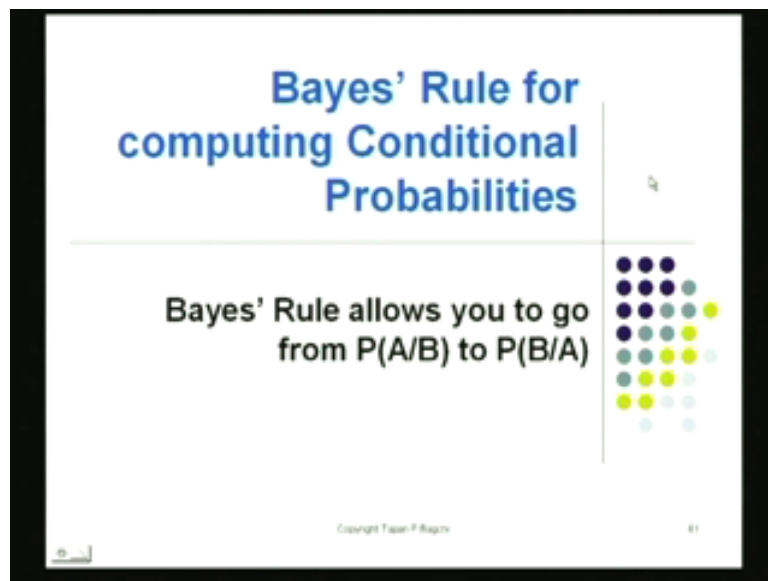


Six Sigma
Prof. Dr. T. P. Bagchi
Department of Management
Indian Institute of Technology, Kharagpur

Module No. # 01
Lecture No. # 08
Review of Probability and Statistics-IV

(Refer Slide Time: 00:27)



Now, we start again and this will be the last class talk in the area of probability and statistics. And I will try to up compress just the highlights and I will try to bring them up here. Something you got to remember as we see in the slides is we have been discussing conditional probability and we have been saying how to get $P(A \text{ given } B)$. And that turns out to be $P(A \text{ times } B, \text{ d r by } P(B \text{ given } A))$ that turns out to be $P(A \text{ given } B)$.

I can actually use this rule which is given by bayes. Bayes was a priest this gentlemen he was also a mathematician and over a 100 years ago he came up with this idea of going from $P(A \text{ given } B)$ to $P(B \text{ given } A)$ and that is a very simple straight forward application of a little bit of algebra.

(Refer Slide Time: 01:07)

Bayes' Rule for finding $P(B/A)$

- Given two events A and B and suppose that $\Pr(A) > 0$. Then

$$\Pr(B|A) = \frac{\Pr(AB)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

Bayes Definitions:

- $\Pr(B) = a \text{ priori}$
- $\Pr(B|A) = a \text{ posteriori}$

Copyright Team P-Baptist 64

Let me show you what is formula looks like? Notice here on the left hand side I have got P B given A, which is equal to by definition of condition probability P A B divided by P A. Now, you can breakup that P A B as P A given B multiplied by P B. That of course, you know it is just the denominator, numerator and if the denominator up got the old stuff preserved there.

So, now on the left hand side I have P B given A on the right hand side I have got P A given B. Say if I know P A given B and if I know P B and P A; I can compute P B given A. This is a very very useful relationship it is called the Bayes rule and so by fault one of again one of the more powerful results in probability theory.

(Refer Slide Time: 01:56)

Bayes' Rule

$P_i(W R)$	R	-R
W	0.7	0.4
-W	0.3	0.6

Events:
R: It rains
W: The grass is wet

Information
 $P_i(W|R)$

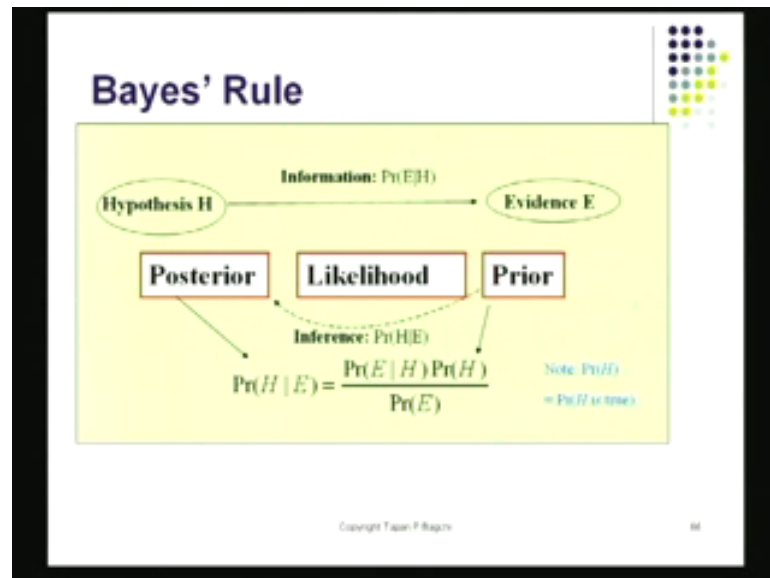
Inference
 $P_i(R|W)$

Copyright Team P. Rajan 88

Or let me give you one example. I am not going to be working this example in great detail, but, I am going to give you a hint as to how to work this example? It is the chance that you walkout and you find the grass is wet. Could it have been caused by rain or was it the sprinkler that was on maybe an hour back or something, so the grass is wet. At based on that you may decide to carry the umbrella or not to carry the umbrella because the grass is wet. The grass is wet when you find in the morning.

The grass is wet you do not necessarily look at the sky. You could have you have rush in you are out going out in your car so you grab the umbrella and run. So, should you or should you not grab that umbrella. That is the question and there this bayes rule can be used and so on and so forth.

(Refer Slide Time: 02:45)



Proper things I would like to make sure you understand is you will run into terms like prior and probability posterior. Prior is which is like probability of there being what we call the hypothesis or the evidence and evidence is E hypothesis is H.

(Refer Slide Time: 03:09)

Solution to the medical test

$P(\text{cancer}) = P(c) = 0.001$, $P(\text{healthy}) = 0.999$
 $P(\text{test +ive} / c) = 0.92$, $P(\text{test +ive} / \text{healthy}) = 0.04$

$$P(c / \text{test +ive}) = \frac{P(+ive / c) P(c)}{P(+ive / c) P(c) + P(+ive / \text{healthy}) P(\text{healthy})}$$

Verify that the answer is 0.0225

Managerial question:
Will you rely on this test to get a treatment?

Copyright Tarek P. Bagchi 67

Let me try to see if I could give you another version of the same. We took the medical test of course, there we use the same formula and the base formula can also be used in more complex situations.

(Refer Slide Time: 03:18)

A More Complicated Example

$\Pr(R) = 0.8$

R It rains
W The grass is wet
U People bring umbrella

$\Pr(UW|R) = \Pr(U|R)\Pr(W|R) \rightarrow ①$
 $\Pr(UW|\neg R) = \Pr(U|\neg R)\Pr(W|\neg R) \rightarrow ②$

$\Pr(W R)$	R	$\neg R$
W	0.7	0.4
$\neg W$	0.3	0.6

$\Pr(U R)$	R	$\neg R$
U	0.9	0.2
$\neg U$	0.1	0.8

Q. What is the probability that people will bring umbrella when they see that grass is wet = $\Pr(U|W) = ?$

Let us get to the point when you got that rain situation again rain, wet grass and umbrella. And let us take a look at what the problem is that we would like to tackle. There are three possibilities one it has rained or not rained or not. I carried the umbrella, I do not carry the umbrella the grass is wet or it is not wet. So, there are three wets and they are actually influenced by each other and what is giving to us is? There are certain relationships there are independent. I am just going to leave this with you. I am not going to solve this problem R is the event that it rains, W is the event that grass is wet; U is the event that people carry umbrella with them. And then some condition probabilities are given as two tables here.

Those condition probabilities are given here and the question that is being asked is if these are the condition probabilities given will you or will you not bring an umbrella with you that day? And to solve this, what I have done is I have basically worked out the thing and I am going to displaying this with a couple of minutes. So, I am going to show you first what this sitsures cells? I am going to be indicating to you two results that we will be using. One result is this one probability of you are carrying the umbrella and grass is wet given that it has rained. If U and W are independent this can be verified of course, we can verify that later on. If these are independent I can break this up into this probability of U given R multiplied by probability of W given R.

U is the event if I am that I am carrying an umbrella and W is the event of the grass is wet. If these two are independent I could write it like this and I could write the same thing for the non rain situation. This is the situation with rain and this is the situation without rain and there again I could write this formula. These are given to you these are provided to you. So, we assume them to be true then I am given this conditional probability and also this conditional probability. Will I really carry umbrella that day.

(Refer Slide Time: 05:37)

SOLUTION

$$P(U/W) = ?$$

From (1) $P(UW/R) = P(U/R) P(W/R) = 0.9 \times 0.7 = 0.63$

From (2) $P(UW/\bar{R}) = P(U/\bar{R}) P(W/\bar{R}) = 0.2 \times 0.4 = 0.08$

Now $P(W) = P(W/R) P(R) + P(W/\bar{R}) P(\bar{R})$
 $= 0.7 \times 0.8 + 0.4 \times 0.2 = 0.56 + 0.08 = 0.64$

Hence $P(\bar{W}) = 1 - P(W) = 0.36$

And $P(UW) = P(UW/R) P(R) + P(UW/\bar{R}) P(\bar{R}) = 0.61$

$$\therefore P(U/\bar{W}) = \frac{P(UW)}{P(W)} = \frac{0.61}{0.64} = 0.953$$

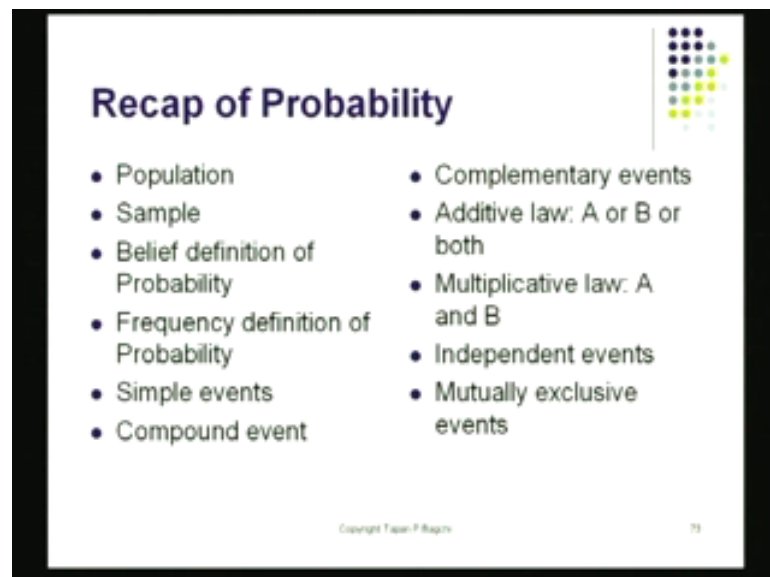
So, to do this what I do? I bring my solution and I will adjust basically need to have a look at the solution I have worked out the steps here. And I have worked out some of the numericals also. Some calculated here for example, $P(U/R)$, U and W given R that is like you carry the umbrella and the grass is wet given that is there is rain. What is the chance for that? I broke that up and I had these quantities from the table. From the matrix table I have these two quantities.

So, I plug those in and I got my probability there. One is this probability the other is the other probability. Then of course, I also worked out the total probability of there being grass being wet, which could happen because of rain or even because of no rain because the sprinkler comes on when there is no rain and that probability also I work out. Therefore, from the second part of the probability of there the grass being dry that probability I work out.

And then of course, I tried to work out the probability of our interest pitches like probability carrying umbrella. Given that the grass is wet and that turns out to be about 95 percent. I am going to leave this with you. I am not going to solve this problem in detail. But, you are supposed to work this through and you are supposed to check out the numbers which are there and you can let me know if the numbers are not correct.

So, this is like something that would like to be able to do and this is possible when you have your the bayes rule in effect. Just doing a quick recap of what we have done so far in the area of probability. Mind you we have not used real data.

(Refer Slide Time: 07:07)

A slide titled "Recap of Probability" with a decorative graphic of colored dots in the top right corner. The slide lists various probability concepts in two columns. The left column includes: Population, Sample, Belief definition of Probability, Frequency definition of Probability, Simple events, and Compound event. The right column includes: Complementary events, Additive law: A or B or both, Multiplicative law: A and B, Independent events, and Mutually exclusive events. At the bottom, there is a small copyright notice and the number 79.

Recap of Probability

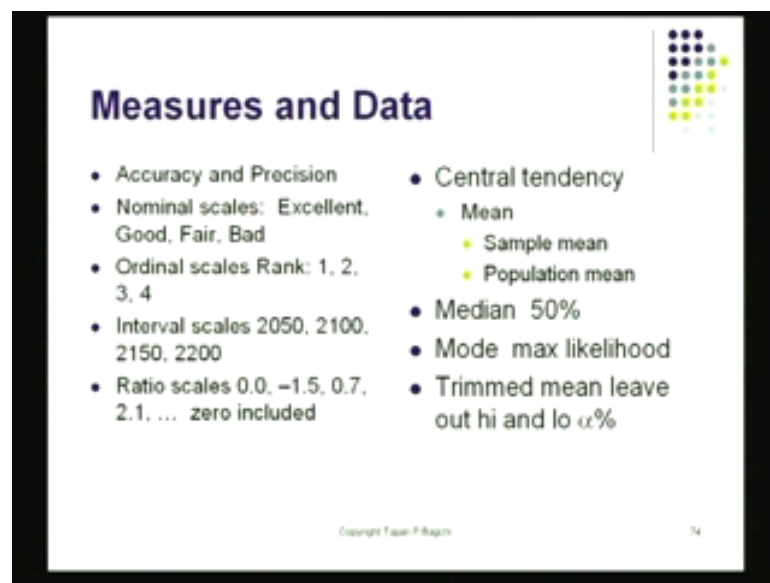
- Population
- Sample
- Belief definition of Probability
- Frequency definition of Probability
- Simple events
- Compound event
- Complementary events
- Additive law: A or B or both
- Multiplicative law: A and B
- Independent events
- Mutually exclusive events

Copyright Tapan P. Bhargava 79

So, we have not really dived into a statistics here. We have just being playing with probability, various principles of probability. What are solve those principles? First of all we have this notion of population which is the total extent of the area of interest in which you have got some interest. It could be the total quantity of production done last year or last month or this week. That is the total population. From that I collect a samples I have got the idea of sample. I have that then I gave you a big definition of probability. Then of course, I gave here subjected motion of probability and any more objective motion of probability. This subjective motion comes from belief in probability. The belief that on July 15th is going to rain in Kharagpur. That is a matter of belief. Frequency estimation of course, will go back into real data.

So, we will look at real metrological data and from that it will try to figure out how often on July 15th has it rained in Kharagpur? Then we discussed simple events and we discussed compound events. And we looked at conditional probabilities and we looked at complementary events. We looked at the additive law and we looked at the multiplicative law. That also we looked at, we looked at independent events, we looked at mutually exclusive events. These we looked at.

(Refer Slide Time: 08:37)



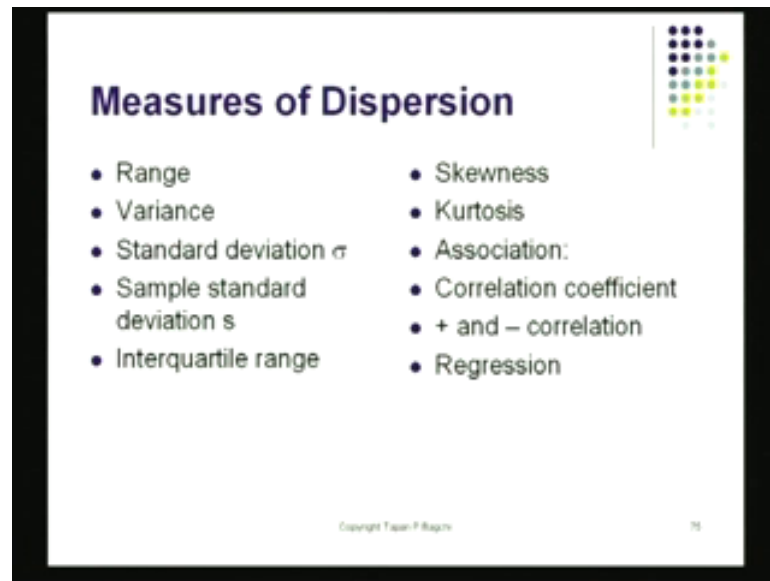
Measures and Data

- Accuracy and Precision
- Nominal scales: Excellent, Good, Fair, Bad
- Ordinal scales Rank: 1, 2, 3, 4
- Interval scales 2050, 2100, 2150, 2200
- Ratio scales 0.0, -1.5, 0.7, 2.1, ... zero included
- Central tendency
 - Mean
 - Sample mean
 - Population mean
 - Median 50%
 - Mode max likelihood
 - Trimmed mean leave out hi and lo $\alpha\%$

Copyright Tarek P. Maghrebi 74

Then of course, I gave you some hint of how data is measured. All this sum up we will be getting into a little bit not a great deal.

(Refer Slide Time: 08:45)



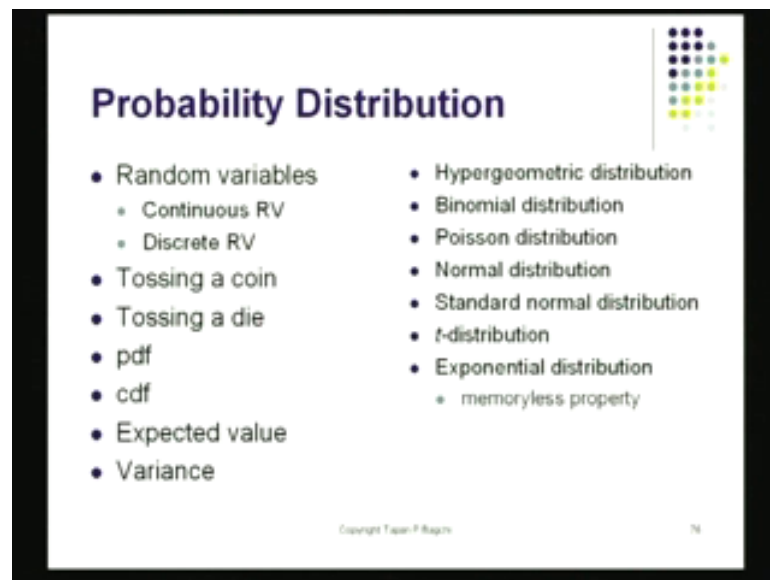
Measures of Dispersion

- Range
- Variance
- Standard deviation σ
- Sample standard deviation s
- Interquartile range
- Skewness
- Kurtosis
- Association:
 - Correlation coefficient
 - + and – correlation
 - Regression

Copyright Tapan P. Bagchi 76

But, we will be getting into a little bit into this measure of central tendency and the measure of dispersion.

(Refer Slide Time: 08:50)



Probability Distribution

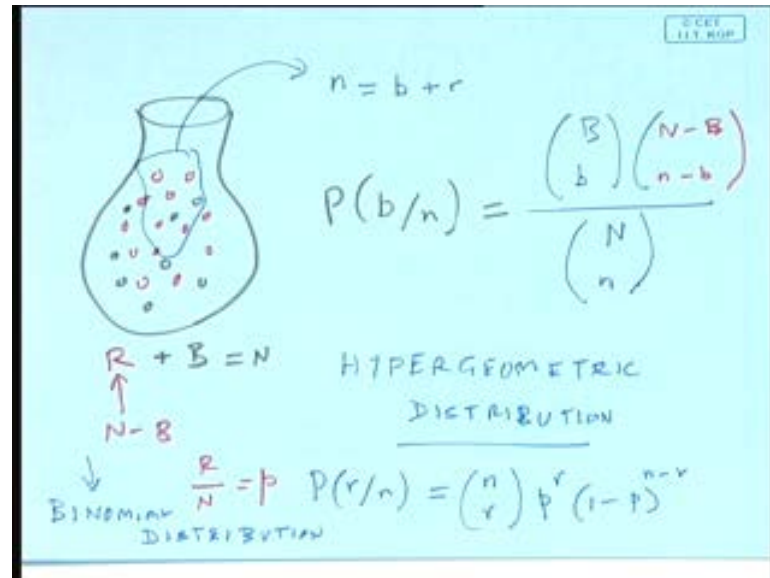
- Random variables
 - Continuous RV
 - Discrete RV
- Tossing a coin
- Tossing a die
- pdf
- cdf
- Expected value
- Variance
- Hypergeometric distribution
- Binomial distribution
- Poisson distribution
- Normal distribution
- Standard normal distribution
- t-distribution
- Exponential distribution
 - memoryless property

Copyright Tapan P. Bagchi 76

The aspect we will be doing there. There are various types of probabilities rather indicated there are available. The moment you start talking about distribution you are talking about data. You are talking about if I collect a large firm volume of data large amount of data. What kind of distributions will I end up with? There is something that

we got to keep in mind and that we will be using when I go into high level of theory. For example, there is certain there is a certain thing called the hypergeometric distribution.

(Refer Slide Time: 09:24)



Let me give you an idea of that. I have a jar and I am going to be drawing it here. I have a jar and the jar contains some red balls and some black balls. (No Audio Time: 09:39 09:47) Let us say this jar is our total production, the red ones are the defective items and the black ones are the ok items. Now, this being too large, this quantity being too large. I am not able to really sample everything. I am getting not really be able to test everything. So, what I do is? I pull out a sample and that sample comes out in here. Before I took the sample of there were there were R red balls and B black balls inside and R plus B this is equal to N plus the total number of balls in the system there. And when I bring out a sample, the sample comes out of this. I have a sample of size N it has got in that case B black balls and R red balls that is in the sample the sample is just a part of it.

So, just that is the sample that is what I pulled out. What I would like to work out is, what is the probability of there being B black balls in a sample of N? So, the question that I am asking is the following probability or there being B black balls in n balls pulled out of the thing. This is the probability that I would like to write down and let me show you how this probability can be written down. It is not very difficult. Look at the situation I have a total of R plus B balls which is like a total of N balls and out of N balls I have picked little N balls. So, really in terms of combination these are the different

ways I could pick N balls out of big N balls. That is the count of total number of ways by which I could pull out N little balls, little N balls out of the total jar. The jar consists of N big balls.

Now, the black balls would have come out. Black balls in the sample would have come out of these black balls. So, I have here B and b . And the red balls would have come out now of the remaining balls which is N minus B . So, I am going to put N minus B in red color. That many red balls are now that is equal to R actually, this is equal to N minus B . That is what I have got there and then how many red balls are there? Now, there are little N minus B . This is the formula that give tells you what is the chance of my finding exactly a little b black balls in n balls picked in as a sample out of this thing. This distribution is very important in quality assurance.

When you are sampling from a finite jar, finite population. It could be a basket, it could be a box, it could be a certain number of items that have been placed on your table. And you are sampling a few out of them and this particular distribution has a name which is called the hypergeometric distribution. (No Audio Time: 13:28 to 13:43). It is the hypergeometric distribution. It is a very important distribution in sampling theory.

Suppose, this jar was very big and the jar being big would mean that I can only talk in terms of fraction defective. I cannot really count the balls and say how many red balls are there? How many b black balls are there? And in that case I will be talking in terms of fraction defective which I could probably denote by a quantity p . p is the fraction defective fraction of red balls in this full jar there. So, really speaking this P is equal to R divided by N that is for basis. If this is the story and I again take out n balls what is the chance that I have picked so many red balls out of the jar or so many black balls out of the jar?

In this case let us assume that the red balls are the defective ones. So, what is this formula now? I will just say this probability P , that there are I have picked a total of little n . So, I will really call it here r red balls in n . That is a sample. A sampled n balls out of them r turned out to be defective. This can be written as following n choose r p raise to the power r 1 minus P raise to the power n minus r . That is the formula of that is to be utilized if you want to find out the number of the probability of there being r defective balls in n that we pulled out. I am calling the red balls now this turn to be defective. This

is the formula. What is this formula? Where is this from? This is from the binomial distribution. So, this is the binomial distribution.

This is also a very important distribution like the hypergeometric distribution in probability theory. What is the difference between these two? This applies when your sample size is when your **when your** jar is finite size. It can all be put on the table for example, and this formula applies the binomial formula applies when production quantity is very very large. And this is generally true benefit of full truck and you are sampling a few items. So, when you will look at sampling plans. Sampling plans are either based on this or they are based on this. Most of the sampling plans that are used in the industry they are based on the binomial formula. They are based on this. So, I just wanted to give you an idea of what it would be if I were doing this?

Now, suppose this P was quite small, P was very very small. That means most of your items were ok. What sort of formula will you use? You will be using the poisson formula and the poisson formula is listed here. See here poisson this again I am going to show you in a little place in a little while I am going to show you. Where did the formula come from? So, let us move on to that.

(Refer Slide Time: 17:16)

**Descriptive Measures of the Population:
Mean, Variance, & Std. Deviation**

- **Population:** a hypothetical set of N observations from which the sample of observations actually obtained can be imagined to come (typically N is very large)
- measure of location: **population mean**

$$\mu = \sum y_i p_i(y_i) \text{ for discrete } Y$$
$$\mu = \int y f_y(y) dy \text{ for continuous } Y$$

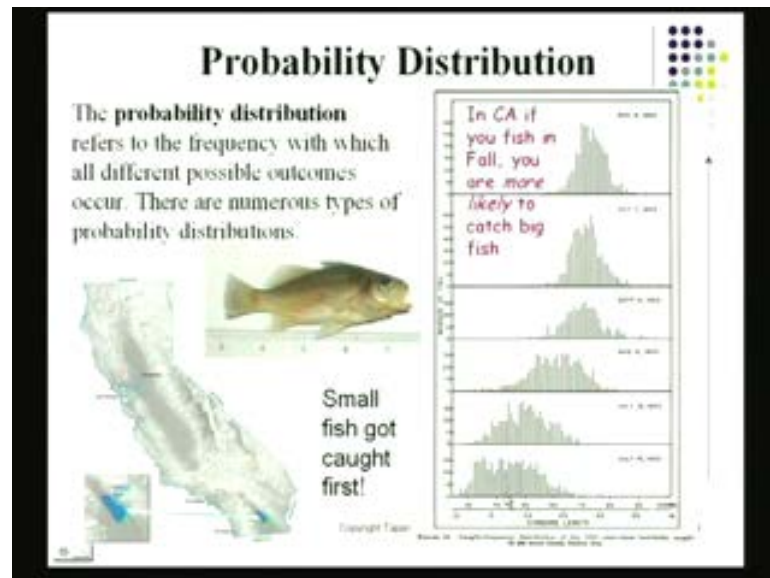
A population does not necessarily refer to people but to any total set of observations. For basketball, a population of scores is all of the scores that any team might get. For a single coin toss it is (head, tail).

Copyright Team P-Biz 2015 83

Let us move on to that some basic ideas are shown in the beginning and you get an idea of mu. Mu is the average, mu is the mean and there is a way to calculate mean and the

formula is given there. And there is a way to calculate mean when the situation is discreet that also is given there.

(Refer Slide Time: 17:31)



And here is the situation let me show you what is happening here. This is fishing of a particular type of fish. Fishing for a particular type of fish in California and the fishing season starts in summer to at the end of summer like July. And then as people start fishing what you find is this is the length of the fish caught. This side is the length of the fish caught. So, if you are more on this side you have got large fish, when you move on this side you got small fish. So, probably when the fish has spawning. Most of the fish are small, many are small. They slowly grow and they grow to larger size and because there are many people who are excited fishing. They go and catch a lot of fish. This is the distribution of fish caught in July. This is like a after about mid July. It is sort of like this then later on, in some more time it moves up like this. Then it moves up like this September is like this, October it is like this. And about November the distribution of the length of the fish which is like the length between the tail and the head. That moves up to this height.

So, what is happening here? I am seeing a shift in the population of the fish that is there. The distribution itself is shifting, the couple of things shifting here. One notice that the mean is shifting. This is a very important thing to notice. The mean length is shifting and certainly the max is shifting the mean is shifting but, certainly the central tendency of the

distribution or the fish caught that is moving up and also the various is changing. Here the various was white. So, you have the lot of dispersion in this area and height is become tight.

So, actually a more fish, more or less of the mean size, not many there are really small not many that are really big. Around here you had that issue there but, they are all seemed to grow up into adults and they are like this. This is like it just gives you an idea; it just gives you an idea of the hard likely to catching a big fish as time goes on. If you go and fish in November your more like to catch a big fish and if you fish in July your more like to catch a catch a small fish. And this is happened because of the shift in the distribution and of course, growth and the amount of fishing that is going on and so on. All those things would influence this.

(Refer Slide Time: 19:57)

Expectation

- A random variable X - $\Pr(X=x)$. Then, its expectation is

$$E[X] = \sum_x x \Pr(X=x)$$
- In an empirical sample, x_1, x_2, \dots, x_N ,

$$E[X] = \frac{1}{N} \sum_{i=1}^N x_i$$
- Continuous case:

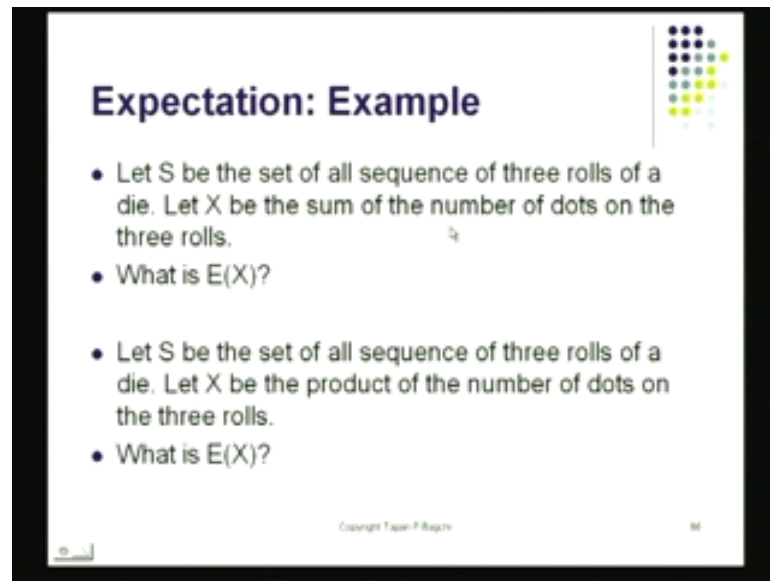
$$E[X] = \int_{-\infty}^{\infty} x p_X(x) dx$$
- Expectation of sum of random variables

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

Copyright Team P. Raghav

Let us take a look at some of the measures now. Clearly the expectation which is the average of the few average size of the random variable that turns out to be x times $P r$ in the discrete case. And the continuous case it turns out to be this formula and I am pretty sure you are familiar with this. If I have two variables X_1 and X_2 and I found to look at this some of those random variables. If I look at want to look at their average value or their expected value that turns out to be $E X_1$ plus $E X_2$ nothing very fancy.

(Refer Slide Time: 20:27)



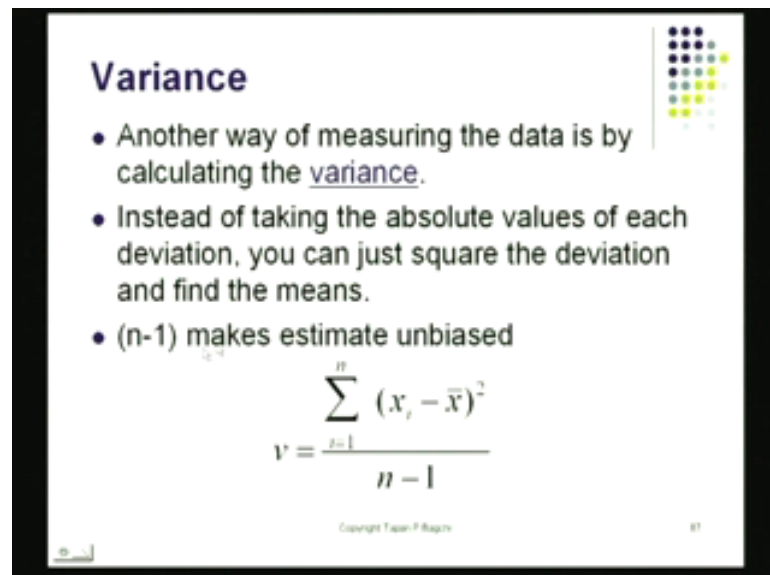
Expectation: Example

- Let S be the set of all sequence of three rolls of a die. Let X be the sum of the number of dots on the three rolls.
What is E(X)?
- Let S be the set of all sequence of three rolls of a die. Let X be the product of the number of dots on the three rolls.
What is E(X)?

Copyright Tarek P. Baheti 86

And of course, you can have many different examples and so on. And all you have to do in this case is you have to figure out these P X and P Y. If you have the P X and P Y known to you P X 1 P X 2 and so on. You can easily calculate the expected value of any this thing any of these things.

(Refer Slide Time: 20:44)



Variance

- Another way of measuring the data is by calculating the variance.
- Instead of taking the absolute values of each deviation, you can just square the deviation and find the means.
- (n-1) makes estimate unbiased

$$v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Copyright Tarek P. Baheti 87

Then there is this measure called variance and of course, I have plotted here this sample variance. Sample variance has a degree of freedom which is like less than the total number of items. There is less by 1 because I have used the data once already. I have

calculated \bar{x} only then I can calculate this variance. When I am calculating sample variance I will lose 1 degree of freedom. So, this guy it turns out to have 1 degree 1 less degree of freedom as compared to the original number of data points there.

(Refer Slide Time: 21:19)

• Taking the square root of the variance which results in the standard deviation.

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

• The standard deviation can also provide information about the **relative spread** of a data set.

• **Range** ($X_{\max} - X_{\min}$) can also show spread

Copyright Tamer P. Bagchi

And you see the formula for standard deviation of which is at the square root of what we had before. That is what we have here. Then there is a very important quantity called the range. Range is actually a very important quantity, range also gives you an idea of the dispersion of the data and range is actually a very simple value to calculate as posed to this formula here. This is the gigantic formula it takes a lot of effort to do it unless you are using excel. If you are using range all you have to do is look at the largest number in your sample and look at the smallest number in your sample, but difference between those two turns out to be the range of the data value is there.

(Refer Slide Time: 22:03)

The standard deviation for a grouped distribution is calculated from

$$s = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n - 1}}$$

TABLE 10.5 Standard Deviation Calculation for Each Group

Group A	Group B
$(p - p_m)^2$	$(p - p_m)^2$
400	2500
225	3600
100	12,100
3600	6400
900	14,400
2500	10,000
625	3600
400	22,500
400	8100
1600	400
$\Sigma = 10,750$	$\Sigma = 81,600$
$s = 34.56 \text{ (kg/m)}^2$	$s = 95.22 \text{ (kg/m)}^2$

And of course, again I have got a formula for the average there. That can be found and there is some example there where again you can see this is rho is really the density here and the densities of two different groups. Those are plotted here and they give you an idea of the standard deviation.

(Refer Slide Time: 22:17)

Variance

- The **variance** of a random variable X is the expectation of $(X - E[X])^2$:

$$\begin{aligned} \text{Var}(X) &= E((X - E[X])^2) \\ &= E(X^2 + E[X]^2 - 2XE[X]) \\ &= E(X^2 - E[X]^2) \\ &= E[X^2] - E[X]^2 \end{aligned}$$
- Range** = $\text{Max}(X_i) - \text{Min}(X_i)$

Copyright Taim F. Rajan

So, that is also shown here that is also something that is shown here. It turns out variance is a very good indication of this spread of the data of the dispersion of the data and so, is range. These are all very very important.

(Refer Slide Time: 22:28)

Descriptive Measures of the Population Spread

- measure of spread: **population variance, population standard deviation**

Discrete:

$$\sigma^2 = \sum_i (y_i - \mu)^2 p_i(y_i) = \sum_i y_i^2 p_i(y_i) - \mu^2$$

Continuous:

$$\sigma^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f_Y(y) dy = \int_{-\infty}^{\infty} y^2 f_Y(y) dy - \mu^2$$

Copyright Tapan P. Bagchi 92

And the formulas they are shown here they can be found from any text book. You of course, will not use directly these formulas. You would be using either the function or built in software or something to be able to do it.

(Refer Slide Time: 22:39)

Descriptive Measures of the Sample: Average, Variance, & Std. Deviation

- **Sample**: a set of n observations actually obtained (typically n is relatively small)
- measure of location: **sample average**
- measure of spread: **sample variance, sample standard deviation**

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 / n}{n-1}$$

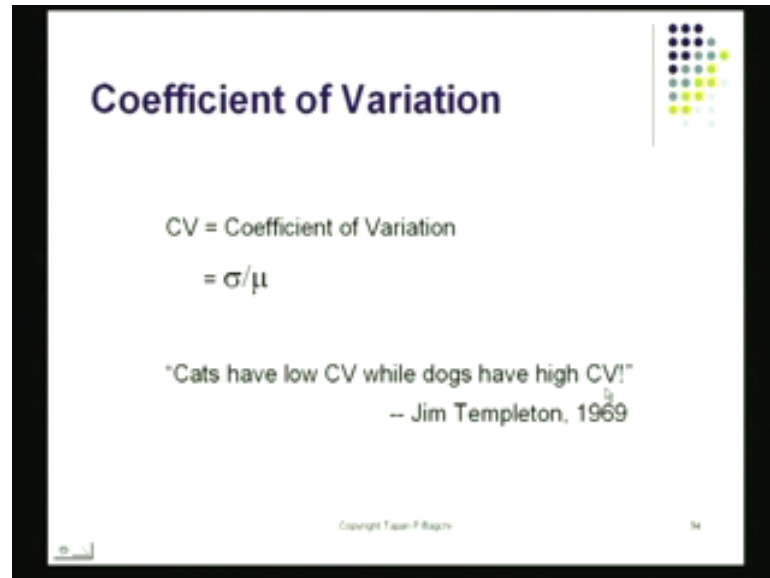
A sample could be the starting salaries of twenty-five fresh MBAs

Copyright Tapan P. Bagchi 93

But, you must have the concept. You must really understand what it really means? The sample is only of course a subset. The sample is not the full population, it is only a subset. So, therefore, anything that I calculate from that sample will be qualified by calling it a sample average or a sample variance or a sample standard deviation. This is

very very important and any quantity that I compute from sample values. Those quantities are called statistics.

(Refer Slide Time: 23:08)



Coefficient of Variation

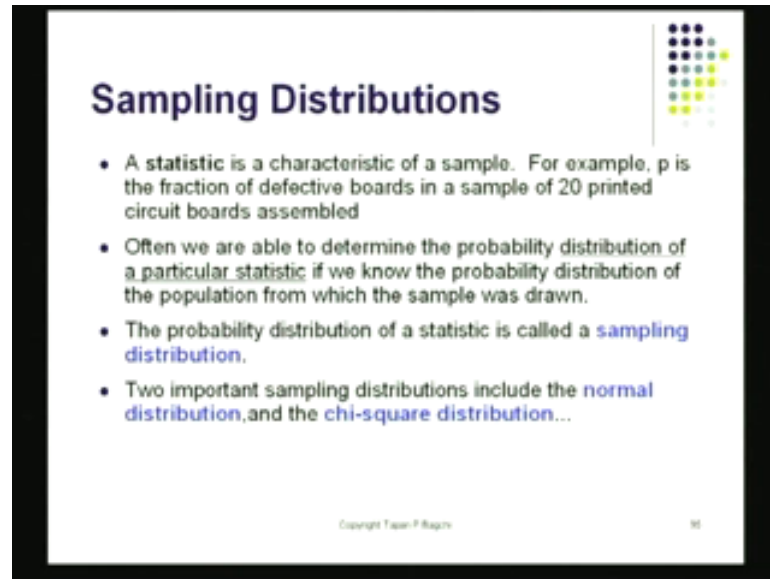
CV = Coefficient of Variation
 $= \sigma/\mu$

"Cats have low CV while dogs have high CV!"
-- Jim Templeton, 1969

Copyright Tapan P. Bagchi 14

And there is another measure of variation that is called the coefficient of variation which is just basically the deviation of sigma, sigma divided by mu. That turns out to be the coefficient of variation and my old professor Jim Templeton, he used to say cats have low CV and dogs have high CV. The reason is this when you workout this ratio for dogs. Dogs actually come in widely varying sizes from little you know teeny weeny once that I can carry in your pocket to the big guys. Dogs come in all sizes. Therefore, dogs sigma is quite a bit high. When you divide that by this the coefficient of variation for dogs turn out to be high. Cats on the other hand they all come more or less of the same size. All cats are about this size. So, there standard division is also is quite small and there CV then turns out to be also low.

(Refer Slide Time: 24:10)



Sampling Distributions

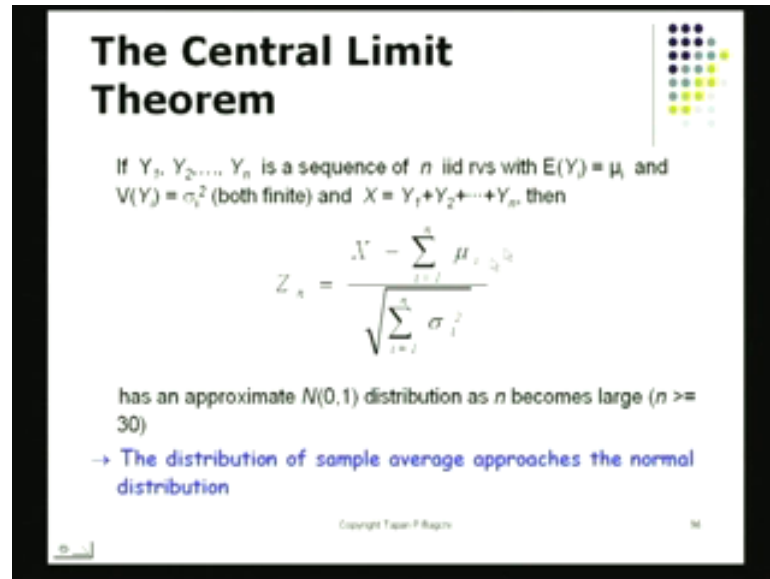
- A statistic is a characteristic of a sample. For example, p is the fraction of defective boards in a sample of 20 printed circuit boards assembled
- Often we are able to determine the probability distribution of a particular statistic if we know the probability distribution of the population from which the sample was drawn.
- The probability distribution of a statistic is called a **sampling distribution**.
- Two important sampling distributions include the **normal distribution**, and the **chi-square distribution**...

Copyright Team P. Rajan 95

So, cats have low CV and dogs have large CV, coefficient of variation. Sampling distributions any data that you calculate using samples those would be called sample distribution. Let me take let me give you an example. See remember we sampled from the diagram here. A sample out of a jar the jar was of a final size and in this case I turned out if you get I turned out to get this into converted into hypergeometric distribution. This probability formula here is based on the hypergeometric distribution.

If the jar becomes really large then of course, I have the binomial distribution. So, this is also another sampling distribution. This is a sampling distribution which describes the property of the sample and the binomial distribution also describes the property of a sample. So, this also turns out to be a sampling distribution and this is what we have one the screen here. I have various types of sampling distribution I will give you some example.

(Refer Slide Time: 25:01)



The Central Limit Theorem

If Y_1, Y_2, \dots, Y_n is a sequence of n iid rvs with $E(Y) = \mu$ and $V(Y) = \sigma^2$ (both finite) and $X = Y_1 + Y_2 + \dots + Y_n$, then

$$Z_n = \frac{X - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

has an approximate $N(0,1)$ distribution as n becomes large ($n \gg 30$)

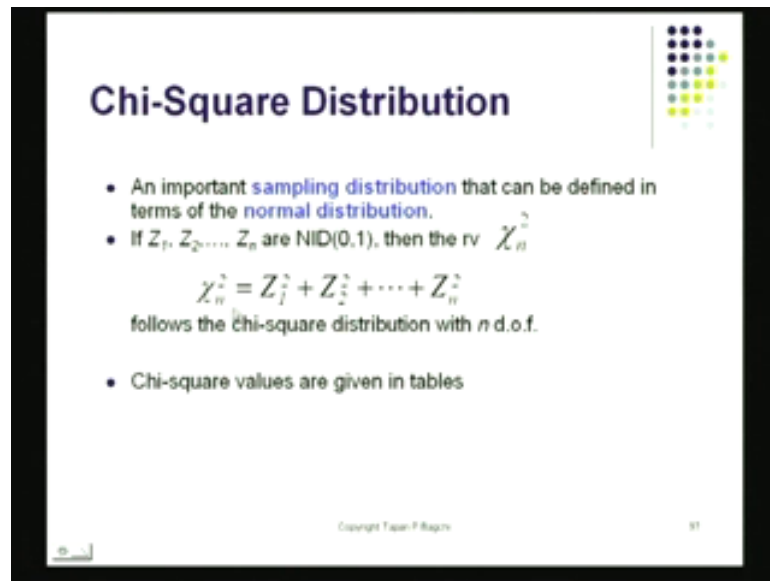
→ The distribution of sample average approaches the normal distribution

Copyright Team P-Maps

It turns out that there is a theorem called the central limit theorem. It is a very very powerful and a very useful theorem. What it says is if you calculate the sum of it does not matter what come in random variables. If you calculate their sum that sum will approach a normal distribution the sample average or the sum both of them they will approach the normal distribution. This is a very useful result. The reason is this I have many tables available for normal distribution.

I have got all kinds of calculations done already by experts who came before us. They have given us tables for the normal distribution and what this theorem is saying, what central you know limit theorem is saying is if you take sums or if you take averages of sample data, the average or the sum they will approach the normal distribution. And then you can use the normal distribution theory to do any calculations that you want to do. So, it is a very very useful result.

(Refer Slide Time: 26:07)



Chi-Square Distribution

- An important sampling distribution that can be defined in terms of the normal distribution.
- If Z_1, Z_2, \dots, Z_n are $NID(0,1)$, then the rv χ_n^2
$$\chi_n^2 = Z_1^2 + Z_2^2 + \dots + Z_n^2$$
 follows the chi-square distribution with n d.o.f.
- Chi-square values are given in tables

Copyright Tarek F. Elmaghrabi 91

If you have Z now remember Z is the Z is been adjusted X_1, X_2, X_3 they have normal distribution and Z had the mean subtracted from it divided by the standard deviation. Subtracted in the mean shifts the mean to 0 shift the mean. So, the it turns out that the mean of the quantity Z will have an will have a 0 mean. The mean of this will be 0 and the standard deviation of Z is going to be 1 by the scaling they are I have done. If I take these quantities, if I take many of them, if I square them, if I square Z_1 . So, I have got Z_1^2 plus Z_2^2 square and so on up to Z_n^2 square. I have got n items in the sample; I convert all of those X_1, X_2 values into Z values. Then I squared them and if I add them. I constructed new summary of the sample statistic. This summary is called the chi-square random variable.

Chi-square is also very very useful property, very very useful distribution and this is utilized many times when I want to check. For example, the nature of the distribution. Nature of the distribution is it normal? Is it exponential? Is it binomial? To be able to do that test. I must use a some sort of summary that I have prepared from the data and that summary generally speaking should be in the form of a chi-square random variable. So, you have got this chi-squared test. Usually tests given a particular set of data. Does it confirm to the pass on distribution or to the normal distribution or to the uniform distribution? Whatever it is I can check that out by doing the chi-square test.

(Refer Slide Time: 27:49)

Example 1: Suppose that Y_1, Y_2, \dots, Y_n is a random sample from a $N(\mu, \sigma^2)$ distribution. Then

sum of squares

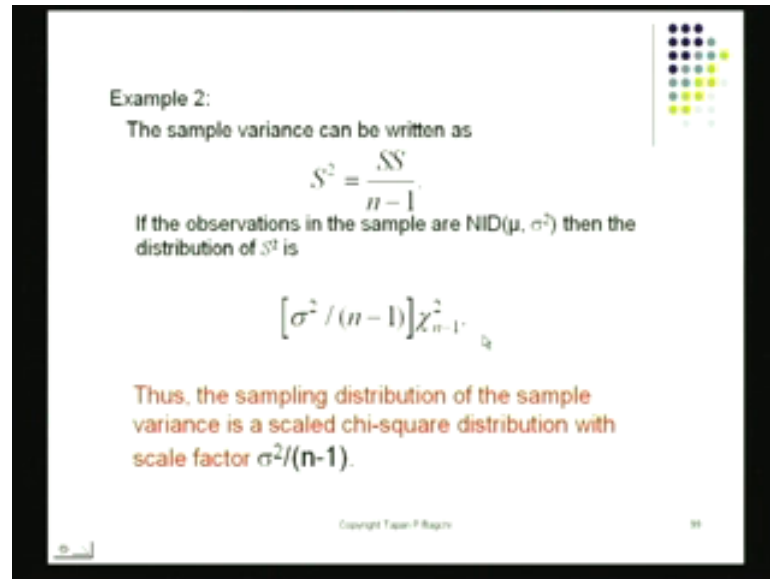
$$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

➤ The sum of squares in normal random variates when divided by the variance follows a **chi-square distribution** with $n-1$ or (v) degrees of freedom.

Copyright Team P-Baptis 18

And it turns out the chi-squared distribution also is found when I have got this quantity and these are utilized whenever I am trying to test. For example, you know you will find out later on that the sample average \bar{X} has a distribution. That is normally distributed. It turns out this sample variance S^2 that has a distribution that is very closely tied to the chi-square distribution. And these are useful when I am trying to estimate, the data estimate the parameters either sigma square or mu. I am trying to estimate them and I am using data that is given to me. In one case I can use normal distribution to guide my calculations. In the other case I can use the chi-squared distribution when it comes to the putting some limits on variance, putting some estimates on variance. Then I can use the chi-squared distribution it is useful there.

(Refer Slide Time: 28:44)



Example 2:
The sample variance can be written as

$$S^2 = \frac{SS}{n-1}$$

If the observations in the sample are $NID(\mu, \sigma^2)$ then the distribution of S^2 is

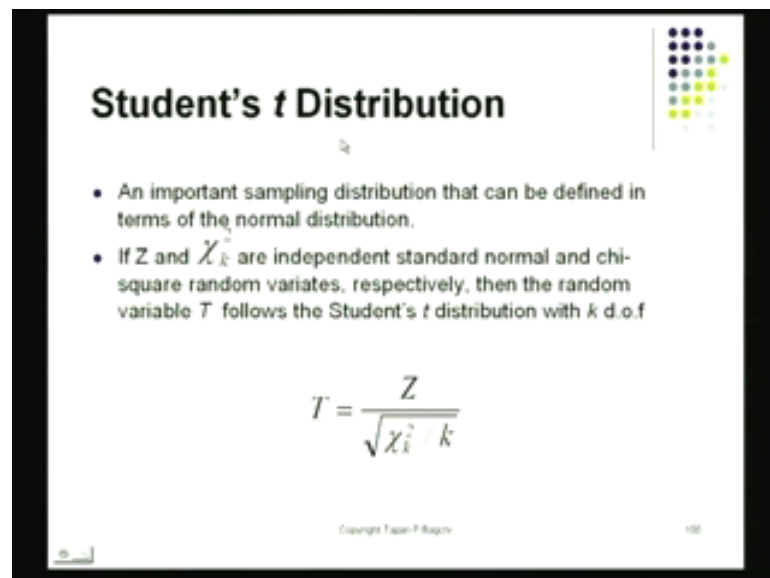
$$\left[\frac{\sigma^2}{(n-1)} \right] \chi_{n-1}^2$$

Thus, the sampling distribution of the sample variance is a scaled chi-square distribution with scale factor $\sigma^2/(n-1)$.

Copyright Tarek P. Maghrebi 99

So, it turns out in that case if I calculate the sample variance, it will have this distribution and this is linked to be chi-squared distribution, a very very useful distribution.

(Refer Slide Time: 28:58)



Student's t Distribution

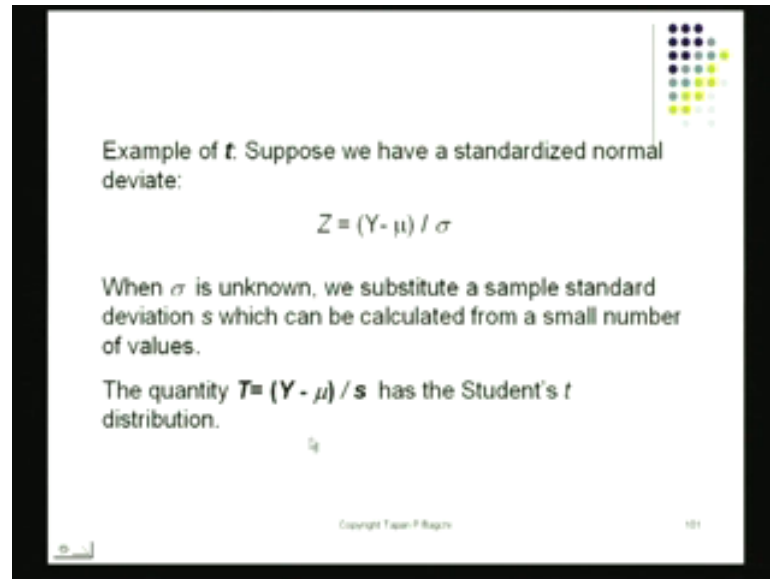
- An important sampling distribution that can be defined in terms of the normal distribution.
- If Z and χ_k^2 are independent standard normal and chi-square random variates, respectively, then the random variable T follows the Student's t distribution with k d.o.f

$$T = \frac{Z}{\sqrt{\chi_k^2 / k}}$$

Copyright Tarek P. Maghrebi 100

Then of course, if I complete this quantity that will have the T distribution. Now, the T distribution is a lot like the normal distribution, like the standard normal distribution. It also rises then it falls like this, but, it is got slightly different shape. It is slightly flat as compare to the standard normal distribution. The T distribution is useful in many different ways.

(Refer Slide Time: 29:23)



Example of t . Suppose we have a standardized normal deviate:

$$Z = (Y - \mu) / \sigma$$

When σ is unknown, we substitute a sample standard deviation s which can be calculated from a small number of values.

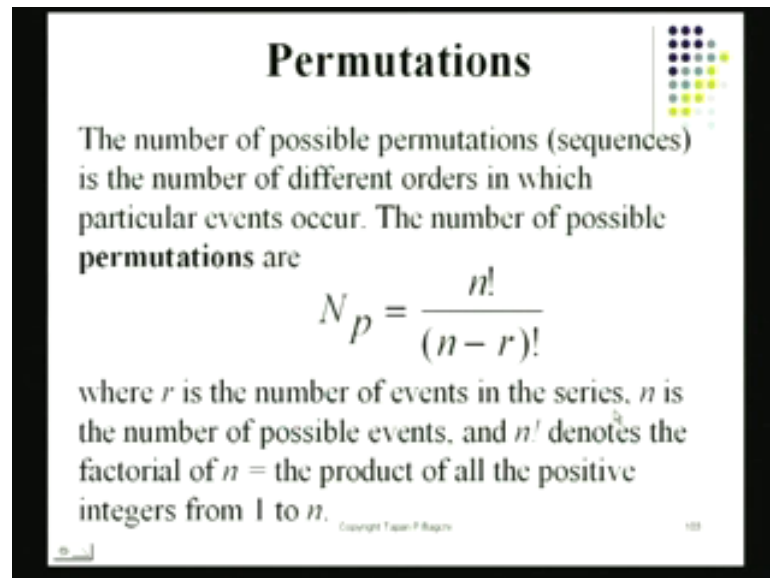
The quantity $T = (Y - \mu) / s$ has the Student's t distribution.

Copyright Team P. Raju 101

If I take this little transformation of my raw data, let us say my raw data was Y and its mean was μ and its standard deviation was σ . And if I converted all the Y values into Z values like this. These values these quantities will have a normal distribution, but, suppose I did not know σ ? I did not know this σ ? I did not know the population, standard deviation? And in place of that I used S . S is my sample standard deviation. If I use this then I get a new quantity, this quantity conforms to the T distribution. So, there is a difference now I am going away from this Z distribution or the zee distribution to a T distribution. And again I have got tables for the zee distribution; I have also got tables for the t distribution. And again my I can do my probability calculation because I know these distributions very well.

The T distribution there is just one difference the Z distribution does not have any degrees of freedom consideration, but, the T distribution will require you to figure out how many data points you began with? You started with and it will give you, it will ask you for the degrees of freedom because it uses this quantity called s and s obviously has a degrees of freedom. S will have a degrees of freedom. S has a degrees of freedom of the n minus 1 and T will have appropriately pick up the same degrees of freedom there.

(Refer Slide Time: 30:55)



Permutations

The number of possible permutations (sequences) is the number of different orders in which particular events occur. The number of possible permutations are

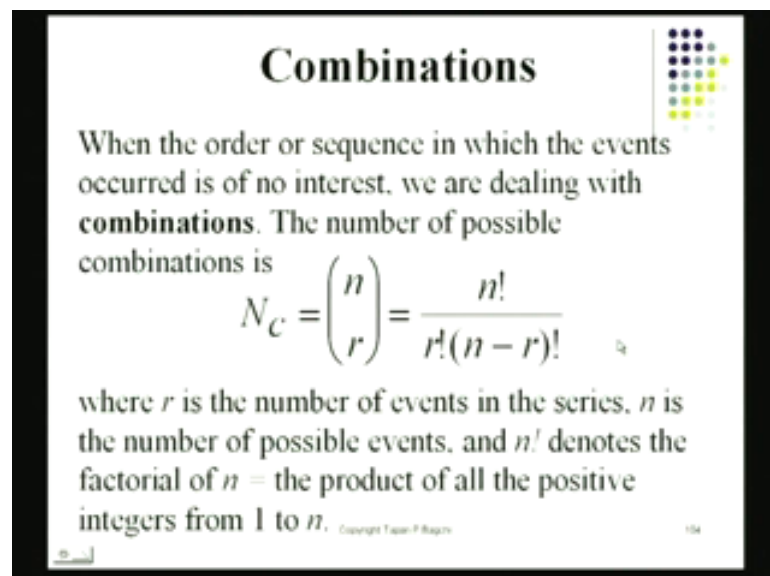
$$N_p = \frac{n!}{(n-r)!}$$

where r is the number of events in the series, n is the number of possible events, and $n!$ denotes the factorial of n = the product of all the positive integers from 1 to n .

Copyright Team P. Bapat

Finding probability by counting events how do I do that? I could do that by doing permutation calculations which is like this. Permutation is important when you worried about the sequence. If I am worried about the sequence like head, tail, head, tail and so on. If you worried about the sequence red, blue, green, red, blue, green. If you are worried about the sequence then you should be using permutation as your calculation bases.

(Refer Slide Time: 31:19)



Combinations

When the order or sequence in which the events occurred is of no interest, we are dealing with **combinations**. The number of possible combinations is

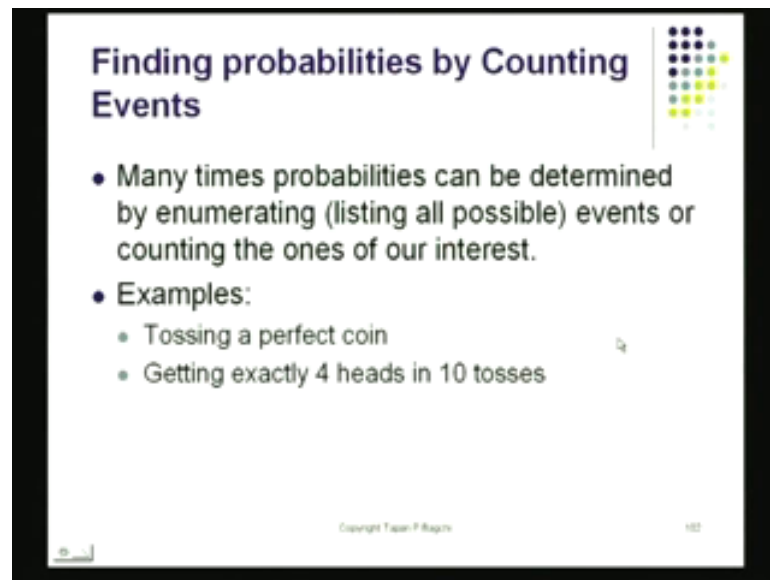
$$N_c = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

where r is the number of events in the series, n is the number of possible events, and $n!$ denotes the factorial of n = the product of all the positive integers from 1 to n .

Copyright Team P. Bapat

If in the other situation, when you are really worried about counting the number of reds and the number of blues and the number of blacks. All you really care is I want to know how many reds are there? How many blacks are there? You will be using combination and the formula changes.

(Refer Slide Time: 31:40)



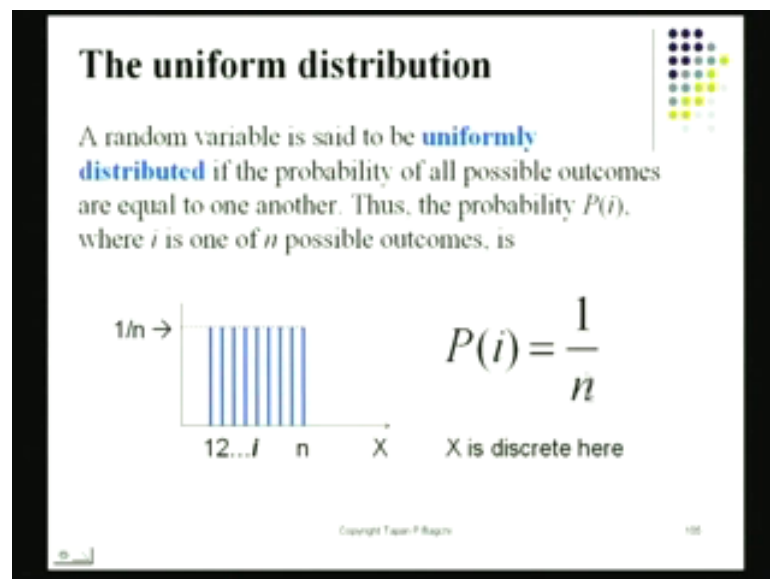
Finding probabilities by Counting Events

- Many times probabilities can be determined by enumerating (listing all possible) events or counting the ones of our interest.
- Examples:
 - Tossing a perfect coin
 - Getting exactly 4 heads in 10 tosses

Copyright Taim F. Rajabi 102

The formula becomes the formula that is shown on the thing there. Why are these calculations important the permutation calculations so on because many times when I am trying to calculate probabilities.

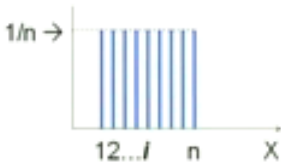
(Refer Slide Time: 31:49)



The uniform distribution

A random variable is said to be **uniformly distributed** if the probability of all possible outcomes are equal to one another. Thus, the probability $P(i)$, where i is one of n possible outcomes, is

$1/n \rightarrow$



$P(i) = \frac{1}{n}$

X is discrete here

Copyright Taim F. Rajabi 103

I will be using either a permutation basis or a combination basis to be able to do it. The simplest case of course, is the uniform distribution in which case I have a discrete random variable which can take values from 1 to n. And because it is uniformed the probability of the same fault finding the values either 1 or 2 or 3 or whatever it is and the probability of finding any one digit to be the realization or the actual number is 1 over n. That is the uniform distribution. You could also have a continuous version of the same distribution.

(Refer Slide Time: 32:27)

Frequency Distribution:

Scores for an Six Sigma class are as follows: 58, 95, 80, 75, 68, 97, 60, 85, 75, 88, 90, 78, 62, 83, 73, 70, 70, 85, 65, 75, 53, 62, 56, 72, 79

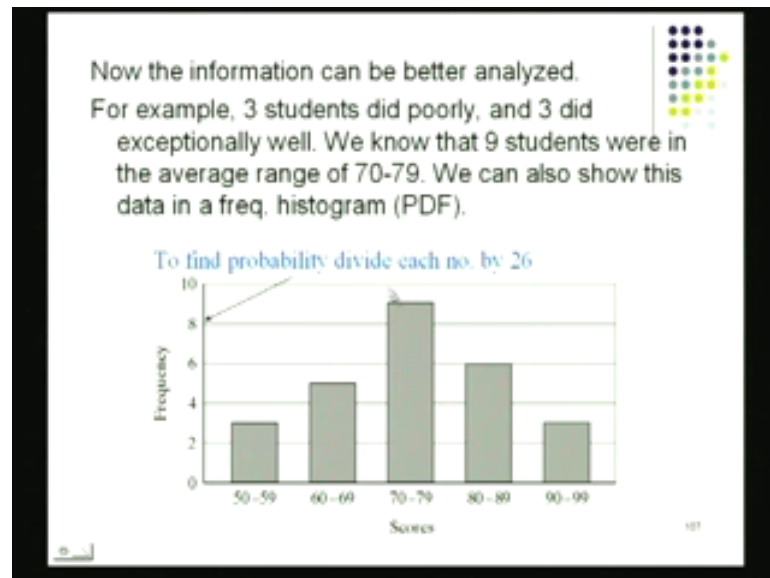
To better assess the success of the class, we make a frequency chart:

TABLE 19.1 Grouped Frequency Distribution for Example 19.2

Scores	Range	Frequency
58, 53, 56	50-59	3
68, 60, 62, 65, 62	60-69	5
75, 75, 78, 73, 70, 70, 75, 72, 79	70-79	9
80, 85, 88, 83, 85, 87	80-89	6
95, 97, 90	90-99	3

In which case you will have the value of f remaining the same going from one to the other end. Our frequency distribution these can be plotted very easily. These can be calculated very easily if I have frequency data given to me and which in some cases is quite possible. I can construct a histogram and in some cases I would like to work out the probabilities.

(Refer Slide Time: 32:38)



And for that I really have to divide the total number of frequencies. Total frequencies by these individual frequency and I will end up with a frequency for this probability, for this a probability for this and so on and so forth. That is the light weight of finding my random variable in this range or this range or this range and so on.

(Refer Slide Time: 33:09)

Cumulative Frequency

- The data can be further organized by calculating the cumulative frequency (CDF).
- The cumulative frequency shows the cumulative number of students with scores up to and including those in the given range. Usually we normalize the data - divide 26.

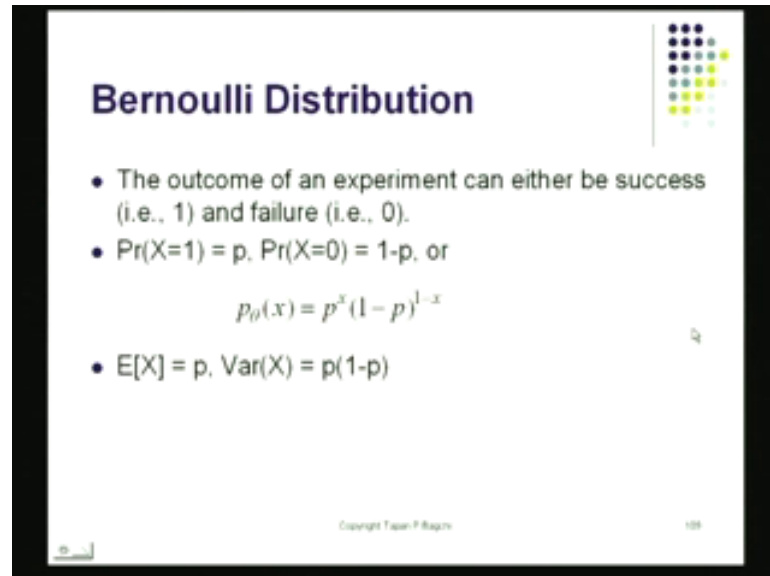
TABLE 19.2: Cumulative Frequency Distribution for Example 19.2

Range	Frequency	Cumulative Frequency
50-59	3	3
60-69	5	3 + 5 = 8
70-79	9	3 + 5 + 9 = 17 or 8 + 9 = 17
80-89	6	3 + 5 + 9 + 6 = 23 or 17 + 6 = 23
90-99	3	3 + 5 + 9 + 6 + 3 = 26 or 23 + 3 = 26

That I could do just by dividing each quantity by the appropriate count of the total count of the thing that could be done. There is something fairly simple and easy to do and it would not be that difficult. And similarly, I can work out the cumulative frequency also.

This we will look if you look up any text book of the statistics of probability. They will give you these ideas.

(Refer Slide Time: 33:23)



The slide is titled "Bernoulli Distribution" in a bold, dark blue font. In the top right corner, there is a decorative graphic of a grid of colored dots in shades of blue, green, and yellow. The main content consists of three bullet points: the first states that an experiment can result in success (1) or failure (0); the second lists the probabilities $\Pr(X=1) = p$ and $\Pr(X=0) = 1-p$; and the third gives the expected value $E[X] = p$ and variance $\text{Var}(X) = p(1-p)$. A mathematical formula $p_0(x) = p^x(1-p)^{1-x}$ is centered on the slide. At the bottom, there is a small copyright notice "Copyright Team P. Bagnoli" and a page number "108".

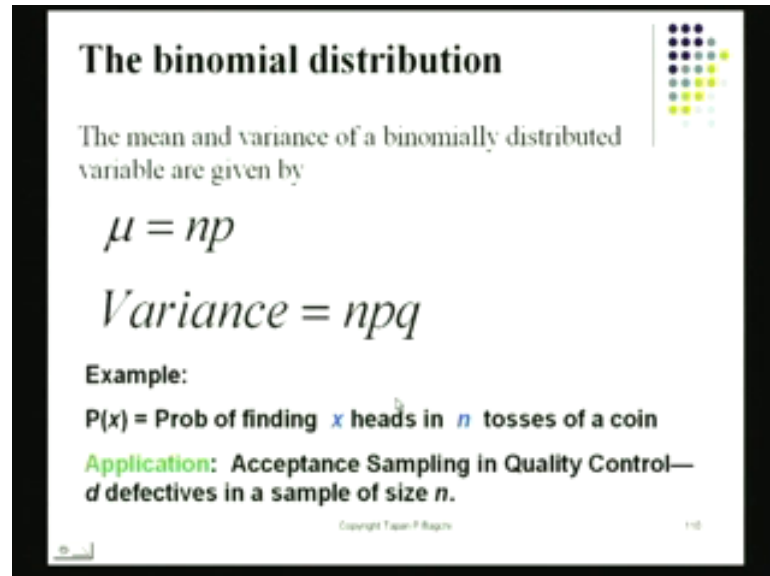
The Bernoulli distribution turns out to be coin tossing situation. Remember we tossed a coin the distribution of coin toss or the outcomes of the coin toss the distribution of that comes out by the Bernoulli distribution. In the case of the Bernoulli distribution we call the chance of success to be P. In coin toss the value of P turns out to be approximately 0.5 that is the only difference.

So, coin toss is certainly not a very general situation. It is a very specific situation when my coin is balanced and when I flip it, I flip it randomly and when I see the data and when I see the outcome of it. The outcome turns out to be precised as a head or tail and there is nothing in between and the chance of either finding a head or finding a tail that turns out to be 0.5 in each case. In the Bernoulli case of course, I have the chance of a success to be P and I define the random variable to be having value 1 would probability P and value 0 would probability P probability 1 minus P. 1 minus P is sometimes also written as q.

So, I have a chance of P or chance of q. The probability of finding either 0 or 1 and that is given by this little formula and this formula comes from the Bernoulli distribution. And the expected value for the bernoulli random variable which can only have a value of

1 or 0. That turns out to be P and the variance turns out to be P times 1 minus P or P times q .

(Refer Slide Time: 35:58)



The binomial distribution

The mean and variance of a binomially distributed variable are given by

$$\mu = np$$
$$\text{Variance} = npq$$

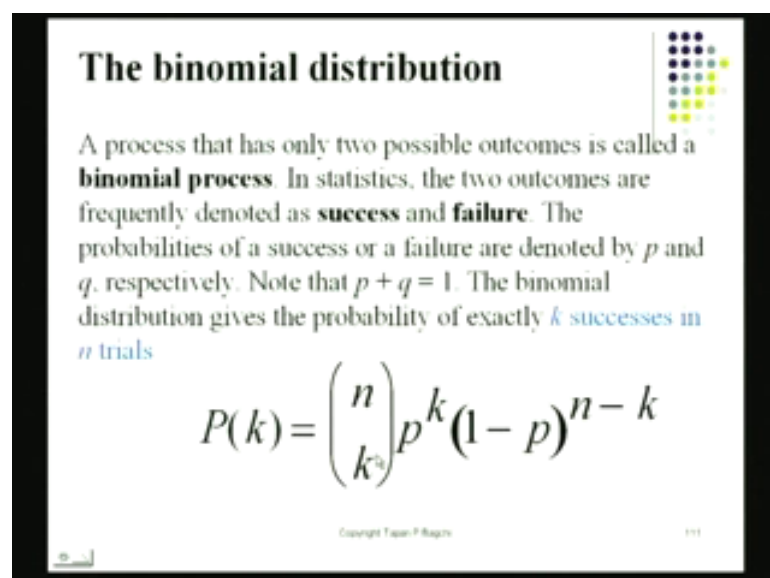
Example:
 $P(x)$ = Prob of finding x heads in n tosses of a coin

Application: Acceptance Sampling in Quality Control— d defectives in a sample of size n .

Copyright Tapan P. Bagchi 110

The binomial distribution is an extension of the Bernoulli distribution. What we are doing there is? We are tossing the coin not only once, but, we are tossing it n number of times and we are counting the number of heads. So, if you look at the Bernoulli distribution I may actually writing down the probability of finding k success in n trials.

(Refer Slide Time: 35:17)



The binomial distribution

A process that has only two possible outcomes is called a **binomial process**. In statistics, the two outcomes are frequently denoted as **success** and **failure**. The probabilities of a success or a failure are denoted by p and q , respectively. Note that $p + q = 1$. The binomial distribution gives the probability of exactly k successes in n trials

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Copyright Tapan P. Bagchi 111

So, I toss the coin n number of times and I count the number of heads. This distribution the number of finding the number of heads if that is equal to k and if I have tossed n times. Then the probability of that probability of finding k heads in n trial that will be given by this formula P_k is n chose k p raise to our k and 1 minus p raise to our n minus k . This is the chance of my finding tails and this is the chance of my finding k heads and this is the different ways I can construct k different heads out of n .

The first k could be heads or anyone of them counting total counting to the total of for there being k heads in n trials that also could be a count of k then or k could be just the last k pieces, last k tosses those also could be k . So, the various ways I could generate k successors or k heads in n trials. That is why I have got this combination. I have got this combination part here. This term actually takes care of those possibilities there.

(Refer Slide Time: 36:29)

Binomial Distribution

- n draws of a Bernoulli distribution
 - $X_i \sim \text{Bernoulli}(p)$, $X = \sum_{i=1}^n X_i$, $X \sim \text{Bin}(p, n)$
- Random variable X stands for the number of times that experiments are successful.

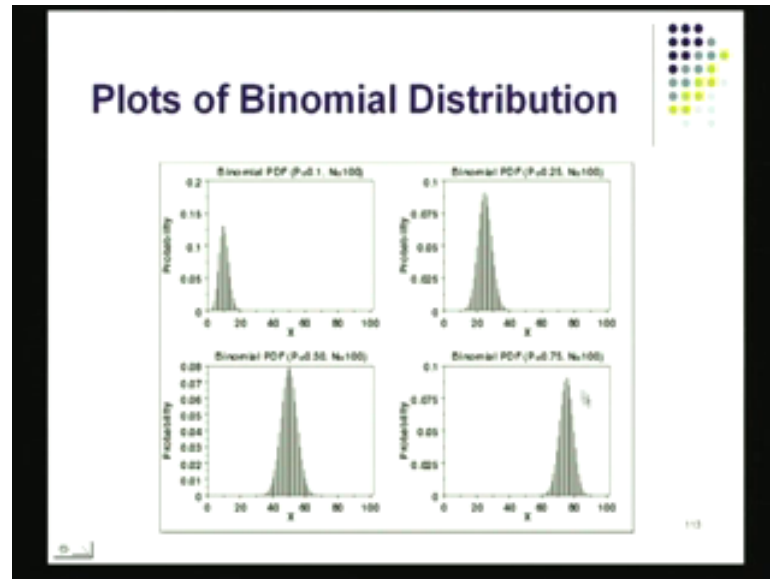
$$\Pr(X = x) = p_x(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] = np$, $\text{Var}(X) = np(1-p)$

Copyright Team P. Rajan 112

And the binomial formula again I am giving you the same binomial formula here again. This gives you more clearly what that formula is for finding a head and notice something here something is very important and very interesting.

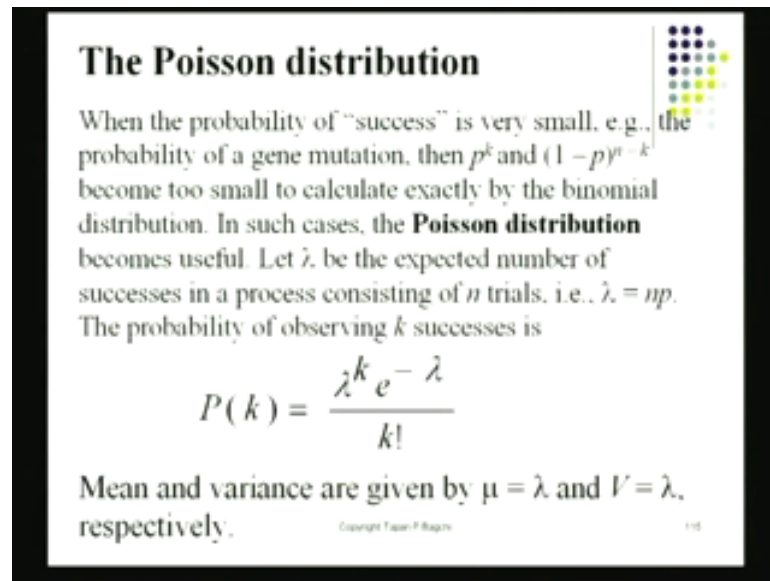
(Refer Slide Time: 36:37)



Under certain conditions can you think of what this shape is now resembling? I started with some shape and I started going to higher and higher values. Look at this shape of the distribution what does this remind you of? Well it reminds me of the normal distribution. It turns out in some cases the normal distribution turns out to be a pretty decent approximation for the binomial distribution. So, if you get tired of calculating all the probabilities for finding a head and tail and so on using the binomial formula. Just use the approximate normal formula.

You will use the same mean which is like np and you will use the same variance sigma square. In this case it will be equal to npq and end up with the same pretty well. The same probabilities that you will calculate by calculating the binomial probability quantity which is like probability of finding k successes in n trials. You could work that out using the normal distribution also. Then there is this distribution called the poisson distribution which is also very useful in statistics.

(Refer Slide Time: 37:57)



The Poisson distribution

When the probability of "success" is very small, e.g., the probability of a gene mutation, then p^k and $(1-p)^{n-k}$ become too small to calculate exactly by the binomial distribution. In such cases, the **Poisson distribution** becomes useful. Let λ be the expected number of successes in a process consisting of n trials, i.e., $\lambda = np$. The probability of observing k successes is

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

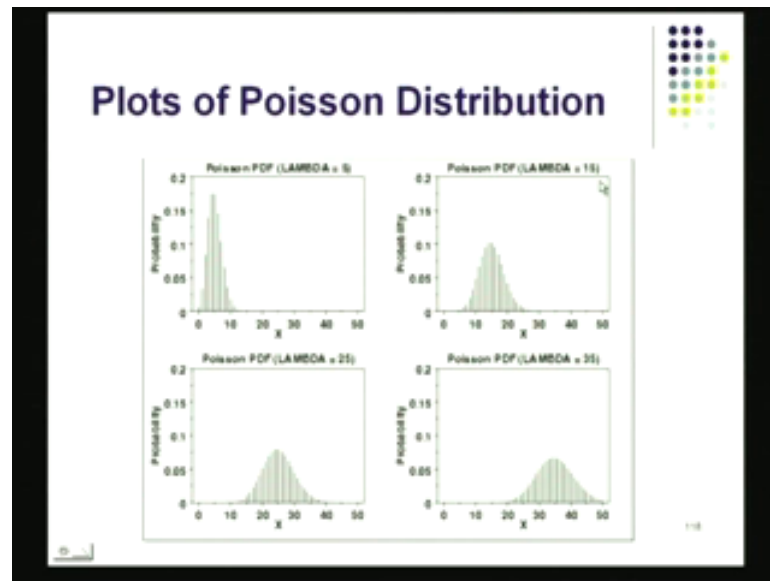
Mean and variance are given by $\mu = \lambda$ and $V = \lambda$, respectively.

Copyright Team P-Maps 148

How do we use this? It turns out if we take that quantity p . Remember that quantity p that we had in the binomial distribution. Suppose, this quantity p became smaller and smaller which is like if p is the fraction defective, fraction of items defective in a lot. And if this p becomes smaller and smaller quality is obviously improving and then suppose I want to set up a system whereby I will be controlling quality by monitoring the number of defectives found in a sample of size n that I collect. In that case I need not really use even the binomial distribution. I can use the poisson distribution which is got the simpler formula and in many books you will find this actually given as a table to you.

The difference between use of the poisson distribution and the binomial distribution is this. In the binomial case you need this you need the size of the sample n . In using the poisson formula you do not need that. So, you can get very close to what we had before? You can get more or less the same numbers in terms of probability estimates using the poisson distribution.

(Refer Slide Time: 39:06)



As you take a look at the Poisson distribution, when it comes to the shape of the distribution, Oh my god again the last distribution begins to resemble the normal distribution. And let me show you something else that is very interesting with the normal distribution with the Poisson distribution. In the Poisson case the expected value is λ and the variance is also λ .

(Refer Slide Time: 39:15)

Poisson Distribution

- Coming from Binomial distribution
 - > Fix the expectation $\lambda=np$
 - > Let the number of trials $n \rightarrow \infty$

A Binomial distribution will become a Poisson distribution

$$\Pr(X = x) = p_x(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- $E[X] = \lambda, \text{Var}(X) = \lambda$

So, in fact, for the Poisson distributions one of the few distributions, which has the same variance as the mean? It is the mean and the variance; they are numerically the same. If I

have this I can go out and I can use the normal distribution because, I have an estimate of the normal of the mean which is lamda and also I have an estimate of the variance which is again lamda. From variance sigma square I can estimate sigma which is basically a square root of lamda. For mean I have already got lamda. So, in this if I have to work out a normal distribution approximation with a poisson distribution. I will use a mean of lamda. And I will use a standard deviation of square root of mean and I will be able to use the normal tables now. That is like a big big jump in terms of efficiency.

(Refer Slide Time: 40:18)

The Normal Distribution

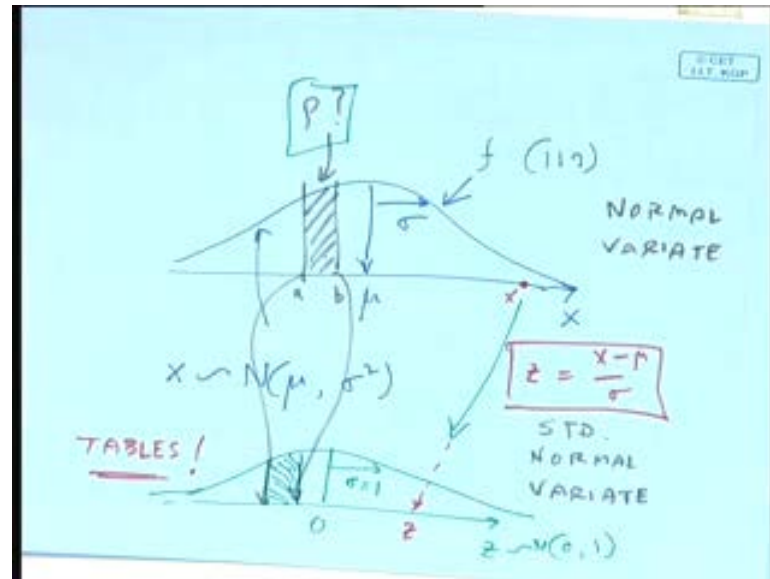
- "It is a symmetric curve with its highest ordinate at its center, tailing off to zero in both directions in a way that is intuitively expected of **experimental error**."
- If Y is a normal random variable, then the pdf of Y

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \quad -\infty < y < \infty$$

where $-\infty < \mu < \infty$ is the mean of the distribution and $\sigma^2 > 0$ is the variance.

Now, let us bring the normal distribution in place. We have we have only talked about the normal distribution. It is got actually a very wide presence in the world almost any measurement that you do it actually has this normal distribution. I just found that this formula here, pretty closely in our resembled that bell curve, you remember the bell curve I have been drawing. I have been just drawing the bell curve all over the place; I have been drawing the bell curve like this.

(Refer Slide Time: 40:47)



This shape and the formula that I show here, this formula actually describe this shape. This shape will describe that f formula that I showed you in on slide 119, slides 119. If you look at slide 119 that slide shows you the formula and f is the same as f has this shape. When this was found, it was found that I could really work out the mean of this, mean is μ as σ is here. These two are the typical parameters of a normal distribution which is designated by this μ σ^2 . This is basically a short hand notation of a normal distribution which is like this.

So, if this is the distribution of random variable X . I will denote it like this, X has this distribution, which is normal which is like this. The normal distribution has some very very decent properties and I might be able to tell you a few of them. First of all let me tell you, if I have X distributors as normal with mean μ and variance σ^2 . I cannot always find a table that will give me the probabilities for this μ and this σ^2 . I may not be able to find it.

Instead, what I can do is? I can convert this into another distribution which is the standard normal distribution like this. And this distribution is called the Z distribution, which is a mean of 0 and there is standard deviation σ equal to 1. So, this Z is now distributed as $N(0, 1)$. That is the issue you have Z and this Z random variable is called the standard normal variate. I will write this down. This is just a normal variate (No Audio Time: 42:41 to 42:49) and this guy is the standard normal variate.

The good thing is that any of these can be converted into a standard normal variate with a simple transversion. Which really says z is equal to x minus mu by sigma. If you use this formula any of these points from here can be converted to a correspondingly Z value here, any value of X can be converted into this by this formula.

The beauty of this is once I have converted into this standard normal variate form, I have tables for it. So, there are tables, my god there are tables. There are tables all over for the normal distribution. And these tables provide you the c d f and c d f is the cumulative distribution function. With that you can find for example, the probability of some quantity falling in this range. If you want to say there are values here a and b and what is this probability P? That you can find very easily convert a to z value and convert b to z value. And the just basically look up this area, look up this area from the table and that area and this area will be the same. And we will end up finding the answer for this. This is very easy, this is not very difficult. Once you do a few examples you would be able to do this almost any time.

(Refer Slide Time: 44:27)

Normal (Gaussian) Distribution

- $X \sim N(\mu, \sigma)$

$$p_D(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$\Pr(a \leq X \leq b) = \int_a^b p_D(x) dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

- $E[X] = \mu, \text{Var}(X) = \sigma^2$
- If $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$, $X = X_1 + X_2$?

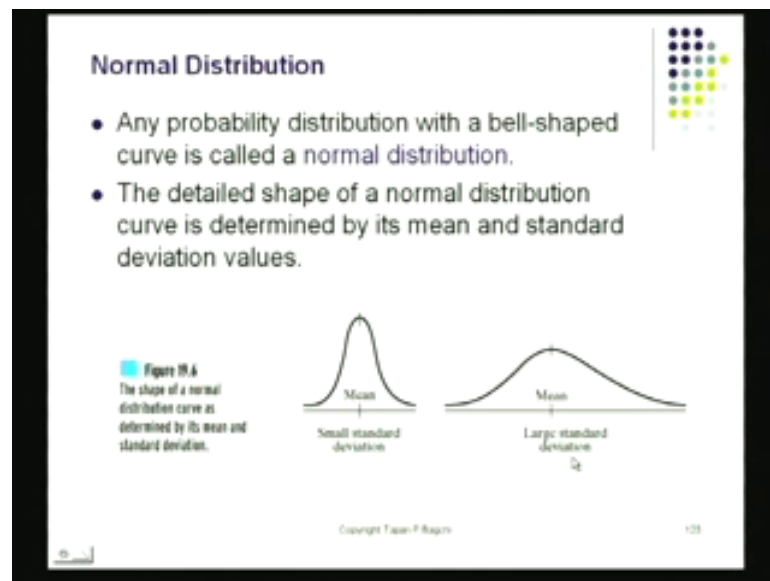
Copyright Tarek P. Magry 121

So, the normal distribution it turns out to be this way I have shown you. First I have shown you the density. This is the density of the normal distribution. And I have shown you the range; this is now the area that I showed you. This area is equal to this quantity here; this quantity here is the same as that area there. And of course, I have got my expectation I have got my standard I have got my variance those are there. And it turns

out to two random variables, one is X_1 distributed like this and X_2 is distributed like this. What does this guy turned out to be?

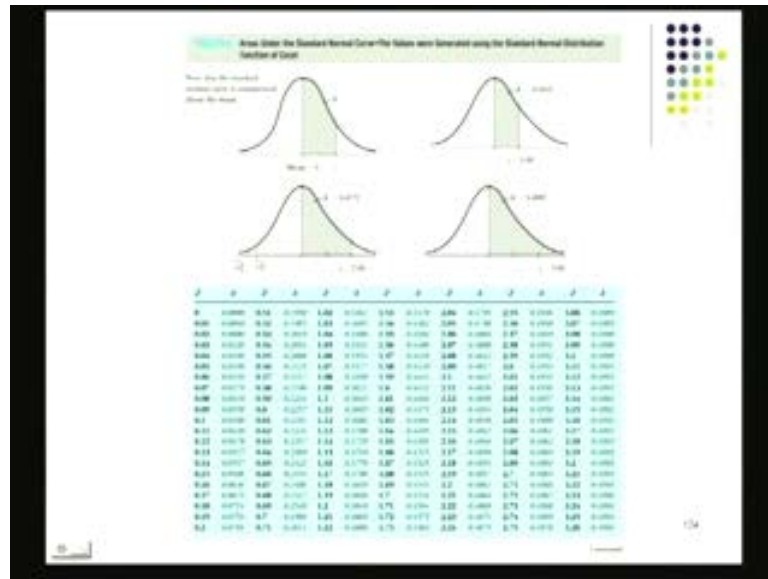
It turns out to be also a random variable which is also normal. It will have $\mu_1 + \mu_2$ as its mean and the variance will turned out to be $\sigma_1^2 + \sigma_2^2$. And that is if X_1 and X_2 are independent, if they are not independent I will end up with the covariance term. That is something that we will study later that we will try to do.

(Refer Slide Time: 45:29)



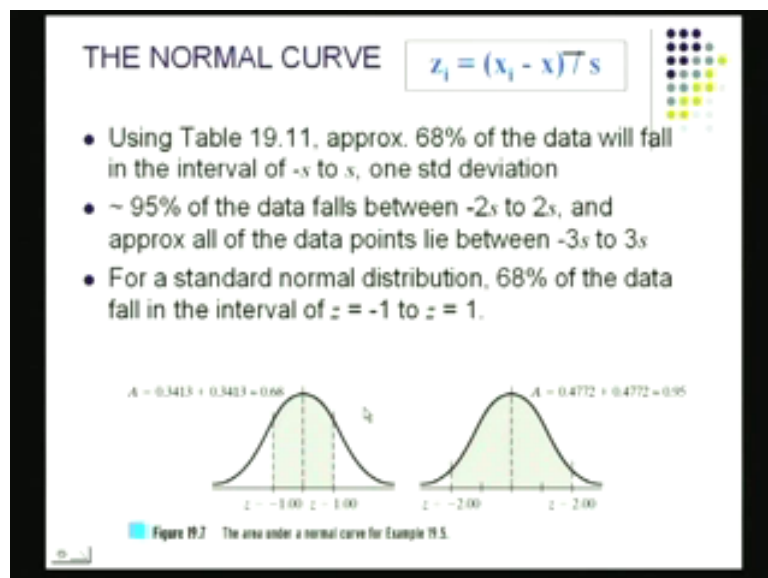
So, the normal distribution turns out to be very important distribution and it is really the basis for many of the test that we had done that are done. Using for example, estimates of parameters and so on and so forth. They always come back to the normal distribution and there tried to work with this and tables are available mentioned. I remember I mentioned tables to you.

(Refer Slide Time: 45:44)



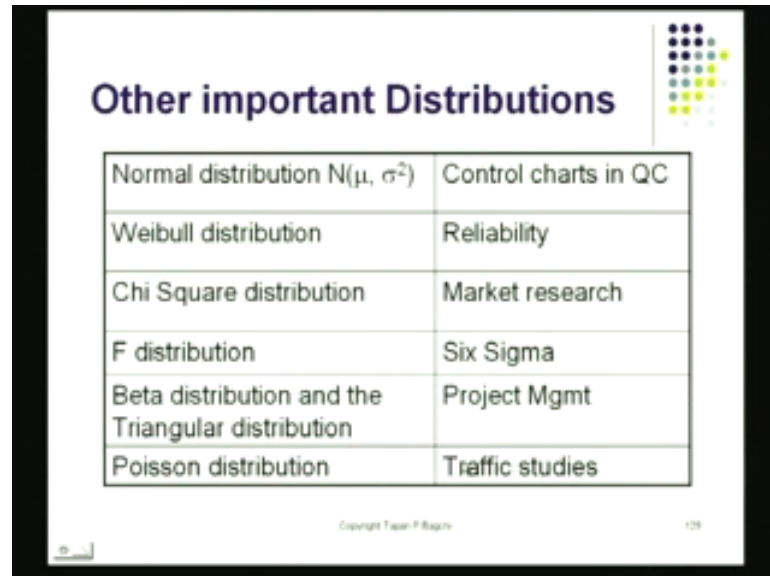
These are the kind of tables that you will see once you get into textbook. You go to at the back of the textbook it will give you the cumulative distribution. It will give you the area, A is the area and it will give you the corresponding value of Z and sometimes this z is calculated from the middle sometimes it is calculated from the left. But, the book will tell you how we are calculating these things? The book will tell you exactly how to get the correct area?

(Refer Slide Time: 46:11)



So, these are all available and with this you can very easily workout the areas of your interest or the probabilities of your interest that can be done quite easily.

(Refer Slide Time: 46:23)



Other important Distributions	
Normal distribution $N(\mu, \sigma^2)$	Control charts in QC
Weibull distribution	Reliability
Chi Square distribution	Market research
F distribution	Six Sigma
Beta distribution and the Triangular distribution	Project Mgmt
Poisson distribution	Traffic studies

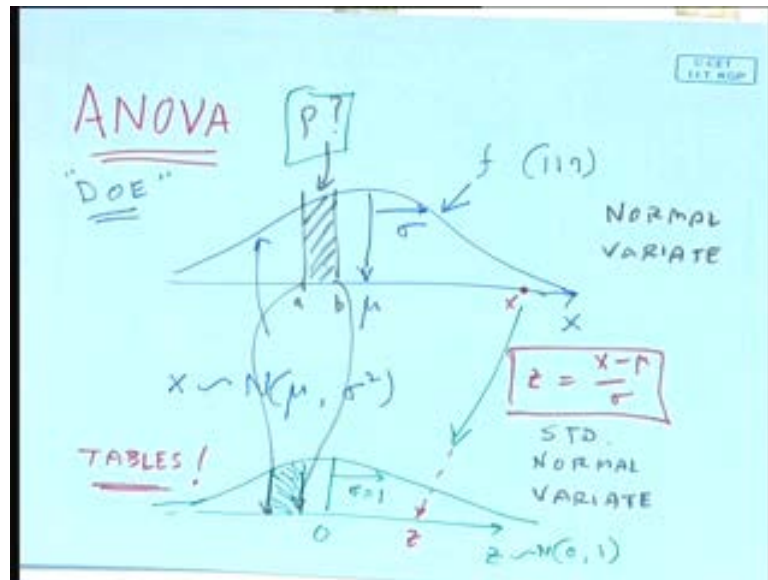
What are some of the other distributions that are of interest to us and that is quality assurance people? The normal distribution is used very heavily in controlled chart, controlled chart development. Controlled chart work which is like the \bar{x} chart and the r chart and many other charts. Many very often the normal distribution is used there by the evoking the central limit theorem. So, that is like something where the controlled limits are calculated using the normal distribution assumption.

Then we have another distribution called the weibull distribution. This is a bit more complicated and this is used when we described the failure times of objects. For example, the tube lights in this room or the P C or the highdry, hard drive of the P C or even some objects which is running on the road or the failure time of for example, tires on your car. For example, if it track down the failure times of these things. Many of these failure times they will tend to appear to be weibull distributed. And weibull has been the basis now for a modeling failure times distribution in large variety of cases. And in liability theory the weibull distribution is used a lot. So, that is also very important distribution the weibull distribution is used in liability.

Then the chi-square distribution I mentioned to you earlier. That is used in market research and some for checking independence and bunch over the stuff distributions and

so on and so forth. That is the actually done quite often in statistics when you collect data and you want to do some tests on it. Many times the chi-square test is done. Then there is a very important test called the f test and the f test is used alongside another little new concept called ANOVA analysis of variance.

(Refer Slide Time: 48:12)



And this I am going to discuss with you later on. Much later I will be discussing a technique called designed of experiment DOE. DOE is a statistical technique which is used and it uses a data analysis method which is called ANOVA by we check finds out if there are multiple factors affecting a particular process. Which of these factors really has a significant effect on the final process? That is done using the DOE scheme and the ANOVA data analysis preceded.

Then in project management when you got job times and I am coming back to the slide here again. In project management many times people you will ask people about when you are trying to for example, workout the critical path network. In the critical path network you will require job completion times. So, you will have task times task a task b task c and you will need there completion time.

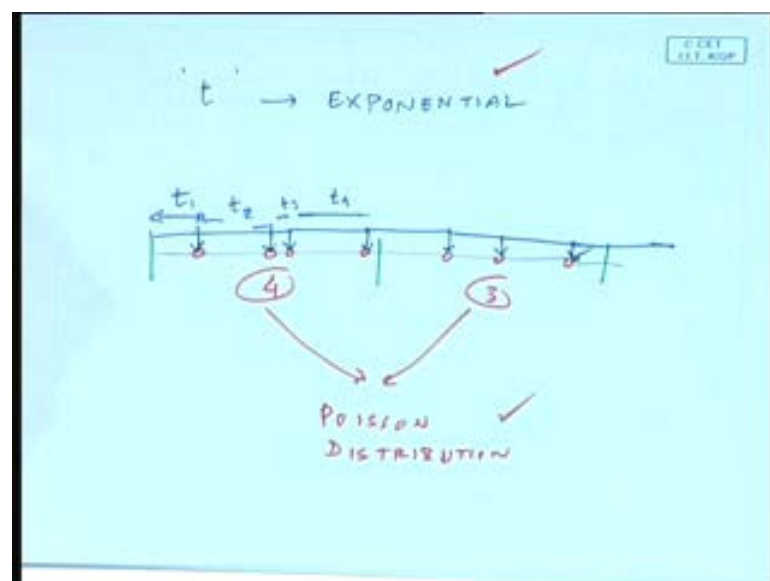
If you ask subcontractors, if you ask contractors and other people including people who do software coding or software testing. If you ask them how much time do you think this particular job is going to do? I have some job in my hand I am going to give it to you please give the time estimate to complete this task. They will probably not be able to

give you an exact time, but, if you ask them. They will say sir it looks like optimistically I can finish it in two days. Pessimistically when things really fall up I may take 7 days for it.

But most likely I will be able to complete it in 4 days. See I have got 2 days, 4 days and 7 days as three estimates. This leads to a different distribution, that distribution is called a beta distribution. It is also called the triangular distribution and this is used in project management in doing porter analysis and so on. We use the beta distribution or the triangular distribution. Then of course, in traffic studies we already saw the poisson distribution. We also saw the formula for it.

In traffic studies for example, for designing you know facilities in a hospital or you controlling traffic on the road and so on and so forth. Or in queuing your applications, queuing theory applications we use a distribution that is called the poisson distribution. The poisson distribution and there is very nice distribution. That is used also very very commonly that is called the exponential distribution. They have a correspondence, the exponential distribution leads to time between events, which is exponentially distributed and if you count the number of events number have spent per hour. That count turns to the passion distribution. So, there are two things we are talking about if you have a time axis.

(Refer Slide Time: 51:13)

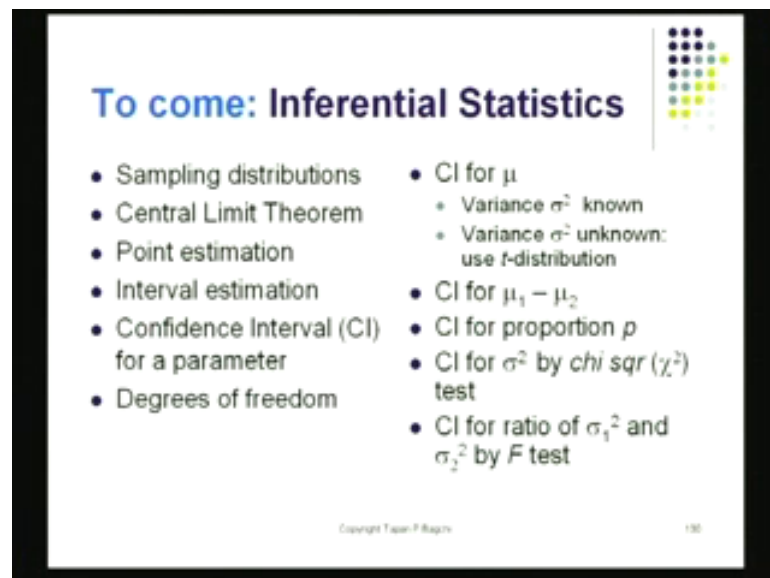


And if you have a time axis and time axis goes like look for say the time axis goes like this. And what I am doing here is? I am counting the number really event taking place and I have got some marks, time marks and I have got time marks like this. In that case suppose I am this is a traffic situation I am waiting for patients to arrive. First patient arrive there, second arrive there, third arrive there, fourth arrive there, fifth arrive there and so on. They are arriving randomly at different times.

If I look at the time between people t_1 , t_2 then t_3 , then t_4 and so on. If I look at the distribution of these times this distribution is exponential. On the other hand if I count the number of events taking place first event, second event, third event, fourth event then here again 1 2 3. So, here I have 4 events, here I have 3 events and so on. This also is the random variable and these will have a poisson distribution.

So, there is a correspondence now between these two and this also we exploit when we get into our quality assurance situations or other situations we get into this. Some other places where we utilize statistics it will be probably into drawing inferences.

(Refer Slide Time: 52:56)



And that shown in the slide here the inferential statistics that is like one place where we use statistics very heavily and we use a lot of distributions like this. For example, for finding the confidence interval for a mean. We will be using for example; we will be using *d l zee* distribution, the *Z* distribution. If I am trying to find the confidence interval for a variance I will be using the *chi-square* distribution. Those are like special

applications where we will be using the appropriate distributions there. Inference is basically saying something about the population based on some sample data that I have collected, that is inferencing.

(Refer Slide Time: 53:31)

To come: Test of Hypothesis

- Null hypothesis H_0
- Alternate hypothesis H_1 (or H_a)
- Test statistic
- Single-tailed test
- Two-tailed test
- Type I error
- Type II error
- Test for difference of means
- t -test
- χ^2 test
- F test

Copyright Tamer P. Bagchi 191

There is another situation when I want to say something about the process or I want to say something about the process parameter.

(Refer Slide Time: 53:43)

$t \rightarrow$ EXPONENTIAL ✓

t_1, t_2, t_3, t_4

④ ③

$H_0: \lambda = 5 ?$
 $H_1: \lambda \neq 5 ?$

POISSON DISTRIBUTION ✓

And I say that the value of lamda is equal to 5 or the value of lamda is not equal to 5. Lamda is the let us say the mean arrival rate which is what I what leads to my poisson

distribution. If this is the question I have got one hypothesis here, another hypothesis here. If I am testing one hypothesis which is a guess against another guess I will be using a principle called the test of hypothesis which I am going to describe to you later on.

This again uses some statistical distributions and those are now dependent on the kind of assumptions you make and the kind of test you construct. And that test could be a t test when you are really test the difference between the two sample means or if you are looking at the difference between two sample variances you would be using the chi-square test. If you look in the ratio of two variances you will be using the f test.

If you are looking at this situation when you want to check out whether two events are independent you have collected a lot of data. You would be using the chi-square test. That is the kind of test we will be using and in whenever you are doing a test like this there is the possibility of committing an error. It could be a when the hypothesis is actually true, but you reject it. You are saying no, it is false. It is true but, you are rejecting. You are rejecting a good hypothesis. That is a type 1 error and if you accept a false hypothesis that is a type 2 error.

Generally speaking in test of hypothesis we try to keep the error percent, the fraction of errors committed as type 1 or type 2 as low as possible. Then only we can get good sound decision using probability theory or using the application of statistics. We actually have to count on the sizing on the sample and also we have to look at something called the significance of the test. Significance is given by alpha interest of hypothesis and the power of the test which is like the power of rejecting the false hypothesis that comes through a concept called one minus beta. That is also done in the same way in doing this.

So, we will be looking at those as we move into this. We will have really not much difficulty doing that, but, when I come to those applications stages as we move through this six sigma lecture. These six sigma lectures we will be reviewing those I will be bring up some numerical examples also for doing them. Thank you very much. I am going to be see you next time in the next session. Thank you.