

Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 21
Normality Test/Odds Ratio

Welcome to the course on Biostatistics and Design of Experiments. We will talk about something different today, that is called the Normality test. Because many of these tests you know, assumes that the distribution which you are having is a Normal distribution.

That means, if I take 100 students in my class and measure their heights and then plot it, in mostly it will follow a Normal distribution, and you know what is Normal Distribution? Mean, Median, Mode, 3 Ms. Mean, Median, Mode should be equal. And if I take the average performance of the class and plot it with that of all the students, mostly it should follow a normal distribution. So many of this statistical test assumes that the distribution which you are having is a Normal distribution. So definitely we need to check whether the data which we have is a Normal distribution. So, that is what is called the Normality test.

There is something called the Skewness that means, the way the distribution has moved to the right or left and there is also something called the Kurtosis, which is an indication the Peakedness of the data, is it too much peak or it is very flat. So, these are indications of Non-normal distribution. Let us look at some points.

(Refer Slide Time: 01:41)

Test for Normality

The null hypothesis that the data come from a normally distributed population

The test results indicate whether one should reject or fail to reject the null hypothesis

Types of normality tests

Anderson-Darling test compares the empirical cumulative distribution function of the sample data with the distribution expected if the data were normal. If the observed difference is adequately large, reject the null hypothesis

The slide contains four plots: a histogram with a normal curve, a histogram with a non-normal curve, a Q-Q plot with points on a diagonal line, and a Q-Q plot with points deviating from the diagonal line. The NPTEL logo is in the bottom left, and a URL is at the bottom right.

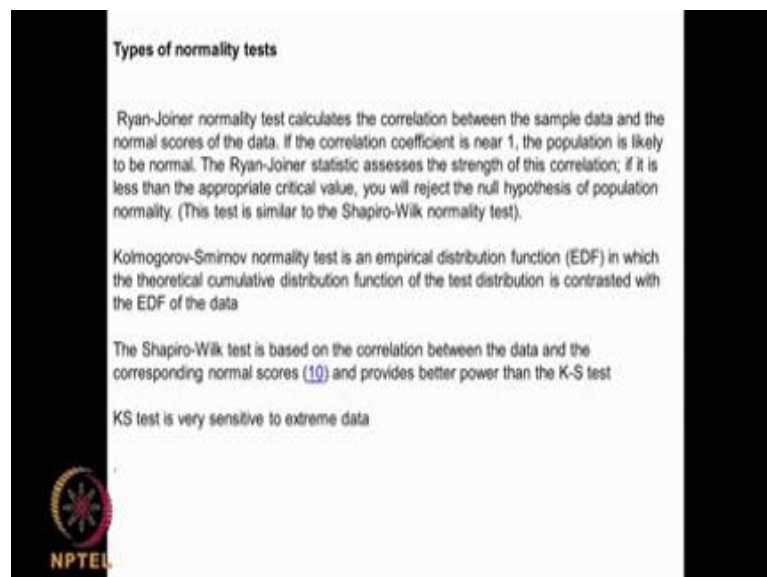
When you are testing for normality, the null hypothesis is the data is normally distributed from the, they are normally distributed population and alternate hypothesis is it is a Non-normal distribution. So you can from the test, accept the null hypothesis or reject the null hypothesis.

There are many types of normality test. Many of them, one is called the Anderson-Darling test. So what it does is, it compares the empirical cumulative distribution function of the sample data with the distribution expected the data were normal. The observed difference is inadequately large then reject the null hypothesis. So if we have a data set, from the data set you calculate a mean and a standard deviation and then you plot that with the data which you have. If it follows normal mostly the expected and the actual will follow on that line, then we could say it follows a Normal distribution. Like this you know, you will have a max and then on both sides you will have data falling down. So, we can say this data set follows a normal distribution.

Whereas we look at this data set, so you have ends in the lower region, large frequency of data or upper region you do not have. So, when you plot, you calculate this mean and standard deviation and you plot assuming a Normal distribution, what is expected and what is observed. So, in most of the cases the observed will be much less than what is expected, that is why these data points are below. The observed will be what is much less than expected. In fact, this example was taken from this particular reference here

actually. So, we can say because it does not follow what is expected, the observed is much, the expected is much below the observed because, most of the data is on the left-hand side here as you can see, obviously, this is a Non-normal distribution. Where as in this particular case we have maximum data in the center that is, maximum frequency on either side of the frequency goes down. That is a bell-shaped curve that is a uniform distribution. So, when you calculate mean and a standard deviation and plot what is expected on one axis and what is observed on another axis they seem to follow almost on that straight line, where as in this particular case they do not seem to follow, fall on that straight line. So obviously, we reject the null hypothesis, we can say that data is non-normal. So, as I said this was taken from this particular reference paper, very interesting.

(Refer Slide Time: 04:44)




Types of normality tests

Ryan-Joiner normality test calculates the correlation between the sample data and the normal scores of the data. If the correlation coefficient is near 1, the population is likely to be normal. The Ryan-Joiner statistic assesses the strength of this correlation; if it is less than the appropriate critical value, you will reject the null hypothesis of population normality. (This test is similar to the Shapiro-Wilk normality test).

Kolmogorov-Smirnov normality test is an empirical distribution function (EDF) in which the theoretical cumulative distribution function of the test distribution is contrasted with the EDF of the data

The Shapiro-Wilk test is based on the correlation between the data and the corresponding normal scores (10) and provides better power than the K-S test

KS test is very sensitive to extreme data



Now, in addition to this Anderson-Darling test, there are other tests like Ryan-Joiner test, a Kolmogorov-Smirnov test, Shapiro-Wilk test and so on. So, in Ryan joiner test, what it does is, it calculates a correlation between the sample data and the normal data. Sample data could be normal or non-normal and the normal data. And then if the correlation coefficient is very near 1, then we can say this particular data set comes from a normal population. As you know correlation coefficient is a measure of how closely the data is correlated to the normal data. If it is very less, obviously we can say generally correlation coefficient lies between 0 and 1, if it is very high and closer to 1, we can say that the sample data is normal population, otherwise we would say it does not come under the normal population.

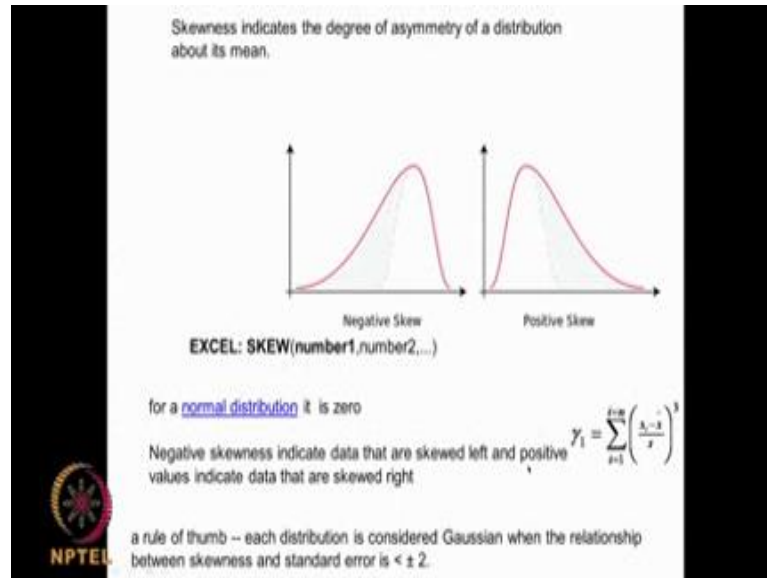
So, it basically looks at the strength of the correlation coefficient. So, you will reject the null hypothesis for normality if the correlation coefficient is large. Now you have the Kolmogorov-Smirnov normality test, it is an empirical distribution function in which the theoretical cumulative distribution function of the test distribution is contrasted with the EDF of the data. So, it calculates an empirical distribution function, for this sample and there is an empirical distribution function for the normal and then it compares actually.

This Shapiro-Wilk is almost similar to Ryan-Joiner and basically you are having a correlation between the data and the corresponding normal, so it is a better test than this particular. So, we have different approaches by which we could do. So, in the Anderson-Darling test what we do is from the data set, we calculate here mean and a standard deviation and we predict how the normal curve will look like and then we compare it, we plot them together with the actual data and that will give you an observed and expected. If they do not fall on that line, then we can say that the data does not come from the normal population. This, as this example shows which is taken from this particular reference paper. Otherwise, the Ryan-Joiner or the Shapiro-Wilk test, what it does, it calculates a correlation coefficient between the sample data and the normal scores of the data and then if the correlation coefficient is high, we can say there is, it is very strongly correlated to be a normal distribution.

So, the another method called Shapiro-Wilk, that also is similar to this Ryan-Joiner method which is based on a correlation. The another method Kolmogorov-Smirnov, it actually calculates something called empirical distribution function, for the data as well as it compares with some theoretical EDF and then tries to predict or try to sum up saying whether, the data is a normal population or not actually. So, these are the different test and many commercial softwares have all these methods available, so you just click a button. This is the easiest of the Anderson-Darling, because you can see from the figure very clearly whether the data is falling under the normal or not. So, the observed and the empirical or expected will fall nicely on the straight line. We can tell it is normal and if the $p > 0.05$, then of course there is no reason for you to reject the null hypothesis. If the at data expected and the observed do not fall in this straight line, it falls all over the place like in this particular example, then obviously, it is a non-normal and the p value will be less than 0.05. So, you reject the null hypothesis and accept the alternative hypothesis.

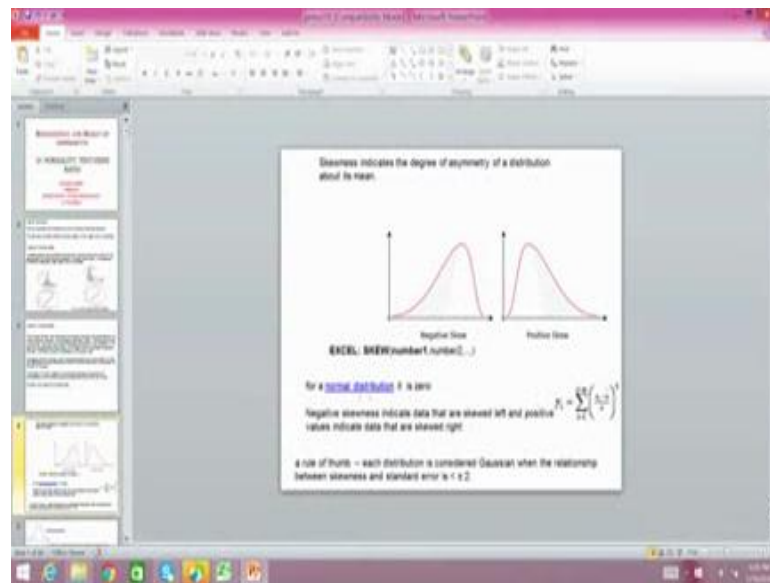
And then the other tests which is mentioned the Ryan Joiner test which calculates something called a correlation coefficient, then Kolmogorov-Smirnov method which something calls empirical distribution function.

(Refer Slide Time: 09:20)



Then, we also have some parameters called Skewness and Kurtosis, they are also indication of the normality of your data. For example, you can have a negative skew that means on the negative side you may have a less data at the left-hand side. Positive skew that means you have skewed towards the right, this is called a negative skew this called a positive skew. So, excel also as function called SKEW.

(Refer Slide Time: 09:59)

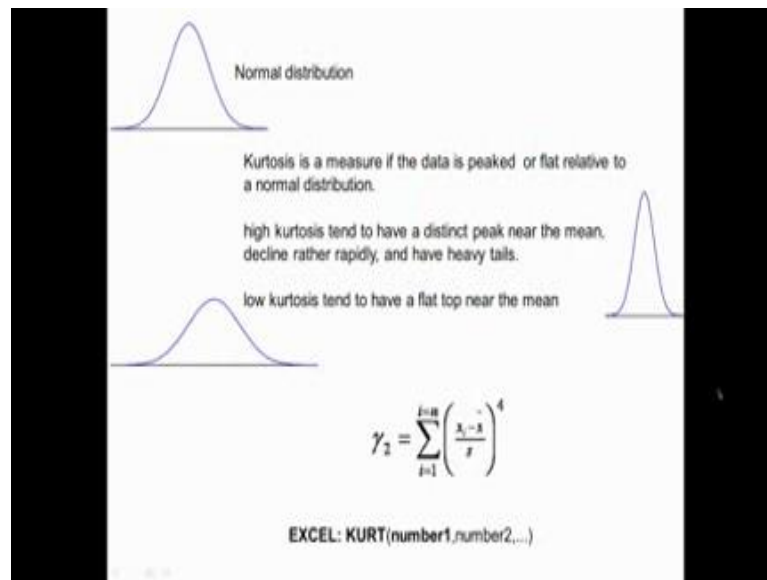


So, given a data set, it calculates what is the skewness of the data. So, I have some data, so I do a skew, s k e w, skewness is an indication 0.7065. So, it is plotting something like this. So, excel function is there, for a Normal distribution this skewness factor will be 0, for the Normal distribution the skewness factor will be Zero, for an negative skewness means you have a skewed to, skewed to the left and positive values indicate data that are the skewed to the right. So, positive value will indicates the data is skewed to the right actually. So, this is the formula for skewness, it is called the Third moment here, so you

have x_i that is the data point, each of the data point minus \bar{x} is the average \div the standard deviation cube actually, and it is calculated for all these data points. So, excel as I showed you can calculate your skewness and it tells you whether it is positively skewed or whether it is negatively skewed.

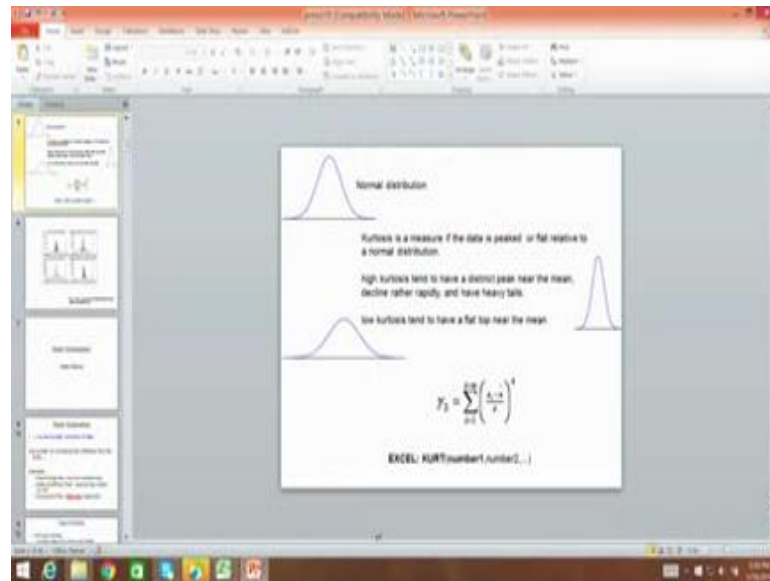
For example, if had more of these points, it is getting pushed as you can see it becoming bigger and bigger. So, skewness tells you how the shape of the curve is. It is sort of leaning towards the right that means the negative skew, it is leaning towards the left that means it is the positive skew. And for a normal distribution the skewness will be Zero.

(Refer Slide Time: 12:38)



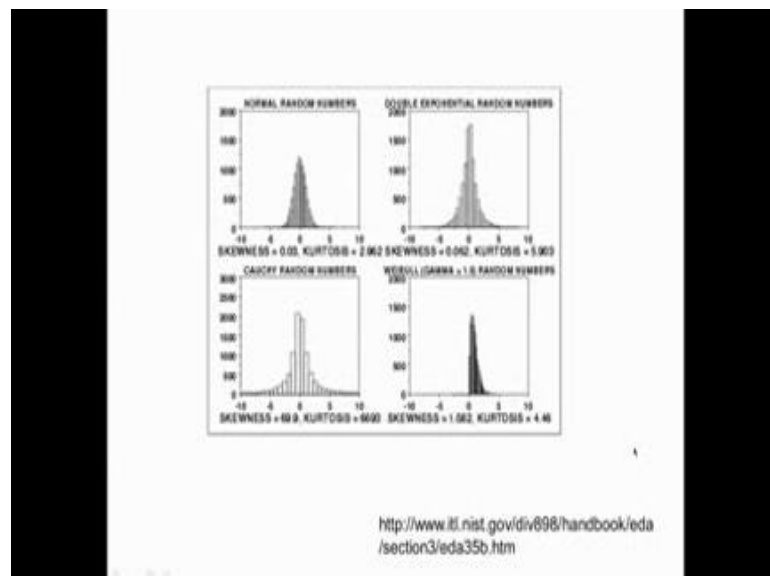
Skewness is called 3rd moment and we have another term that is called Kurtosis. Kurtosis is a measure of Peakedness, suppose you have a normal distribution like this, may have very peak key curve and falling down very fast or we may have a flat curve. So, high kurtosis leads to a very distinct peak near the mean and it is falling down very rapidly on both sides actually, very heavy tail. Low kurtosis that means flat. So, kurtosis is a Fourth moment, as you we can see you know Third moment is skewness, fourth moment is Kurtosis. So, x_i that means each of the data points minus \bar{x} that is mean \div standard deviation raise to the power 4. So, excel also has this term called Kurtosis.

(Refer Slide Time: 13:26)



Let us look at the kurtosis also here for your data set, so we can say minus, sorry, equal to KURT so you get some value for Kurtosis actually. So, if the kurtosis value is very high you will get a very sharp peak **falling** on both sides, if it is low kurtosis the top will be very flat actually. So, the Skewness and Kurtosis determines how the curves look like and obviously, they lead to Non-normally distributions. Now this also, I have taken it from a reference here, which is given here.

(Refer Slide Time: 14:10)

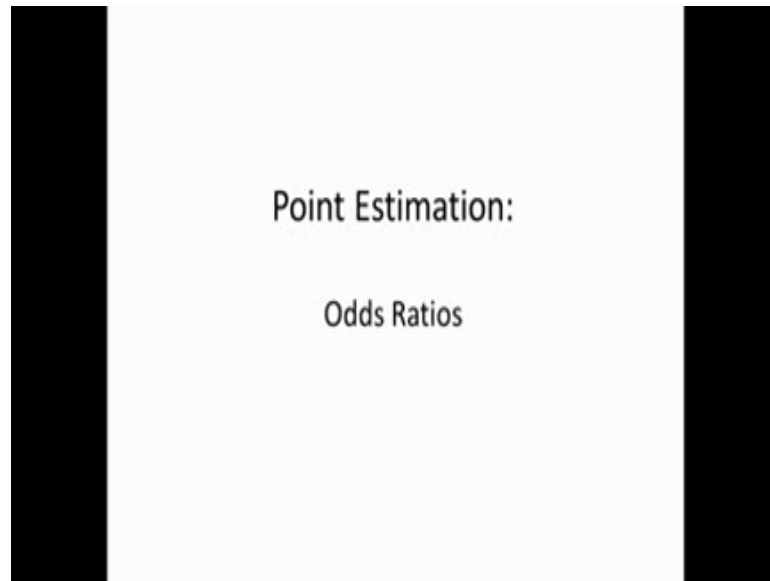


So, this is a normal random numbers, normal distribution. So basically, you do not see any skewness that means movement towards left or right, there is no sharp peak. Now look at this, this particular case skewness is very small, kurtosis is 2.9, this skewness is very small, skewness 0 means it is very normal and will uniformly distributed. In this particular case as you can see skewness is 0.06, but kurtosis is very high 5.9, that is why we get like peak like this, sorry, we get a peak like this and both side it is falling. Now skewness is also high, kurtosis is also high, so you get a distribution like this.

A skewness is low, kurtosis is 4.4 look at this, so you get different types of pictures depending upon this skewness and kurtosis and each one is called some distribution, this is a normal distribution, this called a double exponential distribution, quasi random distribution, this Weibull distribution and so on actually. This was taken from this particular reference actually is very nice because, a normal distribution a bell shaped, there is no leaning towards left or right, whereas when I have very high kurtosis it becomes very peaked. When you have a larger skewness, it is becoming more bent as you can see the kurtosis is very, very large, as you can see in this particular example. So, we can have different types of shapes of the distribution depending upon the skewness and kurtosis and so we talked about various types of factors which determine the shape of your curve, normal distribution curve, this skewness and kurtosis and how important it is to determine whether your data set is normal because in most of the problems we assume normal distribution, most our calculation we assume normal distribution. So, you need to know whether the distribution which you are handling is normal distribution and easiest of it is calculate the mean and the standard deviation and then we calculate what is expected and observed is your sample and see whether they fall in the straight line, if they fall with the large p value then we do not reject null hypothesis, if they do not fall on that and the p value is much less than 0.05, then we reject the null hypothesis and accept the alternate hypothesis ok.

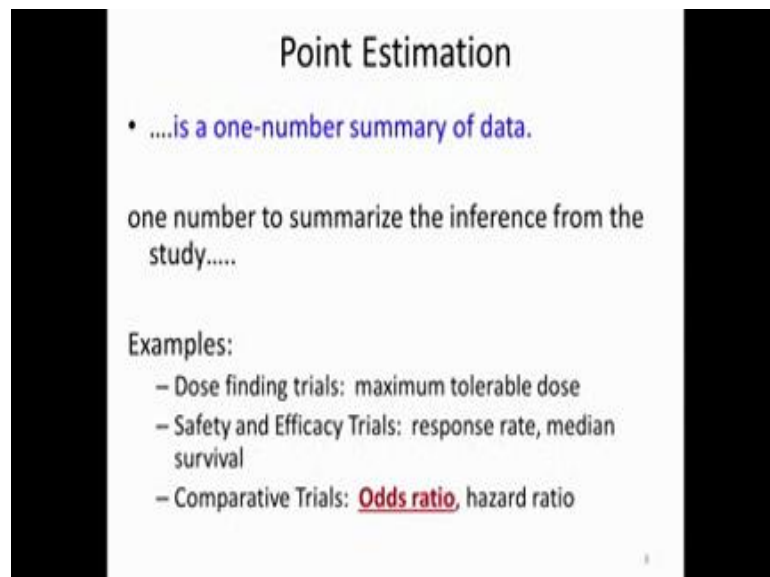
So, this is what is all about when you are testing for normality. Now there is another statistical calculation that is called odds ratios.

(Refer Slide Time: 17:16)



It is a point estimation, its ratio of 2 important numbers which gives you some indication of performance of some drugs or some the effect of certain a treatment and so on, but it is only one point data. How do you make out sense from this one point data? And that is what this odds ratio is all about actually. Basically, it is a one number summary of data.

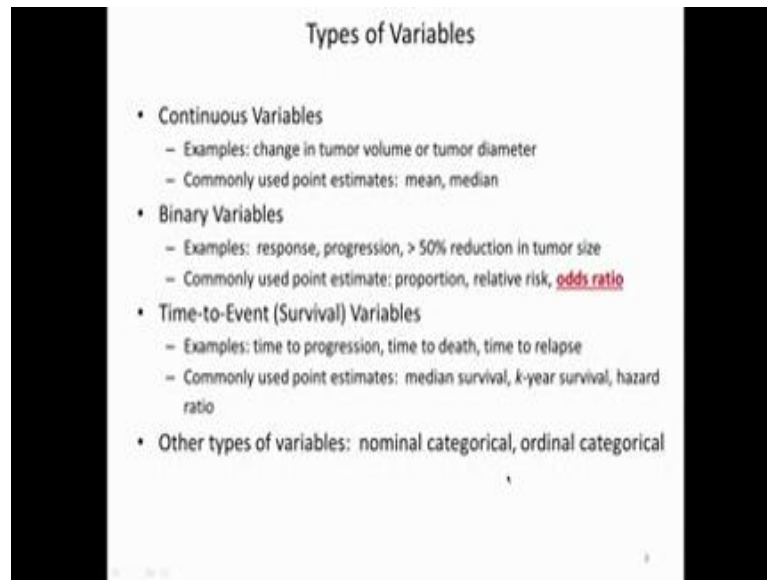
(Refer Slide Time: 17:45)



It is used to infer from the study, it could be like dose finding trials: maximum tolerable dose. What is a maximum tolerable dose of this particular drug on human? Or what will be when you are doing a safety study or efficacy trials, what is a median survival? What

is a response rate? These are all point estimation and they come under the concept of odds ratio or (Refer Time: 18:21) also are called hazard ratio also odds ratio, hazard ratio.

(Refer Slide Time: 18:24)



Let us look at some examples actually, when you are looking at say change in tumor volume or tumor diameter, that is a continuous variable, we can calculate mean, we can calculate median for that, if your looking at how the tumor volume changes in 1 month so that is a continuous data. Binary data we can have responses, progression, 50 % reduction in tumor size we all know about this actually right. Here, where we use this odds ratio where as if the data is continuous we can use mean, median, mode, range, 1st quartile, 2nd quartile, 4th quartile, inter quartile range, all those things but when you having a binary data then odds ratio come into this picture actually.

Proportion of male to female having a particular disease, proportion of observed changes in a certain parameter in 2 different species and so on actually. In such situations, we use this odds ratio. Let us not go too much into other types of variables. Time to event variables like time of progression, time to death, time to replace, then we used median survival, k year survival, hazard ratio, then other type of variable is nominal, categorical, ordinal, let is not go into these things actually.

(Refer Slide Time: 19:55)

The slide is titled "Example" and contains the following text:

Is gender associated with use of standard adjuvant therapy (SAT) for patients with newly diagnosed stage III colon cancer?

- 53% of men received SAT*
- 62% of women received SAT*

How do we quantify the difference?

* adjusted for other variables

"Age, Sex, and Racial Differences in the Use of Standard Adjuvant Therapy for Colorectal Cancer"; Potosky, Harlan, Kaplan, Johnson, Lynch. JCO, vol. 20 (5), March 2002, p. 1192.

10

So, let us look at an example, is gender associated with use of standard adjuvant therapy for patients with newly diagnosed stage three colon cancer. Now 53 % of men received this standard adjuvant therapy, 62 % of women received standard adjuvant therapy. This was taken from this particular reference actually, j c o volume 20 March 2002. So, 53 % of men received this therapy, 62 % of women received this therapy, who were diagnosed with stage 3 colon cancer. Now how do we quantify this difference? Because this looks like can I just take 53, 62 and then I will say 62 - 53 calling it 9 % of men sorry 9 % of more women received this therapy that is wrong. This is where odds ratio come into picture.

(Refer Slide Time: 20:52)

Odds and Odds Ratios

Odds = $p/(1-p)$

The odds of a man receiving SAT is
 $0.53/(1 - 0.53) = 1.13$.

The odds of a woman receiving SAT is
 $0.62/(1 - 0.62) = 1.63$.

Odds Ratio = $1.63/1.13 = 1.44$

Interpretation: "A woman is 1.44 times more likely to receive SAT than a man."

So, odds ratio we can calculate $p/(1-p)$

is for a, for the men 53 %. We will calculate a $p/(1-p)$ is 0.53, $\div 1 - 0.53$, that means 0.47, 0.53, but 0.47 is 1.13. Similarly, for female 62 % that is $0.62 \div 1 - 0.62$, so here $1.62 \div 0.38$ is 1.63. So, we divide this by this, so odd ratio comes out to be 1.44. So, a woman is 1.44 times more likely to receive standard adjuvant therapy (SAT) than man. We do not say that 9 % of women more women receive standard adjuvant therapy treatment, we need to say a woman is 1.44 times more likely to receive standard adjuvant therapy than a man, using this concept of odd ratio, odds ratio as it called. So how do we calculate? $p/(1-p)$, that gives you for the man, $p/(1-p)$ that gives you for woman and then you take the ratio, that is 1.44 understand?. Is a odds ratio is very useful in this type of situation point data.

(Refer Slide Time: 22:16)

Odds Ratio

Odds Ratio for comparing two proportions

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$
$$= \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

OR > 1: increased risk of group 1 compared to 2
OR = 1: no difference in risk of group 1 compared to 2
OR < 1: lower risk ("protective") in risk of group 1 compared to 2

In the example,
 p_1 = proportion of women receiving SAT
 p_2 = proportion of men receiving SAT

Let us look at so the concept is odds ratio **OR** is

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$
$$= \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

. So, if **OR is > 1**, increase risk of group 1 when compared to group 2, if **OR is = 1**, there is no difference in risk between group 1 and 2, **OR is < 1, lower** risk in group 1 when compared to group 2. See in this particular example, p_1 is the proportion of women who received the standard adjuvant therapy, p_2 is a proportion of men who received the standard adjuvant therapy understand?.

(Refer Slide Time: 22:57)

Odds Ratio from a 2x2 table

	SAT	No SAT	
Women	a = 298	b = 252	550
Men	c = 202	d = 248	450
	500	500	1000

$$\begin{aligned} OR &= \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{(298 / 550) / (252 / 550)}{(202 / 450) / (248 / 450)} \\ &= \frac{298 / 252}{202 / 248} \\ &= \frac{298 * 248}{252 * 202} \\ &= \frac{ad}{bc} \end{aligned}$$

$$OR = \frac{p_1(1 - p_2)}{p_2(1 - p_1)} = \frac{ad}{bc}$$

So again, we can use it in the form of a table. So, SAT, no SAT, women, men is, suppose 298 women received SAT, 250 did not receive, 550 is the total, men is 202 received SAT, 248 did not receive that comes to 450 understand?. So how do you calculate odd ratio p_1 ? $298 \div 1 - p_1$ is

$$\frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

$$\frac{(298 / 550) / (252 / 550)}{(202 / 450) / (248 / 450)}$$

So, that comes to

$$\frac{298 / 252}{202 / 248}$$
$$\frac{298 * 248}{252 * 202}$$

So, we get $a * d \div b$ into c , that is the formula for these odds ratio understand. So how do you calculate p_1 ? p_1 is nothing but, $298 / 550$. How do you calculate $1 - p_1$? $252 \div 550$. How do you calculate p_2 ? $202 \div 450$. How do you calculate $1 - p_2$? $248 \div 450$.

So, we have we can cancel all of them so it comes to $298, \div 252, 202 \div 248$. So, $292 * 248, 202 * 252$. So, 298 is nothing but a and 248 is nothing but d , 252 is nothing but b , 202 is nothing but c . So, the odd ratio when you have a table is $a d \div b c$. So, it comes out to be the same. So,

$$OR = \frac{p_1(1 - p_2)}{p_2(1 - p_1)} = \frac{ad}{bc}$$

very simple, so odds ratio from a 2 by 2 table. So, this is a important formula we need to remember.

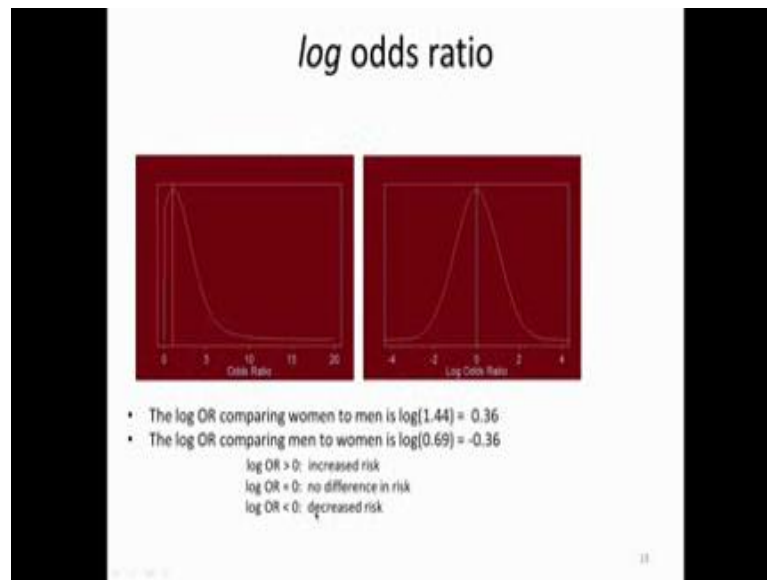
(Refer Slide Time: 25:37)

Odds Ratio

- Ranges from 0 to infinity
- Tends to be skewed (i.e. not symmetric)
 - "protective" odds ratios range from 0 to 1
 - "increased risk" odds ratios range from 1 to ∞
- Example:
 - "Women are at 1.44 times the risk/chance of men"
 - "Men are at 0.69 times the risk/chance of women"

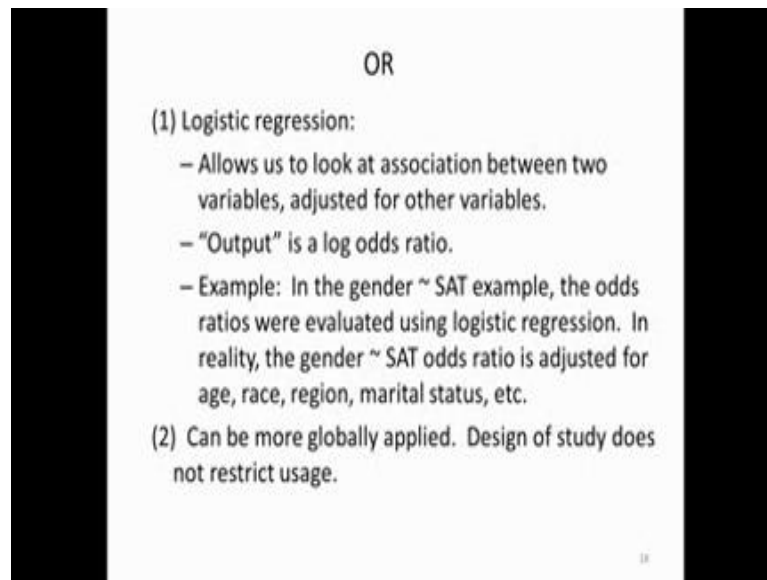
So, it ranges from 0 to infinity, in this particular case we got 1.44. So, the numbers can be very very large, it is not symmetric. So, when you have 0 to 1, it is protective, >1 it is a increased risk. So, women are 1.44 times the risk divided by chance of men. Men are 0.69 times the risk irrespective chance of women right, in that particular example of that actually.

(Refer Slide Time: 26:09)



When we take log logarithm of odd ratio, when you take **logarithm of 1.44** it comes to 0.36, logarithm for men to women, women to men is 1.44 for the inverse will be men to women that is 0.69. So, when you take it becomes **- 0.36. So, when log or odds ratio is > 0 is increased risk or = 0 no difference, logarithm or < 0 is decreased risk ok?. So just like this, OR > 1 increased risk, OR = 1 no difference, OR < 1 lower risk. So, when you take a logarithm, it become 0, > 0, = 0, < 0, that** is what this table is all about, ok?. When we take the logarithm of that it is sometimes is better to take logarithm especially, when you are dealing with very large number of, very large data points. So, log the or the odds ratio helps us to look at association between 2 variables, in this particular case we looked at male and female gender with when you adjust the other variable actually.

(Refer Slide Time: 27:23)



OR

(1) Logistic regression:

- Allows us to look at association between two variables, adjusted for other variables.
- "Output" is a log odds ratio.
- Example: In the gender ~ SAT example, the odds ratios were evaluated using logistic regression. In reality, the gender ~ SAT odds ratio is adjusted for age, race, region, marital status, etc.

(2) Can be more globally applied. Design of study does not restrict usage.

18

So, in this particular example, we are looking at the gender standard adjuvant therapy example. So, we looked at gender standard adjuvant therapy odds ratio is adjusted for ofcourse, you have to assume age, when you talk about men and women you should adjust for age, race they could be coming from different continents region, northern region, southern region, marital status, whether they are married so you need to adjust all these things then only you can do a proper odd ratio otherwise, there these values may be confounding your results. So, you need to remember that and it can be used very globally. So, that is the beauty of this concept of odd ratio, we can use it as a term if you adjust globally and make it a general statement actually.

(Refer Slide Time: 28:10)

Example 2

Groups: paclitaxel (n = 47) versus
CAP (n = 47)

"14 patients in the CAP group
and 8 patients in the paclitaxel
group had complete
responses"

$p_1 = 14/47 = 0.30$
 $p_2 = 8/47 = 0.17$
 $OR = (0.30/0.70)/(0.17/0.83)$
 $= 2.1$

Randomized Controlled Trial of Single-Agent Paclitaxel Versus Cyclophosphamide,
Doxorubicin, and Cisplatin in Patients with Recurrent Ovarian Cancer Who Responded to First-
line Platinum-Based Regimens. Cantu, Parma, Rossi, Florian, Bonazzi, Dell'Anna, Torri,
Colombo. JCO, vol. 20 (5), March 2002, p. 1232

Now let us look at another example, which is related to Paclitaxel, as you know Paclitaxel is a drug which is used for an anti cancer treatment and there is another type of treatment that is called a CAP treatment. Where they give a Cisplatin, Cisplatin is a platinum based material anti cancer drug. So, 14 patients, 8 patients were given complete responses. So, some patients are given Paclitaxel, some patients were given this Cisplatin type of drugs. So, 14 patients in the CAP group had complete responses, whereas 8 patients in the Paclitaxel group had complete responses. So, if you are doing for a CAP, 14 will take p_1 as cap $14 \div 47$ is 0.3, p_2 is $8 \div 47$ is 0.17. So, odds ratio will be $0.3 \div 0.7 \div 0.17 / 0.83$ that comes to be 2.1. So, what does that mean? That means 2.1 of the cap group and 8 when compare to the Paclitaxel groups. So, we can you do the same thing using the table also.

(Refer Slide Time: 29:31)

Odds Ratio via 2x2 table

"14 patients in the CAP group and 8 patients in the paclitaxel group had complete responses"

	CAP	Paclitaxel
CR	14	8
No CR	33	39
	47	47

"Patients in the CAP group are twice as likely to have a CR as those in the paclitaxel group."

CR- complete response

2x2 Table approach:

Odds ratio = ad/bc
= $(14 \cdot 39)/(8 \cdot 33)$
= 2.1

So, we do 14 total is 47. So, we say 14 people complete response, so no response is 33 because 47 is the total, Paclitaxel 8 response, no response is 39. So, if you want to calculate odd ratio 14 into this divided by this into this, right?. So, 14 in the CAP group had response, 8 in Paclitaxel group had complete response. So obviously, if you look at the odds ratio 2. What how do you name it? Patient in the CAP group are twice as likely to have a complete response as those in the Paclitaxel group. And you get the same answer $14 \cdot 39 \div 33 \cdot 8$ actually, $39 \div 8 \cdot 33$, that will give you 2.1. So, the statement you make is patients in the CAP group are twice as likely to have a complete response when compared to the Paclitaxel group.

So, we will talk about these things more in the next class, because odds ratio gives you a single point statement for comparing certain data, mostly binary data, live-dead or a 50 % recovery and so on actually.

Thank you very much for your time.

Key words: Normality test, Distribution types, Skewness, Kurtosis, Peakedness, Non-normal distribution, Anderson Darling test, Mean, Standard Deviation, correlation coefficient, Empirical Distribution Function, Point estimation, odds ratio.