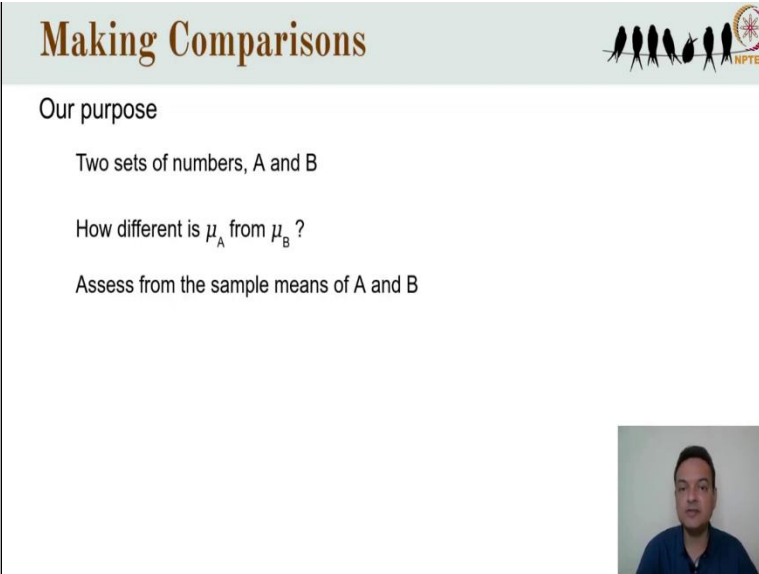


Basic Course in Ornithology
Dr. Suhel Quader
Nature Conservation Foundation

Lecture -27
Introduction to Data Visualization and Analysis Part 2

Now that we have explored how to measure and present results on a single quantity – a measure on one entity in one place at one time, let's move on to the next kind of common task, which is to compare the same quantity across multiple categories. In this section, we will largely focus on the simplest possible comparison the comparison between two categories. We may want to compare Peafowl train length between northern and southern India, for example, or compare chick feeding behaviour between male and female Magpie Robins or compare species richness between highly polluted and less polluted wetlands.

(Refer Slide Time: 00:59)



Making Comparisons

- Our purpose
 - Two sets of numbers, A and B
 - How different is μ_A from μ_B ?
 - Assess from the sample means of A and B

NPTEL

NPTEL

Depending on the situation and the study design, we measure multiple sampling units in each of these categories and so we end up with two sets of numbers, one for each category. Typically, our purpose is to estimate how different the two underlying populations are from each other. Our best guess or estimate of this is the difference in the sample means, but we need to pay careful attention to the study design.

(Refer Slide Time: 01:23)

Making Comparisons



Paired design - mean

Pair No.	F	M	F - M
1	F ₁	M ₁	F ₁ - M ₁
2	F ₂	M ₂	F ₂ - M ₂
3	F ₃	M ₃	F ₃ - M ₃
4	F ₄	M ₄	.
5	F ₅	M ₅	.
6	F ₆	M ₆	.
7	F ₇	M ₇	.
8	F ₈	M ₈	.
9	F ₉	M ₉	.
10	F ₁₀	M ₁₀	.

$$\bar{x}_{F-M} = \frac{\sum(F_i - M_i)}{N}$$



We might have a study design that is matched or paired such that we measured, for example, the chick feeding rates of both individuals of a pair – male and female – or we found pairs of wetlands near each other – one highly polluted and other not – and then the resultant data are also paired. In this case, the quantity of interest is the mean difference within each pair, and so we first take the *difference within a pair* and then we *average across pairs*.

$$\bar{x}_{F-M} = \frac{\sum(F_i - M_i)}{N}$$

(Refer Slide Time: 01:53)

Making Comparisons



Unpaired design - mean

A	B
A ₁	B ₁
A ₂	B ₂
A ₃	B ₃
A ₄	B ₄
A ₅	B ₅
A ₆	B ₆
A ₇	B ₇
A ₈	
A ₉	
A ₁₀	

$$\bar{x}_A - \bar{x}_B = \frac{\sum A_i}{N_A} - \frac{\sum B_i}{N_B}$$

\bar{x}_A \bar{x}_B



The unpaired study design is perhaps a more common situation to be in. You run 40 transects in coffee plantations and 30 other transects in intact forest, and your purpose is to ask what is the

difference in species richness between these two habitats. Or you follow 50 Red-vented Bulbuls and note how they spend their time over 10 minutes and do the same for 80 Red-whiskered Bulbuls, your purpose being to ask what extent the two species differ in the time they spend in, let's say, aggressive behaviour.

Now unlike the situation in paired data, here there is no requirement that the sample sizes be the same, as there is no one-to-one relationship between any of the data in the first group and the data in the second group. Therefore, unlike the paired situation, we cannot take differences within pairs and then average across pairs. Rather, we *average across all data within each group* and then examine the *differences in the means of the two groups*.


$$\bar{x}_A - \bar{x}_B = \frac{\sum A_i}{N_A} - \frac{\sum B_i}{N_B}$$

Our goal is to calculate as best we can the true population difference in the means of the groups as estimated by the differences in the means of the samples we have collected from both these groups.

So, now we have our estimate of the population difference. How precise is this estimate?


(Refer Slide Time: 03:11)

Making Comparisons



Paired design - precision

Pair No.	F	M	F - M
1	F_1	M_1	$F_1 - M_1$
2	F_2	M_2	$F_2 - M_2$
3	F_3	M_3	$F_3 - M_3$
4	F_4	M_4	.
5	F_5	M_5	.
6	F_6	M_6	.
7	F_7	M_7	.
8	F_8	M_8	.
9	F_9	M_9	.
10	F_{10}	M_{10}	.



Answering this is pretty straightforward when your data are paired because you are dealing with just a single set of numbers: the differences within each pair. And so you can treat those numbers just as you did when you are estimating a single quantity, and you can go ahead and calculate the precision of your estimate in the way we have discussed earlier. We can do the usual thing – find

the confidence interval by bootstrapping the original sample, drawing many samples with replacement again and again, looking at the distribution of bootstrap means, and finding the appropriate quantiles say for a 95% confidence interval.

Or we can take our familiar shortcut – calculating the standard error of our estimate and finding the 95% confidence interval by using the multiplier 1.96 +/- the mean.

(Refer Slide Time: 03:59)

Making Comparisons

Unpaired design - precision

A	B
A ₁	B ₁
A ₂	B ₂
A ₃	B ₃
A ₄	B ₄
A ₅	B ₅
A ₆	B ₆
A ₇	B ₇
A ₈	
A ₉	
A ₁₀	
\bar{x}_A	\bar{x}_B

A and B come from populations with equal variances

$$s_{pool} = \sqrt{\frac{(N_A - 1)s_A^2 + (N_B - 1)s_B^2}{N_A + N_B - 2}}$$

$$SE_{pool} = s_{pool} \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}$$

$$df = N_A + N_B - 2$$

A and B come from populations with unequal variances

$$SE_{pool} = \sqrt{\frac{s_A^2}{N_A} + \frac{s_B^2}{N_B}}$$

$$df = \text{smaller of } (N_A - 1) \text{ and } (N_B - 1)$$

When the data are unpaired, things are a little different. Our measure of difference is the difference in the means of the two groups. We imagine this is one of many possible such differences that we could potentially have got. Using bootstrap, in each step you would re-sample A and separately re-sample B, calculate the means of each, subtract one mean from the other and store the difference of means. You do this many times, perhaps a thousand times or ten thousand times, and then examine the distribution of the differences in the means. That was bootstrap.

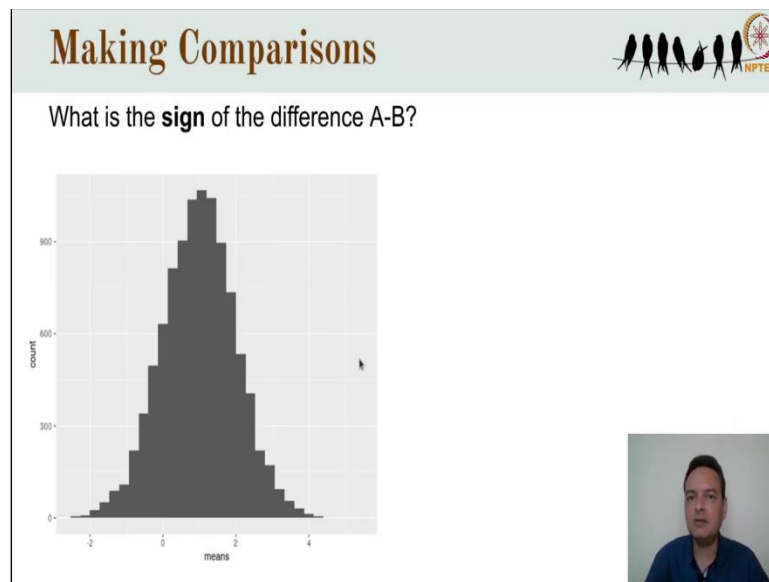
We may instead want to use the properties of the normal distribution to find the confidence interval. We would then need to calculate the standard deviation of the distribution of possible differences. In other words, we need to know the *standard error* of the differences. The complication here is that the standard error cannot just be the standard deviation by the square root of the sample size, because there are two groups of data each with their own standard deviation and their own sample size.

So, instead we need to calculate a combined or a pooled standard error. Now, the details of how to calculate a pool standard error are rather complicated and depend on whether the population variances of group A and group B can be assumed to be identical or not. For both these cases the separate formulas are on the slide, and also the corresponding degrees of freedom, which we know, of course, we need in order to look up the t distribution.

You can pause the video and examine the formulas, but the main message I want to convey is that, as before, we want to ask about the distribution of possible sample means, or in this case difference in means, and to find out what the 95% confidence interval is by identifying the central 95% of that distribution through taking the standard error and multiplying it with the appropriate number.

If the sample size is large, that number would be 1.96, but if that sample size is low – below 30 let's say – that number would depend on the shape of the t distribution which in turn depends on the degrees of freedom.

(Refer Slide Time: 06:10)

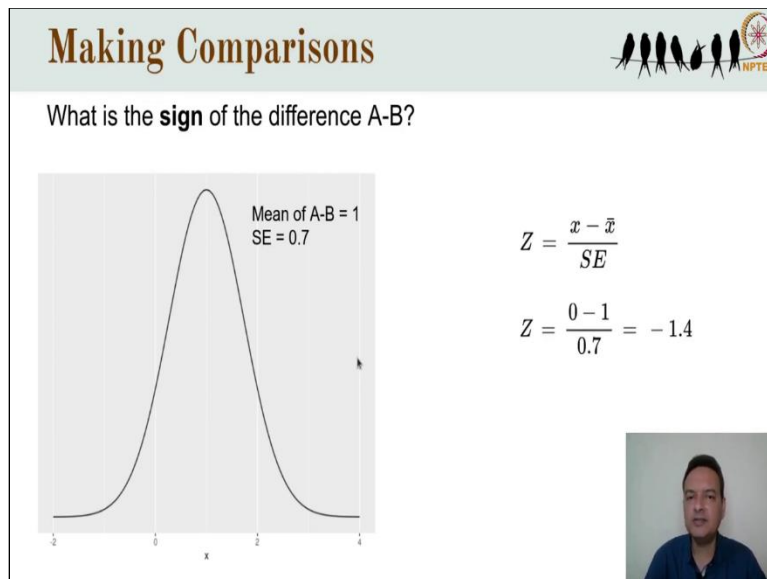


So, far our main goal has been to estimate the *magnitude* of the difference and how confident we can be in this but in some cases, you might be most interested in whether the difference is positive or negative, and the magnitude is of secondary interest. Now, if our sample estimate is of the mean of A minus B to be +3 then our best guess is that the difference is positive. In other words, A has

a larger clutch size say than does B. But what degree of confidence do we have in this conclusion that A has a larger clutch size than B?

Again, we need a distribution, which can be generated in two ways as we have spoken about many times by now – through bootstrap or by assuming a normal distribution. For paired data the distribution is that of possible mean differences, for unpaired data the distribution is of possible differences in the means. In both cases, we ask what proportion of possible sample outcomes fall above 0 -- that is differences are positive, A greater than B? And what proportion of possible sample means fall below 0 that is differences are negative, B greater than A? We can find this out by counting up what proportion of outcomes are below 0 in our bootstrap results. So, here's a distribution of 10000 bootstrap means of differences. By counting up how many of these are less than 0, we find that 1626 are negative. This means that 16.26% of the means of differences are negative.

(Refer Slide Time: 07:48)



Or we can do this from the properties of the normal distribution using a z table. In this example, the mean is one and the standard error is 0.7. So, the normal distribution corresponding to this would look like this (figure). What fraction of the area under such a normal distribution falls to the left of 0? We use our familiar formula to get the z score

$$Z = \frac{0 - 1}{0.7} = -1.4$$

and see that 0 is 1.4 standard errors to the left of the mean. So, that means now we have to look up a z table.

(Refer Slide Time: 08:21)

Making Comparisons

z p

-1.41	0.07927
-1.40	0.08076
-1.39	0.08226
1.39	0.91774
1.40	0.91924
1.41	0.92073

Here is an example of a z table, I know the numbers are small but there are two columns of numbers. The first is a z value or a series of z values, and the second is the area under the normal distribution to the *left* of that value. For example, we need to know the area to the left of a z score of -1.4 and the table tells us that this area is 0.08 or 8%. In other words, minus 1.40 is at the 0.08 quantile.

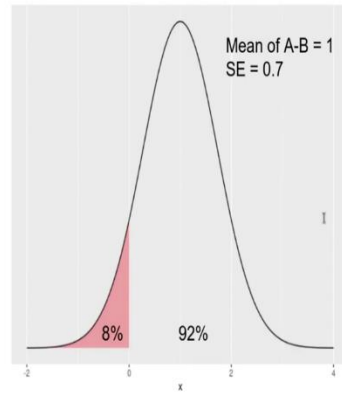
By contrast, if in the distribution of means the overall mean was below 0 such that the z score was actually +1.4, then the table would tell us that the area to the left of +1.4 is 0.919 or 91.9%. So, +1.40 lies at the 0.919 quantile.

(Refer Slide Time: 09:17)

Making Comparisons



What is the **sign** of the difference A-B?



$$Z = \frac{x - \bar{x}}{SE}$$

$$Z = \frac{0 - 1}{0.7} = -1.4$$



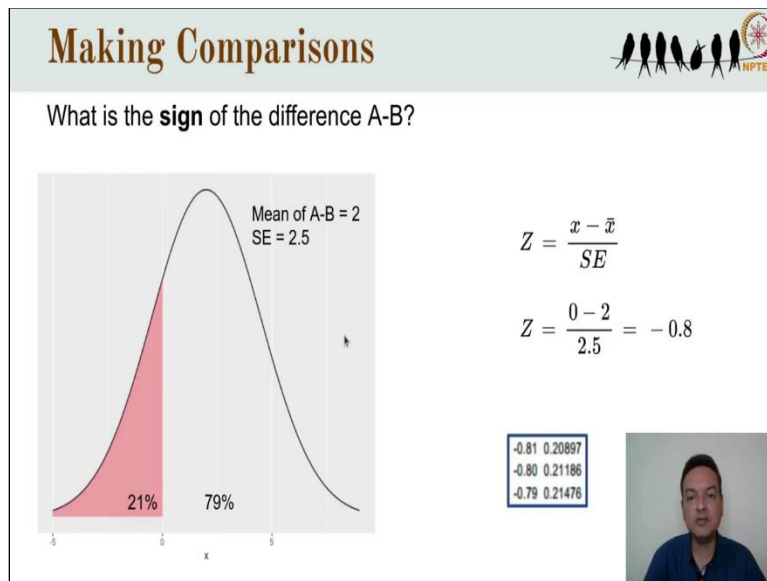
So, now we know that in this distribution, 0 lies at the 0.08 quantile, in other words, it divides the lower 8% of the distribution from the upper 92% of the distribution. What this means is that in our distribution of potential sample mean differences, 8% of those means are expected to be negative and 92% of those means are expected to be positive and this means that although our best guess is that the population mean is +1, and therefore the difference between A and B is positive, there is an 8% chance that we are wrong and that the true population mean is actually negative which is what happens when B is actually greater than A.

So, if we conclude that the difference between A and B is positive (that is the population difference between A and B is positive) we estimate that the chance that this is a wrong conclusion is 8%. If that happens, then the error that we would be making is an error in sign where the sign of the difference is actually negative rather than positive. This is known as an *error of type S* where S stands for sign.

In this case, the probability of committing type S error is 8%. Because the probability of type S error is small in this example, we perhaps are pretty confident in our conclusion that A is greater than B and the difference between A and B is positive. Now please note a potential confusion. Type S error is an error in the same way as we mean it in everyday language it denotes a mistake in our conclusion. This is different from the word error in 'standard error' which denotes variation or precision rather than mistake. I am sorry that the words are being used the different meanings;

if it were up to me I would rename standard error to avoid this confusion, but we have to get used to the same words being used in different meanings.

(Refer Slide Time: 11:17)



So, that was one example. In another example, say the mean difference is 2 and the standard error is 2.5 and hence the sample mean is 0.8 standard errors above 0.

$$Z = \frac{0 - 2}{2.5} = -0.8$$

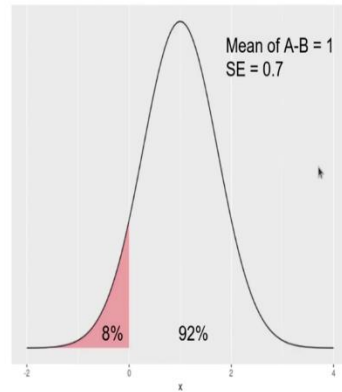
What is the area to the left of 0 in this case? Looking up a z table again, we find that a z score of -0.8 corresponds to the 0.21 quantile. So, the area to the left of -0.8 is 21% of the total. And therefore, if we conclude that the mean difference is positive, we run the risk of committing an error of type S with a probability of 0.21. And because that probability is very large we cannot be very confident in our conclusion.

(Refer Slide Time: 11:58)

Making Comparisons



What is the **sign** of the difference A-B?



$$odds = \frac{0.92}{0.08} = 11.5$$



We can go one step beyond and ask what is the relative probability that the true population difference is positive versus negative? We can do this by calculating what is known as the *odds*. The odds is the ratio between the two probabilities involved. For example 1, the area under the curve above 0 was 92% and the area below 0 was 8%. So, the corresponding probabilities are 0.92 and 0.08. The ratio between these two is called the odds, and in this case it is 11.5.

$$odds = \frac{0.92}{0.08} = 11.5$$

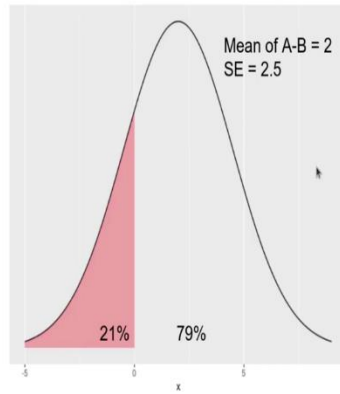
And we can interpret this to mean that given the data we have, the true population difference between A and B is 11.5 times as likely to be positive as it is to be negative. So, there is a lot of evidence towards it being positive, that is towards the population difference being positive.

(Refer Slide Time: 12:50)

Making Comparisons



What is the **sign** of the difference A-B?



$$odds = \frac{0.79}{0.21} = 3.7$$



For example 2, we get an odds of 3.7

$$odds = \frac{0.79}{0.21} = 3.7$$

which means that our evidence is only enough to say that the true population difference between A and B is a little over three times as likely to be positive than it is to be negative. So, when evaluating whether a difference in means is positive or negative it is useful to routinely calculate the probability of type S error and also the odds between positive and negative and make your conclusion based on this.

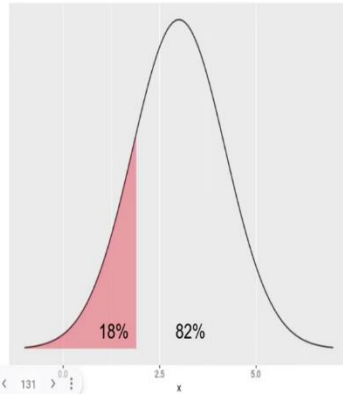
By the way this can also be used when you are estimating a *single* quantity and want to compare that against some threshold, which we will see just now.

(Refer Slide Time: 13:27)

Single Quantity Again



Comparing against a threshold



Hatchling per pair for steady state: 1.9

Sample average = 3

Standard Error = 1.2

$$Z = \frac{x - \bar{x}}{SE}$$

$$Z = \frac{1.9 - 3}{1.2} = -0.91$$

$$odds = \frac{0.82}{0.18} = 4.5$$

For example, say you want to find out whether the average number of eggs that pairs of Great Indian Bustards hatch is greater or less than the number of hatchlings needed to maintain the population at steady state. Now, I do not actually know what that number is, but suppose you know that given the mortality rate of bustards, for the population to be stable, each pair must hatch 1.9 eggs per year. This would exactly balance birth and death, let's say.

So, you go to Rajasthan, you study bustard breeding and find that the average number of hatchlings in your sample of bustard pairs is 3 with a standard error of 1.2. Now remember that, if the true population average is above 1.9 hatchlings per pair, the population will increase but if it is below 1.9, the population would decrease. What can we conclude about what will happen to the population? Our best guess is that the population will increase, because 3 is greater than 1.9. But we also know that there is such a thing as type S error.

And because the standard error is 1.2 and therefore 1.9 lies 0.91 standard errors to the left of 3

$$Z = \frac{1.9 - 3}{1.2} = 0. -91$$

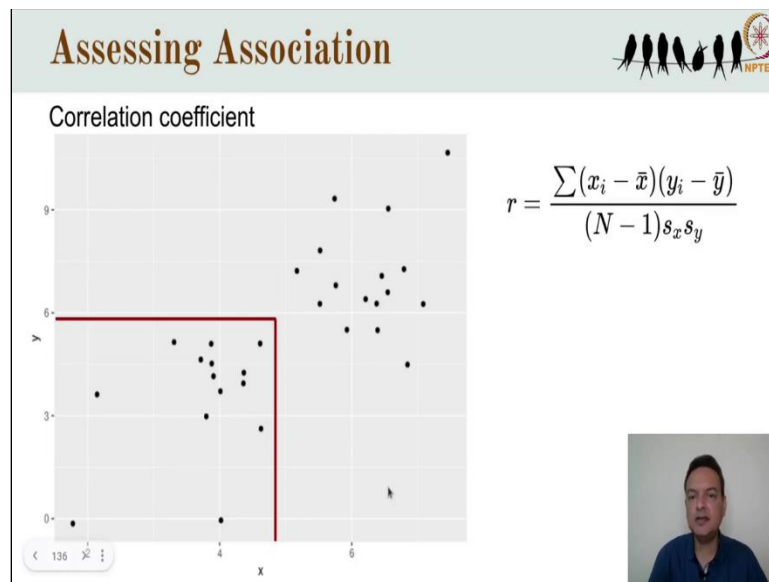
we can look up a z table and find that a z score of minus 0.91 corresponds to the 0.18 quantile which means that the probability of committing type S error is 18 percent and the odds between a true population mean of above 1.9 versus below 1.9 is

$$odds = \frac{0.82}{0.18} = 4.5$$

And so we say that our best guess is that the true population number of hatchlings per pair is above 1.9. But we are not all that confident about this -- we would be much happier if the odds were say 20, so that the probability that the mean number of hatchlings per pair was above replacement rates was 20 times that of the probability that the real number -- the true population number -- is below replacement rate. So, our result is somewhat unsatisfactory. To avoid such a vague result in which you do not have much confidence, the one thing you can and must do is to make sure your study has sufficient *precision*, which if you remember from before, you can achieve through planning the study such that your sample size is large enough.

And please always remember that although I am using the standard normal or z distribution in these examples, if the sample size is less than 30 it is really the t distribution that should be used and as we have talked about, unlike the z distribution, the properties of the t distribution depend on sample size, or more accurately the degrees of freedom, which in this case would be the number of pairs minus 1. We subtract 1 from the sample size because we have taken up one degree of freedom in estimating the mean of the population in the number of hatchlings.

(Refer Slide Time: 16:17)



Let's now move on to our final task, which is to assess the association between two numeric variables. We can do this if each entity that we sample has at least two variables measured. For example, we may have a number of plots within which we have measured canopy cover as well as

number of bird species or we may have a number of individual birds, say Magpie Robins, for which we have measured both body size as well as territory size.

We want to know whether there is an association between canopy cover and number of bird species, or between body size and territory size. In each case, we ask what relationship exists between the two variables. We usually should not and in fact cannot infer anything about cause and effect without a lot of further work. So, when I say association or relationship or correlation, it is just describing the pattern and not implying anything about cause.

Now visualizing these relationships is straightforward. We can draw scatter plots by spreading out one variable on the x axis and the other on the y axis and putting a circle or cross or other symbol representing each sampling unit in our data. And just like in the earlier part of this video, we would like to summarize this visual pattern in some way. For single variables, for example, we were interested in central tendency and we could use the mean or median. But for *relationships* between two variables, we are interested in a measure of *association* and the simplest measure of association is called the *correlation coefficient* (r). So we first identify the means of x and y indicated here by the vertical and horizontal lines. Then for each point on the graph, we take the deviation from the mean and then multiply the deviation in x with the deviation in y and sum up all of these.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N - 1)s_x s_y}$$

You may see that, if there is no relationship between x and y, then positive deviations in x will be associated sometimes with positive deviations in y and sometimes with negative deviations in y and the overall sum will therefore be near 0. By contrast, if there is an increasing relationship between y and x as shown here, then positive deviations in x will be associated with positive deviations in y and negative deviations in x will be associated with negative deviations in y.

In both cases, because we are multiplying them, we will get large positive numbers. On the other hand, we do not have many examples of the converse situation where positive deviations in x are associated with negative deviations in y and vice versa. So, the overall sum of all these products will be a positive number.

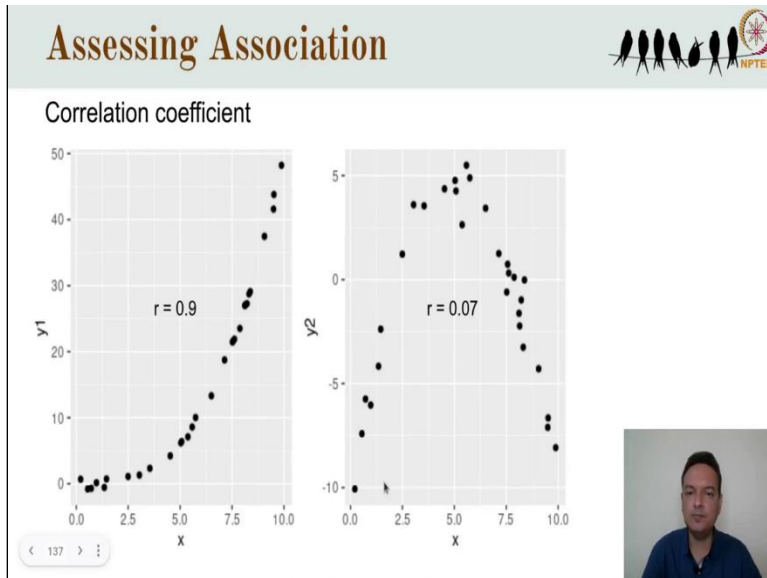
And if there is a *decreasing* relationship between y and x , then positive deviations in one variable will tend to be associated with negative deviations in the other and the overall sum will be negative.

We also need to divide by the degrees of freedom. So, that we get the average of these joint deviations and we also divide by the product of the standard deviations of x and y in order to scale the correlation coefficient within specific bounds. So, what can we tell from the correlation coefficient? We can say something about the *direction* of the association – whether it is positive, negative or no association – and we can say something about the *strength* of the association – whether it is a tight association or loose.

Remember as always, we are deriving the data and the scatter plot and the correlation coefficient by sampling from an underlying larger population. And we hope that the correlation we get in our *sample* is similar or close to the *population* correlation coefficient. So we denote the sample correlation coefficient by a Latin letter, in this case r , and the true population correlation coefficient by the Greek letter ρ which I have not shown here.

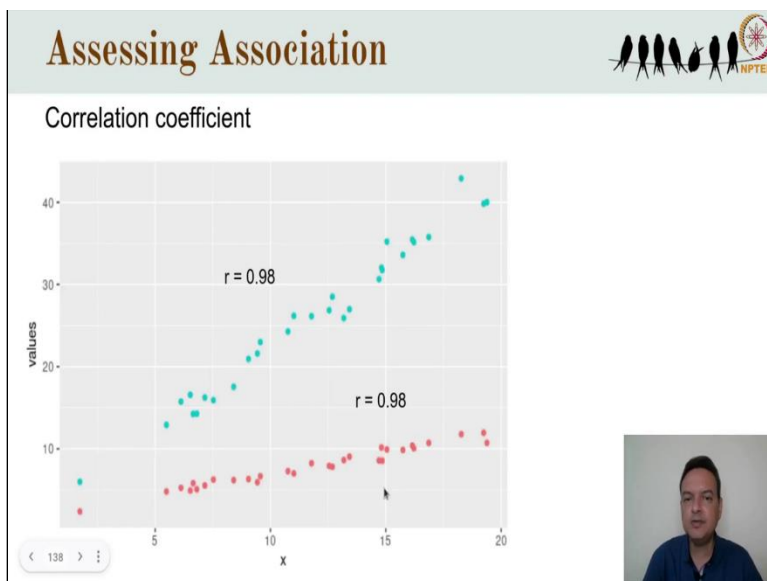
Now the correlation coefficient is fine for an initial look. It varies between -1 and $+1$. The sign of the correlation coefficient tells you the direction of the relationship, and the farther away from 0 the correlation coefficient is, the tighter the relationship. But the correlation coefficient has some limitations. Firstly, it measures only the linear or straight line relationship between the two variables. If the relationship is anything other than a straight line, the correlation coefficient is inadequate and in fact can be quite misleading.

(Refer Slide Time: 20:56)



You can see this from two examples, the points on the left are very tightly arranged and we might expect them to have a correlation coefficient of close to 0.99 but since the relationship is not a straight line the value is lower. The example on the right is more extreme, here y increases with x up to a point, and then decreases; but despite the very clear pattern, the correlation coefficient is near 0, which is of course because the correlation coefficient breaks down in situations like this.

(Refer Slide Time: 21:27)

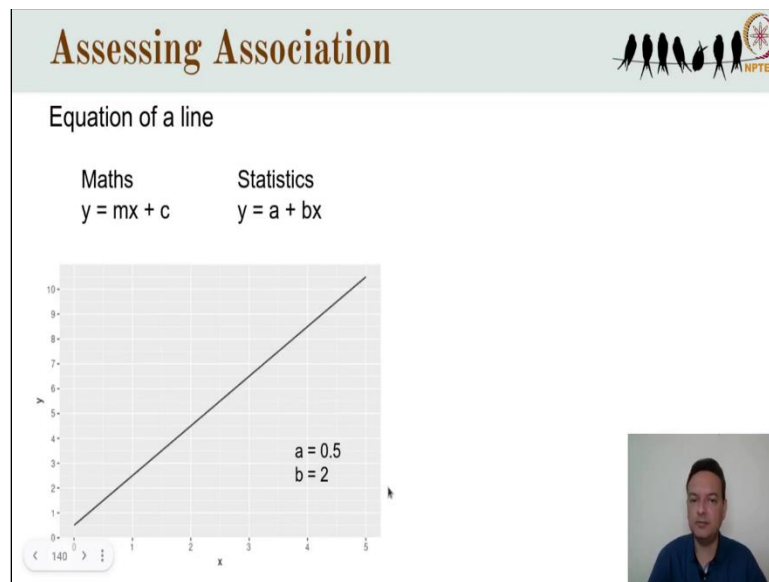


Second consider these two relationships. Both have the same correlation coefficient, but in one case y increases rapidly as x increases and in the other case y barely increases at all as x increases. We often want to understand the magnitude of change of y with x -- for example do bird species

increase greatly as forest patch size increases or only by a little; or do larger Magpie Robins have much bigger territories than small Magpie Robins or are their territories only slightly bigger?

To assess this, we need to draw a line describing the relationship between y and x and examine the properties of that line.

(Refer Slide Time: 22:11)



The simplest line of course that we can draw to describe the relationship between two variables is a straight line. I have said earlier that it may not in fact be a straight line at all in real life, in the data that we have. But let us start with the simplest possible situation. You may have learned in mathematics that the equation of a line is

$$y = mx + c$$

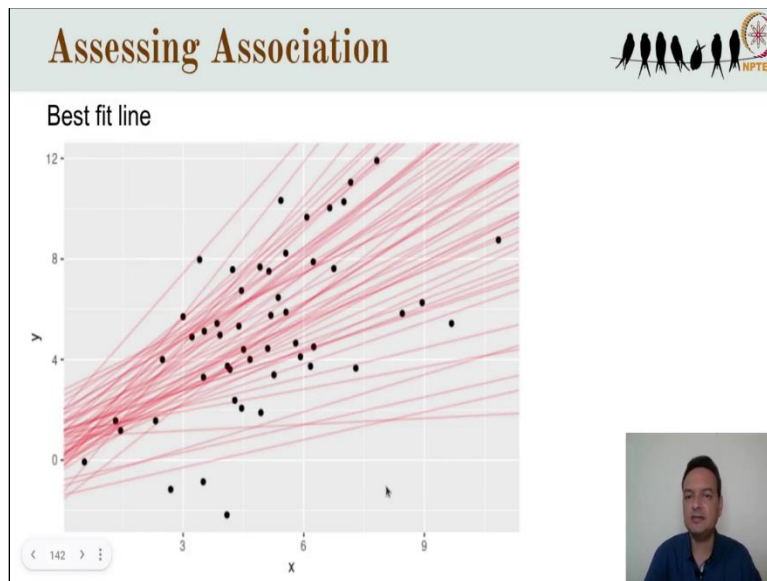
where y and x are the two variables, m is the slope and c is the intercept. In statistics we use a slightly different formulation, placing the intercept before the slope and renaming it. So, we say

$$y = a + bx$$

Here, a is the intercept and b is the slope. Now as you know the intercept is the value of y when x is 0. You can substitute 0 for x in the equation and see that this is true. And b is the slope, which is the rate at which y increases for a unit increase in x. Let's see if this is correct, we set $a=0.5$ and $b=2$. For $x=1$, y is then 2.5 and if we increase x by 1 unit from 1 to 2, y then becomes 4.5 which is 2 units more than earlier and that is the slope. So, it works.

Of course, neither a nor b is restricted to be a positive number. If a is negative it means that when x is 0, y is negative. And if b is negative, it means that as x increases, y decreases. And as before, please do not let this kind of language lead you to believe we are only talking about causal relationships between x and y . For these purposes here we are agnostic about cause and effect. We do not require that: we are only looking at the association between x and y .

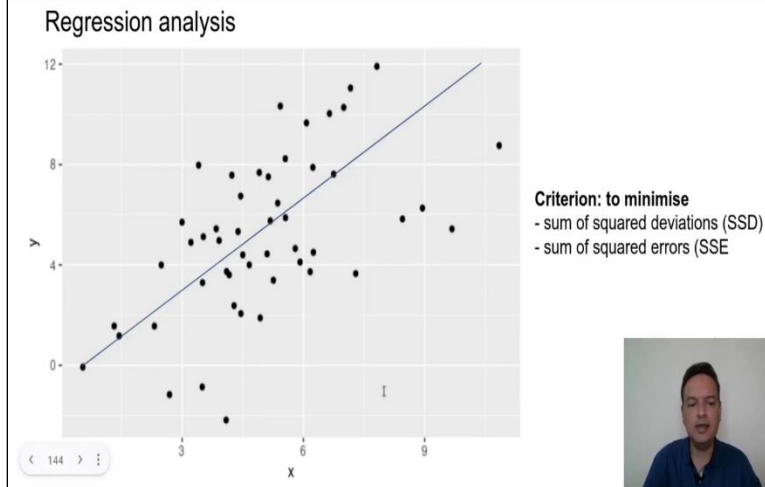
(Refer Slide Time: 24:01)



So, now that we are familiar with the equation of a straight line, the question is: what equation best describes the relationship between y and x ? There are an infinite number of lines that can be drawn; which one should we choose? The answer is that we need to define some criterion, evaluate all possible lines according to this criterion and then choose the one that fits best. So, what criterion should we follow? Now I am sure you can think of multiple possible criteria that could be used, but for now let's imagine that our main purpose is to predict y values from x .

(Refer Slide Time: 24:38)

Assessing Association



In other words, we want to draw a line such that if we know the say the body size of a Magpie Robin, we are able to come up with the best prediction for the size of its territory; and that prediction is the point on the line corresponding to that particular x value.

We will not be able to predict perfectly, of course, because the actual y will usually be some distance from the predicted y, perhaps larger perhaps smaller than predicted. We can call this the *deviation* or *error* in the prediction. If the observed y is greater than predicted we get a positive deviation or error; if it is less than predicted we get a negative deviation or error. Now, whenever we predict we want to do that with the greatest overall accuracy. So, one possible criterion we can think of is to find the line that minimizes the sum of all deviations or errors. This makes sense, but we run up against a familiar problem which is that the sum of the positive deviations might balance out the sum of the negative deviations and this won't do; so we apply a familiar solution to this problem which is to square all the deviations so that all become positive, and their sum will always be 0 or greater. Now, our task is to find the line that minimizes the sum of square deviations also called SSD or the sum of squared errors which we can call SSE, but both mean the same thing. Now you can try out various combinations of intercept and slope, say a million combinations, and for each of them calculate the SSE and then find that intercept and slope which results in the smallest SSE.

This is called a brute force approach, and as you can imagine, requires considerable computational power. Luckily, for a simple problem such as finding the equation of a line that minimizes SSE there is an analytical solution – meaning we can *directly* calculate the answer, rather than rather than having to explore lots of possibilities using a computer. I won't show you how to calculate the solution because you can easily look it up.

But the point is that what we have now is called the *regression* line, which is the line that minimizes the sum of squared error and therefore is also called the least squares line. More generally, the procedure we have just followed is called *regression*, a term that you may have heard before. And so, from now on, we will call this assessment of the association between two variables, x and y , as *regression analysis*.

(Refer Slide Time: 27:07)

Assessing Association

Regression analysis – terminology

y
Outcome
Dependent
Response

Causal
Independent
Predictor
x

145

Let's remind ourselves of some notation, and also bring up some terminology so that we can use it later. Although we do not assume cause and effect, typically, we put a possible causal variable on the x -axis and the outcome variable on the y -axis. Other terms used are independent or predictor variable for the x variable, and dependent or response variable for the y variable. Some of these words have a very causal connotation, but again, resist falling into that trap. The main thing is that we are trying to predict y from x whether or not the relationship between them is causal.

(Refer Slide Time: 27:44)

Assessing Association



Regression analysis – equation

$$\begin{aligned}\hat{y}_i &= a + bx_i \\ y_i &= \hat{y}_i + \epsilon_i \\ y_i &= a + bx_i + \epsilon_i\end{aligned}$$

< 146 > ⋮



Now we can write an equation relating y to x . We use \hat{y} to denote predicted values of y . \hat{y} is described by the equation of a line and is equal to

$$\hat{y}_i = a + bx_i$$

where a and b (that is, the intercept and slope) are estimated through the regression analysis. But of course, the individual y values deviate to some degree from the predicted line. So, we can say that the individual y values are the predicted values plus some error; and of course, because the individual values, the predictions, and the errors are different for each value of x , we can make this explicit by putting the subscript i for y , \hat{y} , x , and e . The intercept and slope are not subscripted, because they apply to the entire relationship. In other words, if y and x are variables, a and b are called constants.

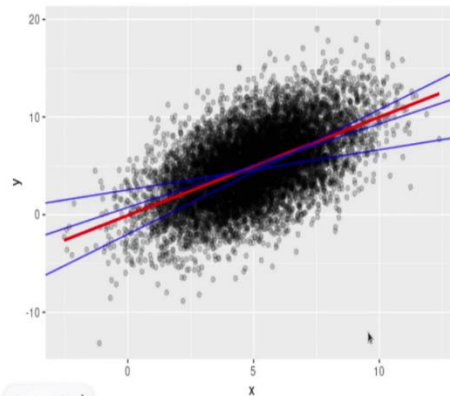
Now, as we have discussed several times already, just as the sample mean is an estimate of the true population mean, here the sample intercept and slope are estimates of the true population intercept and slope.

(Refer Slide Time: 28:55)

Assessing Association



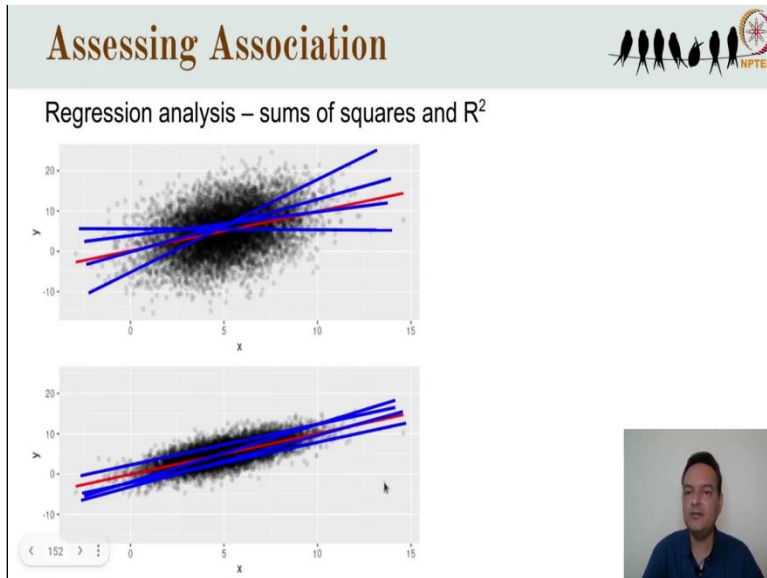
Regression analysis – sampling from a population



Let me depict this visually. Say the scatter describes the entire population, and the entire population shows a mild positive relationship between y and x with a fair bit of scatter. Overlaid onto this is the true population regression, calculated as if we were all-knowing. But of course, we are not all-knowing. So, from this population we sample 30 entities at random and then do the regression analysis, and find the regression line now in blue.

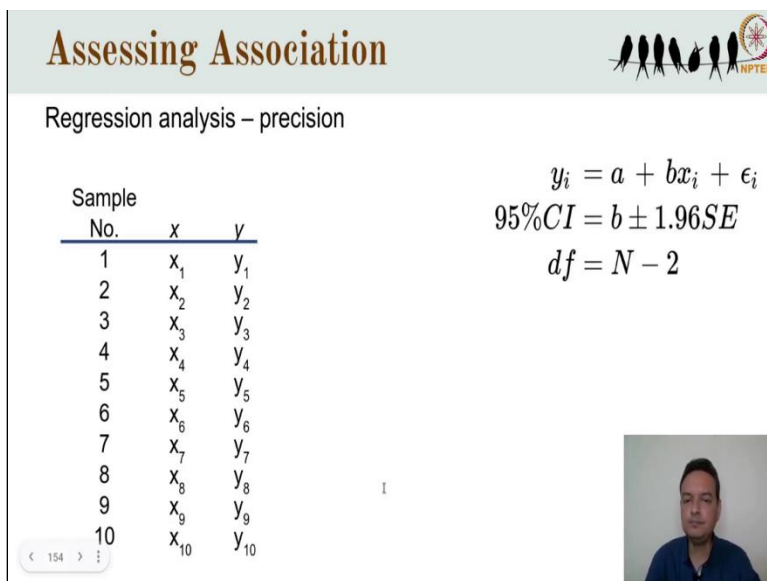
Notice that the intercept and slope of the sample regression are not the same as that of the population regression. If we sampled a different 30 entities from the population at random then we get another pair of values for intercept and slope, again different from the population regression, and so it is with another sample of 30. So, this all is just to reiterate that the sample estimates may be our best guess for the population parameters, but they could be a close guess or they could be a poor guess; and this depends on the precision of our study, which in turn depends on the variability in the population and our sample size.

(Refer Slide Time: 30:02)



To illustrate this, above we have a population relationship with high variability -- we can also call it high scatter – and below is an example of low variability or low scatter. You can imagine that in the first case each set of samples can result in quite different regression lines shown in blue but in the second case the regression lines for different sets of samples are unlikely to be all that different from each other.

(Refer Slide Time: 30:29)



So, just as before we want to understand the precision of our estimates and in particular, we typically focus on the precision or the *slope*, because that is the quantity of interest in most cases, whereas the intercept is rarely of much significance in our analysis. Now, just as before, we want

something like a 95% confidence interval for the sample slope such that we have an idea about the range within which the true population slope is likely to be.

Again, just as earlier, we have two broad ways of arriving at such a confidence interval. We can make minimal assumptions and use a bootstrap method. In this case, we can resample the entities we have measured, with replacement, each time generating a data set of the same sample size as the original. And each time we calculate and store the intercept and slope of the regression line. We then find the 2.5% percentile and 97.5% percentile, and that interval is our 95% confidence interval because it contains 95% of the possible sample slopes.

Or we can use the other method which assumes that the sample slopes follow a normal distribution – or a t distribution of course at low sample sizes. Then, as before, we need to find the standard error of the slope and multiply the standard error with some number – which is 1.96 when the sample size is greater than 30, and we can assume that the distribution of sample size slopes is normal – or that multiplier is something else (not 1.96) if the sample size is less than 30 and we need to use the t distribution instead.

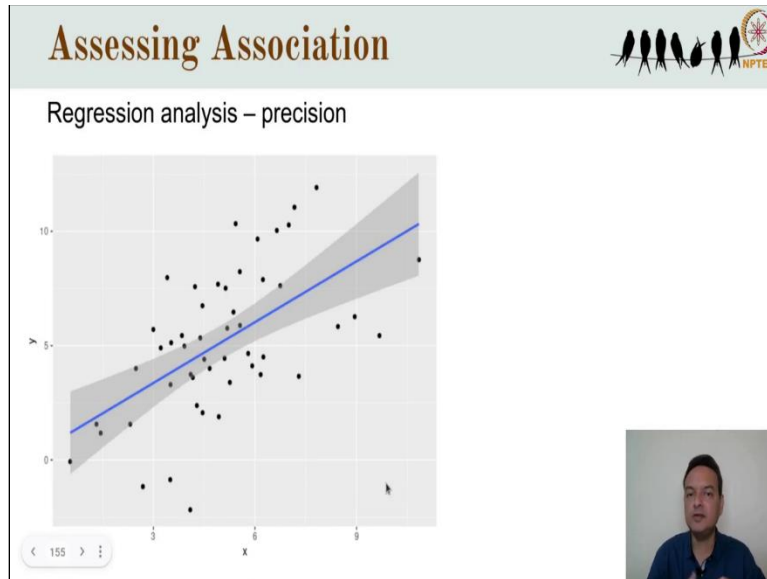
Now the standard error of a regression slope is not quite as easy to calculate as the standard error of the mean of a single variable and so I won't show it here. Suffice it to say for now that the software you use will give you the slope of the regression line as well as its standard error. Once you have that, we find the appropriate number to multiply with the standard error to give the 95% confidence interval around the slope. And just remember, if you use the t distribution, you have to calculate the degrees of freedom as sample size minus 2

$$df = N - 2$$

because you have lost one degree of freedom in estimating the intercept and a second degree of freedom in estimating the slope.

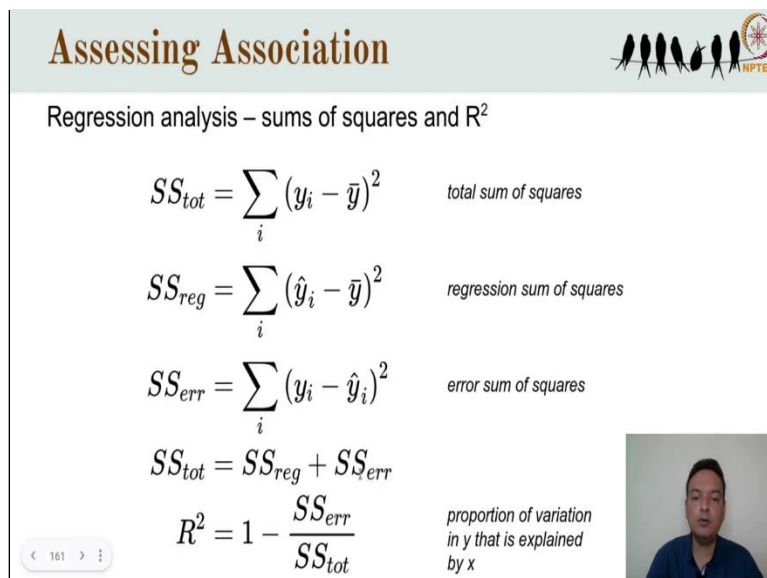
So now you have the confidence interval, and you can also go ahead and find out the probability of type S error – that is the probability that the true slope is in fact of the opposite sign than your best guess; and you can calculate the odds of the true slope being of the sign of your best guess rather than the opposite sign.

(Refer Slide Time: 33:02)



Now when visualizing a regression, you can imagine that you plot the response variable on the y axis and the predictor on the x as discussed earlier. This is called a scatterplot. On top of the scatterplot, you would overlay the regression line, and then in order to depict the uncertainty in the regression line it is common to show the 95% confidence interval in the intercept and slope which appears as the dark grey region over here. This gives a visual idea of the range of possibilities for what the true population regression might look like. So always be sure to plot the confidence interval and not only the regression line.

(Refer Slide Time: 33:42)



Now, one additional measure that you will often see accompanying regression results is something called R^2 . It is important that you know what it is, so let us find out. The variability in y (the quantity that we are trying to predict) can be described by different quantities of what are called “sums of squares”. Before you carry out any regression analysis, your best guess for the y value of any particular entity is simply the mean of y . So, the sum of the squared deviations of all the y 's from their mean is called the *total sum of squares*.

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

Now, once we have carried out the regression, our best guess for the y value of a particular entity is the *predicted y* from the regression line. By how much do these predicted values differ from the mean? If we sum the squared difference between the predicted values and the mean of y , we call that the *regression or explained sum of squares*.

$$SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$$

Finally, even after carrying out the regression, the predictions are not perfect, and the individual y values deviate from the predictions to some degree. The deviations from predicted are also called residuals or errors and the sum of the squared residuals is called the *residual sum of squares* or the *error sum of squares*.

$$SS_{err} = \sum_i (y_i - \hat{y}_i)^2$$

Now, the interesting thing is that the total sum of squares is equal to the regression sum of squares plus the error sum of squares.

$$SS_{tot} = SS_{reg} + SS_{err}$$

From this it follows that the larger the regression sum of squares, the smaller the error sum of squares and if the error sum of squares is small it means that the points are close to the regression line. In this situation, the predictions are more accurate and the line fits the data better. So, if we were able to calculate the fraction: residual sum of squares divided by total sum of squares (SS_{tot}), then the higher that number, the better the fit of the regression line.

In practice, this is most easily calculated using the error sum of squares. So, if you divide the entire equation by the total sum of squares on both sides and rearrange the terms you will see that the quantity we want is equal to one minus the error sum of squares divided by the total sum of squares.

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

This quantity is called the coefficient of determination or R^2 . R^2 can range only from 0 to 1.

Now if there were absolutely no relationship between y and x then the slope of the regression line would be 0 and the intercept would be the mean of y . So, the regression sum of squares would be 0 and the error sum of squares would be the same as the total sum of squares (SS_{tot}), making the R^2 0. At the other extreme if all values of y fall exactly on the regression line then the error sum of squares would be 0 and the R^2 would be 1.

So this property means that the R^2 can be interpreted as the proportion of variation in y explained by x . Recall that the slope does not tell you this, you can have two data sets with the same slope but with different scatter. So, the slope and the R^2 together tell you about both the *average rate of change of y with x* as well as the *proportion of variation in y explained by x* .

And as a small aside, it turns out that the R^2 is actually just the square of our old friend the correlation coefficient. So, everything is related to one another.

(Refer Slide Time: 37:05)

Assessing Association



Regression analysis – assumptions

Overall

- Validity - are you measuring what is meaningful?

Deterministic

- y is linearly related to x

Stochastic

- Errors are independent of each other
- Errors are drawn from normal distribution with constant variance

$$y_i = a + bx_i + \epsilon_i$$

deterministic
stochastic



It is now time to examine some of the key assumptions of regression analysis. The first and most important assumption is that of *validity* – that what you are measuring is meaningful to your research question. If you are interested in body size, is the best measure wing length or leg length or body mass or some combination of the three? Does the total number of species really measure the aspect that you are interested in, which is conservation value of a patch of habitat? If your measurement does not correspond relatively closely to what you care about in your research then no amount of fancy statistics can help.

For further assumptions of regression, we need to look at the regression equation and divide it into two parts – what is called the *deterministic* or fixed part of the equation ($a + bx_i$) and what is called the *stochastic* or random part of the equation (ϵ_i).

$$y_i = a + bx_i + \epsilon_i$$

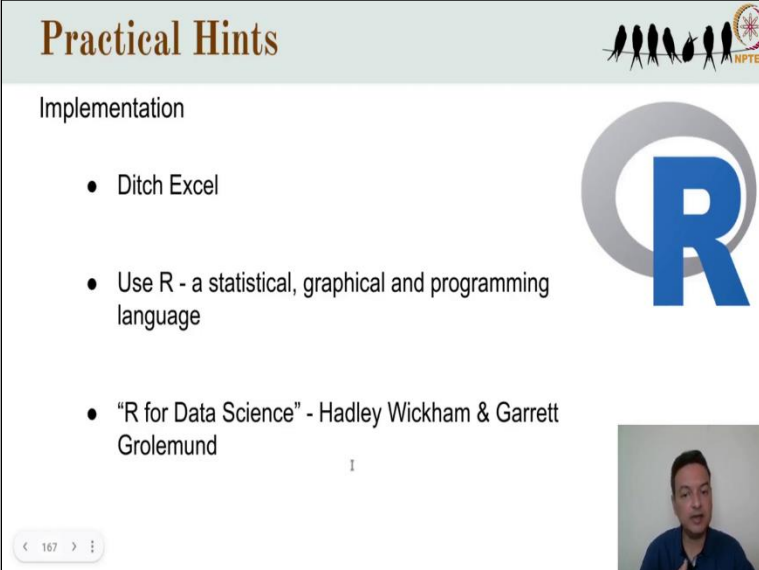
The deterministic part specifies how exactly we expect y to be related to x and the stochastic part tells us about the error or deviations or residuals -- all these three referring to the same thing.

So the next most important assumption is about the deterministic part of the equation. The assumption is that y is indeed related to x in the manner specified. In a simple regression the relationship is a straight line. So, the assumption is that y is linearly related to x. This may seem reasonable to you, but consider that many variables of interest cannot go below 0 including number of species, population size, time spent on a particular activity and so on. On the other side, some

measures cannot go above 1 or 100, like probabilities or percent of area under forest, or proportion of time spent foraging. And in other cases, we may expect a curved or saturating relationship like between the area of a forest and the number of species it hosts. This is likely to grow to a point and then stop. A more extreme example is when the slope of the relationship changes sign for example in the relationship between number of species and habitat disturbance: we often see the most species and places that have intermediate levels of disturbance. In all these cases, simple linear regression will be at best misleading, and at worst it will be utterly wrong. Now there are ways to identify and deal with all these sorts of situations but for now I want to encourage you to not simply assume that the relationship between y and x is linear but rather think about it carefully first and also make sure you plot the data to check for deviations from linearity.

Lastly, we have some assumptions about the *stochastic* part of the regression equation. The assumptions are that the error or residuals are independent of one another and are drawn from a normal distribution with constant variance. These assumptions are particularly important when we want to use the shortcut method to calculate standard error and confidence intervals from the properties of the normal distribution or t distribution; but the assumptions about the stochastic part of the equation are less important if we use the bootstrap method to find the precision of our estimate.

(Refer Slide Time: 40:17)



The slide is titled "Practical Hints" in a brown serif font. In the top right corner, there is a logo for NPTEL featuring a stylized bird and a gear. The main content is under the heading "Implementation" and consists of three bullet points:

- Ditch Excel
- Use R - a statistical, graphical and programming language
- "R for Data Science" - Hadley Wickham & Garrett Golemund

To the right of the text is a large blue "R" logo. At the bottom left, there is a navigation bar with a left arrow, the number "167", a right arrow, and a vertical ellipsis. At the bottom right, there is a small video inset showing a man speaking.

Now there are many more aspects of regression that I would leave you to explore, including how to detect points that lie outside the main cluster of points, and therefore have a disproportionate effect on the results, and what to do them if anything should be done. I also leave you to explore more complicated regression models with more than one predictor variable including predictors that are categorical and not only numeric.

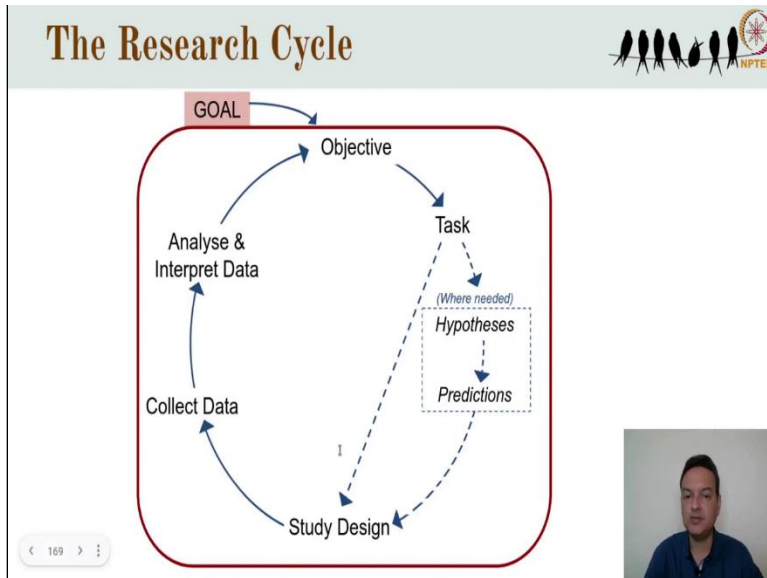
But even while exploring these more complicated versions, keep in mind that most of the fundamental aspects discussed here remain the same. Remember that we are sampling from a population, and that our goal is to come up with some conclusion about the population parameter

– and of course, that conclusion is meant to provide information needed to answer the research question we started with.

Now having discussed visualization and analysis of single quantities comparisons as well as associations between two variables you may be wondering how one actually implements all these ways of visualizing and analyzing data. If you are serious about research and data analysis the one main piece of advice I would give you is to move away from using spreadsheet programs like excel and learn a statistical and programming language like R. All the graphs and analysis I have shown you in this video were done in R – it is one of the most powerful tools you can learn for all manner of graphics and analysis. And because it is not just a software for statistics and graphics but also a programming language, it is very flexible and you can implement just about any analysis that you want using R.

There are many online resources through which you can learn R and I would say once you have a basic understanding of R, I would suggest you find and work through the free online book called R for data science. And if you manage to work through to the end of that book you will be very well placed to tackle any kind of graphical or analytical challenge that you may encounter in your research.

(Refer Slide Time: 42:17)



And finally going back to where we started this video, I wanted to say that although each of the different aspects of the research cycle we have spoken about are often treated separately with, for example, separate books on research design and separate books on statistical analysis, in reality the different boxes here should not be thought of as independent of one another. Ideally you would view all the steps as a whole, and plan and design your study having thought about each step.

For example, after framing your objective it is best to think about the tasks needed to be carried out to meet the objective. Then think about the design of the research -- for each task, imagine what kind of data you might collect, how you might visualize, analyze and interpret it, and then assess whether that will meet the objective. Plan out different combinations of tasks, different research designs, and different analyses, so that you can compare them and choose which best meets your overall need. Often, you might start by deciding that a particular kind of graph or comparison is what is needed to meet your objective, and from that starting point you might design your study. So, to summarize, before you actually embark on your study, it is useful to scribble out the entire process in a notebook, sketching out various possibilities and then deciding which to follow, making sure that you follow good practices in each of these steps, and that the logic has no flaws that you can see. In fact, it is a good idea to explain your plan to several friends and colleagues, and ask them to pick holes in it -- ask them to find problems in it. Better to identify opportunities, and even deficiencies, *before* you embark on your study, rather than have them pointed out to you after you have spent six months or three years on your work.

So, this brings us to the end of a rather long lecture, if you are already familiar with many of these concepts. I hope this has been a useful refresher. If much of what I have said is new, do not expect it all to sink in immediately. You may have to watch this video multiple times, pausing here and there and sketching things out with pen and paper, or crunching numbers on a computer. Understanding data and dealing with it to extract the kinds of meaning we are looking for is not a simple task, nor does it come instinctively to most of us.

But for much of our research in which quantitative analysis is a key element, it is worth trying to develop a sort of an intuition about numbers -- about basic mathematics, statistical inferences and data visualization. This will not come immediately or easily, but rather through a combination of reading and learning, as well as experience and handling data. And in the meantime, please try to resist the temptation of viewing data analysis as a recipe that can be followed from a textbook of statistics. Instead, while you carry out your analysis, try to understand every step of what you are doing well enough to feel it in your bones, so to speak. Now this is easier said than done, I know, but I hope that this lecture helps start you off on this journey.