left(\frac{N}{2} + 1 \right) Z_i + \sum_{i=\frac{N}{2}+1}^N \left(\frac{N}{2} + 1 \right) Z_i \quad \sum_{i=1}^N \left(\frac{N}{2} + 1 \right) Z_i \\
 &= m \left(\frac{N}{2} + 1 \right) = 2\mu \quad \text{Hence Symmetric}
 \end{aligned}$$




$$\begin{aligned}
 &T_N(Z) + T_N(Z') \\
 &= \sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{i=1}^{\frac{N}{2}} i Z_{N-\frac{N}{2}+i} + \sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_{i - \frac{N}{2}} \\
 &= \sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{j=\frac{N}{2}+1}^N \left(j - \frac{N}{2} \right) Z_j + \sum_{j=1}^{\frac{N}{2}} \left(\frac{N}{2} - j + 1 \right) Z_j \\
 &\quad \therefore i = j - \frac{N}{2} \quad \therefore j = i + \frac{N}{2}
 \end{aligned}$$


Which is equal to $\sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{j=1}^{\frac{N}{2}} (N - j + 1) Z_j$ that is, we are combining this and this. In a similar way, we combine these 2 and therefore, we get this following $\sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{j=\frac{N}{2}+1}^N (j - \frac{N}{2}) Z_j$ is equal to $\sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{j=1}^{\frac{N}{2}} (\frac{N}{2} - j + 1) Z_j$.

Here we can make a change of variable instead of j if I call it i , then what we are getting is $\sum_{i=1}^{\frac{N}{2}} i Z_i + \sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i$ and when we combine then, i and $-i$ cancels therefore, we get $\sum_{i=1}^{\frac{N}{2}} (N - i + 1) Z_i + \sum_{i=\frac{N}{2}+1}^N i Z_i$ is equal to $\sum_{i=1}^{\frac{N}{2}} (N - i + 1) Z_i + \sum_{i=\frac{N}{2}+1}^N i Z_i$. In a similar way, when I am calling it instead of j equal to, we call it i , then we will get that it is equal to $\sum_{i=\frac{N}{2}+1}^N (N - i + 1) Z_i + \sum_{i=1}^{\frac{N}{2}} (\frac{N}{2} - i + 1) Z_i$. Now, this cancels with this.

Therefore, we get $N - \frac{N}{2}$ is equal to $\frac{N}{2} + 1$ times Z_i . So, together what we are getting? Together we are getting it is that, $\sum_{i=1}^N \frac{N}{2} + 1$ times Z_i . Now, we can take it out of the summation and $\sum_{i=1}^N Z_i$ is equal to 1 to N is equal to m . Therefore, what we are getting is twice μ . Therefore, it is symmetric.

(Refer Slide Time: 57:13)



With this background we look at:

Mood Test Statistic



NPTEL



MOOD Statistic

$$M_N = \sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 Z_i$$

where Z_i is the indicator variable such that $Z_i = 1$ if i^{th} rank corresponds to X sample and $Z_i = 0$ otherwise

Under the null hypothesis,

(1) $E(M_N) = \frac{m(N^2-1)}{12}$ &

(2) $Var(M_N) = \frac{mn(N+1)(N^2-4)}{180}$

When $m = n \left(= \frac{N}{2} \right)$,

(3) the null distribution of M_N is symmetric about the mean,

But the symmetry property does not hold for unequal sample sizes.



NPTEL

With this background, we look at the Mood Test. In the Mood test, the statistic is a M_N is equal to $\sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 Z_i$. Where Z_i is equal to 1 if i is rank corresponding to an X sample, that means, we are looking at the ranks of X values. We want to show that expected value of a M_N is equal to $\frac{m(N^2-1)}{12}$. Variance of M_N is equal to $\frac{mn(N+1)(N^2-4)}{180}$. And if m is equal to n and both are equal to $N/2$, then the null distribution of M_N is symmetric about the mean.

(Refer Slide Time: 58:06)

Substituting this in the formula for the mean and variance of T_N are
As follows:

$$\begin{aligned}
 E(M_N) &= m \sum_{i=1}^N \frac{a_i}{N} = \frac{m}{N} \sum_{i=1}^N a_i = \frac{m}{N} \sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 \\
 &= \frac{m}{N} \sum_{i=1}^N \left(i^2 - 2 * \frac{N+1}{2} * i + \left(\frac{N+1}{2} \right)^2 \right) \\
 &= \frac{m}{N} * \left[\sum_{i=1}^N i^2 - (N+1) * \sum_{i=1}^N i + \left(\frac{N+1}{2} \right)^2 \sum_{i=1}^N 1 \right] \\
 &= \frac{m}{N} * \left[\frac{N(N+1)(2N+1)}{6} - (N+1) * \frac{N(N+1)}{2} + \left(\frac{N+1}{2} \right)^2 * N \right]
 \end{aligned}$$

NPTEL

So, let us put the first result, expected value of a M_N is equal to m summation a_i over N , this we all know is equal to this m into $\sum a_i$ is equal to m by N $\sum$ is equal to 1 to N , i minus N plus 1 by 2 whole square, is equal to m by N multiplied by summation of i square minus 2 into N plus 1 by 2 times i plus N plus 1 by 2 whole square by opening up this square.

Is equal to m by N times $\sum i$ square minus n plus 1 times $\sum i$ plus n plus 1 by 2 whole square times $\sum 1$, because there is no variable involved is equal to N into N plus 1 into $2N$ plus 1 by 6 minus N plus 1 times N into N plus 1 by 2 summation over i plus N plus 1 by 2 whole square times N ; this is straightforward.

(Refer Slide Time: 59:22)



$$\begin{aligned}
 &= \frac{m}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4} \right] \\
 &= \frac{m}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} \right] \\
 &= \frac{m}{N} \cdot \frac{N(N+1)}{12} [2 \cdot (2N+1) - 3 \cdot (N+1)] \\
 &= \frac{m(N+1)}{12} [4N+2-3N-3] \\
 &= \frac{m(N+1)(N-1)}{12} = \frac{m(N^2-1)}{12} \\
 \therefore E(M_N) &= \frac{m(N^2-1)}{12}
 \end{aligned}$$


Now, let us simplify it is equal to N into N plus 1 into $2N$ plus 1 by 6 minus a N into N plus 1 whole square by 2 plus N into N plus 1 whose square by 4 is equal to the first term comes as it is these 2 gives us minus N into N plus 1 whose square by 4 . Therefore, as you simply it, we take out N into N plus 1 by 2 as common, what we get is 2 into $2N$ plus 1 minus 3 into N plus 1 . Which is further simplified as $4N$ plus 2 minus 3 , which is equal to N minus 1 . Therefore, the expected value comes out to be m into N square minus 1 by 12 .

(Refer Slide Time: 60:13)



In a similar way one can compute the variance of $M_N =$

$$Var(M_N) = \frac{mn(N+1)(N^2-4)}{180} \checkmark$$

$$Var(M_N) = \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N a_i^2 - \left(\sum_{i=1}^N a_i \right)^2 \right]$$

$$= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 - \left(\sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2 \right) \right] \checkmark$$


In a similar way I can compute the variance of M_n 's. M_n into N plus 1 into N square minus 4 by 180. I have to go and expand it in this following way by putting the value of a_i but, as you can understand that it is very laborious arithmetic, I am leaving it as an exercise for you. You please try and check that actually you could arrive at this value for the variance.

(Refer Slide Time: 60:43)



(3)
Using the result on symmetry of null distribution of
linear rank statistic,
When $m = n \left(= \frac{N}{2} \right)$
the null distribution of M_N is symmetric about the mean



NPTEL



The third 1 was that if m is equal to n then the null distribution of M_N is symmetric about the mean this already, we have proved that for equality of M_n , all the linear Rank statistics are going to be symmetric. Therefore, it comes naturally. Ok friends. I stop here today, in the next class. I shall describe a few tests on goodness of fit. Thank you very much.