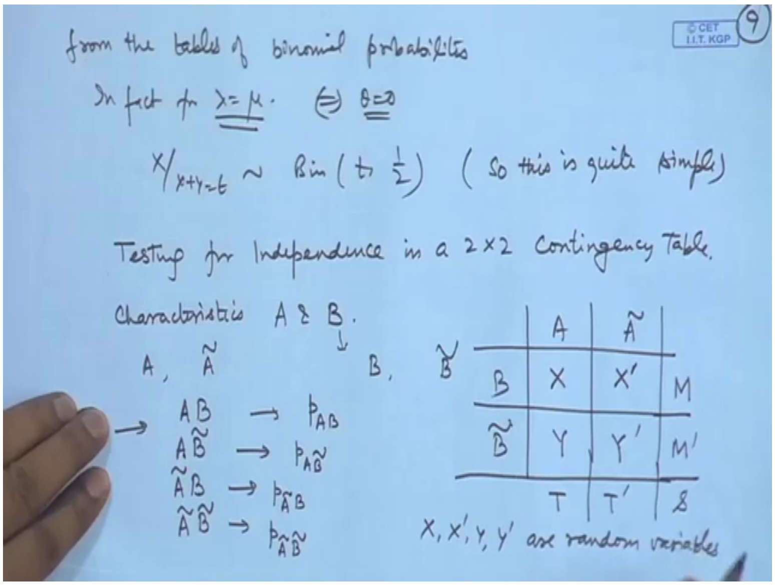
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 46
Applications of UMP Unbiased Tests - II

We will further show the applications to comparing the means etcetera for normal populations and variances also testing for the means and variances in normal populations. So, these are applications I will be giving. However, let me take a slightly different kind of example, we are all familiar with the contingency tables; contingency tables is for example, we take a random sample of a population.

We record the number of people according to two different characteristics. For example, we may record the people according to their say health status. So, we may define some health status and we may then say that people are having good health or bad health and another criteria then we may fix up as their level of education. So, once again, whether they are having good education or they are not having good education. And in a given population then we correlate the two things, we like to see whether there is a dependence on in among these two characteristics. For example, whether better education status leads to good health among the people or not, so, this is like the data in the contingency table.

(Refer Slide Time: 01:49)



from the tables of binomial probabilities

In fact for $\underline{\underline{\lambda = \mu}}$. $\Leftrightarrow$ $\underline{\underline{\theta = 0}}$

$X /_{X+Y=t} \sim Bin(t, \frac{1}{2})$ (So this is quite simple)

Testing for Independence in a 2×2 Contingency Table.

Characteristics A & B.

$A, \tilde{A}$ $\quad$ $B, \tilde{B}$

$AB \rightarrow p_{AB}$

$A\tilde{B} \rightarrow p_{A\tilde{B}}$

$\tilde{A}B \rightarrow p_{\tilde{A}B}$

$\tilde{A}\tilde{B} \rightarrow p_{\tilde{A}\tilde{B}}$

	A	$\tilde{A}$	
B	X	X'	M
$\tilde{B}$	Y	Y'	M'
	T	T'	S

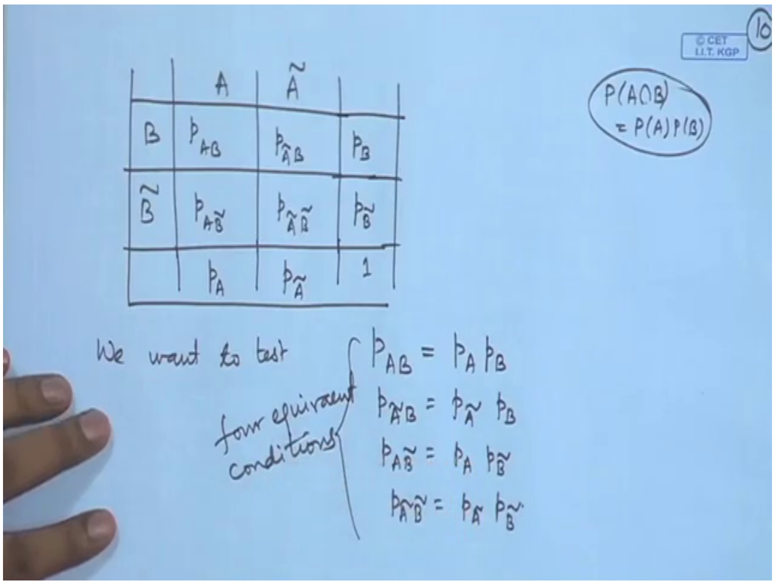
X, X', Y, Y' are random variables

So, let me write here: testing for independence in a 2 by 2 contingency table. So, we may have say characteristics A and B so; that means, a person may have characteristic A or he may not have. So, A and not A, we will write as a tilde. Similarly having characteristic B or not having characteristic B, we will denote by B tilde. So, now, when we take a data of say n number of persons then among n number of persons, there will be people we will have both the characteristics A and B.

They may have A, but not B they may not have A or have B or they may not have both, ok. And then the probabilities of this for example, what is the probability that a person will have both the characteristic A and characteristic B, this probability we are denoted by p A B. What is the probability that a person has characteristic A, but not B, then we denoted the probability by; what is the probability that the person will not have characteristic A, but he will have characteristic B then similarly A tilde B tilde. So, we represent the data in the following form as usual in the contingency table.

A A tilde B B tilde X is the number of people who are having both the characteristics say X prime, let me say the number of people who do not have A, but have B, Y is the number of people who do not have B, but have A and Y prime is the number of people who are having the characteristic, who do not have the characteristic A and B both. Now these will have certain totals here let me write here M M prime, here it will be say T T prime and the total is say S.

(Refer Slide Time: 04:19)



	A	$\tilde{A}$	
B	p_{AB}	$p_{\tilde{A}B}$	p_B
$\tilde{B}$	$p_{A\tilde{B}}$	$p_{\tilde{A}\tilde{B}}$	$p_{\tilde{B}}$
	p_A	$p_{\tilde{A}}$	1

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

We want to test

four equivalent conditions

$$p_{AB} = p_A p_B$$
$$p_{\tilde{A}B} = p_{\tilde{A}} p_B$$
$$p_{A\tilde{B}} = p_A p_{\tilde{B}}$$
$$p_{\tilde{A}\tilde{B}} = p_{\tilde{A}} p_{\tilde{B}}$$

© CET
I.I.T. KGP
10

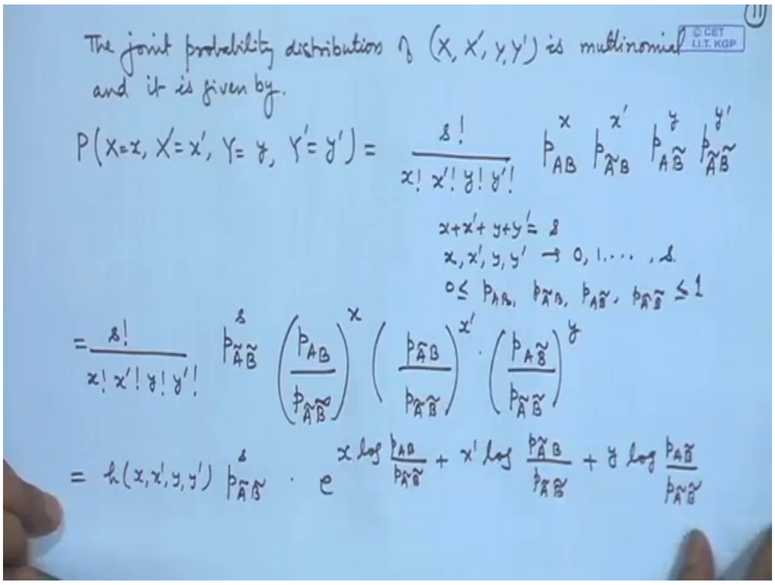
The corresponding probability distribution is described by A A tilde B and B tilde. So, p AB p A tilde B p A B tilde p A tilde B tilde. See, if we add this we will get p B that is the proportion of the people having characteristic B or the probability of people having characteristic B probability of the people having characteristic B tilde; that means, not having characteristic B. Similarly, p A p A tilde and this is one here. So, our interest is to check the independence.

So, in the independence, what we are going to check? We want to test whether the characteristics are independent. Now the characteristics independent, means see the usual condition of independence of two events. How do we write? Probability of A intersection B is equal to probability of A into probability of B. So, here this will translate to p A B is equal to p A into p B and equivalently, we may check these conditions also. So, they are all equivalent, one of them will be enough any of these four equivalent conditions.

So, any one of them that means we should write down the distribution of the observables in such a way that my parameter theta to be tested will yield me one of these conditions, ok. Now, what is the random variable here, we have taken a data on S persons or the populations. And the sample size is S now among S, we do not know how many will fall into each. So, basically X X prime Y, Y prime these are the random variables. Let me point out here, we may have different sampling schemes I have considered the sampling scheme where the total sample size is fixed here, but how many of observations of in each category that is random. So, this are the random we may have other situations.

For example, we may fix this and this and then this may be random, we may fix these, this may be random and so on. There are 4-5 different varieties for details one may again look at the book of Lehmann, they have discussed in detail each of these cases. So, here we write down the distribution of X, X prime Y and Y prime. So, what will be the distribution this will be a multinomial distribution.

(Refer Slide Time: 07:41)



11
© CET
I.I.T. KGP

The joint probability distribution of (X, X', Y, Y') is multinomial and it is given by.

$$P(X=x, X'=x', Y=y, Y'=y') = \frac{s!}{x!\,x'!\,y!\,y'!}\, p_{AB}^{x}\, p_{\tilde{A}B}^{x'}\, p_{A\tilde{B}}^{y}\, p_{\tilde{A}\tilde{B}}^{y'}$$

$x+x'+y+y'=s$

$x, x', y, y' \to 0, 1, \dots, s.$

$0 \le p_{AB}, p_{\tilde{A}B}, p_{A\tilde{B}}, p_{\tilde{A}\tilde{B}} \le 1$

$$= \frac{s!}{x!\,x'!\,y!\,y'!}\, p_{\tilde{A}\tilde{B}}^{s} \left(\frac{p_{AB}}{p_{\tilde{A}\tilde{B}}}\right)^{x} \left(\frac{p_{\tilde{A}B}}{p_{\tilde{A}\tilde{B}}}\right)^{x'} \cdot \left(\frac{p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}}\right)^{y}$$

$$= h(x, x', y, y')\, p_{\tilde{A}\tilde{B}}^{s} \cdot e^{x \log \frac{p_{AB}}{p_{\tilde{A}\tilde{B}}} + x' \log \frac{p_{\tilde{A}B}}{p_{\tilde{A}\tilde{B}}} + y \log \frac{p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}}}$$

So, the joint probability distribution of X X prime Y Y prime, this is multi nominal and it is given by So, we write it as probability of X is equal to small x X prime is equal to small x prime Y is equal to small y Y prime is equal to small y prime that will be X factorial divided by X factorial. X factorial prime X prime factorial Y factorial Y prime factorial then probability.

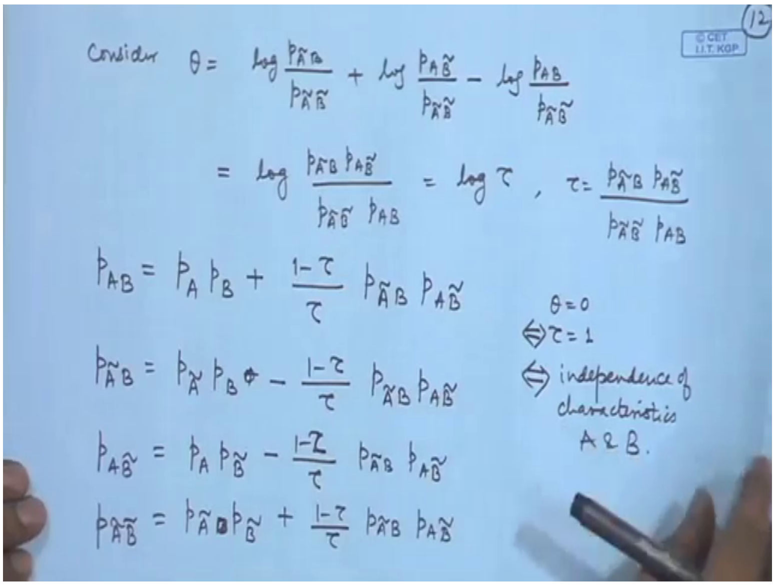
So, p A B to the power X p A tilde B to the power X prime p A B tilde is equal to Y A tilde B, tilde is to the power Y prime. And this is X plus X prime plus Y plus Y prime is equal to S where each of these X X prime Y, Y prime they take values from 0 1 to S. And each of these probabilities, probability of this is A B A tilde B A B tilde and A tilde B tilde, they lie between 0 and 1.

Now, we in order to consider the testing of this statement, we need to rewrite this and the form of a multi parameter exponential family. So, the first thing that we can do here we express it as S factorial divided by X factorial X prime factorial Y factorial Y prime factorial and p A B. Now, in order to bring in other terms I drew little bit of adjustment of the terms. Why consider multiplication and division by A term p A tilde B tilde to the power S here.

So, I will get here p A B by p A tilde B tilde to the power X p A tilde B divided by p A tilde B tilde to the power X prime p A B tilde divided by p A tilde B tilde to the power Y. Now these are the terms which involve the random observations. So, I write it as some H of XX prime Y Y prime. And then there is a term which involves the parameters and here I write E to the power X log of p A B divided by p A tilde B tilde plus X prime log off p A tilde B divided by p A tilde B tilde plus Y log of p A B tilde divided by p A tilde B tilde.

You can see that this is now in the form of A 3 parameter exponential family you have this as say theta this has nu 1 nu 2 u T 1 T 2. Naturally this does not suggest that how we can test about the independence here, because for independence I have written the condition. And how this condition will come, it is not clear here of course, what is clear here is that I may test about p A B being equal or not equal to p A tilde B tilde or I may test about p A tilde B being equal to this or not and so on. I may be able to do these things; however, we do further re adjustment of the terms here, let me show how that can be done.

(Refer Slide Time: 12:41)



(12)

© CET I.I.T. KGP

Consider $\theta = \log \frac{p_{\tilde{A}B}}{p_{\tilde{A}\tilde{B}}} + \log \frac{p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}} - \log \frac{p_{AB}}{p_{\tilde{A}\tilde{B}}}$

$= \log \frac{p_{\tilde{A}B}\, p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}\, p_{AB}} = \log \tau \,, \quad \tau = \frac{p_{\tilde{A}B}\, p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}\, p_{AB}}$

$p_{AB} = p_A p_B + \frac{1-\tau}{\tau} p_{\tilde{A}B}\, p_{A\tilde{B}}$

$p_{\tilde{A}B} = p_{\tilde{A}} p_B - \frac{1-\tau}{\tau} p_{\tilde{A}B}\, p_{A\tilde{B}}$

$p_{A\tilde{B}} = p_A p_{\tilde{B}} - \frac{1-\tau}{\tau} p_{\tilde{A}B}\, p_{A\tilde{B}}$

$p_{\tilde{A}\tilde{B}} = p_{\tilde{A}} p_{\tilde{B}} + \frac{1-\tau}{\tau} p_{\tilde{A}B}\, p_{A\tilde{B}}$

$\theta = 0$

$\Leftrightarrow \tau = 1$

$\Leftrightarrow$ independence of characteristics A & B.

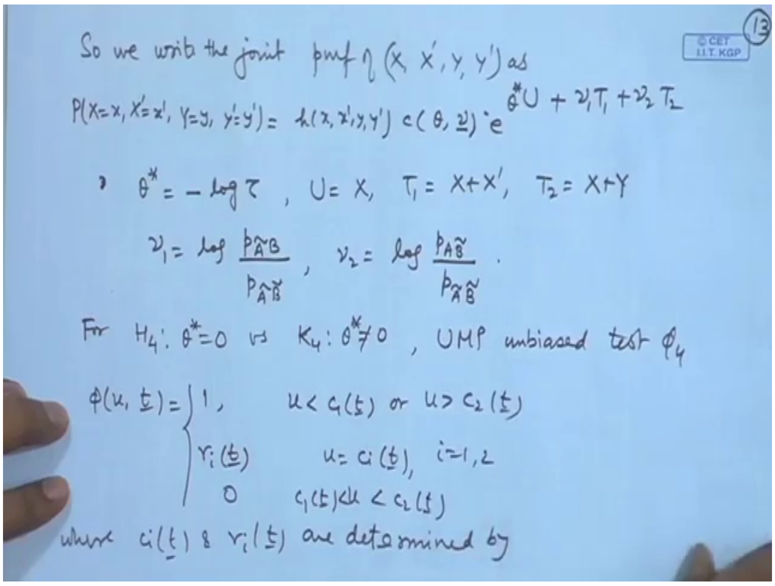
Consider say theta is equal to log of p A tilde B divided by p A tilde B tilde plus log of p A B tilde divided by p A tilde B tilde minus log of p A B divided by p A tilde B tilde that is equal to log of. So, these two will be multiplied so, I will get p A tilde B p A B tilde divided by p A tilde B tilde p A B.

Now, this looks like a parameter which will give me the correct information. Let me write it as say a log of some tau where tau is actually this p A tilde B p A B tilde divided by p A tilde B tilde p A B. In fact, I can write here p A B as p A p B plus 1 minus tau divided by tau p A tilde B p A B tilde and various alternative forms. Let me write it here p A tilde B can be similarly written as p A tilde p B plus sorry, minus 1 minus tau by tau p A tilde B p A B tilde p A B tilde is equal to p A p B tilde minus 1 minus tau divided by tau p A tilde B p A B tilde.

And lastly p A tilde B tilde is equal to p A tilde B tilde into p B tilde plus 1 minus tau by tau p A tilde B p A B tilde; that means, if I test tau is equal to 1 or theta is equal to 0, then this is equivalent to independence of characteristics X A and B, if I put tau is equal to 1 then this will come 0. And therefore, this will become p A B is equal to p A into p B.

So, if I can write in the form of multi parameter exponential family then testing for theta is equal to 0 theta is equal to 0 that is H 4 versus K 4 will test for the independence in this particular case. So, now our aim is to write this form into having this theta here. So, we do that thing here with the adjustment of the terms.

(Refer Slide Time: 16:27)



13

So we write the joint pmf of (X, X', Y, Y') as

$P(X=x, X'=x', Y=y, Y'=y') = h(x, x', y, y')\, c(\theta, \underline{\nu})\, e^{\theta^* U + \nu_1 T_1 + \nu_2 T_2}$

$\theta^* = -\log \tau$, $U = X$, $T_1 = X+X'$, $T_2 = X+Y$

$\nu_1 = \log \frac{p_{\tilde{A}B}}{p_{\tilde{A}\tilde{B}}}$, $\nu_2 = \log \frac{p_{A\tilde{B}}}{p_{\tilde{A}\tilde{B}}}$.

For $H_4: \theta^* = 0$ vs $K_4: \theta^* \neq 0$, UMP unbiased test ϕ_4

$$\phi(u, \underline{t}) = \begin{cases} 1, & u < c_1(\underline{t}) \text{ or } u > c_2(\underline{t}) \\ \gamma_i(\underline{t}) & u = c_i(\underline{t}),\ i=1,2 \\ 0 & c_1(\underline{t}) < u < c_2(\underline{t}) \end{cases}$$

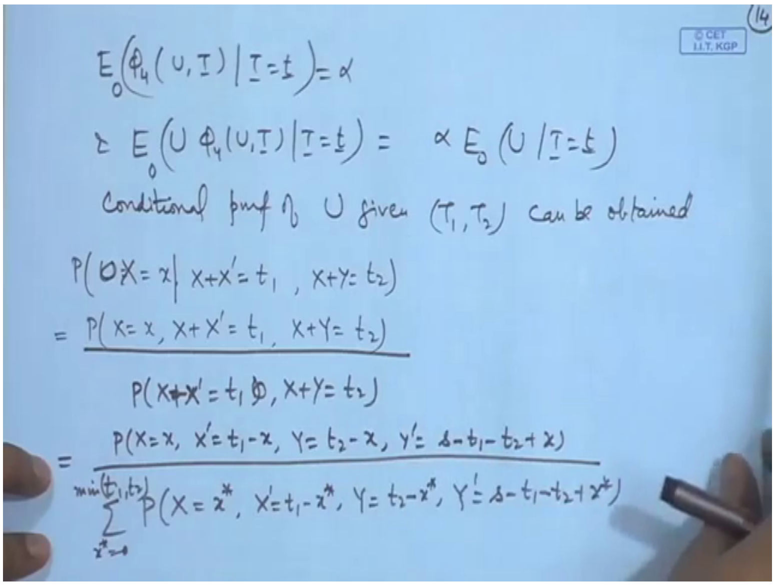
where $c_i(\underline{t})$ & $\gamma_i(\underline{t})$ are determined by

So, we write the joint probability mass function of X, X prime Y Y prime as a function of X X prime Y Y prime and a function of the parameters theta nu here I will get actually more than 1. And then some other terms will also come. So, then E to the power theta U x plus nu 1 T 1 plus nu 2 T 2 where theta I have define there actually theta is equal to.

Let me write it theta is star theta is star is equal to minus log of tau u is equal to X T 1 is equal to X plus X prime T 2 is equal to X plus Y nu 1 is equal to log of p A tilde B divided by p A tilde B tilde nu 2 is equal to log of p A B tilde divided by p A tilde B tilde. So, if we use the theorem 2 which I gave in the previous lecture for H 4 theta star is equal to 0 versus K 4 theta is star not equal to 0 UMP unbiased test phi 4 will be given in the form phi u T is equal to 1 for u that is X is less than C 1 of T R u is greater than C 2 of T, it is equal to gamma I of T.

If u is equal to C i of t for i is equal to 1 2, it is equal to 0. If u is between C 1 and C 2 where C i's and gamma i's are determined by expectation of phi 4 U T given t is equal to alpha at 0 and it is equal to alpha.

(Refer Slide Time: 19:17)



$$E_0\left(\phi_4(U, \underline{T}) \mid \underline{T} = \underline{t}\right) = \alpha$$

$$\& \; E_0\left(U \phi_4(U, \underline{T}) \mid \underline{T} = \underline{t}\right) = \alpha E_0\left(U \mid \underline{T} = \underline{t}\right)$$

Conditional pmf of U given (T_1, T_2) can be obtained

$$P(\cancel{U} X = x \mid X + X' = t_1, X + Y = t_2)$$

$$= \frac{P(X = x, X + X' = t_1, X + Y = t_2)}{P(X + X' = t_1, X + Y = t_2)}$$

$$= \frac{P(X = x, X' = t_1 - x, Y = t_2 - x, Y' = s - t_1 - t_2 + x)}{\sum_{x^* = 0}^{\min(t_1, t_2)} P(X = x^*, X' = t_1 - x^*, Y = t_2 - x^*, Y' = s - t_1 - t_2 + x^*)}$$

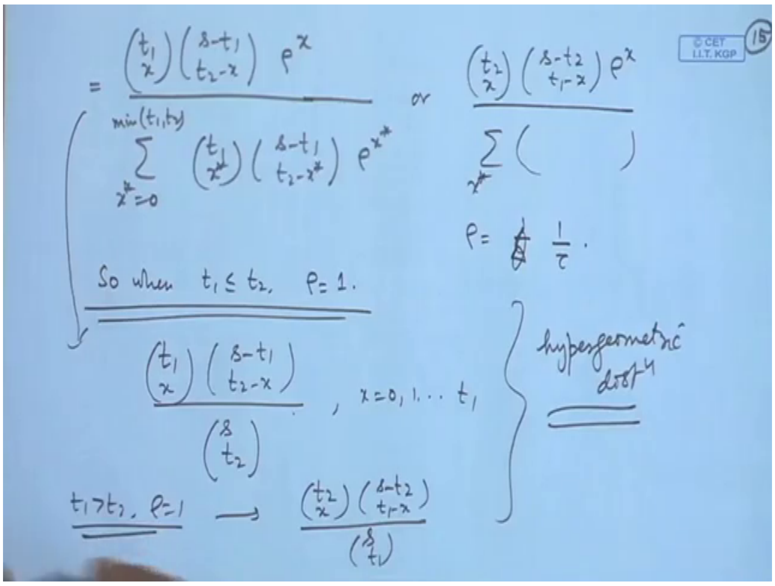
So, we need the conditional distribution of U given T. So, conditional distribution of U given T 1 T 2, this can be obtained. As we have done in the binomial and Poisson examples, let me just give a sketch of this one, see probability of U is equal to something.

So, U is actually X, let me write it in that form just for convenience T 1 is X plus X prime that is equal to t 1 and t 2 is X plus Y that is equal to small t 2. So, that is equal to probability of X is equal to x X plus X prime is equal to t 1 X plus Y is equal to t 2. So, this is divided by probability of X plus X prime is equal to t 1 and X plus Y is equal to t 2. Now as in the binomial and Poisson examples, we substitute here and then represent that thing in the denominator also.

So, this actually becomes probability of X is equal to x probability of X prime is equal t 1 minus x Y is equal to t 2 minus x. Once, X X prime and Y are fixed y prime is also fixed that is equal to s minus t 1 minus t 2 plus x. And in the denominator then we are simply getting probability of X is equal to say x star X prime is equal to t 1 minus x star, Y is equal to t 2 minus x star Y prime is equal to s minus t 1 minus t 2 plus x star for x star is equal to 0 to t. It will go up to minimum of t 1 and t 2.

Now, this joint probability can be written here and I will not write all the steps here. One can substitute as we have done in the binomial and Poisson examples.

(Refer Slide Time: 22:01)



$$= \frac{\binom{t_1}{x}\binom{s-t_1}{t_2-x}\rho^{x}}{\sum_{x^*=0}^{\min(t_1,t_2)}\binom{t_1}{x^*}\binom{s-t_1}{t_2-x^*}\rho^{x^*}} \quad \text{or} \quad \frac{\binom{t_2}{x}\binom{s-t_2}{t_1-x}\rho^{x}}{\sum_{x^*}(\quad)}$$

$$\rho = \cancel{\delta}\ \frac{1}{\tau}.$$

So when $t_1 \le t_2$, $\rho = 1$.

$$\frac{\binom{t_1}{x}\binom{s-t_1}{t_2-x}}{\binom{s}{t_2}}, \quad x = 0, 1, \ldots t_1$$

$t_1 > t_2$, $\rho = 1$ $\longrightarrow$ $\frac{\binom{t_2}{x}\binom{s-t_2}{t_1-x}}{\binom{s}{t_1}}$

hypergeometric distn

© CET
I.I.T. KGP
15

After simplification one can write this is equal to t 1 c x s minus t 1 c t 2 minus x and some rho to the power x divided by sigma x star is equal to 0 to minimum of t 1 t 2 t 1 c x s minus t 1 c t 2 minus x star rho to the power x star. Here I could have written it in a slightly alternative form also I could have written it as t 2 c x s minus t 2 c t 1 minus x rho to the power x divided by sigma again same thing.

So, either of the forms could have been written here rho is equal to 1 by delta actually the term that I defined here as tau. So, this is equal to 1 by tau. So, when we are taking say t 1 is less than or equal to t 2 and say rho is equal to 1 that is at the boundary condition that is when theta is star is equal to 0, we are getting rho is equal to 1. Then this term becomes 1, then this term become simply a hyper geometric sum.

So, this term will then become t 1 c x s minus t 1 c t 2 minus x divided by s c t 2. For x is equal to 0 1 2 t 1 or when t 1 is greater than t 2 and rho is equal to 1. Then this can be written as t 2 c 6 s minus t 2 c t 1 minus x divided by s c t 1. So, these are hyper geometric distributions. So, we can calculate the constants which are there in the test function that is c I s and gamma I s these can be determined from the tables of the hyper geometric distribution.

So, today I have discussed in detail the application of UMP unbiased test theory to deriving the tests for various comparison like in the binomial proportions comparison in the arrival rates of the poison distribution testing for the independence. In the contingency table as I already mentioned here the testing can be done in slightly different way. If the sampling scheme is different here I fix the total sample size as s. If I fix the sites like the total category A or total category B etcetera; then the test will get will slightly modified for details on my look at the book of Lehmann and Romano from where this theory we are following.

In the next lecture, I will be considering applications to the testing for parameters of normal distributions. What we have done earlier that we considered only one parameter. Usually I was considering like normal mu 1; that means, I was considering the variance to be known or normal 0 sigma x square; that means, I was taking the mean to be known. And in that case testing for the other parameter was there and from one parameter exponential family the UMP and UMP unbiased tests exist. Now we will show that even for the two parameters cases, we can derive the UMP unbiased tests for all of those problems. So, I will be covering it in the next lecture.