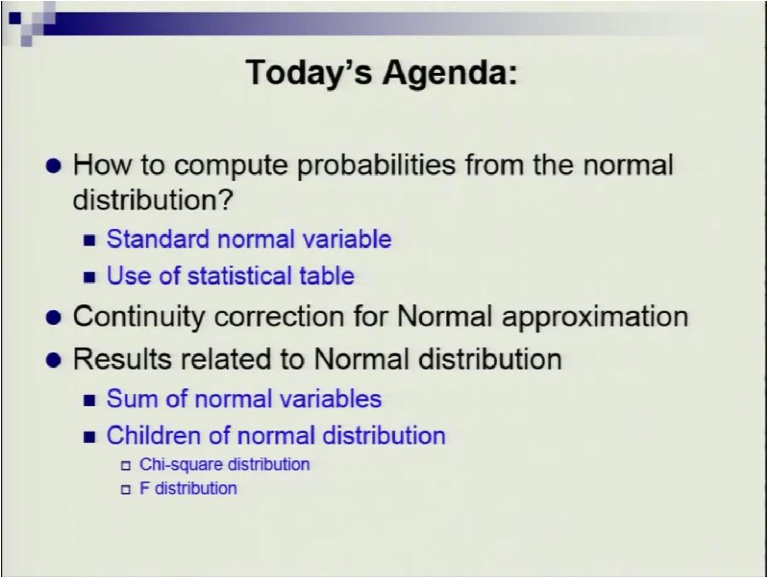


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 7
Normal Distribution

Hello, friends. Welcome back to the lecture series on applied statistics and econometrics. So in last lecture we have just introduced the concept of normal variable and the associated probability distribution and then we ended the lecture with a case of normal approximation to binomial and Poisson these are the discrete random variable distributions. So, before we further proceed, let us have a look at the clear cut agenda items for today's lecture.

(Refer Slide Time: 00:47)



Today's Agenda:

- How to compute probabilities from the normal distribution?
 - Standard normal variable
 - Use of statistical table
- Continuity correction for Normal approximation
- Results related to Normal distribution
 - Sum of normal variables
 - Children of normal distribution
 - Chi-square distribution
 - F distribution

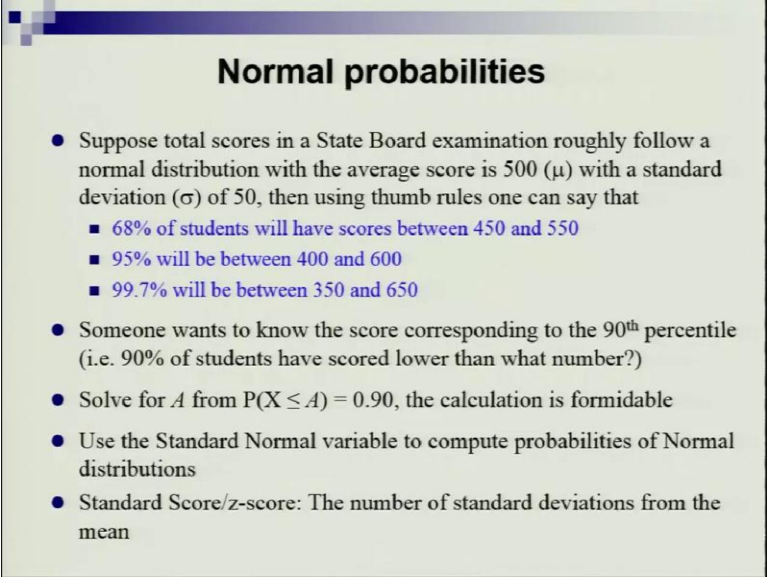
So we are going to start our discussion by answering a particular question. How to compute probabilities from the normal distribution to solve problems which involve normal distributions? If you remember that the PDF and the CDF of normal distributions have complicated integral and they are very difficult to solve and find particular probability numbers for our problems.

So to make our life easy statisticians have developed a concept called standard normal variable and they have made statistical tables available, so that we can refer to, to figure out a particular probability value that we are looking for. So this has made our life very easy and this is a very practical lesson that we must learn to proceed further.

Now, in the next topic of today's lecture, we are going to discuss the continuity correction for the normal approximation. We have already told you that there is a normal approximation available for the binomial and Poisson. But note that, binomial and Poisson are the discrete distributions. So, when you approximate discrete histograms or discrete bars by continuous smooth curve, then there has to be some continuity problem and how to correct for that, that we are going to study the next.

And then we are going to continue our discussion with normal distribution and we are going to highlight some results, I am not going to get into the details or statistical proof of what I am going to show here. But, we are going to discuss sum of normal variables. And then, we are going to finally end today's discussion with some idea about children of normal distribution and we are going to talk about 2 very useful distributions for applied statistical and econometric research, namely chi-squared distribution and F distribution.

(Refer Slide Time: 02:55)



Normal probabilities

- Suppose total scores in a State Board examination roughly follow a normal distribution with the average score is 500 (μ) with a standard deviation (σ) of 50, then using thumb rules one can say that
 - 68% of students will have scores between 450 and 550
 - 95% will be between 400 and 600
 - 99.7% will be between 350 and 650
- Someone wants to know the score corresponding to the 90th percentile (i.e. 90% of students have scored lower than what number?)
- Solve for A from $P(X \leq A) = 0.90$, the calculation is formidable
- Use the Standard Normal variable to compute probabilities of Normal distributions
- Standard Score/z-score: The number of standard deviations from the mean

So, let us assume that a student has got scores obtained by other students in a State Board Examination, and the total scores roughly follow a normal distribution with the average score of say 500 that is a proxy for the population mean. And there is a standard deviation sigma of 50. Now, if you remember, we have ended the last lecture on normal distribution by talking about some thumb rules about the probability or the portion of the data that lies between some intervals

like $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$. So we are going to see an application of that.

So here, if we apply that $\mu \pm \sigma$ rule, then we can say that 68 percent of students will have scores between 450 and 550. Note that the number 450 is derived by deducting σ from the μ value. So $500 - 50 = 450$. Similarly, we get the 550 number by adding σ to the μ value which is $500 + 50$.

Now, similarly, we can say that 95 percent of data will be between the scores 400 and 600; and 99.7 percent of the data or scores will be between 350 and 650. So, if someone wants to know that score corresponding to the 90 percentile, then how do you proceed about that? So that means that 90 percent of students have scored lower than a particular number. So, of course, I have shown you how to find out the percentile values or the quartile values in the last lecture. So, following that trick, you have to solve for an unknown, say capital A from the equation $\text{probability of } X \text{ less than or equal to } A \text{ equals to } 0.9$.

But the calculation looks formidable, is not it? Because it involves integrals, so how to get the value A? So, that is why the statisticians have come forward with a concept called standard normal variable, so that we can compute probabilities of normal random variables easily. And when we talk about standard normal variable, we actually talk about the standard score or it is also known as z-score. So that is basically the number of standard deviations from the mean of the random variable at hand.

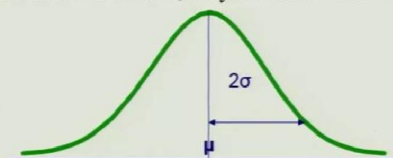
(Refer Slide Time: 05:44)

Standard Normal

- ♣ To standardize a Normal variable, subtract its mean from each observation and then divide the difference by the std. deviation and get

$$Z = \frac{X - \mu}{\sigma} \xrightarrow{\text{PDF}} f(Z) = \frac{1}{\sqrt{2\pi}} e^{-1/2 Z^2}$$

- ♣ The standard normal or Z distribution has mean 0 and variance 1, so it is denoted by $Z \sim N(0,1)$
- ♣ Most of the area under a standard normal curve lies between -3 and $+3$
- ♣ The distribution is the same, only the scale has changed



Original test scores	500	600	X	$(\mu = 500, \sigma = 50)$
Standardized test scores	0	2	Z	$(\mu = 0, \sigma = 1)$

Now here, in this slide, we are going to talk about the standard normal variable. First, we need to standardize a normal variable. What do we mean by standardization? We have to subtract its mean from each observation and then divide the difference by the standard deviation. So we define variable, mu variable capital Z equal to capital X minus mu divided by sigma and that is basically our standard normal variable.

So, one particular realization of that variable capital Z is small z, and that is basically the standard score or z-score. Now from this standard normal variable z, one can get the PDF, and the PDF is written as f of capital Z equals to 1 over root of 2pi times e to the power minus half Z square. so that is a little bit simple compared to the normal case. Actually, this formula of PDF is derived from the normal PDF only. And note that, in some textbooks, you will find that they will represent this PDF of standard normal as small phi, the Greek letter.

Now, the standard normal or the Z distribution has a mean of 0 and variance of 1. So it is basically denoted as Z follows a normal distribution with mean 0 and variance 1. So we can apply the same cherry shaped type result on the standard normal distribution as well. And if we applies, then we get this interesting result that most of the area under the standard normal curve lies between minus 3 and plus 3. So, you can say that more than 99.7 percent of data of standard normal variable lies between the values minus 3 and plus 3.

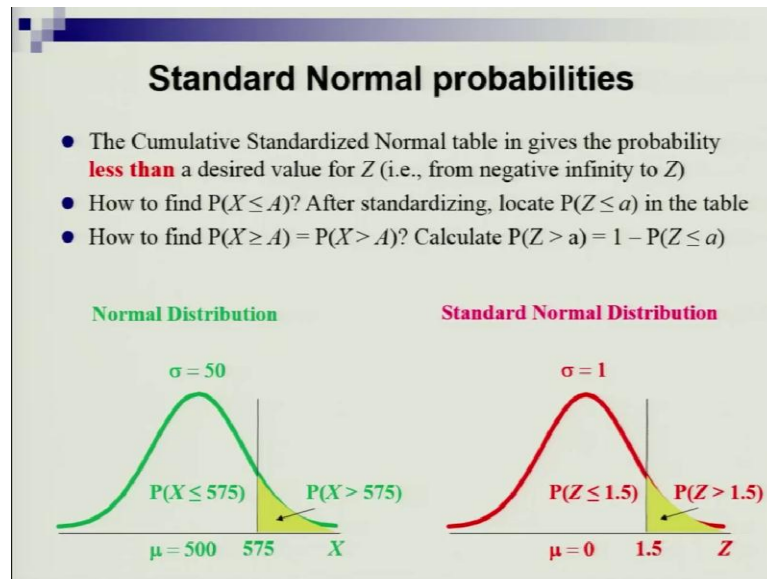
So, you look at the distribution now, there is no change actually, so the distribution is almost same if I want to comment on the shape of the distribution. And our benchmark is of course, the continuous random variable, which is a normal bell curve. Now, you see, only the scale has changed here. So if you look at the normal curve, in this slide, you see here, I am showing you the original test scores and standardized test scores along the x axis and z axis.

So here, first I will describe what is going to happen for the original test score. So we have some data, we plot the relative frequency distribution and then we fit a smooth normal curve and that is what we are going to consider first. So, for that data set, we have the mean of 500, so that is basically my μ . And then, you see σ is of course 50, that is given to us. Now, if we apply this concept of standard normal variable, we need to actually subtract 500 from the original test scores and then divide these difference by the number 50.

So, of course, the numbers will be now both positive and negative. Remember, in the case of original test score, you are not supposed to see a negative number. But here, interestingly, as we are taking difference, the numbers or the variable values of z or small z, they can take negative values. So here, the mean is 0. So, if you look at the standardized test scores, then the mean is 0. And then, what will happen to the variance? As I said that the standard normal variable Z takes variance 1, so variance of this distribution will be 1.

Now, here in this diagram, I am showing you if I get a score say 600, then what is going to be the location of that number? Now, note that if I standardize that particular number 600, then basically, I get a standardized test score or z score as 2 and so that is also marked here along the z axis and you see that actual location or point 2 on the z axis or the actual location of the score 600 along the x axis actually, 2 sigma away from the mean, either in the case of original test score it is 500 or in the case of the standard normal variable case it is 0. But the difference between the mean in both cases and these actual observation is always 2 sigma.

(Refer Slide Time: 10:31)



So, now we are going to study how the standard normal probabilities can be computed. So here, we have seen how we can write the PDF or small phi of the normal variable. Now, we are going to look at the cumulative of that standardized normal variable or the CDF from the previous lecture. And that is written with capital phi in many textbooks. So we actually have to now convert the CDF of the standard normal variable Z .

It gives the probability less than a desired value for the random variables Z , and this Z actually now ranges from negative infinity to a particular value, the desired value that we are talking about. So how to find a particular probability figure, say probability that x takes less than or equal to value capital A . A is an arbitrary number. So we have to first standardize the variable and not only the variable, but also the arbitrary score, the chosen score. And then, we actually rewrite this probability expression as probability of Z less than or equal to small a .

So small a , how it is derived? It is basically derived as the difference A minus μ is to be divided by the standard deviation σ and that is how you get these value small a from capital A . So in that case, once this expression is final, you go to the table, and then figure out the probability. How? That we are going to show next. But also, can we find the probability that X is greater than equal to A ? Yes, that is also possible. That is equivalent of asking the question probability X is strictly greater than A , because the probability for the continuous random variable taking a particular value is 0. So we can write these expressions interchangeably.

And then, we can calculate probability of Z greater than the small value a and that is equal to 1 minus probability of Z less than or equal to small a . So, that is basically from the probability theory. And then again, you go back to the table, the CDF standard normal table and you find the value for probability of Z less than or equal to small a and you are done. You can calculate the number. Now, I am going to show you a comparative picture, how the probabilities are to be calculated when we are dealing with normal distribution or we are dealing with a normal random variable.

So, let us first look at the normal distribution figure that is at the left hand side. And here we are showing the normal distribution with $\mu = 500$ and $\sigma = 50$. And the normal curve is basically a bell shaped curve with green color. And suppose someone is telling me that, well, I am interested to know what is the probability that a student has scored less than or equal to 575? So here, now we are replacing this arbitrary number capital A with 575. And then, note that by the theory of probability, we can say that the probability is basically the area under the normal curve kept at the vertical line, which is drawn at the value 575.

So, the mass or the area under the normal bell shaped curve to the right of this curve is basically the probability that x takes value less than or equal to 575. So, how to get the probability of x less than, equal to 575 from the diagram? Can we mark it in that diagram? Yes, of course. So, the area below the normal curve, which is capped at the value X equal to 575, by drawing the vertical line on that particular number that basically gives me the area. So how to look at the area? So, the area here is basically the white shaded area, it is white color mask that is basically at the right hand side of the curve.

So, the residual part which is marked yellow, and that is basically at the left of these vertical line. So, that basically gives me the probability that takes value greater than 575. So now how to translate this entire story told in terms of the normal distribution to the standard normal distribution, because ultimately we are going to make use of the standard normal distribution to compute the probability to avoid to take integrals.

So, first we have to standardize the variable X , so we define the variable Z ; and now the Z is a standard normal variable with mean μ equal to 0. So, you see here in the diagram, at the left hand side of the slide, you see that there is a red bell shaped curve and that gives you the

standard normal curve. Here, sigma is equal to 1, of course. Now you see that white area, which is the probability of X less than or equal to 575 has been transferred to a new probability expression and that is probability Z taking value less than or equal to 1.5. How did I get this number 1.5?

So, I have to basically take the number 575, then I have to subtract 500 from that number, and then, I have to divide that by 50, the standard deviation. And that is the way I get the number 1.5. And then, similarly, probability of X greater than 575. So, a probability that, so, the probability that a student has scored above 575, that area is translated in the standard normal distribution picture with another area marked by yellow color. And here, this area is indicated as probability that Z value is higher than 1.5. Now, we are going to move to the case of standard normal table.

(Refer Slide Time: 17:52)

Standard Normal Table										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7883	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

The ROW shows the value of Z to the first decimal point

The COLUMN gives the value of Z to the second decimal point

What is the area to the left of 2 in a standard normal curve?

Suppose $X \sim N(500, 2500)$. Find $P(X < 575)$.

Ans: Get $z = 1.50$.
Find $\Phi(1.50) = 0.9332$

Suppose $X \sim N(61, 81)$. Find $P(X > 75)$.

Ans: Get $z = 1.56$.
Find $\Phi(1.56) = 0.9406$.
Find $P(z > 1.56) = 1 - 0.9406 = 0.0594$

So, standard normal tables are very important. And they are very handy as they can actually help us to compute normal probabilities. Now, standard normal tables I am showing you here in this slide. But, of course, you can find that online, there are various good sources, authentic sources from where you can download a standard normal table. And also, I will show you later that the standard normal probabilities can be computed through software like R and Excel as well.

So, for the time being, let us look at excerpt of a standard normal table that came from textbook, so it is in print. So here, you see, it is kind of a matrix design. And the row shows the value of

the standard normal variable Z to the first decimal point. So you see the Z values are given along different rows; 0.0, 0.1, 0.2, 0.3, etc. And in this matrix set up the column gives the value of Z to the second decimal point.

So you see the headings of the columns. So it is 0.00, 0.01, 0.02, so suppose I am interested to deal with Z value of 0.47. So what to do? I have to first come down to the row, which shows me 0.4 and then, I have to move towards the end of this table. And I will stop as soon as I hit the column, which is headed with this number 0.07. So that is the way to find the probability from the table. So the number that you see in the cell, say in this case, 0.47 Z value means that it is 0.6808 that is basically the probability number that you are looking for.

So now, with this introduction to standard normal table, let us look at 2 problems and 2 solutions. So, now, let us look at 2 numerical problems and the corresponding solutions. So, I will start with the test score case, because we have been discussing that case for a while. So here, if you remember that original test score X follows a normal distribution with mean 500 and standard deviation 50. So, the variance will be 2500. So, it is basically normal 500 comma 2500 distribution and we are asked to find out the probability that x is less than 575.

So, first you get the Z score, and that is 1.5, I have already explained how to get that Z score. Then you come to the standard normal table and get the capital Φ value. So, how to get the capital Φ value? So, the Z score is 1.50. So, you have to now come down to the row which is showing you the number 1.5. And then you have to move to the column which shows you the title or heading 0.00. So, the number in the cell that you see here from the table is 0.9332. So, that is basically the probability that I am looking for.

Now, let me move on to another example. So, that you understand it very clear. Now, here I will assume that my X follows a normal 61, 81 distribution. So, the standard deviation σ is square root of 81, which is 9. And then, I am asked to find out the probability of X greater than 75. So, again, we will first get the Z value and that is 1.56 here. How do I get it? So, I have to subtract 61 from 75 and then, I have to divide it by 9 and this is the way I get the Z score.

Now, I am interested to find the corresponding cumulative probability, which is given by the mathematical expression capital Φ of 1.56 from the standard normal table. So, what to do, I

have to come down further row wise so that I find the number 1.5 and then, I have to now move along the columns and I need to stop where I see the column has a heading of 0.06. So, if I come to that cell, I spot a number given and that is 0.9406 that is basically the probability that I am looking for. And this is not the end of the story, because note that, here I am asked to find out the probability that X is higher than some number.

So, I have to now subtract this particular phi value or the cumulative probability value from 1 and I get the number 0.0594 that is my probability that X will be higher than 75 in this case. So, now, we are going to talk about the case of normal approximation to the binomial and Poisson type discrete distributions. So, we are going to first show you a diagram to explain the concept and then we are going to look at an illustration so that this concept of continuity correction becomes clear to you through an example.

(Refer Slide Time: 23:40)

Normal Approximation

- When using the *continuous* normal distribution to approximate a *discrete* Binomial distribution, move 0.5 unit to the left and right of the midpoint to include all possible x -values in the interval.

$P(x = c)$

$c - 0.5 \quad c \quad c + 0.5$

- A survey report tells that 31% of the class XII students in certain rural districts plan to attend college. If 50 Class XII students are randomly chosen from high schools in a rural district, find the probability that less than 14 students plan to attend college.
- Variable X is approximately normally distributed with mean $np = 15.5$ and std. dev. 3.27. Find $P(X < 13.5) = P(Z < -0.61) = 1 - P(Z < 0.61)$

Standard Normal Table

The COLUMN gives the value of Z to the second decimal point

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7883	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Suppose $X \sim N(500, 2500)$. Find $P(X < 575)$.
 Ans: Get $z = 1.50$.
 Find $\Phi(1.50) = 0.9332$

Ans: Get $z = 1.56$.
 Find $\Phi(1.56) = 0.9406$.
 Find $P(z > 1.56) = 1 - 0.9406 = 0.0594$

So, when we are using continuous normal distribution as an approximation to a discrete binomial distribution, we have to move 0.5 unit to the left and right of the midpoint to include all possible x values in the interval. So, that is basically the statement of the continuity correction, let us look at the diagram below and then probably this idea of continuity correction, will get clear to you.

So, note that as usual along the x axis, I am measuring some values and here let us start with discrete numbers because we are basically saying that x is a discrete random variable. So, for different values of x I can generate different bars or columns and they will give me a histogram. So, note that here I am showing you five bars, but there could be 7, there could be 10 any number of bars possible dependent on the values that x take.

So, basically I want to pass that smooth bell shaped curve through these histograms such that, this histogram is kind of enveloped by the smooth curve. So, of course, here you can see that with only 5 bars there are large gaps between the continuous normal curve and histogram height. But I can tell you that as the number of bars increase, so, then basically this gap between the normal curve and the histogram bars are going to be smaller and smaller. So, in that case continuity will be much more plausible. But, this diagram is just for an illustration, just to show you the entire idea of continuity correction.

Let us now focus on the very fifth bar and you see that the class mark for this particular class is c and suppose, we are interested to find probability regarding this value x, random variable taking

value c . Now, c is basically a realization of a discrete random variable x . Now, if we want to apply the normal approximation, the variable must be continuous and must be taking more values in an interval.

So, to create that interval artificially, what we are doing? So, what are we going to do to do the continuity correction, we are going to first subtract the number 0.5 from that class mark value c and then we are going to add 0.5 to the class mark value c . And that basically creates an interval of length 1 around the class mark c . And now, we assume our continuous random variable can take any value in that interval of width 1.

So, let us now look at one example, with the hope that this example will make this idea of continuity correction much clearer to you. So, let us assume that there is a survey report, which tells us that 31 percent of the class 12 will students in certain rural districts of North India plan to attend college. So, if we now choose 50 class 12 students from 2 or 3 high schools in a rural district, then 1 may be interested to find out the probability that less than 14 students plan to attend college, how to go about the problem?

First, let us see whether we can at all apply normal approximation to this problem or not. So, here X is the discrete random variable. Now, that discrete random variable follows a binomial distribution with parameters in n and p . So if you now multiply the values of n and p , then you will get the mean which is 15.5. And then you can use this n , p and multiply that number with 1 minus p to get the variance and take a square root of that to finally get the standard deviation, which is 3.27.

Now, the question that we have is basically, the, we have to figure out what is the probability of X taking value less than 14? But note that 14 is basically coming from that discrete random variable. So, we have to make a continuity correction before we apply normal. So what we do? We basically subtract number 0.5 from that value 14, so we get 13.5. So, now, the revised problem is to find out the probability of X less than 13.5. So, of course, then we know what we have to do.

We have to basically transform X and that number into standardized scores, so, we get the standard normal variable Z , which is taking value less than minus 0.61. Now, note that, all the

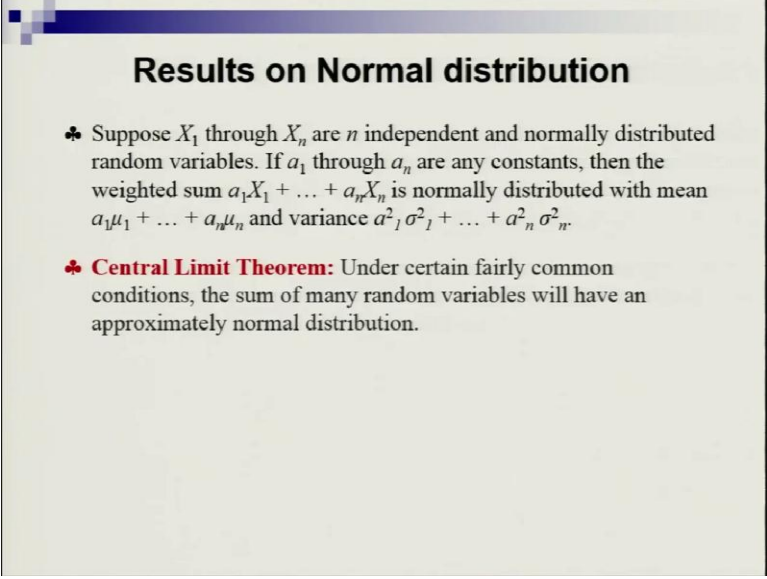
examples we have done before, we never encountered a negative value of Z . And if you remember, the table that I have shown you for the normal distribution that also does not have negative numbers? So how do you find an associated probability that is related to this negative number of Z ?

So, remember the trick, if you have this negative number for Z , then basically, you have to look at the probability value of P of Z less than the absolute value that you were given there. And then, once you find out that probability figure, you have to subtract that number from 1. That is the way actually you get the value of negative Z score. So, if I take you to the probability table, then probably these things will be clear to you. So remember that here the Z value that we need to refer to is 0.61. Keep that in mind.

So here is our famous standard normal table, so we have to figure out the probability for Z score of 0.61. So first, we have to come to row which has 0.6. And then we have to move across the columns. And we have to stop when we figure out the column with titles 0.01. And note that the probability number, it is showing here, it is 0.7291.

So you have to take this number and then, you have to subtract this number from 1. So probably you will get a score or probability value for our problem. So, if you take that probability number from the standard normal table and apply here, then the probability of finding X value less than 14 will be at around 0.27 something. So, we will now move on to the next topic.

(Refer Slide Time: 31:03)



Results on Normal distribution

- ♣ Suppose X_1 through X_n are n independent and normally distributed random variables. If a_1 through a_n are any constants, then the weighted sum $a_1X_1 + \dots + a_nX_n$ is normally distributed with mean $a_1\mu_1 + \dots + a_n\mu_n$ and variance $a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2$.
- ♣ **Central Limit Theorem:** Under certain fairly common conditions, the sum of many random variables will have an approximately normal distribution.

Now, I will focus on two very interesting results related with normal distribution and they are very handy in applied socio-economic research. So, suppose I am dealing with n independent normal random variables x_1 to x_n . And suppose, I also have some constants a_1 to a_n and then, I am interested at the weighted sum a new variable say y that is equal to a_1 times x_1 plus dot, dot, dot a_n times x_n .

So, how will this weighted sum variable behave? How will my y behave? So, the newly created variable y will also be normally distributed with mean a_1 times μ_1 plus dot, dot, dot a_n times μ_n , where the μ_1 and μ_n 's are nothing, but the means of the normal random variables x_1 to x_n respectively. And the variance will become a_1 square times σ_1 square plus dot, dot, dot a_n square times σ_n square.

But note that, here I will make an assumption that here my x_1 to x_n variables are independent in nature, if they are not independent, if they are somehow related to each other, then I may have problem. Actually, we will not have a problem, we will just have to add extra terms to these variance, but that we will discuss later, but not now. So, please be patient, we will discuss the case of independence and covariance, etc. in the later lectures.

Now, we are going to talk very briefly about a very interesting concept called central limit theorem, which plays a very big role in statistics and econometrics when it comes to estimation

and hypothesis testing. So, what is central limit theorem? So, statisticians say that under certain fairly common conditions, the sum of many random variables will have an approximately normal distribution.

Note that, here we are not saying that these random variables that you are adding, they have to follow normal distributions also. So, it is a pretty general result, but it holds for large sample and we are going to come back to central limit theorem again, when we will discuss the case of sampling and sampling distributions. Now, we are going to look at two children of the normal distribution and these two children distributions are very useful in advanced statistical studies, especially in hypothesis testing.

(Refer Slide Time: 33:49)

Children of Normal distribution

- ♣ **Chi Square distribution** (also chi-squared or χ^2 -distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables
$$\chi_{(2)}^2 = \frac{(X_1 - \mu)^2}{\sigma^2} + \frac{(X_2 - \mu)^2}{\sigma^2} = z_1^2 + z_2^2$$
- ♣ **F distribution** is the ratio of two chi-square distributions, where each chi-square has first been divided by its degrees of freedom
$$F = \frac{\chi_{(v_1)}^2 / v_1}{\chi_{(v_2)}^2 / v_2}$$
- ♣ **Degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary

35:46 44:02 35:46/44:02

So, first, we will talk about chi-square distribution, which is denoted by this Greek symbol chi-square also or sometimes it is also called chi-squared distribution, anyway. So, this is a special distribution with k degrees of freedom. And that is basically defined as the sum of the squares of k independent standard normal random variables. So, suppose, initially we are given two random variables and they are independent to each other. So, they are denoted by x_1 and x_2 and suppose they have common mean, μ and the common variance σ^2 . And now, how to define the chi-squared distribution?

So, you see that, you have to first and get the standard normal variables z_1 and z_2 and then you need to square them, if you sum them, you get a chi-square distribution with 2 degrees of freedom. Now, what is a degree of freedom? That we will talk in this lecture only after maybe a minute or so, so, please keep patience. So, the next distribution that is in line is F distribution. And F distribution is the ratio of two chi-square distributions, where each chi-square has first been divided by its degrees of freedom.

So, if I want to show you mathematically, you see that suppose I have 2 chi-square distributions, 1 with v_1 degrees of freedom and the other 1 is with v_2 degrees of freedom. So, in that case, what will happen? I need to divide the first chi-square variable by the degrees of freedom v_1 and then, I have to divide the second chi-square variable with respect to its degrees of freedom v_2 and then, I have to take the ratio.

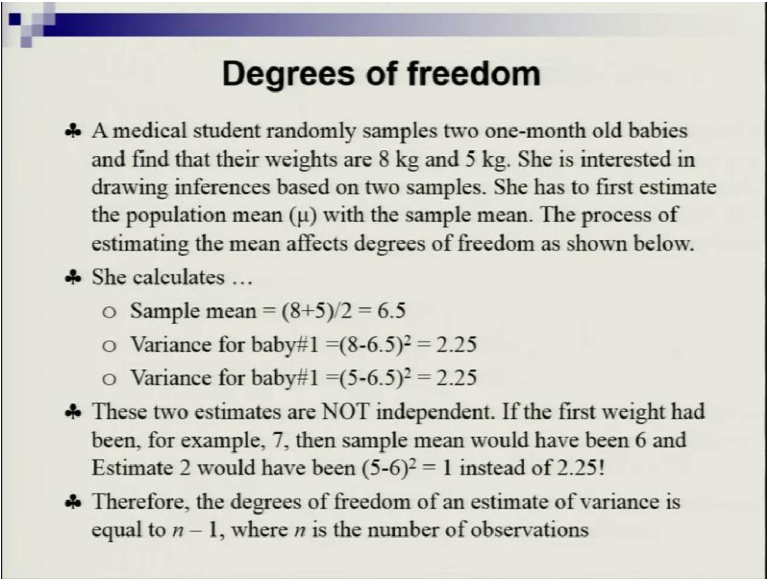
Now, I will end this slide by defining degrees of freedom. So, it is the number of values in the final calculation of a statistic that are free to vary. So, we have heard of degrees of freedom several times so far in the course, now, what is it? I have already defined it using statistical jargon and terminologies. But, if I want to explain it in nutshell, I will talk about degrees of freedom through different examples, two different examples, one is a kind of qualitative and the other one is quantitative.

So, let me first talk about the qualitative story. Suppose, there is a person who is a very colorful person and wants to wear different colored shirts every day. So, he does not want to repeat a particular shirt in 5 working days in a week or 6 working days in a week. And suppose the person has, in his wardrobe he has 6 different colored shirts. And on Monday when he is preparing for office, he picks a yellow color shirt.

So, that shirt he has used on Monday, the next, Tuesday, when he is about to be ready for office and he wants to wear a shirt, then that yellow colored shirt is out of question, because he does not want to repeat that shirt. So, this time probably he will pick a white color shirt. So, on Wednesday day when again, he has to choose one from the wardrobe to wear, then these 2 options are out. So, you see, so, on a particular day when this person is deciding which shirt to wear, his degrees of freedom actually is getting reduced day by day as he moves on in the week.

So, on the Saturday, when he has to choose, then he has no other choice, but to pick there is only one shirt in the wardrobe, which he has not used in this week. So he has to now wear that particular shirt. So in this case, the degrees of freedom will become finally 0. And on Monday again, when he will start the week he will have the degrees of freedom 6, because he can choose any 1 from these 6 that is there in the wardrobe. Now, let me move on to quantitative story and this involves a little bit of statistics, but it involves very basic concepts like mean and variance only.

(Refer Slide Time: 38:25)



Degrees of freedom

- ♣ A medical student randomly samples two one-month old babies and find that their weights are 8 kg and 5 kg. She is interested in drawing inferences based on two samples. She has to first estimate the population mean (μ) with the sample mean. The process of estimating the mean affects degrees of freedom as shown below.
- ♣ She calculates ...
 - Sample mean = $(8+5)/2 = 6.5$
 - Variance for baby#1 = $(8-6.5)^2 = 2.25$
 - Variance for baby#2 = $(5-6.5)^2 = 2.25$
- ♣ These two estimates are NOT independent. If the first weight had been, for example, 7, then sample mean would have been 6 and Estimate 2 would have been $(5-6)^2 = 1$ instead of 2.25!
- ♣ Therefore, the degrees of freedom of an estimate of variance is equal to $n - 1$, where n is the number of observations

So, let us look at these degrees of freedom example. Suppose a medical student is told in class that the mean of the 1 month old babies is around 4.5 to 5 kilograms and he or she wants to test that out using data. So that student, she picked 2 samples from the hospital on a specific day and recorded their weights and these weights are 8 kilogram for baby number 1 and 5 kilograms for baby number 2. Now, she is interested in drawing some inferences on the variance also, based on these 2 samples.

So, the population mean is not known. It is vaguely known, of course, somebody told her that this could be an approximate population mean, but it is not known to her. So, if she wants to now, look at the variance, she has to first estimate the population mean with the sample mean. And the process of estimating these mean affects the degrees of freedom as it will be clear from the calculations below.

So she keeps on calculating. So first, she will calculate the sample mean and that is basically 6.5. And then, she will use the sample mean to calculate the variance for the baby number 1 and that is going to be 2.25. Similarly, the variance for the baby number 2 can be calculated and that is, again going to be 2.25. Now these variances are same 2.25 each, but it is just by fluke. So now can we say that everything is going all right and these estimates for variants for baby number 1 and baby number 2, sorry, there is a typo here. So, the third bullet point or sub-bullet point should have baby number 2.

So these estimates, are they interrelated or are they totally independent? Now, these 2 estimates are actually not independent. Why? Because when the student is computing the sample mean, which is going to be used for the variance calculation, the student is making use of all data points, which in this case are 2. Now, if the first weight had been for example 7, then the sample mean actually could have changed to 6 and estimate 2 for variance for baby number 2 would have become 5 minus 6 square, which is 1 instead of 2.25.

So you see that if you change one number here, then there is an impact on the sample mean and the sample variance as well. So basically, you see that, when you were estimating some sample statistic, then actually you do not have all n independent observations. So basically, if you have some steps to arrive at the final sample statistic, which is your destination, then in this process n route, if you were making calculation for some other sample statistics, you have to take that into account and then you have to deduct that from the original number of observations. And that is called the degrees of freedom.

So, if you start with n number of observations to calculate the estimate of variance, actually is equal to n minus 1, because you have to calculate the sample mean first to come up with the estimate for variance. So now you know, how I have used the degrees of freedom in the chi-square distribution, probably have understood it from the last discussion. But let me spend 1 minute more. So you start with the normal random variable with population mean μ and then population standard deviation σ . But actually, you do not know the value of μ and σ .

So you want to actually use a proxy for the standard deviation σ by say sample standard deviation. But in order to calculate that proxy measures sample standard deviation, you have to first calculate the sample mean. So, in the case of the chi-square variable generation, you have to

save one observation as the dependent observation and rest $n - 1$ are the independent observations. So, we end this discussion by emphasizing the role of different continuous distributions and the concept of degrees of freedom. And we will explore the uses of these concepts in greater detail as we move on in the course. So till then, bye. Thank you.