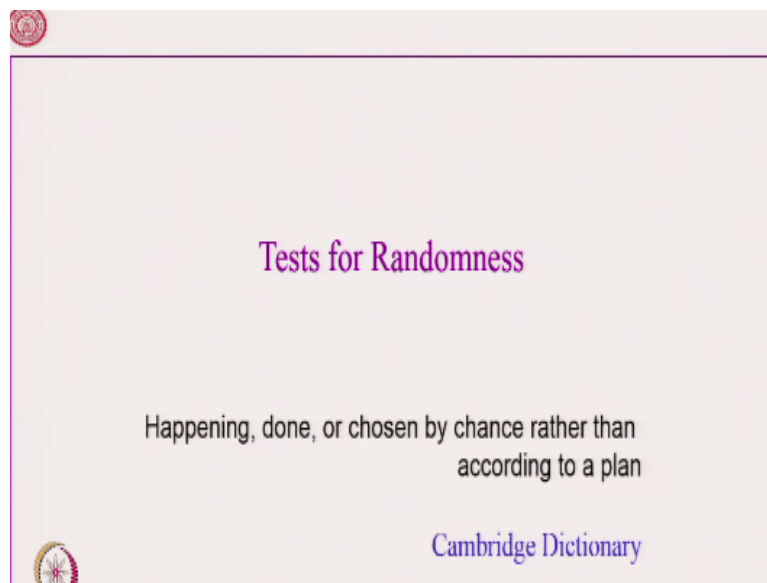


Nonparametric Statistical Inference
Professor. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture No. 07

Welcome students to MOOCs series of lectures on Nonparametric Statistical Inference, this is lecture number 7. As I said in the last class that in this lecture we shall study tests for randomness. Now random process or random number or random graph, random indexing are certain terms which are nowadays very important because of machine learning.

(Refer Slide Time: 00:56)



But from a statistician's perspective most important thing is random sample. Question is what is random? Cambridge dictionary gives this is the definition of randomness that is happening, done or chosen by chance rather than according to a plan.

(Refer Slide Time: 01:12)



In order to arrive at a conclusion about a distribution on the basis of a sample we often need to have a *random sample*, i.e.

sample is taken without any bias or pre-designed rule of selection, or to support some purpose etc.

Hence often we test the randomness of a sample.

Note:

- * It is important to maintain the order in which the samples are chosen.



NPTEL

So, when we talk about drawing inference about a population or about a distribution we often take a random sample. What does it mean? It means that sample is taken without any bias or predesigned rule of selection or to support some purpose etcetera. Therefore, when we have a data we need to test the randomness of the data. How to do that? In order to check whether something is random it is very important to maintain the order in which the samples are chosen.

If we consider the data points only as a set then we lose the flavor of randomness because the order in which the data points have been chosen that is very important as I am going to illustrate now.

(Refer Slide Time: 02:12)

 Illustration

Suppose you are picking 12 balls from a basket with equal & large number of red and blue balls.

Consider the following selection orders:

- BBBBBBRRRRRR
- BRBRBRBRBRBR
- RRRBBBBBBRRR

But consider

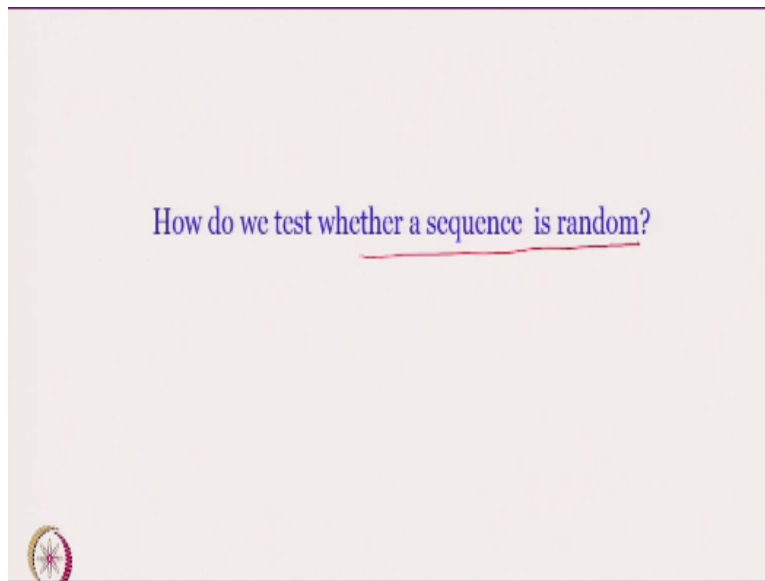
- RBBRRBRBBBR



Suppose, you are taking 12 balls from a basket with a large number of red and blue balls. Now we have taken 6 red balls and 6 blue balls. Suppose, the order of selection is this first 6 are blue balls and the second set of 6 are red balls. Will you consider it to be a random selection? No. What about this if we choose it alternatively blue, red, blue, red like that then do you consider this to be a random selection?

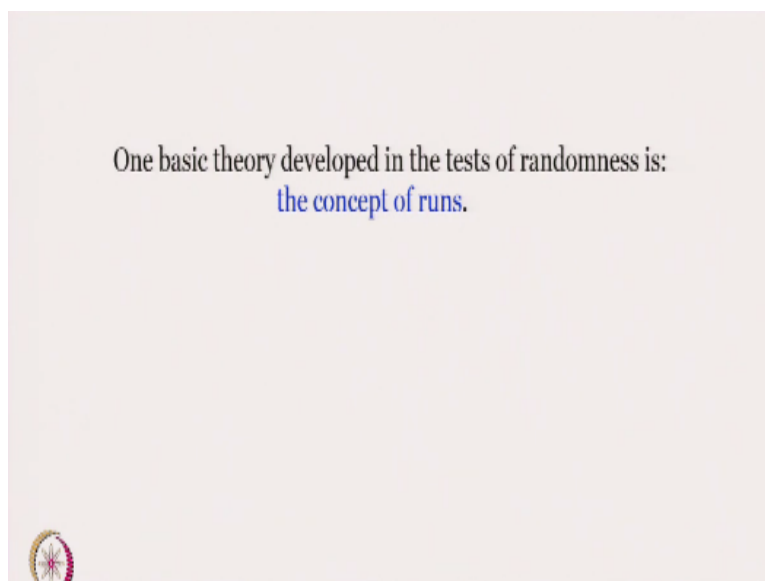
Again we understand that it is not random. Let us consider a third example. If the pattern of selection is like this: first 3 Rs then 6 Bs then again 3 Rs we understand that there is a specific way of selecting the balls from the basket. Therefore, this also is not random. However, if we consider this selection red, blue, blue, red, red then blue then again red, red then blue, blue, blue and red, then it appears that the selection of red and blue ball from the basket is rather random. Question is how do we test because all the four selection orders are having 6 blue and 6 red balls.

(Refer Slide Time: 03:55)



So, that is the basic question how to test whether a particular sequence is random or not.

(Refer Slide Time: 04:04)



The basic theory developed in test of randomness is based on the concept of runs.

(Refer Slide Time: 04:15)

Given an ordered sequence of two or more types of symbols, a **run** is defined to be succession of one type of symbol that is succeeded and preceded by another type of symbol or no symbol.

For example,

- BBBBBBRRRRRR - 2 runs ✓
- BRBRBRBRBRBR - 12 runs ✓
- RRRBBBBBRRR - 3 runs ✓
- RBBRRBRBBBR - 7 runs ✓



So, what is a run? Given an ordered sequence of two or more types of symbols, a run is defined to be succession of one type of symbol that is succeeded and preceded by another type of symbol or no symbol at all. So, let us consider the 4 selection orders that you have just seen. With respect to the first one, this is one run of blue ball because there is no red ball in between. It is a continuous sequence of observations of 6 blue balls similarly a continuous sequence of observation of 6 red balls.

Therefore, how many runs are there? There are 2 runs. This is succeeded by this run and this sequence of R balls is preceded by the sequence of blue balls. Consider the second one how many runs are there? There are 12 runs as I have explained here this is 1 run blue it is intercepted by a red again that is intercepted by a blue then red so like that there are 12 number of runs in this sequence.

With respect to the third one we can say that this is one run of R, this is one run of blue and this is one run of R again that gives us 3 runs and this one is very similarly first run, second run, third, fourth, fifth, sixth and seventh there are seven different runs in this sequence. What does it tell us?

(Refer Slide Time: 06:08)

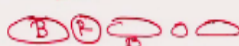
 Test based on Total Number of Runs

It is clear from the above example that if the number of runs is too large or too small then it is unlikely that the points are chosen randomly.

Of course the number of runs depends upon the composition of the data .

For illustration:

a) With 12 balls, 6 B's and 6 R's, maximum number of runs can be 12. *✓ of course the minimum will be 2*

b) However, if there are 10 B's and 2 R's the maximum number of runs is 5. *✓* 

So, it is clear from the above example that if the number of runs is too large or too small, then it is unlikely that the points are chosen randomly. However, what is too large and what is too small that depends upon the composition of the data. For example, we have in the previous illustrations 12 balls of which 6 B's and 6 R's that means 6 are blues and 6 are reds. Therefore maximum possible run can be 12 of course, the minimum will be 2, one sequence of red and one sequence of blue.

However, suppose there are 10 B's and 2 reds. then the maximum number of runs can be 5. One sequence of blue, one sequence of red may be one red, one sequence of blue, one red and another sequence of blue. This is the maximum number of runs that can be achieved from 10 blues and 2 Rs. Thus, the maximum number of runs is 5 although the minimum will still remain 2. Therefore, when we are talking about if the number of runs is too large we have to be very careful about the composition of the data as well.

(Refer Slide Time: 07:52)

Preliminaries

- * Runs test are applicable to either quantitative or qualitative data.
- * In case of numeric data the dichotomy is usually effected by comparing each number with a focal point, e.g. a median or mean.

Illustration: Picking 12 real numbers between 50 and 100.

Sample 1. 95.6 88.3 75.8 90.5 79.2 84.7 ✓
69.8 60.6 65.2 55.0 52.5 59.2
Median = 72.8

Sample 2. 59.2 88.3 84.7 55.0 95.6 52.5 ✓
90.5 65.2 69.8 79.2 75.4 60.6 ✓

Now the tests based on runs are applicable to either quantitative or qualitative data. It is not necessary that the data has to be numeric. It can be symbolic as we have already seen sequence of R's and B's in our previous illustrations. However, in the case of numeric data typically the dichotomy is effected by comparing each number with a focal point which can be median or mean. What do we mean by that?

Dichotomy means that when there is a sequence of numbers how do we split it into two different sets like R and B. In sign test we have seen plus and minus so how do we do that? So, that is done by comparing with median or mean. For example, suppose we have taken 12 real numbers between 50 and 100. So this is one sequence 95.6, 88.3, 75.8 like that 59.2. Another sequence is 59.2, 88.3 and up to 60.6, which one of them is random and why?

To test that what we do we first compute the mean or median. In this case median is coming out to be 72.8 because this is sort of the average of these two then what we do as we did in the case of sign test we consider how many of these observations are bigger than the median and how many of them are smaller than the median. The same we do with respect to this sequence of observations as well and we find the following.

(Refer Slide Time: 09:50)

By converting the data like what we did for Sign Test, we get the Following:

Sample1: +++++- ----- # of Runs = 2

Sample2: -++-+- -++- # of Runs = 9



With respect to the first one we get that there are 6 plus followed by 6 minus. Therefore, number of runs is 2. However, with respect to the second sample we get minus, plus, plus another minus plus then 1 minus plus minus, minus, plus, plus minus that means the 6 plus and 6 minuses are very nicely intermingled among themselves and we get that the number of runs is equal to 9. Now if I ask you which one of them is random we can easily see that this is definitely random and this is definitely not.

So, while working on numeric data this is one we resolve the dichotomy and thereby we can solve the randomness problem and we can check if the data is actually random or not in the following way. However, if we have a numeric data there is another important test which is called runs up and down test which we will do later, but before that let us just study the common what is called run test.

(Refer Slide Time: 11:18)



So, the run test are based on the total number of runs in the sequence.

(Refer Slide Time: 11:23)

A slide with a light pink background and a purple border on the left. It contains three bullet points with handwritten red notes and checkmarks. The first bullet point states: '* Assume an ordered sequence of N elements of two types, A & B with n_1 of them being A and n_2 are B , where $n_1 + n_2 = N$.' Handwritten notes include $n_1 = 6$, $n_2 = 6$, and $N = 12$. The second bullet point states: '* In order to derive a test of randomness we need to obtain the distribution of $R = r_1 + r_2$, where r_1 is the number of runs of A and r_2 is the number of runs of B .' The third bullet point states: '* Total number of distinct arrangements of N elements having n_1 elements of type A and n_2 elements of type B , is i.e. $\frac{(n_1 + n_2)!}{n_1! n_2!}$ or equivalently $\binom{n_1 + n_2}{n_1}$.' Handwritten notes include $\frac{(n_1 + n_2)!}{n_1! n_2!}$ and $\binom{n_1 + n_2}{n_1}$. A small decorative wheel icon is in the bottom left corner.

Assume an ordered sequence of N elements of two types A and B with n_1 of them being A and n_2 of them are B such that $n_1 + n_2 = N$ it is the total number of data points. With respect to the previous example n_1 was 6, n_2 was 6 and N was 12. In order to derive a test of randomness we need to obtain the distribution of R what is R , R is the total number of runs which is $r_1 + r_2$ where r_1 is the number of runs of A and r_2 is the number of runs of B .

Now what is the total number of possible arrangements of N elements when there are n_1 of them are of type A and n_2 of them are type B . The number of possible arrangements is

$n_1 + n_2 C_{n_1}$ or equivalently $n_1 + n_2 C_{n_2}$. Why? Some people often asked because arrangement is a permutation problem, but this is a combination. So, how do we interpret a permutation problem using a combination?

The answer is simple because n_1 plus n_2 many elements can be arranged among themselves in n_1 plus n_2 factorial. However, n_1 of them are same and n_2 of them are same. Therefore, we need to divide it by n_1 factorial n_2 factorial and this is nothing, but n_1 plus n_2 c n_1 which we have written here. Ok

(Refer Slide Time: 13:25)

* The number of distinct ways of distributing n like objects into r distinguishable cells with no cell empty is $\binom{n-1}{r-1}$ $n \geq r$

The diagram illustrates the distribution of 7 balls into 3 boxes. The top part shows a partition (2, 2, 3) with 7 balls and 2 dividers. The bottom part shows a partition (1, 3, 3) with 7 balls and 2 dividers.

Now the number of distinct ways of distributing n like objects into r distinguishable cells with no cell empty is $n-1 C_{r-1}$ when $n \geq r$. This is a very simple combinatorics problem. So, suppose for example I have 7 balls and I want to put them in 3 boxes such that no box is empty. In how many ways one can do that? So, let us first understand that in order to put them in 3 boxes we need to draw some partitions there.

So, let us consider I draw these two partition therefore in the first box there are 2 balls, in the second box there are 2 balls, in the third box there are 3 balls. So, this we can write it as say 2, 2, 3. Another example suppose I have again 7 balls and I choose the following lines then what do I get? We get 1, 3, 3.

Therefore, each selection of two lines out of the 6 empty places between the 7 balls 1, 2, 3, 4, 5, 6, the number of ways of choosing 2 out of 6 that is going to give me 3 partitions each one

is non empty. In general, therefore when there are N objects to be put in R boxes when n is greater than equal to r number of ways of doing it is ${}^{n-1}C_{r-1}$.

(Refer Slide Time: 15:35)

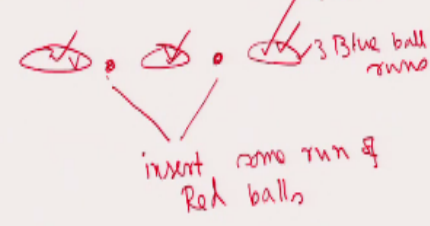
* Thus, the total number of distinct arrangements starting with runs of type A is $\frac{(n_1-1)}{(r_1-1)} \frac{(n_2-1)}{(r_2-1)}$

$n_1 \rightarrow r_1$ runs $\frac{(n_1-1)}{(r_1-1)}$

$n_2 \rightarrow r_2$ runs $\frac{(n_2-1)}{(r_2-1)}$

3 Blue ball runs

insert some run of Red balls

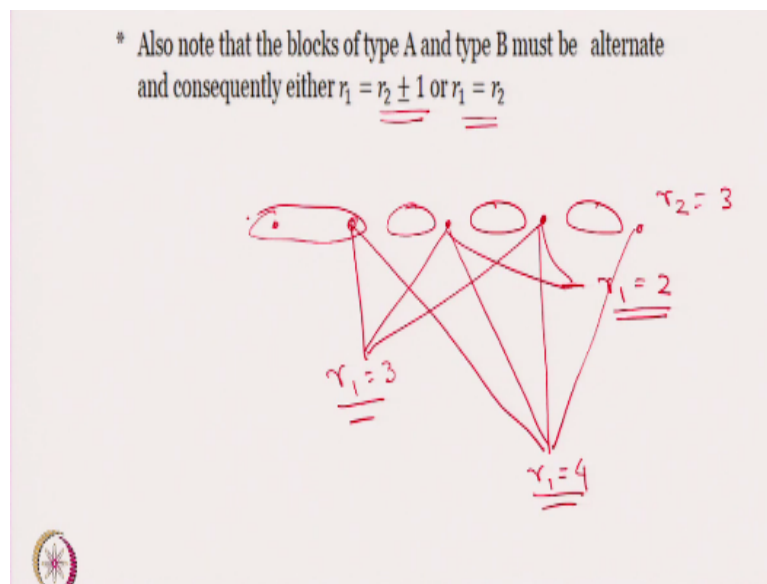


Now we had n_1 objects split into r_1 runs and n_2 objects split into r_2 runs. Therefore, number of ways of doing it is ${}^{n_1-1}C_{r_1-1}$. Number of ways of doing this is ${}^{n_2-1}C_{r_2-1}$. Therefore, total number of combinations ${}^{n_1-1}C_{r_1-1}$ multiplied by ${}^{n_2-1}C_{r_2-1}$. Why you are multiplying? So, let us understand that suppose this is one sequence of blue balls.

Now in order to make these three to be runs in each of this we have to fill in some objects of the different kinds. So, if these are blue balls 1, 2, 3 so to make this run separate we have to insert some run of red balls and if these partitions of 2 runs can be done in k different ways so we can fill it in k different ways. Therefore, the number of ways of making 3 blue ball runs multiplied by the number of ways of making 2 red ball runs will give me 5 different runs.

That is why we see that this comes to be ${}^{n_1-1}C_{r_1-1} \times {}^{n_2-1}C_{r_2-1}$.

(Refer Slide Time: 17:45)



Also note that very interesting thing that r_1 and r_2 cannot be very arbitrary. Given r_2 r_1 can take values $r_2 + 1$ or $r_2 - 1$ or r_1 can be equal to r_2 . Say for example, if these are the runs of r_2 3 runs then there can be 3 possibilities for r_1 .; r_1 can be 2, r_1 can be 3 or r_1 can be 4. So, if r_2 is equal to 3 r_1 can take either 2 or 3 or 4 it has no other possible choice because if there are more element of the first type then actually we will combine them into 1 run.

(Refer Slide Time: 19:03)

Thus, we make the following observations

So, with these observations we get the following.

(Refer Slide Time: 19:07)

To obtain the probability that R takes a value r
 (Note that r can be either even, say $2d$, or odd, say $2d + 1$)

1. If $R = 2d$, then both r_1 and r_2 are equal to d , and the sequence may start with either A or B .
2. If $R = 2d + 1$ then either,
 - i. $r_1 = d + 1$ and $r_2 = d$ and the sequence starts with A
 - Or,
 - ii. $r_1 = d$ and $r_2 = d + 1$ and the sequence starts with B

Then we find $P(R = r)$ as follows:

To obtain the probability that R takes a value small r where R is the total number of runs. Therefore, small r can be either even say $2d$ or odd that is $2d + 1$. Now if the total number of runs is $2d$ that is even then both r_1 and r_2 have to be equal to d and the sequence may start with either A or B . Similarly, if R is equal to $2d$ plus 1 that is odd numbers then r_1 can be $d + 1$ and r_2 can be d . Or alternatively, r_1 is equal to d and r_2 is equal to $d + 1$.

In the first case the sequence starts with A and also ends with A . In the second case it starts with B and also ends with B . Therefore, probability that R is equal to small r is computed as follows.

(Refer Slide Time: 20:22)

The number of ways that r_1 can take a value d is $\binom{n_1-1}{d-1}$ ✓
 and r_2 can take the value d is $\binom{n_2-1}{d-1}$ ✓

∴ The number of ways that $R = 2d$ is $2 * \binom{n_1-1}{d-1} * \binom{n_2-1}{d-1}$

Hence, $P(R = 2d) = \frac{2 * \binom{n_1-1}{d-1} * \binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}}$ ✓

In a similar way,

$P(R = 2d + 1) = \frac{\binom{n_1-1}{d} * \binom{n_2-1}{d-1} + \binom{n_1-1}{d-1} * \binom{n_2-1}{d}}{\binom{n_1+n_2}{n_1}}$ ✓

Number of ways that r_1 can take value of d is $^{n_1-1}C_{d-1}$ just sometime back we studied this. Similarly, r_2 can take value d in $^{n_2-1}C_{d-1}$ ways. Therefore, the number of ways that R can be two times d is $2 \times ^{n_1-1}C_{d-1} \times ^{n_2-1}C_{d-1}$. Why the 2 is coming it is coming because it may start with A and end with B or it may start with B and end with A .

Therefore, probability R is equal to $2d$ is equal to

$2 \times ^{n_1-1}C_{d-1} \times ^{n_2-1}C_{d-1}$ divided by $^{n_1+n_2}C_{n_1}$ as this is the total number of possible runs. In a similar way probability R is equal to $2d + 1$ is equal to $^{n_1-1}C_d \times ^{n_2-1}C_{d-1}$ that is this is the case when there are $d + 1$ runs of the first type and d runs for the second time plus

$$^{n_1-1}C_{d+1-1} \times ^{n_2-1}C_d.$$

That means d runs of the first type and $d + 1$ runs of the second type divided by the total number of possibilities that is $^{n_1+n_2}C_{n_1}$.

(Refer Slide Time: 22:05)



Let us now consider for illustration a case with $n_1 = n_2 = 5$.

(Refer Slide Time: 22:14)

$$\begin{aligned}
 P(R = \underline{2}) &= \frac{2 \cdot \binom{n_1-1}{d-1} \cdot \binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}} \\
 &= \frac{2 \cdot \binom{5-1}{1-1} \cdot \binom{5-1}{1-1}}{\binom{10}{5}} = \frac{2 \cdot \binom{4}{0} \cdot \binom{4}{0}}{\binom{10}{5}} = \frac{\underline{2}}{252} = \underline{0.008}
 \end{aligned}$$

A A A A A
 B B B B B
 B...BA...A

$$\begin{aligned}
 P(R = \underline{10} (= 2 \cdot 5)) &= \frac{2 \cdot \binom{n_1-1}{d-1} \cdot \binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}}
 \end{aligned}$$

A B A B ... A B
 B A B A ... B A

$$\begin{aligned}
 &= \frac{2 \cdot \binom{5-1}{5-1} \cdot \binom{5-1}{5-1}}{\binom{10}{5}} = \frac{2 \cdot \binom{4}{4} \cdot \binom{4}{4}}{\binom{10}{5}} = \frac{\underline{2}}{252} = \underline{0.008}
 \end{aligned}$$

Therefore, probability that R is equal to 2 that is number of runs is equal to 2 that is coming out to be from this formula 2 upon 252 which is 0.008. Now what is 2? 2 is equal to the number of arrangements that satisfy that total number of run is 2 and we know that it can be only in two ways that is A, A, A, A and A followed by B, B, B, B and B or it can be B followed by A.

There are only two possibilities to have total number of run is 2. In a similar way probability R is equal to 10 that is also coming out to be 2 upon 252 by applying the formula and here also we can justify the same that the number of run is 10 if we get AB, AB up to AB or BA, BA up to BA. So, this formula whatever it is giving is correct from practical justification.

(Refer Slide Time: 23:51)

$$P(R = 6 (= 2 * 3)) = \frac{2 * \binom{n_1-1}{d-1} * \binom{n_2-1}{d-1}}{\binom{n_1+n_2}{n_1}}$$

$$= \frac{2 * \binom{5-1}{3-1} * \binom{5-1}{3-1}}{\binom{10}{5}} = \frac{2 * \binom{4}{2} * \binom{4}{2}}{\binom{10}{5}} = \frac{72}{252} = 0.2857$$

*∴ No. of possible arrangements = 2 * 6 * 6 = 72*
5 A's can be arranged into 3 runs in 6 different ways. Similarly 5 B's can be arranged into 3 runs in 6 different ways.

What about the probability of R is equal to 6. So, the same formula by applying the value of d is equal to 3 we get it is giving us 72 upon 252 which is 0.2857. Question is how do we get 72 many arrangements to give us 6 runs? So, 5 A's can be arranged into 3 runs in 6 different ways. Similarly, 5 B's can be arranged into 3 runs in 6 different ways. Therefore, number of possible arrangements is equal to 2 times 6 into 6 is equal to 72. Thus, from here also we get the correct answer.

(Refer Slide Time: 25:43)

Justification: R = 6. Therefore $r_1 = r_2 = 3$

Now 3 runs of 5 A elements can be in the following six ways:

(3 1 1), (2 2 1), (2 1 2), (1 3 1), (1 2 2), (1 1 3)

ooo . . *oo oo 1* *!* *.*

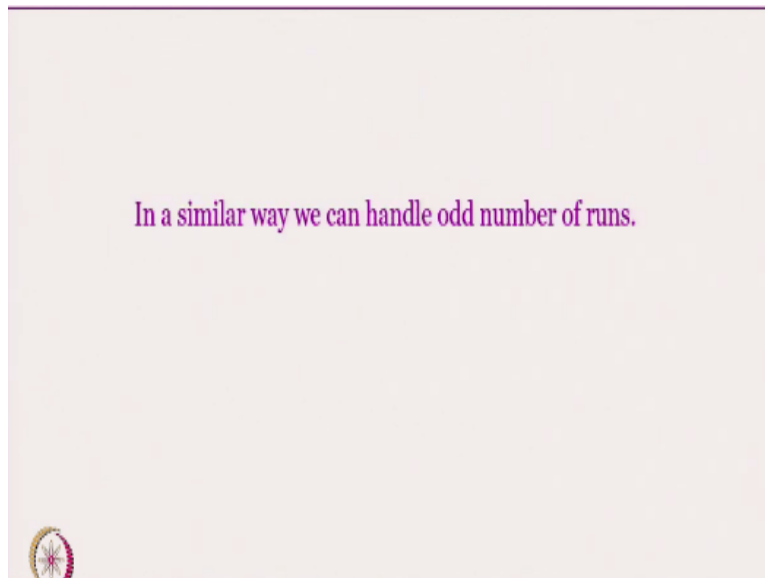
In a similar way 3 runs of 5 B elements can be in six ways.

Hence total possibility = 6 * 6 * 2 = 72

Now the question is what are the 6 different ways if we look at we have shown it to you that 5 elements can be put into 3 runs. The possibilities are 3, 1, 1 or 2, 2, 1 and similarly 1, 3, 1

and from here we get 1, 1, 3. So, these are the only 6 ways of splitting 5 elements into 3 runs. So, that is what we have applied there for both A and B and our ultimate answer came out to be 72.

(Refer Slide Time: 26:40)



In a similar way we can handle odd number of runs as well.

(Refer Slide Time: 26:44)

$$\begin{aligned}
 P(R = 7 (= 2 + 3 + 1)) &= \frac{\binom{n_1-1}{d} \binom{n_2-1}{d-1} + \binom{n_1-1}{d-1} \binom{n_2-1}{d}}{\binom{n_1+n_2}{n_1}} \\
 &= \frac{\binom{5-1}{3} \binom{5-1}{3-1} + \binom{5-1}{3-1} \binom{5-1}{3}}{\binom{10}{5}} = \frac{2 \cdot \binom{4}{3} \binom{4}{2}}{\binom{10}{5}} \\
 &= \frac{48}{252} = 0.190
 \end{aligned}$$

For example consider R is equal to 7 therefore by applying the formula we get the number of such runs is 48 so that probability of 7 runs is 0.190.

(Refer Slide Time: 27:04)

Justification: $R = 7$

Number of ways of getting 3 runs of five elements is 6 ✓

Number of ways of getting 4 runs of five elements is 4: ✓

2111, 1211, 1121, 1112.

.....

∴ Total no. of runs is
 $2 \times 6 \times 4 = 48$



Now the question is how do we know that there will be 48 runs? That comes from this calculation. Just now we have seen that probability of 3 runs of 5 elements is 6. Now getting 4 runs out of 5 elements there are 4 possibilities. We can have them 2 then 1, 1, 1 or 1, 2 then 1, 1, 1 and similarly from the other two you can understand the partition into 5 runs. Therefore, 6 from here and 4 from here.

Therefore, total number of runs is 2 into 6 into 4 is equal to 48. The 2 is coming because we may have 4 runs of A and 3 runs of B or alternatively 4 runs of B and 3 runs of A. So, that is why the factor 2 has come, but at the end this justifies why the formula gave us that there are 48 many different possibilities of arranging the elements such that total number of run is equal to 7.

(Refer Slide Time: 28:34)

Thus we have : $P(R = 7) = 0.190$

In similar way we have:

$$P(R = 8) = \frac{2 * \binom{5-1}{4-1} * \binom{5-1}{4-1}}{\binom{10}{5}} = \frac{2 * \binom{4}{1} * \binom{4}{1}}{\binom{10}{5}} = \frac{32}{252} = 0.127$$
$$P(R = 9) = \frac{\binom{5-1}{4} * \binom{5-1}{4-1} + \binom{5-1}{4-1} * \binom{5-1}{4}}{\binom{10}{5}} = \frac{2 * \binom{4}{4} * \binom{4}{3}}{\binom{10}{5}} = \frac{8}{252} = 0.032$$
$$P(R = 10) = \frac{2 * \binom{5-1}{5-1} * \binom{5-1}{5-1}}{\binom{10}{5}} = \frac{2 * \binom{4}{4} * \binom{4}{4}}{\binom{10}{5}} = \frac{2}{252} = 0.008$$

Therefore, we have probability R is equal to 7 is 0.190. In a similar way, we can calculate that probability R is equal to 8 is 0.127 probability R is equal to 9 is 0.032 probability R is equal to 10 that we have already obtained is 0.008.

(Refer Slide Time: 29:04)

Hence we have the following Right Tail probabilities:

1. $P(R \geq 10) = 0.008$ ✓
2. $P(R \geq 9) = 0.008 + 0.032 = 0.040$ ✓
3. $P(R \geq 8) = 0.008 + 0.032 + 0.127 = 0.167$ ✓
4. $P(R \geq 7) = 0.008 + 0.032 + 0.127 + 0.190 = 0.357$ ✓

These values are needed for acceptance and rejection of H_0 .

Now if we look at the right tail probabilities we can see that by adding the terms probability $R \geq 10$ is 0.008, $R \geq 9$ is 0.040, $R \geq 8$ is 0.167 and $R \geq 7$ is 0.357. Why did we calculate that because we need these values to accept or reject the null hypothesis.

(Refer Slide Time: 29:44)

Right-tail probabilities															
n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
2	2	4	.333	4	8	9	.071	5	11	11	.068	6	12	12	.075
2	3	5	.100			8	.212			10	.154			11	.217
		4	.500			7	.467			9	.374			10	.395
2	4	5	.200	4	9	9	.098	5	12	11	.075	6	13	13	.034
2	5	5	.286			8	.255			10	.181			12	.092
2	6	5	.357	4	10	9	.126	5	12	9	.421			11	.257
2	7	5	.417			8	.294	5	13	11	.092			10	.439
2	8	5	.467	4	11	9	.154			10	.208	6	14	13	.044
3	3	6	.100			8	.330			9	.465			12	.111
		5	.300	4	12	9	.181	5	14	11	.111			11	.295
3	4	7	.029			8	.363			10	.234			10	.480
		6	.200	4	13	9	.208	5	15	11	.129	7	7	14	.001
		5	.457			8	.393			10	.258			13	.004
3	5	7	.071	4	14	9	.234	6	6	12	.002			12	.025
		6	.286			8	.421			11	.013			11	.078
3	6	7	.119	4	15	9	.258			10	.067			10	.209
		6	.357			8	.446			9	.175			9	.383
3	7	7	.167	4	16	9	.282			8	.392	7	8	15	.000
		6	.417			8	.470	6	7	13	.001			14	.002
3	8	7	.212	5	5	10	.008			12	.008			13	.012
		6	.467			9	.040			11	.034			12	.051
3	9	7	.255			8	.167			10	.121			11	.133
3	10	7	.294			7	.357			9	.267			10	.296

And for that we need to look at a table with right tail probabilities. So, here is the table and we are considering number of A and number of B is to be 5 and 5. So, if we look at we can see that the right tail probabilities for 10 is 0.008 for 9, 0.04 for 8 it is 0.167 and for 7 it is 0.357. Therefore, we see that we have obtained the same values through our calculations using the formula for obtaining the run probabilities.

(Refer Slide Time: 30:33)

The above table giving the Right Tail probabilities is taken from

Non-Parametric Statistical Inference, 4th Edn
by J.D. Gibbons and S. Chakraborti)

The above table giving the right tail probabilities is taken from the book nonparametric statistical inference written by Gibbons and Chakraborti.

(Refer Slide Time: 30:45)

Large sample Approximation.

When either n_1 or n_2 are greater than 20 or both are greater than 10, one can approximate with Normal distribution with the following

Mean and Variance:

$$E(R) = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1 \quad \checkmark$$
$$Var(R) = \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)} \quad \checkmark$$

The above test is known as Wald-Wolfowitz Run Test



Now large sample approximations when either n_1 or n_2 are greater than 20 or both are greater than 10 then one can approximate with normal distribution with the following parameters. The mean is coming out to be

$$\frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1$$

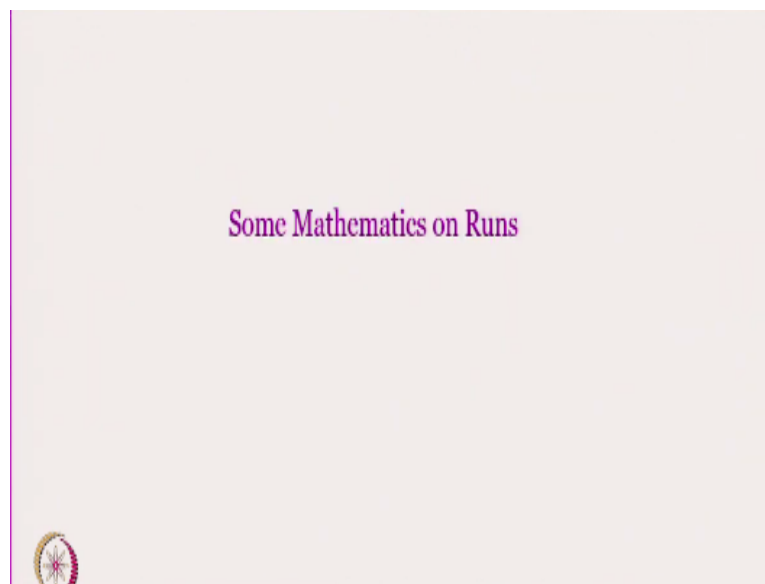
Note that we are using the notation m and n for the number of elements for the two populations here it is given n_1 and n_2 .

So m is actually the n_1 and n is actually the n_2 . Therefore, variance of R is given as

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

This is a complicated expression we are not going to derive it here in this class. Let us accept this formula the test that we have described so far based on the runs is called Wald Wolfowitz run test.

(Refer Slide Time: 32:08)



So, some simple mathematical observations on the number of runs.

(Refer Slide Time: 32:15)

A slide with a light beige background and a purple border, containing handwritten notes in red ink. The text is as follows:

Q1. What is the marginal distribution of R_1 ? min 1
max = n_1

Given n_1 and n_2 the random variable R_1 can take integral values in the set $\{1, 2, \dots, n_1\}$

To compute $P(R_1 = r_1)$ when $r_1 \in \{1, 2, \dots, n_1\}$

The slide also features a small circular logo in the bottom-left corner and the NPTEL logo at the bottom center.

What is the marginal distribution of R_1 where if you remember R_1 is the number of runs of the first element. We know that the minimum value for R_1 is 1 and the maximum value of R_1 is n_1 that is the number of elements. Therefore, given n_1 and n_2 the random variable R_1 can take value integral values in the set 1, 2, 3 up to n_1 . To compute probability R_1 is equal to small r_1 we do the following.

(Refer Slide Time: 32:59)

Hence one can compute the marginal probability $f_{R_1}(r_1)$ as follows:

$$f_{R_1}(r_1) = f_{R_1, R_2}(r_1, r_1) + f_{R_1, R_2}(r_1, r_1 - 1) + f_{R_1, R_2}(r_1, r_1 + 1)$$

$$f_{R_1}(r_1) = \frac{2 \cdot \binom{n_1-1}{r_1-1} \binom{n_2-1}{r_1-1}}{\binom{n_1+n_2}{n_1}} + \frac{\binom{n_1-1}{r_1-1} \binom{n_2-1}{r_1-2}}{\binom{n_1+n_2}{n_1}} + \frac{\binom{n_1-1}{r_1-1} \binom{n_2-1}{r_1}}{\binom{n_1+n_2}{n_1}}$$

If R_1 of small r_1 is equal to the joint distribution of r_1, r_1 in r_1, r_2 that means the number of runs of the first element and number of runs of the second element are same. We have already seen that this value can be same or ± 1 of R_1 . Therefore, we also add to it the probability of $f_{R_1, R_2}(r_1, r_1 - 1)$ and similarly r_1 and $r_1 + 1$. We know all these values this we have already computed.

This is ${}^{2n_1-1}C_{r_1-1} \times$ by ${}^{n_2-1}C_{r_1-1}$ similarly from this we get this and from here we get this. This we are getting by replacing the value of r_2 with the corresponding parameter in the relevant expression. Now we need to simplify this. How we simplify?

(Refer Slide Time: 34:14)

$$\begin{aligned}
 & \therefore \binom{n_1+n_2}{r_1} f_{R_1}(r_1) \leftarrow \text{marginal probability} \\
 &= 2 * \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1-1} + \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1-2} + \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1} \\
 &= \binom{n_1-1}{r_1-1} * \left[\binom{n_2-1}{r_1-1} + \binom{n_2-1}{r_1-2} + \binom{n_2-1}{r_1-1} + \binom{n_2-1}{r_1} \right] \\
 &= \binom{n_1-1}{r_1-1} * \left[\binom{n_2}{r_1-1} + \binom{n_2}{r_1} \right] \\
 &= \binom{n_1-1}{r_1-1} \binom{n_2+1}{r_1}
 \end{aligned}$$

$nCr + nCr-1 = n+1Cr$
 $\checkmark \binom{n+1}{r}$

So, let us look at it. So $\binom{n_1+n_2}{r_1} f_{R_1}(r_1)$ that is this is the marginal distribution or marginal probability. Therefore, it is writing as

$$2 * \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1-1} + \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1-2} + \binom{n_1-1}{r_1-1} * \binom{n_2-1}{r_1}$$

Now let us take $\binom{n_2-1}{r_1-1}$ common and we write two times this as summation of this. Therefore we are getting $\binom{n_2-1}{r_1-1}$. Another one we write it here $\binom{n_2-1}{r_1-1} + \binom{n_2-1}{r_1-2} + \binom{n_2-1}{r_1}$. Now summation of $\binom{n_2-1}{r_1-1}$ and $\binom{n_2-1}{r_1-2} = \binom{n_2}{r_1-1}$. So, we know that $nCr + nCr-1 = n+1Cr$. So, $\binom{n_2-1}{r_1-1} + \binom{n_2-1}{r_1-2}$ by applying this formula is $\binom{n_2}{r_1-1}$.

In a similar way this is giving us $\binom{n_2}{r_1}$. Therefore the whole thing comes out to be

$\binom{n_1-1}{r_1-1} \binom{n_2+1}{r_1}$. Again we are applying the same formula we get $\binom{n_1+n_2}{r_1}$.

(Refer Slide Time: 36:34)

$$\therefore f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1} \binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}}$$

Thus we get the marginal probability of R_1 taking the values r_1 , where possible values for r_1 are 1, 2, .. n_1 .

In a similar way, marginal probability of R_2 can be computed for $r_2 = 1, 2, \dots, n_2$

We accept the following theorem without proof.



Therefore

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1} \binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}}$$

Thus, we get the marginal probability of r_1 taking the values small r_1 in the set 1, 2, 3 up to n_1 .

In a similar way, the marginal probability of r_2 can be computed for r_2 is equal to 1, 2 up to n_2 and it does not take much to understand that it is going to be

$$f_{R_2}(r_2) = \binom{n_1+1}{r_2} \times \binom{n_2-1}{r_2-1} / \binom{n_1+n_2}{n_1}.$$

Now we accept the following theorem without proof.

(Refer Slide Time: 37:51)

Theorem: The probability distribution of R, the total number of runs with n_1 objects of Type A and n_2 objects of Type B in a random sample is given by:

$$f_R(r) = \begin{cases} \frac{2 \binom{n_1-1}{r/2-1} \binom{n_2-1}{r/2-1}}{\binom{n_1+n_2}{n_1}} & \underline{\underline{r \text{ is even}}} \\ \frac{\binom{n_1-1}{(r-1)/2} \binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2} \binom{n_2-1}{(r-1)/2}}{\binom{n_1+n_2}{n_1}} & \underline{\underline{r \text{ is odd}}} \end{cases}$$

The probability of distribution of R that is the total number of runs with n_1 objects of type A and n_2 objects of type B in a random sample is given by this complicated formula when r is even and this is when r is odd. It looks complicated, but we are familiar with this $n_1-1 \text{ C } r \text{ by } 2$ minus 1 because it is even say $2d$. Therefore, this is actually the d that we are talking about and here also this is the d that we are talking about. In a similar way when r is odd it is $2d$ plus 1 so one has to be d other has to be d plus 1 and by replacing $2d$ plus 1 by r we can get this formula. I want you to verify that because all the necessary mathematics we have already studied.

(Refer Slide Time: 38:49)

Observation

Suppose we define $n_1 + n_2$ random variables:

$R_{11}, R_{12}, \dots, R_{1n_1}$ and $R_{21}, R_{22}, \dots, R_{2n_2}$ ✓

Where, R_{ij} denotes the number of runs of type i , (where $i = 1, 2$) of length j , $j = 1, 2, \dots, n_i$

Note that

Similarly for B

$R_{21} = \# \text{ of runs of length 1}$

And,

$R_{2,n_2} = \# \text{ of runs of length } n_2$

And,

$\sum_{j=1}^{n_1} j R_{1j} = n_1$ ✓

$\sum_{j=1}^{n_2} j R_{2j} = n_2$ ✓

And,

$\sum_{j=1}^{n_1} R_{1j} = r_1$ ✓

$\sum_{j=1}^{n_2} R_{2j} = r_2$ ✓

Possible values for R_i

$= 1, 2, \dots, n_i$

$R_{11} = \# \text{ runs of length 1}$

$R_{12} = \# \text{ runs of length 2}$

$\vdots$

$R_{1,n_1} = \# \text{ runs of length } n_1$

This is w.r.t A

Also there is another interesting observation with respect to the runs is that if we define n_1 plus n_2 random variables R_{11}, R_{12}, R_{1n_1} and R_{21}, R_{22}, R_{2n_2} where R_{ij} denotes the number

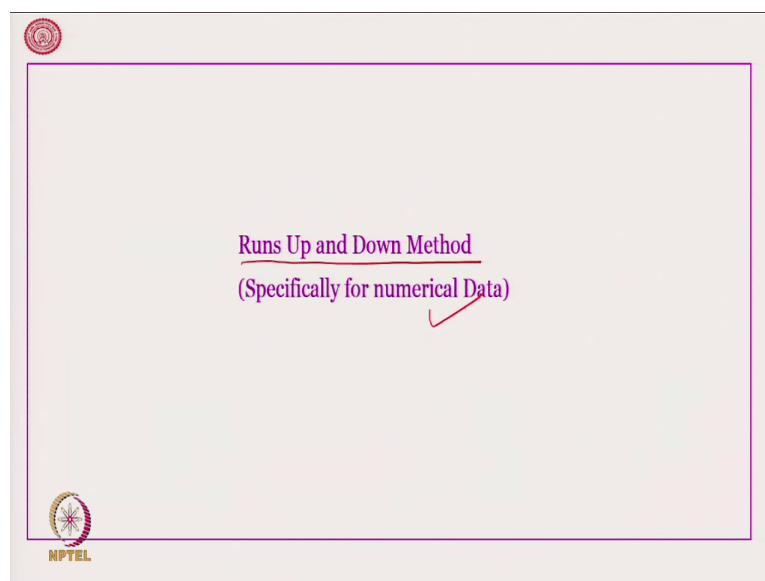
of runs of type i where i is equal to 1, 2 of length j that is j is equal to this should be $j = 1, 2, 3$ up to n_1 . What does it mean? That means the number of possible values for R_1 is equal to 1, 2, 3 up to n_1 .

So, R_{11} is number of runs of length 1 R_{12} is equal to number of runs of length 2 up to R_{1n_1} is equal to number of runs of length n_1 . This is with respect to element of first type that is A. Similarly, for B what we will get R_{21} is equal to number of runs of length 1 up to R_{2n_2} is equal to number of runs of length n_2 that is how we get R_{11}, R_{12} up to $R_{1n_1}, R_{21}, R_{22}, R_{2n_2}$ so many variables that is n_1 plus n_2 too many variables.

But let us note that if R_{1n_1} it cannot take any value more than 1 because if the length of the run is n_1 that means that all A observations are in that run therefore there can be only at most one run of length n_1 . Similarly, there can be at most one run of length n_2 with respect to the elements of type B. Now note that $\sum_{j=1}^{n_1} j r_{1j}$ is equal to 1 to n_1 is equal to n_1 and j is equal to 1 to n_2 $\sum_{j=1}^{n_2} j r_{2j}$ is equal to n_2 . Why? Since r_{1j} is the number of runs of length j therefore basically we are looking at number of runs of length j multiplied by the j that is the length of that run.

Therefore, when you sum over all the possible values we get the total number of elements which is n_1 and in this case which is n_2 . Also note that $\sum_{j=1}^{n_1} r_{1j}$ is the total number of runs of the first element and similarly this expression gives the total number of runs of the second element.

(Refer Slide Time: 42:40)



Let me now describe a variation of the run test which is called runs up and down method this is applicable only for numerical data.

(Refer Slide Time: 42:54)

Suppose there is a time series data. One may like to check if there is any trend in the data.

E.g. Number of cars passing through a corridor every day.
Number of infected cases in Corona Virus on daily basis.

One may be interested whether there is any upward trend or downward trend or is it just random.

Typically in such cases, to test randomness one classifies the values as a sample of two symbols '+' and '-' depending upon whether a sample value is above or below the median.



Suppose, we have a time series data and one may like to check if there is any trend in that data. For example, the number of cars passing through a corridor every day, it can be number of infected cases of Corona Virus on a daily basis. So, if we take the data for different time instance which in this case are the days. Suppose, the data is increasing then we can find a trend like this which is an upward trend or we may find a downward trend which may look like this.

And we are often interested in testing if there is any trend in the data or it is absolutely random. If we just want to test a randomness we can split the data into two parts. Some of them are plus and some of them are minus depending upon whether the sample value is above or below the median. This we have seen.

(Refer Slide Time: 44:13)

Illustration

Consider the following Table

Time	1	2	3	4	5	6
Sample Value	8	13	1	3	4	7
Symbol	+	+	-	-	-	+

(Sample median is 5.5)

But in this way some information is lost which can be used to some extent in identifying pattern in the time -ordered observation.

Runs Up and Down uses this information for more detailed analysis.

For example, consider the table if we have 6 time instance and the observed values are like this therefore the sample median is coming to be 5.5 and therefore what we are looking at that these 3 are above the sample median and these 3 are below the sample median. However, in this type of test we lose the information about the trend because we have done similar thing with respect to sign test. Runs up and down test preserves the information about the trend.

(Refer Slide Time: 44:58)

Here, given n time-ordered sample $x_1, x_2, \dots, x_n$ we construct an $n - 1$ element sequence D_{n-1} of + and - whose j^{th} element is,

$$\begin{cases} + & \text{if } x_{j+1} - x_j > 0 \\ - & \text{if } x_{j+1} - x_j < 0 \end{cases}$$

Thus for the above data we get the following table:

Time	1	2	3	4	5	6
Sample Value	8	13	1	3	4	7
Sequence of Symbols (D_5)	+	+	-	-	-	+

How it works. So, here what we do given n time ordered sample $x_1, x_2, \dots, x_n$ we construct an n minus 1 element sequence which you call D_{n-1} or plus and minus, but whose j th element is plus if $x_{j+1} - x_j$ is greater than 0 and it is minus if $x_{j+1} - x_j$ is

less than 0. Therefore, from the same data now we get a sequence of 5 symbols, but it is different from the one we get for the sign test.

Now why this is plus because at time instance 1 the value is 8 at time instance 2 the value is 13 therefore there is an increment hence it is plus from 2 to 3 the value has come down from 13 to 1 therefore it is minus. From 3 to 4 value goes from 1 to 3 therefore it is plus here it is 3 to 4 plus and here it is 4 to 7 therefore another plus. Therefore, we get a different sequence here.

(Refer Slide Time: 46:16)

Under the null hypothesis of randomness H_0 the + and - are expected to be nicely intermingled, so that there will not be a longer run of any of the two symbols.

Thus here focus is NOT on the number of runs. 

Rather here we look at the length of the runs.

A long run of + (or -) implies that the data is showing some clear trend. ✓

Hence H_0 will be rejected if there are r runs of length t or more, where r and t are decided on the basis of α .

Therefore, under the null hypothesis of randomness the plus and minus are expected to be nicely intermingled so that there will be not be longer run of any of the two symbols because if there is a longer run of plus, plus, plus like that what we understand that for the entire time period the value is increasing that means there is an upward trend and therefore the value is not random.

Therefore, here we look at the length of the runs not the total number of runs that we have just seen with respect to Wald-Wolfowitz runs test. If there is a long run of plus or minus it implies that the data is showing some clear trend. Therefore, H_0 will be rejected if there are r runs of some length t or more where r and t are decided on the basis of α that means if there are large number of runs of certain length or more than that then we are going to reject the randomness of the data.

(Refer Slide Time: 47:41)

The Analysis

The method considers the number of runs of length j ,
 $j = 1, 2, \dots, n-1$

Let R_j be the random variable denoting the no. of runs,
 either up or down of length j in the sequence $D_{(n-1)}$

For illustration, let $n = 3$.

Therefore, there can be runs of length 1 or length 2.

Now three numbers can be arranged among themselves
 in $3! = 6$ possible ways each one is equally likely.

For each of them D_2 will contain runs of + and/or -

$++- - - +$

So, the method considers the number of runs of length j , j is equal to 1, 2, 3 up to n minus 1 because the length of a run can be 1 can be 2 or maximum it can be n minus 1 because if the entire data is showing an upward trend there will be one run of length n minus 1 all plus. For illustration let n be equal to 3. Therefore, there can be runs of length 1 or 2. Now the three numbers can be arranged among themselves in 6 possible ways each one of them is equally likely.

Therefore, since n is equal to 3 you are looking at D_2 and D_2 will contain runs of plus or minus or maybe mixture of that one.

(Refer Slide Time: 48:38)

S_3	D_2	r_1	r_2
$(a_1 < a_2 < a_3)$	$(+, +)$	0	1
$(a_1 < a_3 < a_2)$	$(+, -)$	2	0
$(a_2 < a_1 < a_3)$	$(-, +)$	2	0
$(a_2 < a_3 < a_1)$	$(+, -)$	2	0
$(a_3 < a_1 < a_2)$	$(-, +)$	2	0
$(a_3 < a_2 < a_1)$	$(-, -)$	0	1

$r_1 = \# \text{ runs of length 1}$
 $r_2 = \# \text{ runs of length 2}$

two instances of $(0, 1)$
 four instances of $(2, 0)$

So, let us illustrate suppose we have three elements a_1, a_2, a_3 . A_1 is the value obtained at the first instance, a_2 is the value obtained at the second instance and a_3 is the value obtained at the third instance. Now if a_1 is less than a_2 less than a_3 then D_2 is going to be plus because a_2 is greater than a_1 and another plus because a_3 is greater than a_2 . Therefore we get a run of length 2.

Therefore, what we are saying r_1 is equal to 0 r_2 is equal to 1 so what is r_1 ? R_1 is equal to number of runs of length 1 R_2 is equal to number of runs of length 2 and therefore we get 0, 1. If a_1 is less than a_3 less than a_2 therefore from a_1 to a_2 we get a plus from a_2 to a_3 we get a minus therefore we get 2 runs of length 1 therefore r_1 is equal to 2 and r_2 is equal to 0. We know that there will be 6 possible permutations. For each of them we calculated the value of r_1, r_2 and we are getting that two instances of 0, 1 and 4 instances of 2, 0 as it is clear from here.

(Refer Slide Time: 50:34)

Corresponding to each arrangement we can compute the value of r_2 and r_1 .

Here r_2 and r_1 represent the number of runs of length 2 and 1 respectively.

In this example (r_1, r_2) can have only 2 possible values: (0, 1) and (2, 0) with 2 and 4 instances, respectively.

Therefore $f_3(0, 1) = \Pr((0, 1)) = \frac{1}{3}$ and $f_3(2, 0) = \Pr(2, 0) = \frac{2}{3}$.



Corresponding to each arrangement we can compute the value of r_2 and r_1 . Hence, r_2 and r_1 represents the number of runs of length 2 and 1 respectively as I have indicated before and therefore $f_3(0, 1)$ this is the notation which is actually the probability that we get a 0, 1 sequence that probability is 1 by 3 and $f_3(2, 0)$ that means we get 2 runs of length 1 and 0 runs of length 2 that probability is going to be 2 by 3 as we have seen just now.

(Refer Slide Time: 51:20)

In general one can compute $f_n(r_1, r_2, \dots, r_{n-1})$, which gives the probability that there are $r_1, r_2, \dots, r_{n-1}$

- * exactly r_1 many runs of length 1,
- * r_2 many runs of length 2
- ...
- * r_i many runs of length i ,
- ...
- * r_{n-1} many runs of length $n - 1$.

Now let us generalize it if there are n samples we can think of f_n that is the probability of r_1, r_2, r_{n-1} where r_1 is the number of runs of length 1, r_2 is the number of runs of length 2 and r_{n-1} is the number of runs of length $n-1$.

(Refer Slide Time: 51:44)

Observations

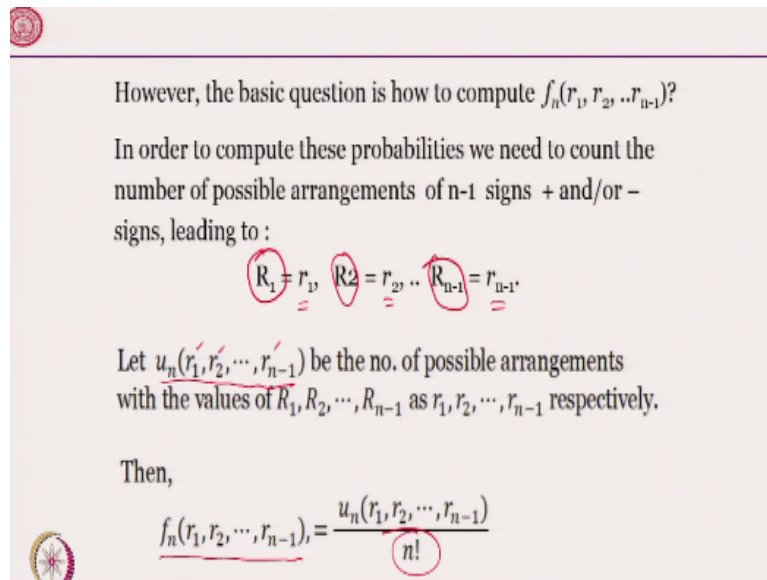
It is obvious that,

- r_{n-1} cannot be more than 1.
- If $r_{n-1} = 1$, then all other values are 0. ✓
- If $r_j > 0$ for any $j \neq n-1$, then $r_{n-1} = 0$
- If $r_1 = n-1$ then all others are 0. + - + - - -
- r_1 can never be equal to $n-2$. ✓ - + - + - -
- $\sum_{j=1}^{n-1} j r_j = n-1$

It is obvious that r_{n-1} cannot be more than 1 because if there is run of $n-1$ many elements whether plus or minus that means it has taken care of all the transitions and therefore there cannot be any other element in the sequence therefore there can be only one run of length $n-1$, but if r_{n-1} is equal to 1 then all other values are 0. However, if r_j is greater than 0 for any j not equal to $n-1$ then r_{n-1} is equal to 0.

Because then it can have a run of length n minus 1 if r_1 is equal to n minus 1 where all others are going to be 0 because we have got n minus 1 many runs of length 1 that means it has to be plus, minus, plus, minus or something like minus, plus, minus, plus like that. R_1 can never be equal to n minus 2 which is pretty obvious and $\sum_j jr_j$ where j is the length of the run and r_j is the number of runs of length j . So, that gives us the total number of symbols in the run and that is going to n minus 1.

(Refer Slide Time: 53:15)



However, the basic question is how to compute $f_n(r_1, r_2, \dots, r_{n-1})$?

In order to compute these probabilities we need to count the number of possible arrangements of $n-1$ signs + and/or - signs, leading to :

$$R_1 = r_1, R_2 = r_2, \dots, R_{n-1} = r_{n-1}.$$

Let $u_n(r_1, r_2, \dots, r_{n-1})$ be the no. of possible arrangements with the values of $R_1, R_2, \dots, R_{n-1}$ as $r_1, r_2, \dots, r_{n-1}$ respectively.

Then,

$$f_n(r_1, r_2, \dots, r_{n-1}) = \frac{u_n(r_1, r_2, \dots, r_{n-1})}{n!}$$

However, the basic question is how to compute this probability $f_n, r_1, r_2, \dots, r_{n-1}$. So, in order to compute these probabilities we need to count the number of possible arrangements of n minus 1 sign of plus or minus and which leads to definition of n minus 1 random variables where R_1 is the random variable denoting the number of runs of length 1. Similarly R_2 and R_{n-1} and we are looking at the probabilities of these variables taking the value small r_1 small r_2 up to small r_{n-1} .

So, let $u_n, r_1, r_2, \dots, r_{n-1}$ be the number of possible arrangements so that we can get the runs of these numbers. Therefore, $f_n, r_1, r_2, \dots, r_{n-1}$ is going to be $u_n, r_1, r_2, \dots, r_{n-1}$ divided by n factorial because that is the number of way of arranging n many elements.

(Refer Slide Time: 54:34)

For illustration, with respect to the earlier example:

$$u_3(0,1) = \text{No. of arrangements with no run of length 1 and one run of length 2} = 2$$

and

$$u_3(2,0) = \text{No. of arrangements with 2 runs of length 1 and no run of length 2} = 4$$

Therefore $f_3(0,1) = 2/6 = 1/3$ and $f_3(0,2) = 4/6 = 2/3$

As we have already seen here:

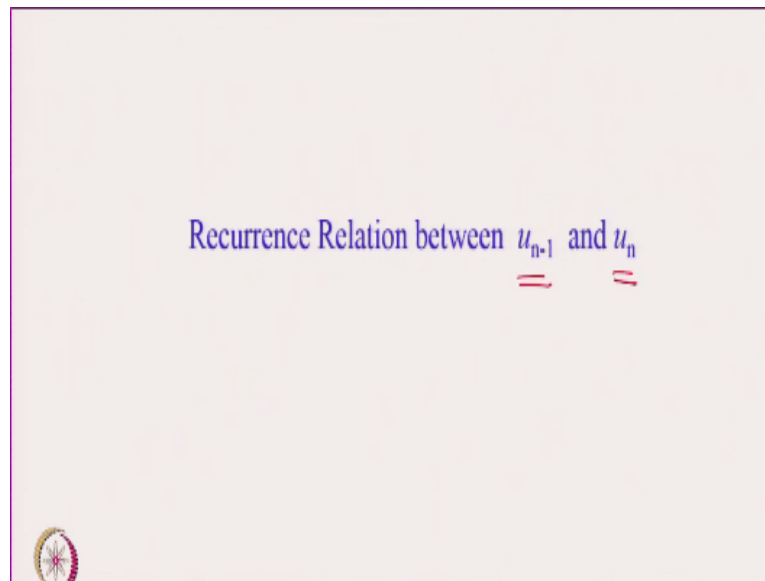
For illustration with respect to the earlier example $u_3(0,1)$ was 2 $u_3(2,0)$ was 4 therefore $f_3(0,1)$ was coming out to be $1/3$ $f_3(0,2)$ was coming out to be $2/3$.

(Refer Slide Time: 54:54)

S_3	D_2	r_1	r_2
$(a_1 < a_2 < a_3)$	$(+, +)$	0	1
$(a_1 < a_3 < a_2)$	$(+, -)$	2	0
$(a_2 < a_1 < a_3)$	$(-, +)$	2	0
$(a_2 < a_3 < a_1)$	$(+, -)$	2	0
$(a_3 < a_1 < a_2)$	$(-, +)$	2	0
$(a_3 < a_2 < a_1)$	$(-, -)$	0	1

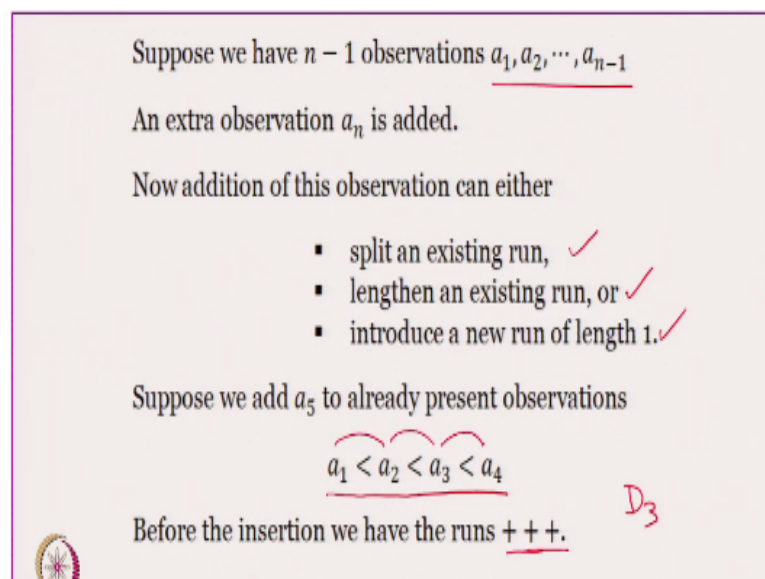
As we have seen from this table sometime back.

(Refer Slide Time: 54:59)



Now question is how do we compute all these u_n 's basically there is a recurrence relation which helps us to get u_n from u_{n-1} .

(Refer Slide Time: 55:12)



So, suppose we have $n - 1$ observations $a_1, a_2, \dots, a_{n-1}$ and extra observation a_n is added. Now addition of this observation can either split an existing run or increase the length of a run or it can introduce a new run of length 1. For illustration, consider 4 data points a_1, a_2, a_3, a_4 and such that it is in increasing order therefore in D_3 we will have plus, plus, plus because from here to here plus, here to here plus and from here to here there is another plus.

(Refer Slide Time: 56:00)

Now suppose a new element a_5 is considered.

Then following possibilities arise:

a_5 is the largest element

	S_5	D_4	r_1	r_2	r_3	r_4
1	$(a_1, a_2, a_3, a_4, a_5)$	$++++$	0	0	0	1
2	$(a_5, a_1, a_2, a_3, a_4)$	$-+++$	1	0	1	0
3	$(a_1, a_5, a_2, a_3, a_4)$	$+ - ++$	2	1	0	0
4	$(a_1, a_2, a_5, a_3, a_4)$	$++ - +$	2	1	0	0
5	$(a_1, a_2, a_3, a_5, a_4)$	$+++ -$	1	0	1	0

So, in case 1, the insertion lengthened an existing run

In cases 2, 5, the insertion introduced a new run of length one.

In cases 3, 4, an existing run of length 3 is split.

Now, suppose the new element a_5 is such that it is the largest of all then what will happen we will get another plus here on the top of that 3 plus and therefore it is increasing the length of this run of plus. Instead if the fifth element comes at the very first instance then here we will get a minus because a_1 is less than a_5 , but after that there is going to be plus, plus, plus. Similarly, if a_5 comes at the second instance then we will get plus, minus, plus, plus.

And similarly for all the 5 cases note that here we have assumed that a_5 is the largest element. Therefore, introduction of a new element will change the pattern of the runs of plus and minus. So, as we have written in case of 2 and 5 the insertion introduced a new run of length one. As you can see here we get a minus, here we get a minus, but in case of 3 and 4 an existing run is split into two parts.

Because here it was plus, plus, plus that is split into plus, minus, plus, plus and accordingly you can see that the values are changed as we have indicated now it is taking as 2 runs of length one, one run of length 2 and no runs of length 3 and 4. So, the entire calculation is run here let me not elaborate that.

(Refer Slide Time: 57:54)

Note that

If the initial arrangement of a_1, a_2, a_3 and a_4 is different then the effect of introducing a_5 will also be different.

Moreover, the relative magnitude of a_5 in comparison with a_1, a_2, a_3 and a_4 will determine the final sequence of arrangements of the 4 symbols + and/or -.

Hence we need to look for generalization. In fact we identify four mutually exclusive and exhaustive cases.

Let me go further. Now if the initial arrangement of a_1, a_2, a_3 and a_4 is different instead of that they are increasing and then the effect of introducing a_5 will also be different as the arrangement of the plus and minus will depend upon the relative order of a_1, a_2, a_3, a_4 and also a_5 . Hence, we need to look for a generalization in fact we can identify for mutually exclusive and exhaustive cases as mentioned below.

(Refer Slide Time: 58:32)

Let S_{n-1} denote the number of arrangements with $n-1$ many + and/or -

- 1) An additional run of length 1 can be added in the arrangement S_n ✓
- 2) A run of length $i-1$ in S_{n-1} can be changed into a run of length i in S_n for $i = 1, 2, \dots, n-1$.
- 3) A run of length $h = 2i$ in S_{n-1} can be split into a run of length i , followed by a run of length 1, followed by a run of length i for $1 \leq i \leq \left\lfloor \frac{n-2}{2} \right\rfloor$

++++ → +-++

So, suppose this n minus 1 denote the number of arrangements with n minus 1 many elements of plus and minus then an additional run of length 1 can be added in the arrangement of S_n because of the introduction of the new element that we have already seen or a run of length i

minus 1 in S_n minus 1 can be changed into a run of length i in S_n that means that an existing run is now increased by length 1 or it can be split into two parts.

Here we have divided that also into two different segments. Suppose h is $2i$ that means it is an even number and the new element comes at the middle so it may split it into run of length i followed by a run of length 1 followed by a run of length i that means suppose we had something like this and the new element converts it into plus, plus, minus, plus, plus. So, this is the case that is taken care of in 0.3,.

(Refer Slide Time: 59:52)

4) A run of length $h = i + j$ in S_{n-1} , where $h > i > j, 3 \leq h \leq n-2$ can be split up into

a) A run of length i , followed by a run of length 1, followed by a run of length j .

b) A run of length j , followed by a run of length 1, followed by a run of length i .

The overall recurrence relation between u_{n-1} and u_n can be obtained from this formula:

Handwritten notes and diagrams illustrating the splitting of a run of length $h = i + j$ into two parts. The diagrams show runs of pluses and minuses, with arrows indicating the split. For example, a run of 4 pluses is split into a run of 3 pluses and a run of 1 plus, or a run of 1 plus and a run of 3 pluses. The overall recurrence relation is given as $u_n = u_{n-1} + u_{n-2}$.

But more generally if the length is h which is i plus j i and j are different numbers then it can split it into a run of length i followed by a run of length 1, followed by a run of length j that means the split is not equal for example we can have plus, plus, plus, plus divided into three parts plus followed by a minus followed by three plus. So here, i is equal to 1 j is equal to 3. On the other hand it can also split it like plus, plus, plus, plus it goes to plus, plus, plus, minus, plus.

Therefore, i is equal to 3 j is equal to 1. The overall recurrence relation between u_{n-1} and u_n can be obtained from the following formula.

(Refer Slide Time: 60:59)

$$\begin{aligned}
 & u_n(r_{n-1}, r_{n-2}, \dots, r_h, \dots, r_i, \dots, r_j, \dots, r_1) \\
 &= 2u_{n-1}(r_{n-2}, \dots, r_1 - 1) \\
 &+ \sum_{i=2}^{n-1} (r_{i-1} + 1) u_{n-1}(r_{n-2}, \dots, r_i - 1, r_{i-1} + 1, \dots, r_1) \\
 &+ \sum_{\substack{i=2 \\ (h=2i)}}^{\lfloor \frac{n-2}{2} \rfloor} (r_h + 1) u_{n-1}(r_{n-2}, \dots, r_h + 1, \dots, r_i - 2, \dots, r_1 - 1) \\
 &+ 2 \sum_{i=2}^{n-3} \sum_{j=1}^{i-1} (r_h + 1) u_{n-1}(r_{n-2}, \dots, r_h + 1, \dots, r_i - 1, \dots, r_j - 1, \dots, r_1 - 1)
 \end{aligned}$$

As you can understand this is pretty complicated we are not deriving it here neither I am asking you to check the correctness of this things. It is only for you to conceptualize how runs up and down works.

(Refer Slide Time: 61:19)

Levene and Wolfowitz have computed the means and variances of number of runs of length t (or more).

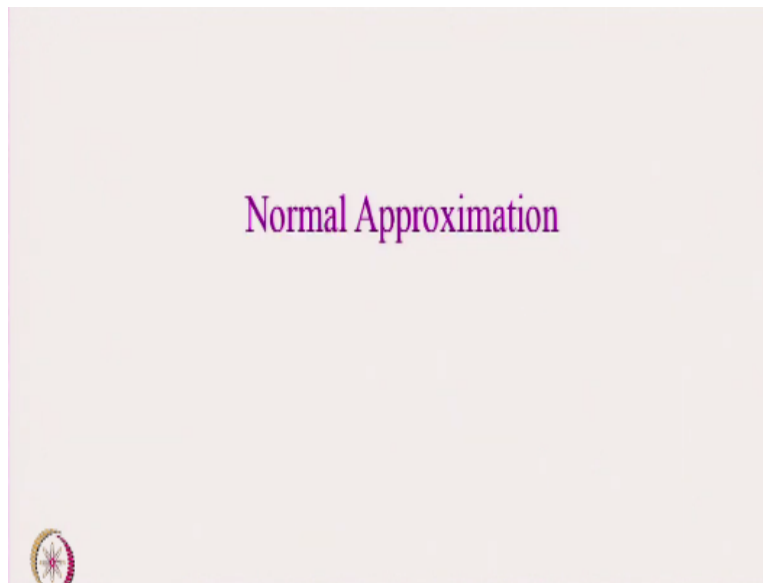
For $n \leq 14$ the exact probabilities of getting at least r runs of length t or more are also available.

From there critical regions can be constructed and H_0 can be tested against suitable alternatives.

The asymptotic distribution can be used for large n .

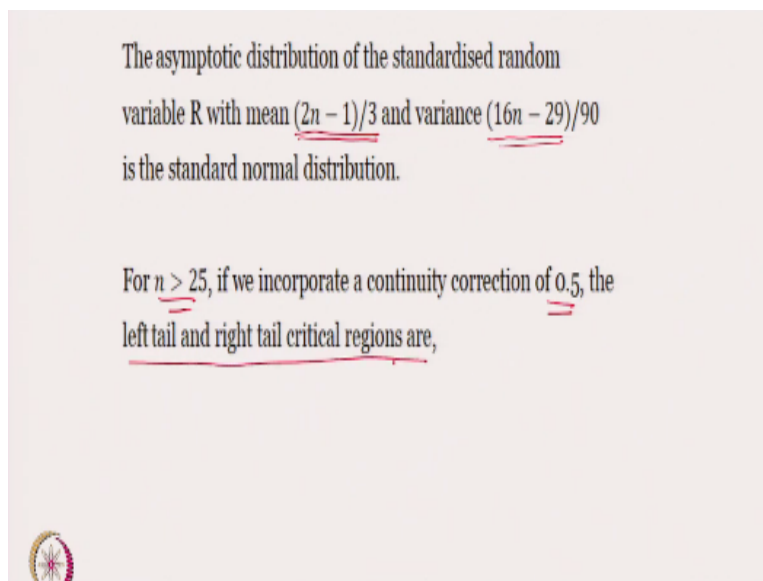
Now Levene and Wolfowitz have computed the means and variance of number of runs of length t or more. For n less than equal to 14 the exact probabilities of getting at least r runs of length t or more are also available and they are tabulated and from there the critical regions can be constructed and H_0 can be tested against suitable alternatives.

(Refer Slide Time: 61:54)



However, when the number of values is large we can go for normal approximation for the same and the calculation is the following.

(Refer Slide Time: 62:02)



The mean is coming out to be $2n$ minus 1 upon 3 and the variance is coming out to be $16n$ minus 29 upon 90. As you can understand that these are little bit complicated so I am not going to derive the formulae, but if n is greater than 25 and if we incorporate the continuity correction then we can get the left tail and right tail critical regions by using normal approximation as follows that.

(Refer Slide Time: 62:34)

$$\frac{\left(R + 0.5 - \frac{(2n-1)}{3}\right)}{\sqrt{\frac{16n-29}{90}}} \leq -z_{\alpha}$$

And,

$$\frac{\left(R - 0.5 - \frac{(2n-1)}{3}\right)}{\sqrt{\frac{16n-29}{90}}} \geq z_{\alpha}$$

The two tailed critical region is a combination of the above two with z_{α} replaced by $\frac{z_{\alpha}}{2}$

For the left tail it is R plus 0.5 minus the mean divided by the standard deviation has to be less than equal to minus z_{α} where Z_{α} is the critical value. For normal distribution similarly for right hand tail it is going to be $R - 0.5$ minus $2n - 1$ which is the mean divided by the standard deviation. So, this is for one sided and if you want to do it two sided then replace Z_{α} by $Z_{\alpha/2}$.

Okay friends I stop here today. I think you have understood the intuition behind using runs for testing whether a given data is random or not. I stop here and in the next class I shall study some more test based on two samples. In particular I will look at Kolmogorov Smirnov two sample test and median test for testing the equality of distribution. Okay friends. Thank you.