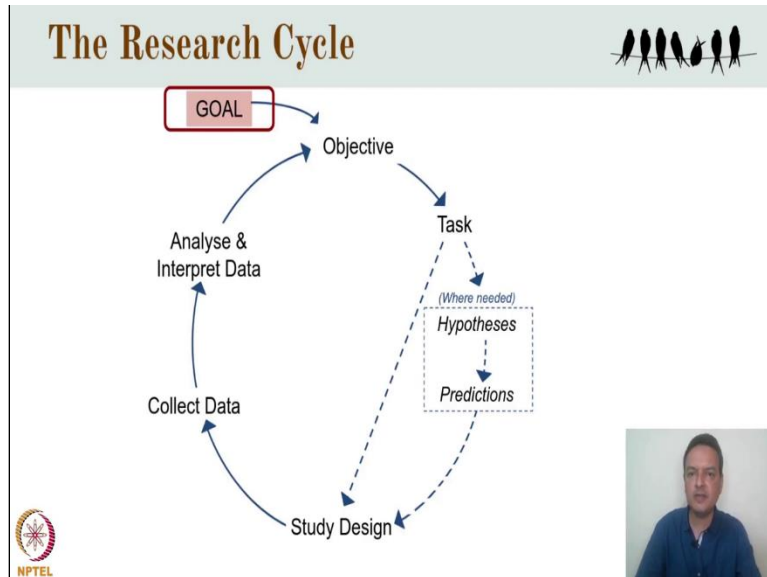


**Basic Course in Ornithology**  
**Dr. Suhel Quader**  
**Nature Conservation Foundation**

**Lecture -19**  
**Basics of Research Design**

(Refer Slide Time: 00:26)

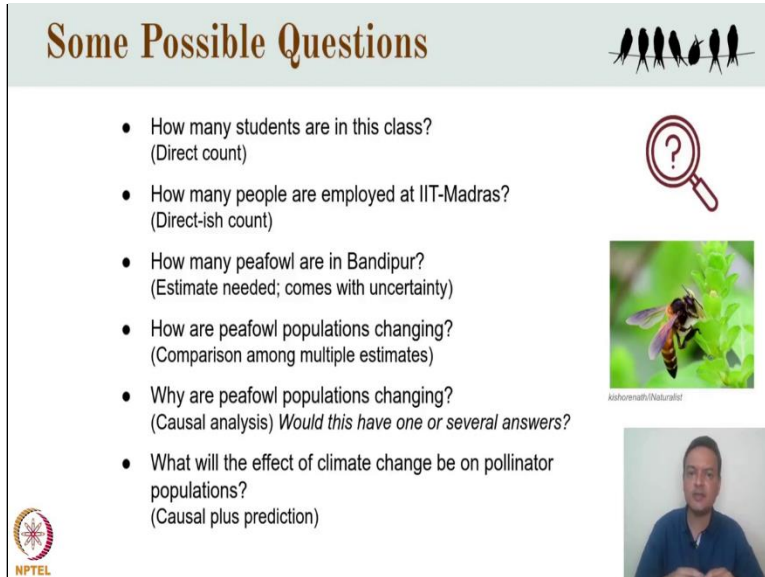


Welcome to a brief outline of research design. Here is a reminder of the generalized research cycle. After having thought carefully about the motivation underlying a study, we need to articulate a clear goal for the study that flows from the motivation. The goal in turn leads to one or more objectives, often phrased as questions. From those objectives are derived different tasks which might also be phrased as questions.

If answering the task-level question is not possible through direct observation, we have to articulate one or more hypotheses from which we deduce observable predictions that can be tested through observation or experiment. Of course, if the phenomenon of interest is directly observable then hypotheses and the predictions are not needed. In this lecture, we will talk about the next phase, once we have determined what the key tasks are in order to meet the objectives. We then need to carefully design our studies such that we can learn as much as we can,

and so that we can come to the strongest possible conclusion about the question that we are asking. Even if a study starts out with the most interesting and important questions, its success depends on the care and attention that you pay to the design of the study. Poor study design can lead to an inability to answer the original question or worse: to a confident but wrong conclusion.

**(Refer Slide Time: 01:47)**



**Some Possible Questions**

- How many students are in this class?  
(Direct count)
- How many people are employed at IIT-Madras?  
(Direct-ish count)
- How many peafowl are in Bandipur?  
(Estimate needed; comes with uncertainty)
- How are peafowl populations changing?  
(Comparison among multiple estimates)
- Why are peafowl populations changing?  
(Causal analysis) *Would this have one or several answers?*
- What will the effect of climate change be on pollinator populations?  
(Causal plus prediction)

The slide includes several icons: a row of birds at the top right, a magnifying glass with a question mark, a photograph of a bee on a flower, and a small video inset of a man speaking. The NPTEL logo is in the bottom left corner.

Let's start by sketching out a few different kinds of questions and the sorts of tasks they imply need to be done. We start simple and increase the complexity of.. complexity of the questions as we go along. If the task before us is to understand how many students are in this class, it is rather straightforward: all we need to do is to conduct a direct count. If we need to know how many people are employed at IIT Madras, things become a little more complicated.

We need to define employment and whether it does or does not include contractual, temporary and part-time staff and then we need to find some means of counting them up.

Answering the question - how many peafowl are in Bandipur tiger reserve, adds more complexity - it is impossible to count each peafowl one by one. And so we need to estimate the answer by sampling some fraction of the sanctuary perhaps through a technique like line transects, and then carefully extrapolating to the whole sanctuary. This is called an estimate, since we have no access to the true total number, but we hope that our study design and analysis brings us somewhere close to that true number. Estimates are always accompanied by a measure of uncertainty, a measure of how confident we are that our estimate is close to the true number.

If we now need to know how peafowl populations are changing, we usually need multiple estimates of the population over time, and our measure of change over the years also comes with uncertainty.

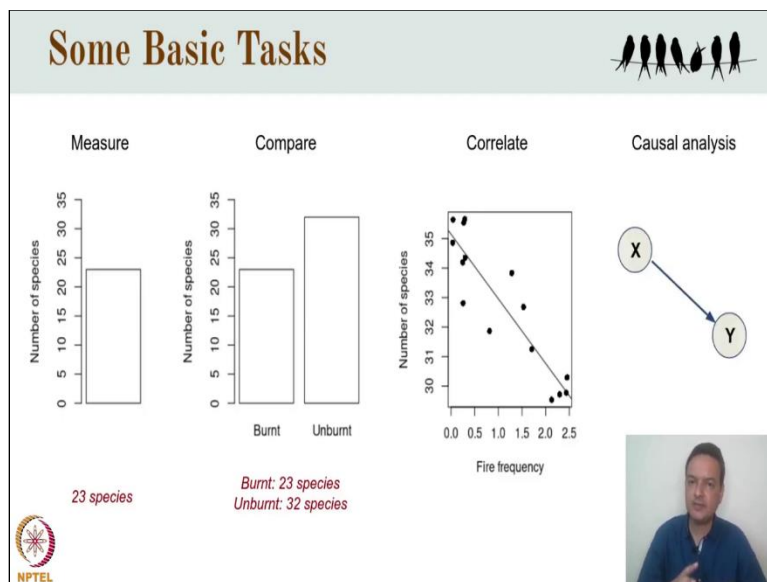
Asking why populations are changing now shifts us from the realm of description into the realm of explanation, and in this case, we are interested in cause and effect. Because processes are usually not observable, we would typically think of multiple possible hypotheses that might underlie the population change; derive predictions from each one and then collect information to see which predictions match with the evidence.

An additional complication is that it is not necessary that only a single factor is the cause of population change - instead multiple factors can add act together.

And finally in our list is a question that involves predicting or forecasting what might happen in the future, and this might require some explanatory or causal understanding plus some means of extending that understanding into the future.

So, you can see that there is a range of complexity in the questions we may ask and therefore in the corresponding tasks as well.

**(Refer Slide Time: 04:23)**



Basic tasks in research usually fall into one or more of the following types. We may need to measure a single thing in a single place. For example, the number of bird species in a particular grassland. We devise a method to do so and come up with an estimate. Here it is 23 species. More often our need is to make a comparison. In this example we want to compare the number of species in burnt and unburnt grassland to see which category of category of grassland has more species and by how many.


Here we see that unburnt grassland has nine more species than does burnt grassland. There may be more than two categories to compare and they can be different kinds of comparisons, for example, from one time period to another.

Another common task is to look for associations or correlations between two measures. In this example, we see that low fire frequency in a grassland corresponds to more bird species and higher fire frequency to fewer species.


Now of course it is tempting to conclude from this that fire has the effect of reducing species richness but you know quite well that correlation does not necessarily imply causation. And so, the last kind of task I have listed here is causal analysis, where we are interested in understanding whether x affects y in a causal manner. Tackling this requires careful thinking and design which we will talk about later in this video.

**(Refer Slide Time: 05:59)**



## Questions in Study Design



- What should I measure?
- What are the sampling units or replicates?  
*(On what entities should I conduct measurements?)*
- How should I choose sampling units and at what scale?
- How many replicates should I have?



Subhadra Devi / Macaulay Library at the Cornell Lab



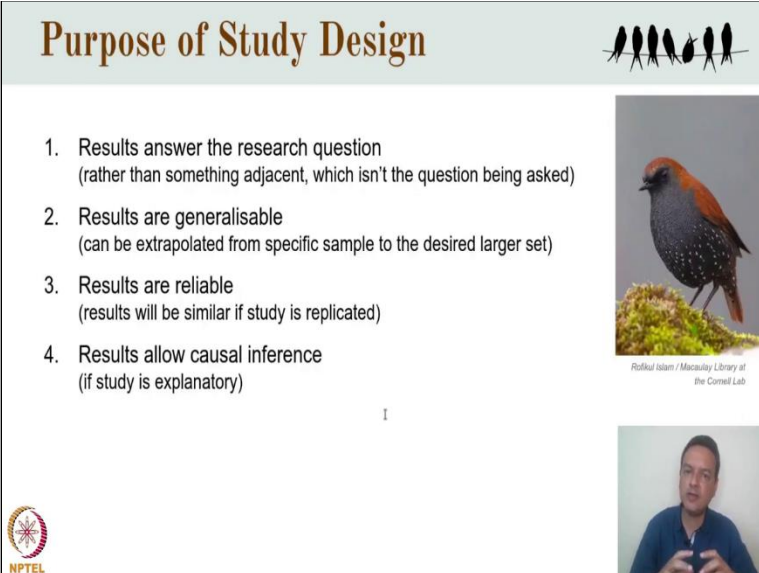
When it comes to the actual design of a study, several questions need to be answered. First, what should be measured? Is it the number of species or the number of individuals of a particular species, or some aspect of the behaviour of individuals, and so on.

Second, on what entities should these measurements be made? Number of species might be measured along a transect and number of individuals perhaps in a similar manner, but behaviour often needs to be measured on individuals, and so the sampling unit might be an individual bird. If you are measuring nesting success then nests would be your sampling units.

Third, given that there are a large number of potential transects or possible individuals to measure or nests to monitor, how do you choose which ones to actually measure and across what scale in space and time should they be chosen from? And finally, we need to think of how many replicates we should have:

How many transects to walk, how many individual birds to follow, how many nests to monitor, and so on. To try and answer these questions in our study design, it helps to understand the main purposes of good study design.

**(Refer Slide Time: 07:14)**



**Purpose of Study Design**

1. Results answer the research question  
(rather than something adjacent, which isn't the question being asked)
2. Results are generalisable  
(can be extrapolated from specific sample to the desired larger set)
3. Results are reliable  
(results will be similar if study is replicated)
4. Results allow causal inference  
(if study is explanatory)

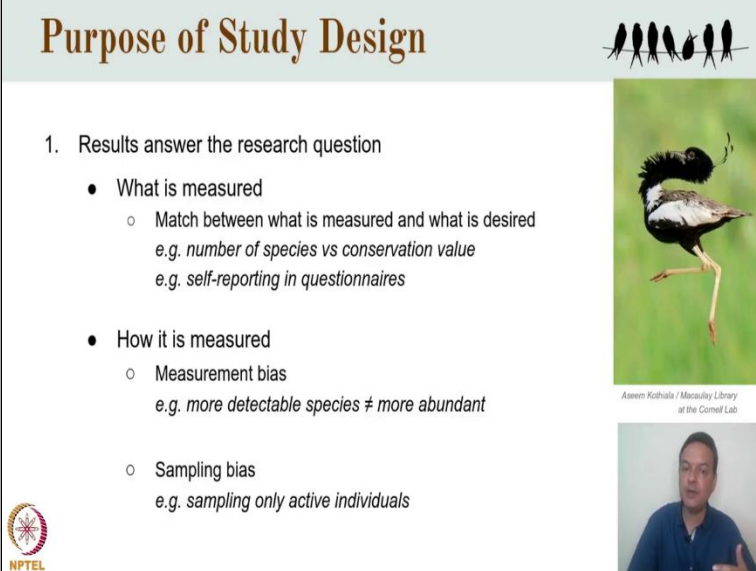
Rodkui Islam / Macaulay Library at the Cornell Lab

NPTEL

In other words, what are we trying to achieve through good study design? Well, we want our result to actually answer the specific question we have posed. If we are not careful, we may end up

answering something similar to but not quite what we need. We want our results to be generalizable, which means we want to be able to say something about the larger world not just the specific sampling units we have measured. We want to maximize the reliability of our answer which means that if the study is redone by ourselves or by others the result should not be very different. And finally, if the study is explanatory, we want to be able to conclude something about cause and effect and this requires additional effort. So, let's look at these one by one to see what the implications are for the design of our studies.


(Refer Slide Time: 08:04)





**Purpose of Study Design**

1. Results answer the research question

- What is measured
  - Match between what is measured and what is desired  
*e.g. number of species vs conservation value*  
*e.g. self-reporting in questionnaires*
- How it is measured
  - Measurement bias  
*e.g. more detectable species ≠ more abundant*
  - Sampling bias  
*e.g. sampling only active individuals*



  
Aseem Kothalia / Macaulay Library at the Cornell Lab



Whether the results actually answer the specific research question depends on both what is measured and how it is measured. The main thing to ask ourselves when we decide what to measure is how closely that matches with what we need in order to answer the research question. For example, let's say we want to compare two locations in terms of their conservation value for birds. We decide to measure the total species richness which is the total number of species.

But, one can argue about whether this is a good measure of conservation value. Perhaps a better measure might be the number of endemic species or the number of threatened species or the number of habitat specialists. But then again is any kind of species richness actually giving us what we need? Perhaps population densities are more important, or demographic processes like survival and reproduction.

You will see that there can be quite a gap between what we measure and what we might claim or want to conclude. Here is another example: say we want to understand the level of crop damage by parakeets in an area and we decide to ask farmers about whether their fields have suffered damage and by how much. There is again a possible gap between what we want and what we have measured since our results can only tell us about how much damage is *reported*, and this could possibly be quite different from the damage that *actually* occurs.

Next, let's examine *how* we measure things. This can also limit the extent to which our results answer the actual question being asked. And in particular, we can think of various forms of measurement bias where the quantity we are interested in might consistently be over- or underestimated. An example of this is when we go out into the field and want to come up with an assessment of the relative abundance of different species, but of course, the species we see most frequently and in larger numbers are not necessarily the ones that are the most abundant, because of differences among species in detectability. Some species might be large, colourful, active and loud, all of which bias our estimates of their abundance to the high side, compared with small cryptic species. So, our results might be completely off if we do not think about sources of measurement bias and attempt to account for them.


The bias described in this example is obvious and well-known and so are the methods of correcting it but there are likely to be many biases that are more subtle and we need to work hard to uncover them and deal with them.

Another source of bias is sampling bias. Let's understand this through an example. Say we want to understand the time activity budget of a species, which is the proportion of time it spends in various activities like flying, foraging, resting, preening and so on.

We go out find an individual of the species, record what it is doing, and repeat this for many individuals. You can imagine that in many cases, birds that are more active are more likely to be found. And those perched quietly in a bush are less likely to be found. This means that our estimates of the time spent in active behaviour will be biased high compared with our estimates of the time spent in say resting or preening which in turn means that we have to think about ways of reducing or eliminating this bias.

**(Refer Slide Time: 11:34)**

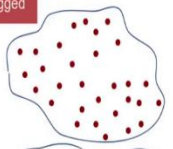
## Purpose of Study Design




2. Results are generalisable


- What is the **sampling frame** (the larger set)?  
(also called 'population')
- How **representative** is the sample?
  - Sampling strategies: haphazard/convenience; random; stratified random, systematic
- Are the replicate samples **independent**?
  - Non-independence and pseudoreplication


A: logged



B: unlogged







Our second requirement in designing studies is that the results can be extrapolated from the individual units we measure to some larger set. For example, we may walk five transects in Bandipur; but of course we want to conclude something about Bandipur as a whole not just those specific transects. Or we may measure the behaviour of 30 birds, but those specific individuals are important only to the extent that they allow us to generalize to a larger set, say all birds of that species in that locality.

So, if generalization from the specific sample to some larger set is important to us as it almost always is. Then the first thing to ask ourselves is what is that larger set? This is also called the population that we want to be able to generalize to do. We want to conclude something about the density of peafowl in Bandipur reserve as a whole or only a single forest range within Bandipur; or all of southern India.

Similarly, are we interested in saying something about the behaviour of House Crows in Dehradun city specifically, or across the entire western Himalayan foothills. Once we have decided this, it should make sense to you that when we choose which entities to measure - be they transects or individual birds or anything else they should be representative of the larger set. In other words, sampling units must be chosen in a manner such that taken together they are representative of the larger sampling frame or population.



So, how can we try and make sure that our sample is representative? In the absence of any particular strategy, people may take a haphazard sample or a convenience sample which is just a way of saying that the researcher measures whichever entity they encounter. But this may easily lead to biases, for example, we might sample only the most active individuals or lay transects in the most accessible terrain.

Note that just wondering about sampling the first bird you see might be casually described as 'random' but it is far from random in the technical sense. The formal definition of random for our purposes here is that whether we are talking about transects or birds or anything else, all sampling units have equal probability of being sampled, and that probability is independent of which other sampling units have been chosen.

So, the wandering-about method should be described as haphazard rather than random.

How do we choose samples at random? Well, if we know all possible sampling units in our population, we can number them and then use a random number generator to select a subset of them to be measured. For example, using a map you could generate a complete list of possible one hectare squares in Bandipur and then choose at random among them, by which I mean truly random using a computer program.

Or if there are 200 nesting trees of storks in my study area, I could number them all and use a random number generator to select say 50 of them at random. Although random sampling is often described as the best way to ensure representativeness this outcome is not guaranteed. It is true that when your sample size is large - that is if you are measuring large numbers of sampling units - then random samples are likely to be representative, .

but when sample size is small then there is scope for trouble. You know for example that when you roll a die the probability of getting a six is one by six. But suppose you do not actually know this probability and instead you have to estimate it by rolling the die and counting up what happens. So, if you roll a die a thousand times and tally up how often you got a six that fraction will be roughly one in six times. But if you roll a die only 30 times then purely by chance you could get a result that is quite different from one by six.

I did it just now and I got a six only twice which means the fraction was 1 in 15, quite different from the true probability. And so, if I relied on that result my answer to the question what is the probability of getting six would be quite wrong.

Similarly, if I had two species on a map like this in red and blue and selected only six individuals at random, I could very easily get four blues and two reds purely by chance and conclude that blues outnumbered reds by two to one. Or the converse - I could get 4 reds and 2 blues and conclude that the red is twice as abundant as blue when the truth is that they are actually equal in abundance in this example. So, what can we do to maximize the chance that our sample is representative? Here are some possibilities. Sometimes, you know that your largest set or population is composed of different *kinds* of entities with different properties these are called strata.

Examples of strata may be sex of birds (male and female) or age (juvenile and adult) or habitat type (riparian forest or dry deciduous forest) and so on. In such cases, you could stratify the population which means divide it up into these strata and then sample at random within each stratum. This ensures that the various importance strata are adequately represented and that you have not left them out by chance.

This strategy is called stratified random sampling.

For example, over here it turns out that there are two distinct habitats - the top left is dry deciduous forest and the bottom right is moist deciduous forest. You can see that the red species is more abundant in the dry forest and the blue species is more abundant in the moist forest. But of course, you do not actually know that in advance. By ensuring that these underlying strata are separated, and then sampling within each stratum, you reduce the possibility of unrepresentative sampling purely by chance.

When you do not know about underlying differences you might instead systematically divide your population into subunits. For example, you might overlay a grid over your study area and ensure that there is at least one sampling unit chosen, at random, within each grid cell and this strategy is called systematic random sampling. When considering the various options available to you, remember that the larger purpose is that the sample is representative of the population, such that your results are generalizable from sample to population.

And finally, we need to think about whether our replicates provide us with information that is *independent* of other replicates. In many ways, the need for independence of replicates flows from the requirement that the sample be representative of the population, but I have listed it in its own point here since it is often discussed separately.

Let us take an extreme example say you wanted to describe the length of the tail of peafowl properly called the *train*. You find one male and measure his train 10 times; the same male.

Then you take an average of these measurements and say that your sample size is 10. This is clearly absurd because although you can say something about that individual male with great precision. From the population point of view, generalizing from only one male is very risky. It is very clear that the 10 replicate measurements are *not* independent from one another and they are actually false replicates or pseudo-replicates.

Whenever replicates are expected to be more similar to each other than the similarity expected in the population as a whole, they are called non-independent or pseudo-replicated. And this can happen if you want to generalize across individuals but actually take repeated measures on the same individuals as in the peafowl example; or if you want to generalize, let's say chick growth across nests, but treat multiple chicks in the same nest as independent of each other or if you want to generalize across ponds, but treat multiple measures within each pond as independent replicates.

Here is an example of how you can easily be misled by pseudo-replication. Say you are interested in the effect of timber extraction on bird species richness. You have two forest patches one that has been logged and one that has experienced no logging. In each forest patch, you conduct 30 point counts where the points are carefully laid out through a stratified random design. So, at the end of your study you have 30 data points in each from which you can see whether logged forests are different from unlogged forests in their bird species richness.

But actually, no, you have only two forest patches A and B. and in your conclusion you can say something about the average difference between A and B, but these are the *only* examples of logged and unlogged forests in your data set. If you wanted to come to a conclusion about logged

and unlogged forests in general, then your data are highly pseudo-replicated. In actual fact you have only *one* logged forest patch and *one* unlogged forest patch.

This is not to say you should not carry out such studies just that you have to be very cautious about what you conclude based on the degree of generalization that is possible from your study design.

**(Refer Slide Time: 21:26)**

The slide is titled "Purpose of Study Design" and features a decorative header with silhouettes of birds. The main content is a list of points under the heading "2. Results are generalisable". The points are:

- What is the **sampling frame** (the larger set)? (also called 'population')
- How **representative** is the sample?
  - Sampling strategies: haphazard/convenience; random; stratified random, systematic
- Are the replicate samples **independent**?
  - Non-independence and pseudoreplication

The slide includes a topographic map of the western Himalayas with three red dots indicating study sites. An inset map shows a grid of yellow dots representing a sampling frame. A small video inset in the bottom right corner shows a man speaking. The NPTEL logo is in the bottom left corner.

Similarly, if you want to study how bird species richness changes with elevation in the western Himalaya, you might take a kind of a transect in Uttarakhand from say Corbett at 350 metres above sea level to Mandal at 1500 meters to Tungnath at 3500 meters. And at each of these sites, you could have 20 plots where you record species richness. But although this seems like a lot of information, if your purpose is to make a statement about the western Himalaya as a whole you actually have only *one* site at each of the three elevations.

So, although you might be able to contrast Corbett, Mandal and Tungnath with considerable confidence, it would be rather risky to generalize to the western Himalayas as a whole: for that purpose the replicates are non-independent.

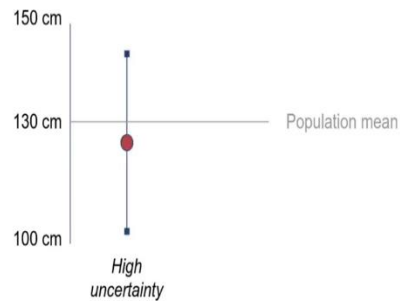
**(Refer Slide Time: 22:22)**

## Purpose of Study Design



### 3. Results are reliable

- What is the precision (or uncertainty) of the result?



Copyright - Birds of the World/Cornell Lab of Ornithology



On to the next purpose of good study design, we want our studies to be reliable which means that if someone, including you, were to repeat your study again the result should not be very different from what you originally found. We ideally want our answer to be as close as possible to the truth - specifically, the estimate you get from your sample should be close to the *true* measure for the population from which you have sampled and that you wish to generalize to.

The tricky thing is that because in most cases you will never be able to measure the whole population the truth is *unknown* and we have to use all our wits to try and get as close as possible to it. For example, say the average length of a peacock train in a population of peacocks in Delhi is 130 centimeters. This by the way is actually unknown and also unknowable because no one will ever be able to measure all the peacocks

in Delhi. But if we sample in a representative and unbiased manner, as discussed earlier, then as we keep measuring more and more peacocks our estimated mean should get closer and closer to the true population value. So, for any study, a researcher has to estimate and present also the *precision* of the result. In practice it is the *uncertainty* (that is the inverse of precision) that is estimated and presented.

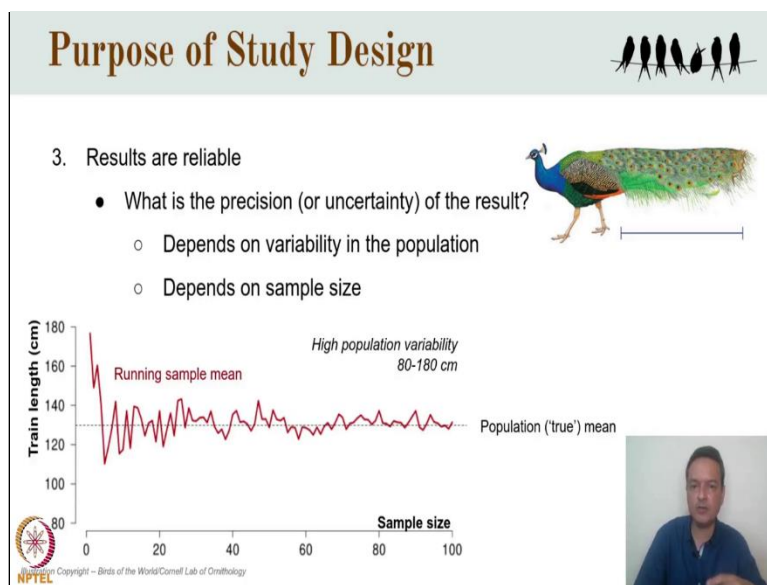
For example, I might sample peacocks in Delhi and estimate a mean train length of 126 centimeters with an uncertainty of plus minus 20 centimeters. The uncertainty tells us how confident we are

that this estimate of 126 centimeters is close to the true value, assuming an unbiased sample. In this case, the confidence is rather low the data we have collected is consistent with a true mean train length of as low as 106 centimeters as well as as high as 146 centimeters - quite a wide span.

And remember that although I have marked in the graph where the population mean lies, in reality, we do not know where it is. All we know is what the data tell us and from this our best guess is that it is somewhere between 106 and 146. By contrast say the uncertainty was plus minus 5 centimeters then we conclude that the true value is likely to be somewhere between 121 centimeters and 131 centimeters, which is a much better situation to be in since the uncertainty in the true value is quite low.

Again, do remember that we do not actually know the true value but rather are trying to estimate it as best we can and in this case, because we have low uncertainty, we have been able to estimate it much better than in the earlier example.

**(Refer Slide Time: 25:23)**



This leads us to the following question - what affects our measure of precision, or its inverse, uncertainty? Firstly, it depends on how variable the population is. At one extreme, if all peacocks had a train length of exactly 130 then all we might need to do is measure one individual and we have our answer. If there was a little variability say train lengths were all between 125 and 135

centimeters then even if we measured only 5 or 10 peacocks, we should still not be far from the correct answer.

At the other extreme, if train lengths were highly variable say between 80 centimeters and 180 centimeters; then you can imagine that by measuring only a handful of individuals we could get an answer that purely by chance is very far from the truth. The amount of variability in the population is just a part of the phenomenon we are studying we have no control over it. This brings us to the second point.

Although we have no control over the variability in the population, we do have control over our sample size - in this case, the number of individuals we measure. For a given amount of population variability as we increase the sample size, we increase the precision and decrease the uncertainty. If you imagine taking more and more samples and calculating a running mean as you go, you will see that the estimate at the start can be quite far from the population mean.

But as you accumulate more samples and move to the right on this graph. The sample mean will converge on the population mean assuming, no measurement bias or sampling bias of course. This example here is for a population of peacocks with high variability - with train lengths varying from 80 to 180 centimeters. And as you can see even after reaching a sample size of about 100 the mean continues to change a bit with each additional sample.

So, even with a relatively large sample size there is some uncertainty in our estimate. By contrast when there is very little variation in the population, with train lengths varying only between 125 and 135 centimeters, then just a handful of samples are enough for the sample mean to stabilize very near the population mean, and we can reach a conclusion about the population mean with high precision, that is, with low uncertainty.

**(Refer Slide Time: 27:55)**

## Purpose of Study Design



### 3. Results are reliable

- What is the precision (or uncertainty) of the result?
  - Depends on variability in the population
  - Depends on sample size



$$\text{Sample size} \propto \frac{\text{Variability}}{\text{Uncertainty}}$$



Copyright - Birds of the World/Cornell Lab of Ornithology



Uncertainty is therefore directly proportional to population variability and inversely proportional to sample size. So if you know the population variability - normally estimated by the variability in your sample - and if you know the sample size, you can calculate what the resultant uncertainty in your study will be. Rearranging terms, you can decide to aim for a particular level of uncertainty, and then calculate what sample size you need to get there.

This is usually an important part of planning a study, you collect some pilot data to estimate the population variability, decide what precision you need - is plus minus 20 okay or do you need plus minus 5 for your purpose - and then calculate the needed sample size to get there. We won't talk about the actual calculations at the moment, but for now if you intuitively understand the relationship between these three aspects, that is excellent.

**(Refer Slide Time: 28:57)**

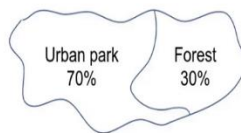


## Purpose of Study Design



### 3. Results are reliable

- What is the precision (or uncertainty) of the result?
  - Depends on variability in the population
  - Depends on sample size
  - *Measurement error adds to variability*
  - *How to distribute samples across strata*



Distribute samples by area?  
Or by variability?



Copyright - Birds of the World/Cornell Lab of Ornithology



Two further things to say about uncertainty, As mentioned before population variability is out of your control but actually measurement error also adds to it. So, if you estimate peacock train length by eye - that is just by looking at it - then the large measurement error, in doing so, will add on to the population variability. And so, to achieve a particular degree of precision you will need a larger sample size compared with if you measure train length with a measuring tape or some other more precise method.

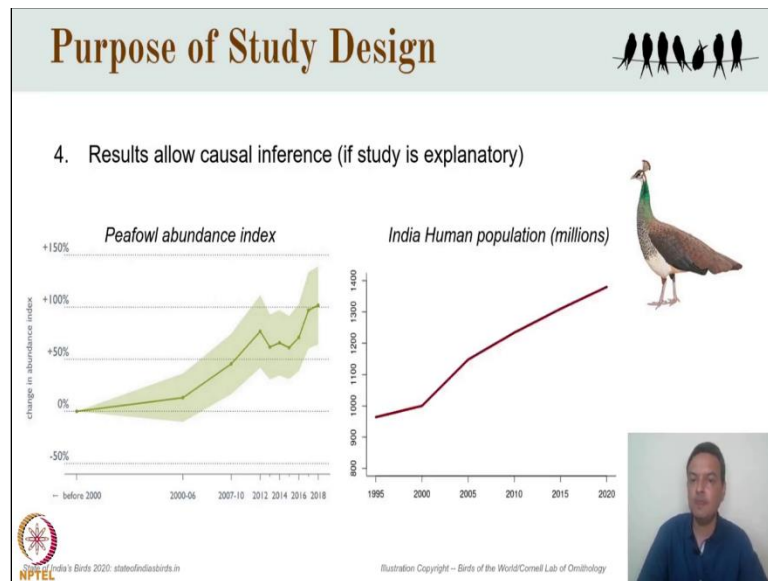
This is separate from the possible bias in visual estimation: if you tend to overestimate train length when viewing by eye then with large sample size the estimate might be quite precise but it will be biased too high. The precision means that if you carry out the study again with a new sample of peacocks you will get a very similar answer. But, any bias will mean that *that* answer would still be quite far from the true population mean.

The second point to add here is that considerations of precision and its relationship to population variability and sample size have implications not just for overall sample size but also for how to allocate sample size across strata. If some strata are more variable than others then we need to sample them more.

Say we want to calculate overall peafowl densities but know that there are two habitats - forests and urban parks.

In the absence of any further information, we might sample by area. If forests account for 30% of the area and parks cover 70% of the area, then we could quite reasonably place 30% of our transects in forests and 70% in parks. But ideally, we would conduct a pilot study to understand variability. And if it turns out that peafowl density is more variable in forests than in parks then our calculations might lead us to place more transects in forests and fewer in parks. Clearly, there are lots to think about when it comes to trying to maximize the reliability of our studies.

**(Refer Slide Time: 31:09)**



A fourth possible purpose of good study design is to understand cause and effect. This is relevant for studies that are not about just describing patterns, but attempt to understand the underlying processes or mechanisms that lead to the patterns we see. Not all studies aim to do this, but it is interesting that even studies that say that they are simply describing patterns tend to begin to use causal language in the discussion section of the resultant article sometimes even in the abstract or title.

Clearly, it is a human trait to seek explanations for the things we see around us. Unfortunately, as we know well, understanding cause is not merely a matter of observing whether there is an association between two measures. For example, the State of India's Birds report shows that the abundance of peafowl has increased over the past 20 years. At the same time, we know that the human population has also increased. But I do not think anyone would claim that peafowl have

increased *because* humans have increased, or indeed the other way around. So clearly, we need to do more to understand cause and effect.

**(Refer Slide Time: 32:21)**

**Purpose of Study Design**

4. Results allow causal inference (if study is explanatory)

- Experiment (eg Randomised Controlled Trial): 'gold standard'
  - Treatment vs Control (ie, some kind of comparison)
  - Allocating subjects to treatments
    - Random allocation
    - Stratification

COVID-19 VACCINE

Random shuffling after stratifying by female and male

400F, 600M      4F, 6M

200F, 300M (40%F)      200F, 300M (40%F)      2F, 3M (40%F)      2F, 3M (40%F)

NPTEL

The slide features a title 'Purpose of Study Design' at the top left, a decorative graphic of silhouettes at the top right, and a 'COVID-19 VACCINE' graphic with a syringe on the right. A small video inset of a speaker is visible in the bottom right corner.

Let's look at two kinds of situations in which we want to infer cause. The first is where an experiment is possible. Experiments are often considered to yield the best possible evidence for inferring cause and effect. You will have heard of RCTs or Randomized Controlled Trials referred to as the gold standard of evidence in various fields of science. A straightforward example is in clinical trials, let's say to test the efficacy of a potential Covid vaccine.

Here, we may inject one group of people with a potential vaccine, and also have another group of similar people who do not receive the vaccine; and then see what fraction of each group contracts Covid. If the vaccine is effective then a smaller fraction of those who receive the vaccine should contract Covid, compared with those who did not receive the vaccine. Now of course conducting an experiment is not quite as simple as that. So, let us look at some of the key aspects that we need to consider.

First, experiments typically involve a comparison between two categories - sometimes called a treatment and a control. But the comparison can be across multiple treatments instead. So, it is best to think of a control as just a special type of treatment. The treatments should differ only in the specific aspect that we are interested in.

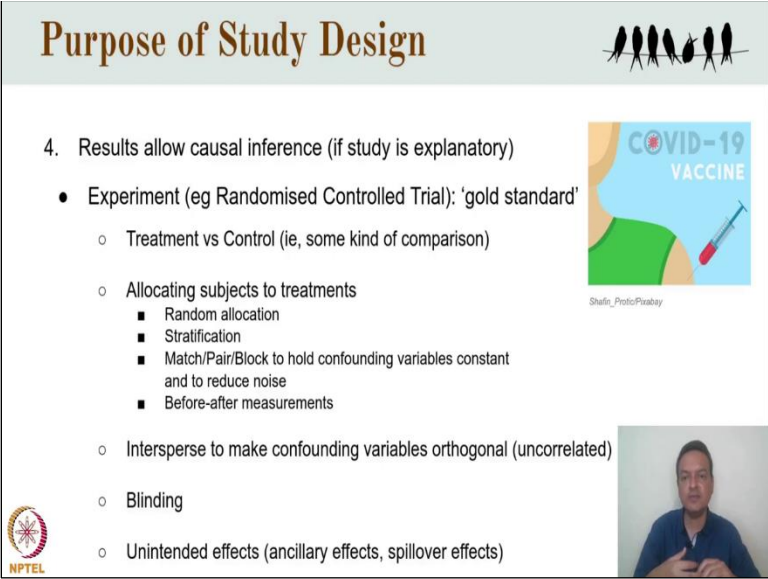
In the Covid vaccine example, we do not want the control to consist of people who have not been injected at all because we know that the very act of being injected, no matter with what substance, can alter a person's physiology and immune system and this is one example of what is called the placebo effect. And instead, by injecting people in both groups, one with a vaccine and the other with something neutral like saline, we allow ourselves to infer that any resultant differences between the groups are the result of the vaccine specifically and cannot be attributed to anything else. Next, we have to worry about how to allocate subjects to treatments. In this case, how do we distribute the volunteers in the clinical trial between vaccine and non-vaccine groups - which are the two arms of the experiment. While doing this we want to ensure that the two groups do not have any pre-existing differences on average between them.

For example, if we leave it up to the subjects to decide which group they want to be in, perhaps those more at risk for contracting Covid might prefer to be in the treatment group and then the resultant infection rates may be related more to *those* risk factors than to the vaccine itself. There are various ways to try and ensure that there are minimal underlying differences between the treatment and control groups. If the number of subjects is large then we can randomly allocate each subject to one group or another. Remember that random has a technical meaning, which in this case is that the probability of a particular person going to one group or the other is equal and is independent of the probability of any other person going to one group or the other. If you put the first 50 people who volunteer into the treatment group and the next 50 people into the control group, that is clearly not random; and I leave it to you to think about what kind of underlying differences they might be in this case which could contaminate the result.

Now, when the number of subjects is small random allocation may not be sufficient to ensure that underlying differences are minimal between the groups. If I have a thousand subjects of whom 400 are female and 600 male and I randomly shuffle them into two groups, it is quite likely that both groups will be around 40% female. But if I have only ten subjects - four female and six male - then when I randomly shuffle them into groups I can quite easily, purely by chance, get a breakup of one is to four or three is to two or something even more extreme. And that could be a problem for my experiment.

One obvious thing we can do is to stratify when there are *known* differences among subjects. So, in this case, we could have two strata, male and female, and ensure equal representation of each sex in the two treatment arms - in this case the treatment and control. So, in this way through stratification we do not allow these chance effects to lead to these very very different outcomes.

**(Refer Slide Time: 37:07)**



**Purpose of Study Design**


4. Results allow causal inference (if study is explanatory)

- Experiment (eg Randomised Controlled Trial): 'gold standard'
  - Treatment vs Control (ie, some kind of comparison)
  - Allocating subjects to treatments
    - Random allocation
    - Stratification
    - Match/Pair/Block to hold confounding variables constant and to reduce noise
    - Before-after measurements
  - Intersperse to make confounding variables orthogonal (uncorrelated)
  - Blinding
  - Unintended effects (ancillary effects, spillover effects)

**COVID-19 VACCINE**

NPTEL

Shafiq\_ProtoParabay



But there might also be more subtle or invisible differences among people such that we do not know how to stratify. One approach we can use is to use something called matching or blocking. For example, suppose subjects from the experiment are drawn from different localities across a city. It is quite possible that these localities are differentially exposed to Covid. So, we can then *match* subjects according to locality; such that locality A sends one subject to the treatment and one to control, and locality B sends one subject to treatment and one to control and so on. And this allows us to contrast the efficacy of the vaccine within each locality and then some across all localities to come up with an overall answer.

A final way of trying to ensure that underlying differences among subjects do not bias an experiment is to measure the feature of interest both before and after the treatment. This does not work so well for the vaccine example we have been using but let us say we want to know the efficacy of a drug to reduce blood pressure. We know that each person differs in their baseline blood pressure. So, rather than implementing the intervention and simply comparing average

resultant blood pressure in treatment versus control groups we also measure blood pressure *before* the intervention.

And our outcome is now calculated as the average before-after *difference* in blood pressure between treatment and control groups.

Now, there may be other factors that we are not interested in but can still affect the outcome and we need to ensure that *they* do not contaminate the results. For example, if the experiment is taking place over an extended duration, we should ensure that early subjects are not differentially allocated to one experimental arm or the other compared with subjects who enter the experiment later. Systematically interspersing treatments among subjects ensures that other confounding variables like time are orthogonal (which means uncorrelated) to the experimental treatments. In practical terms this might mean that as volunteers walk into the clinic, we assign them to treatment and to control arms in alternating manner.

Another common practice in experiments is called blinding in which the experimenter does not know which subjects were allocated to which treatment. This is necessary to ensure that any unconscious bias the experimenter might have does not affect the results. You can imagine that where vaccines cost a lot to develop, and can potentially result in large earnings, pharmaceutical researchers need to be very careful that experiments are blinded, in order to avoid being misled. In the case of human subjects, the subjects themselves are often also kept blind to which experimental arm they have been assigned to treatment or control. This ensures that any expectation that *they* may have about the effect of the treatment does not influence the outcome. If I know that I have been given a drug that is supposed to lower my blood pressure that can well have the unconscious effect of calming me down, and in fact, lowering my blood pressure, even if the drug itself does not work at all or if I know I have been given a drug that is intended as a vaccine I might subsequently engage in more risky behaviour,

and the resultant calculations of vaccine efficacy are contaminated by the altered behaviour of those in the vaccine arm of the treatment.

And finally, there can be unintended effects of experimental treatments which I will describe as part of the next example.

**(Refer Slide Time: 41:03)**

**Purpose of Study Design**

- 4. Results allow causal inference (if study is explanatory)
  - Experiment (eg Randomised Controlled Trial): 'gold standard'
    - Treatment vs Control (ie, some kind of comparison)
    - Allocating subjects to treatments
      - Random allocation
      - Stratification
      - Match/Pair/Block to hold confounding variables constant and to reduce noise
      - Before-after measurements
    - Intersperse to make confounding variables orthogonal (uncorrelated)
    - Blinding
    - Unintended effects (ancillary effects, spillover effects)

NPTEL

Copyright - Birds of the World/ Cornell Lab of Ornithology

Let us go through all these ideas about experiments using a bird example. Say we want to understand whether there is an effect of song playback on the time activity budget of birds. The context is that bird watchers, photographers and tour guides sometimes play bird song through loudspeakers to bring birds out into the open to be better seen and photographed. And one concern is that it might disturb the birds and disrupt their normal behaviour.

So, to examine this we decide to conduct an experiment, in which some individual birds of a species say Magpie Robins are subjected to song playback. And we observe their subsequent behaviour in terms of how much time they spend foraging, resting, singing and so on. Now since we need a comparison, we decide to have a control group that is not subjected to song playback. Remember that treatment and control groups should differ only in the intervention of interest, which is song.

So, the control birds need to experience the same kind of other experimental effects. The experimenter approaches them to the same distance and perhaps even plays back a sound that is *not* song. Some people play white noise at the same decibel level as they play song in the treatment arm, or they may play the call of some other bird species. In this way, we can conclude that any

differences are due to song alone and not due to general human disturbance or attributed to *some* sound being played back versus *no* sound.

Now, we need to decide *which* individual Magpie Robins should be in the treatment arm and which in the control arm. If we plan for a large sample of Magpie Robins then we can label each bird in the population and allocate them at random to one or the other experimental arm.

We can also see if stratification makes sense especially when we are constrained to small sample size but also a good idea with large samples, perhaps the birds are in two different habitats forests and gardens. Then we can stratify accordingly. Or we might worry that their time activity budgets might be affected by smaller scale phenomena, which we do not know about. In this case, we could match or pair them in space such that we take pairs of neighbouring Magpie robins and then within each such duo, we allocate one at random to treatment and the other to control and then after the experiment we look at the difference in behaviour *within* each duo and average across all such duos in our sample. Further Magpie Robins might have *underlying* differences in behaviour - some might be naturally much more active than others for example. To take this into account, we might measure the activity levels of our subjects before the playback intervention and contrast that to their activity after the playback. So, this was about allocating subjects to treatments such that the comparison between experimental arms tells us what we want to find out.

In addition, there could be a number of confounding variables, including time of day and season which might also affect our outcome, which is activity levels. We need to ensure that our experimental replicates are appropriately interspersed across these such that we do not end up with, say, more treatment trials early in the morning and more controls later in the morning; or more controls early in the season and more treatments later in the season. So, here is where some manner of systematic interspersion is needed.



**(Refer Slide Time: 44:40)**




## Purpose of Study Design

4. Results allow causal inference (if study is explanatory)

- Experiment (eg Randomised Controlled Trial): 'gold standard'
  - Treatment vs Control (ie, some kind of comparison)
  - Allocating subjects to treatments
    - Random allocation
    - Stratification
    - Match/Pair/Block to hold confounding variables constant and to reduce noise
    - Before-after measurements
  - Intersperse to make confounding variables orthogonal (uncorrelated)
  - Blinding
  - Unintended effects (ancillary effects, spillover effects)

P.Jaganathan/WikiCommons





It is ideal if the researcher is blind to the treatment when collecting the data on activity levels else, she might be unconsciously biased while observing the birds, since, people often feel strongly about playback in one direction or the other. In a field study such as this you might think that it is impossible for the researcher to be blind to the treatment but there may be clever ways of trying this. For example, the bird could be video recorded and then its behaviour coded back in the lab by someone blind to *which* experimental arm the bird was subjected to.

**(Refer Slide Time: 45:18)**


## Purpose of Study Design

4. Results allow causal inference (if study is explanatory)

- Experiment (eg Randomised Controlled Trial): 'gold standard'
  - Treatment vs Control (ie, some kind of comparison)
  - Allocating subjects to treatments
    - Random allocation
    - Stratification
    - Match/Pair/Block to hold confounding variables constant and to reduce noise
    - Before-after measurements
  - Intersperse to make confounding variables orthogonal (uncorrelated)
  - Blinding
  - Unintended effects (ancillary effects, spillover effects)

Aseem Kothalia / Macaulay Library at the Cornell Lab

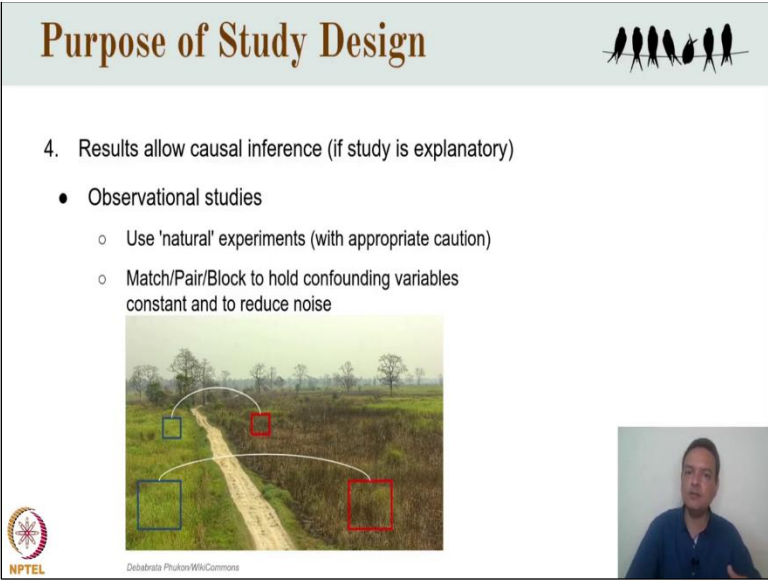


We have spoken about how there could also be effects that are not meant to be part of the experiment itself. For example, disturbance caused by the researcher's presence. Some of these effects can be accounted for by careful experimental design,

in this case setting up the appropriate control. But there could be *other* kinds of unintended effects for example, playing back song might attract rivals to the scene, or predators, and those in turn could affect the subject's behaviour. And then there could be spillover effects, in which the experimental treatment might unintentionally affect not just the designated subject but also other subjects. For example, playing back song in one bird's territory might affect neighbouring birds as well which may be a problem if those effects are long lasting and those neighbours are meant to be experimental subjects afterwards.

So, you will see that there are a variety of possible *unintended* effects of any experimental intervention and researchers have to think carefully about what these might be and what kinds of problems they might pose and how to deal with them.

**(Refer Slide Time: 46:32)**



**Purpose of Study Design**

4. Results allow causal inference (if study is explanatory)

- Observational studies
  - Use 'natural' experiments (with appropriate caution)
  - Match/Pair/Block to hold confounding variables constant and to reduce noise

The slide includes an image of a field with a dirt path and several experimental plots marked with colored boxes (green and red) and arcs, illustrating the concept of matching or blocking in observational studies. A small inset video shows a man speaking. The NPTEL logo is visible in the bottom left corner.

Now in many situations experiments are out of the question, perhaps it is not feasible to go in and change things for ethical or logistical or financial reasons, or in some cases there is simply no way to experimentally alter the potential causal variable like elevation or rainfall or temperature. Now this does not mean that questions of cause and effect do not apply here. On the contrary, often the most important phenomena to understand are those for which experimental manipulation is impossible.

So, what can we do in such cases? Studies in which researchers do not go in and alter potential causal factors are called observational. And although our ability to infer cause and effect through observational studies is limited this does not mean that we are completely helpless. There are various opportunities and strategies that we can take advantage of in order to maximize our ability to draw causal conclusions, even if any single study cannot reach 100 percent certainty.

First, we can keep an eye out for so-called *natural experiments*. For example, we may need to understand the effect of fire on grassland birds, but perhaps for various reasons we cannot experimentally burn grasslands. Now since the grasslands we study are subjected to burning for *other* reasons we can pay close attention and when a section of the area *happens* to burn, we can swing into action with our measurements of grass biomass, insect abundance and bird responses - comparing burnt with unburnt sections of the habitat.

If we are lucky there may be multiple burnt and unburned sections which we can then use as replicates. The major caution with natural experiments is that the *underlying* conditions may be quite different. For example, perhaps the drier part of the habitat is more prone to burning than the wetter part, and so the treatment and controls in our natural experiment have a pre-existing and underlying difference, limiting our ability to detect the effect of burning alone.

So two possible approaches to tackle this might come to your mind. One is to match treatment and control, in this context perhaps that might mean choosing to compare burnt and unburnt plots that are *close* to each other and therefore hopefully with minimal underlying differences. Then you would take the *difference* between burnt and unburnt in each pair, and average across all such pairs.

Another, and ideally additional, scenario is where you also had the relevant measurements before the habitat burned not just after.

By taking the difference before and after the burn, you can potentially subtract away underlying variation and use these differences as your measure of the effect of burning. The before-after difference in unburned sites is used to check for any background changes over time that have

nothing to do with the burn itself. Using these strategies, we may be able to maximize the confidence in any causal inferences we make.

Now the same strategies can be used in many other kinds of observational studies from which we want to draw causal inferences when looking at how eucalyptus plantations affect bird populations compared with forest you can again use matching or pairing to try and minimize underlying differences. Similarly, when comparing the effect of organic versus conventional farming on birds in agricultural landscapes.

Do be aware of possible spillover effects. If you choose sampling units that are too close to one another they may be influenced not only by the habitat that they are located in but also the nearby habitat that we are using as a comparison.

**(Refer Slide Time: 50:22)**

**Purpose of Study Design**

4. Results allow causal inference (if study is explanatory)

- Observational studies
  - Use 'natural' experiments (with appropriate caution)
  - Match/Pair/Block to hold confounding variables constant and to reduce noise
  - Measure Before and After
  - Intersperse to make confounding variables orthogonal (uncorrelated)

Early	F	M
Late	F	M
Early	F	M
Late	F	M

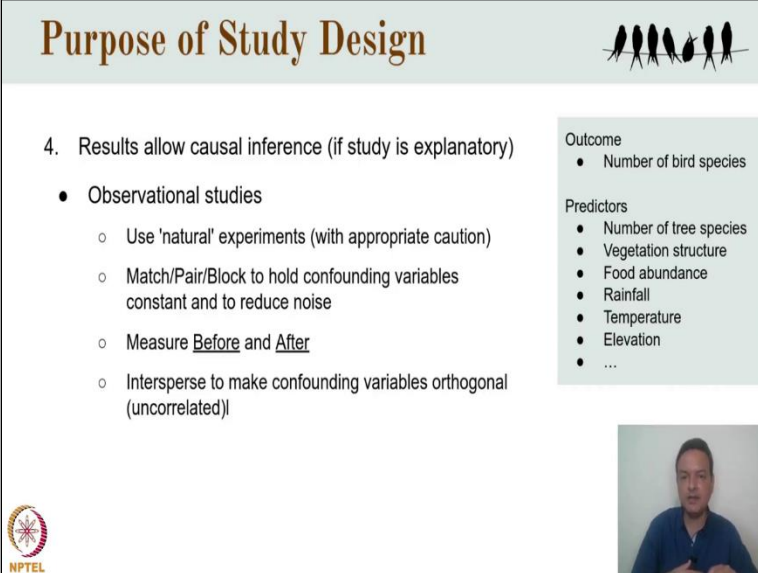
NPTEL

As always make sure you *intersperse* with other possible confounding factors like time of day, season, elevation, rainfall and so on. Interspersing, to remind you, has the effect of making other possible causal variables uncorrelated with each other. Here is an example.

Imagine that you measure females largely in the early morning and males largely in the late morning. Any conclusions you then draw about how the sexes differ in behaviour is confounded by time of day because sex and time of day are correlated in your design. In other words, you may

find that females differ from males in their behaviour but that difference could actually be entirely due to the *time* of day when observations were taken rather than to the sex of the birds. But then if you *intersperse* such that the number of males and females measured are *equally* distributed between the two times of day then sex and time of day have been made orthogonal (or uncorrelated) with each other. And this maximizes our ability to ascribe differences in behaviour to one causal factor versus another.

**(Refer Slide Time: 51:32)**



**Purpose of Study Design**

4. Results allow causal inference (if study is explanatory)

- Observational studies
  - Use 'natural' experiments (with appropriate caution)
  - Match/Pair/Block to hold confounding variables constant and to reduce noise
  - Measure Before and After
  - Intersperse to make confounding variables orthogonal (uncorrelated)


Outcome

- Number of bird species

Predictors

- Number of tree species
- Vegetation structure
- Food abundance
- Rainfall
- Temperature
- Elevation
- ...

NPTEL



Making possible causal factors orthogonal in the design of our observational studies is something we should be doing much more of. Many researchers measure a large number of variables, as what is called *predictors*, trusting that complex statistical analysis will separate among them and allow causal inference about what is called the *outcome*. For example, we might want to understand variation in species numbers which is the outcome and so, we run transects across a variety of locations. At these locations we collect information on various predictors, tree species, vegetation structure, food resources, rainfall, temperature, elevation and more. And we put all of these into a statistical model whose job it is to separate out important and unimportant predictors. This is difficult, and sometimes impossible, when the predictors are correlated among themselves as you can imagine many of these are likely to be.

In such a situation, careful study design in advance might lead you to choose your sampling locations more carefully, such that key predictors are orthogonal - perhaps having equal number of combinations of high and low rainfall combined with high and low elevation, and so on.

Now, unfortunately the results from even the most careful observational study following all the strategies described here are in most cases *still* not enough to confidently infer cause and effect. What else can we do?

(Refer Slide Time: 53:05)

**Purpose of Study Design**

4. Results allow causal inference (if study is explanatory)

- Observational studies
  - Use 'natural' experiments (with appropriate caution)
  - Match/Pair/Block to hold confounding variables constant and to reduce noise
  - Measure Before and After
  - Intersperse to make confounding variables orthogonal
- Try and triangulate with multiple lines of evidence

• Compare across space  
• Compare across time  
• Clinical trials  
• Modelling


NPTEL Copyright - Birds of the World/Cornell Lab of Ornithology

One additional thing we can do is to *triangulate* with multiple lines of evidence. For example, if we suspect that vultures are declining because of diclofenac poisoning, we can compare vulture mortality across different regions that have different prevalence of diclofenac. We can also try and find some before-after data from when diclofenac became widely used to treat cattle. We can feed diclofenac to captive vultures (or a surrogate species) to look at the physiological effect. And we can do some calculations to see whether the prevalence of diclofenac use is sufficient to explain vulture declines. Any single line of evidence would not be considered sufficient but conservation scientists have done *all* of these different things; and taken together, the combined evidence very strongly points to diclofenac as the primary cause of vulture declines in south Asia.

So, in your study, think about all the different angles that can be taken to attack the problem and pursue as many as you possibly can if you want your causal inferences to be as strong as possible.

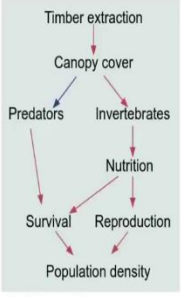
(Refer Slide Time: 54:10)

## Purpose of Study Design





4. Results allow causal inference (if study is explanatory)

- Observational studies
  - Use 'natural' experiments (with appropriate caution)
  - Match/Pair/Block to hold confounding variables constant and to reduce noise
  - Measure Before and After
  - Intersperse to make confounding variables orthogonal
- Try and triangulate with multiple lines of evidence
- Test multiple links in the putative causal chain



```
graph TD; TE[Timber extraction] --> CC[Canopy cover]; CC --> P[Predators]; CC --> I[Invertebrates]; P --> S[Survival]; I --> N[Nutrition]; N --> R[Reproduction]; S --> PD[Population density]; R --> PD;
```



And finally, we know that causal processes are usually not simple but rather follow a cascading chain of cause and effect. For example, timber extraction from a forest most likely does not directly kill birds, thereby reducing their population. Rather, the causal chain may look somewhat like this - timber extraction leads to opening of habitat, which increases the number of predators and decreases the abundance of invertebrates. Increased number of predators decreases the survival rates of small birds, and decreased invertebrates leads to lowered nutrition - thereby decreasing reproduction and also decreasing the survival rates of young and adult birds. And it is the decreased reproduction and survival that eventually leads to a decrease in population densities. Now, if we are to study *multiple* links in this hypothetical causal chain, and find that at least some of these conjectures is true, then we have a much stronger argument about what may be causing population change, than if we only had measured timber extraction on the one side and bird population densities on the other.

So, this was a quick tour of the basics of research design covering what *kinds* of research questions we might ask, what the basic decisions are in research design and how understanding the purposes of good research design can help us more carefully and effectively make those decisions when designing our studies.