ISES Europe Training Series

DoE 9: Statistics and Epidemiology

Module 1: Introduction to Statistics

Hello and welcome to this video lecture on statistics for exposure science.

Just a brief legal notice. We retain copyright for this presentation, so you need permission to use it, and it is for educational, non-commercial purposes.

Today's lecture is part of Domain of Expertise 9, Statistics and Epidemiology. It is the final of the nine Domains of Expertise, and you can access all the videos at this web link on the ISES Europe website.

Domain of Expertise 9 has three modules in it: today's module, Introduction to Statistics; then a follow-on module on statistics specifically for exposure scientists; and finishing up with an introduction to epidemiology.

Today's lecture forms the content for Module 1, on the Introduction to Statistics. It is very much an overview of some core concepts in statistics, which hopefully you'll decide to learn more about.

My name is Ruairí Weiner. I work in University College Dublin in the School of Public Health as a teaching fellow in research methods and statistics.

Just the usual disclaimers: this lecture is providing you with an introductory framework. Some topics have to be simplified in line with the short introductory video. We've tried to keep everything accurate, but there could be some errors. What's presented here doesn't represent the views of the presenter's employer or affiliated organisations.

So what are our learning objectives today? We want to gain an understanding of the statistical methods commonly used in exposure assessment, and to learn how to explore relationships between variables, to summarise data, and to handle uncertainties in the context of exposure assessment using these methods.

Today's presentation is broken into six sections. We'll start with the introduction section.

In this introduction section, I'm going to explain how we summarise statistics. We've collected some data, perhaps on an exposure. How do we summarise that data to understand an overview of what we've collected?

We have two choices of summary statistic: measures of central tendency, and measures of variability.

Central tendency, as the name suggests, is where the middle value in that data is, or the typical value. Measures of variability are about the spread of values around that middle.

Starting with central tendency: this is about describing where most of the data are. If you look at my image there on the right, we have a nice symmetric distribution. If you can imagine, this curve represents where the data are. Most of the data are in the middle, where the curve is highest, and then in the tails, where the curve is lowest, that represents where data are less frequent.

We want a summary statistic that describes where that middle is, where the data are the most frequent. If I don't know anything else about the variable that I've collected—the exposure—my best guess is to use a measure of central tendency.

If I want to guess what your height is, but I've never met you and I know nothing about you, my best guess is the average height. If I know your sex, I could guess the average height for your sex, and then I would make a better guess. But if I'm constantly guessing people's heights, on average I'll be closest to the right answer by using a central tendency measure.

The measure of central tendency I use depends on how the variable is measured.

Numeric data, like height and weight, I can summarise using the mean, which many people are familiar with. If my data aren't exactly numbers like a number of centimetres, but I can rank them in order—like my seniority in the university, ranking assistant professors below associate professors, below full professors—that's an ordinal variable. It's not a number, but I can put it in order. I can't calculate a mean, but I can get a median.

Nominal variables would be categories that I can't rank in order, like the colour of a car. I can't say that one colour is higher or lower than another, but I can report the mode. I can say which colour is the most common.

The mean is what we often think of as the average, where we simply add up all the values and divide by the number of values. The average height of seven people is: I add up all their heights in centimetres and divide by seven.

This is typically best for numeric data because I can use all the information that I have. I use the specific measurement I've taken for each case, each individual. When the data are symmetrically distributed, if you think of that curve from earlier, the mean is in the middle, right down the middle of the data.

But the mean is very sensitive to skew. What if most people are around 180 cm, but a few people are 190–200 cm? Then the data are skewed; they would have a long tail out towards those higher heights. The mean wouldn't be in the middle anymore. If I'm trying to figure out the central tendency of the data, I have to think about whether the mean is appropriate.

Similarly with an outlier: if I have one really, really tall person, or if I think of salary—if there are ten people in the room and most of us earn similar salaries, and then a billionaire walks in and we get the mean salary in the room—that mean doesn't really represent all the people who aren't billionaires anymore. Because of one very large outlier, the mean is very sensitive.

The median is when I rank all the values from the lowest salary to the highest, and then the median is the value in the very middle. If I have seven people in the room, I put the person with the lowest salary first and the highest last, and then that fourth person—the person in the middle of the line—would represent the median salary, the salary in the middle.

It uses less information than the mean. I didn't use your exact salary, only your rank in a sense. But by using the middle value, I'm no longer as sensitive to skew or an outlier when I'm looking for what's the middle of the data—the central tendency.

The mode is simply the value that occurs most often. Again, I'm using less information because I don't tell you anything about any of the other values. If I use the mode to summarise the colour of cars, and the most common car colour is red, then I might tell you the modal car colour is red—it's the most common colour.

Maybe there were six red cars, but I haven't told you anything else. I haven't told you what other colours I observed or how many. So it's not using information about the other cars, other than to figure out that red was the most common. But it's brilliant for data that can't be ranked, because it still lets me give some summary. I can't use the mean for the colour of cars, but the mode still tells me about the typical car.

If all car colours appeared equally often, and then there is one more red car, red is still the mode, even though it's barely more common than the other colours. The mode can be very useful if it's a lot more common than the other categories. It might be misleading to say that red cars are most common if there's only one more red car than all the other colours, and all the other colours appear equally often.

If we had six red cars and six green cars, and all the other car colours occurred once, then we'd have two modal colours, two modes: two colours that stand out for being more common. Then it would be bimodal. We wouldn't have just a single value as our mode.

So the mode is very useful for non-numeric data, but it has its limitations too, and it uses less information in a sense.

This little graphic here just summarises the difference between the mean, the median, and the mode. For the distribution in the middle, this nice symmetric curve, the mean, the median, and the mode converge. Because the data are symmetric, the mean is in the middle. The median is always in the middle—50% of the data are below it, 50% are above—and the mode is at the most frequent value.

You can see in the examples on the left and the right that when the data are skewed, for example on the left, we have most of the data clustered together, but then some very large negative values or some values which are much lower than most of the data, which pulls the mean down in a way that doesn't pull the median down. The mode stays wherever the most frequent value is.

So, in skewed data, the three measures of central tendency diverge. In symmetric data, they converge. Those summary measures give you a clue about the shape of your data distribution, about whether there is skew or not.

Measures of variability tell us how spread out the data are around that centre, around the middle of the data.

I'm going to tell you about four ways we can describe the spread of data. The **range** is simply the distance from the lowest value to the highest value.

The **standard deviation** incorporates information from all the values. It's a summary of how far away each value is from the mean—how spread out the data are around the mean. The closer people are, in this example in height, the smaller the standard deviation. If people's heights are very different, you get a bigger standard deviation, more spread on average in the data.

The **interquartile range** is where I go from the person a quarter of the way up the data—if you think about ranking from the lowest to the highest person again—to the person who's a quarter of the way from the top. So the lower quartile is the value such that 25% of people are below it; the upper quartile is such that 75% of people are below it. I go from that lower quartile to that higher quartile, and that I call the interquartile range. It takes the middle 50% of the data and describes how far apart the bottom and top of that middle range are. That gives us a sense of the spread without outliers. The tallest person might be very unusually tall, but if we just take the middle 50%, we get more of a sense of the range within those more common heights.

For data that aren't numeric and can't be ranked, we can still understand the spread of the data by looking at the **proportion**. Maybe red cars were the modal car, the most common, but then we can report for each of the other colours how relatively common they were. Simply: maybe 50% of cars were red, maybe 30% were green. Now I know how frequent each car colour was, relatively. That tells me about the spread of that kind of categorical data.

The next topic to introduce you to is the concept of **normality** of data.

Normality is an important concept that pops up all around statistics.

In essence, if numeric data vary randomly around a mean, they form a symmetric shape that we call a normal distribution. There's a mean height—we'll say it's 178 cm—and if the spread around that mean height looks random, whether you're higher or lower than that mean, then the data will be symmetric. You're not more likely to be higher nor lower than the average. It's a symmetric distribution.

If we assume data follow this normal distribution, then we can invoke a lot of powerful statistical techniques. A lot of statistical techniques rely on an assumption that data are normally distributed. If data aren't normal, there are different techniques we'll have to use that we'll talk about.

Here's an example of a basic drawing of a normal-ish distribution. It's symmetric. If you think of that vertical line as the mean, and the curve represents where the data are, the data are symmetrically distributed around that mean. Most people are close to the mean, and data are less likely as you get further away from the mean.

The normal distribution is actually very common in nature because it arises from random variation. Where differences between individuals are random, we tend to get this normal distribution. It is very

common in epidemiology. Things like height and weight typically do follow a normal distribution, whether it's for humans or dogs or whatever it is, and many biomarkers and anthropometric measures will follow a nice normal distribution. Most people are similar to each other; some people are quite unusual.

In the case of exposure science, we do have to consider some non-normal data. Particularly, **log-normal distributions** are very common in exposure science.

A log-normal distribution is a distribution that looks normal after applying a log transformation to it.

What does a log transformation do? Without getting into it too much, it's a multiplicative transformation of the data. It changes the data values, but the amount it changes the value depends on how big the value is. If a value is very large, the log transformation will pull it in a lot more than it will for a small value. A large value will be changed a lot; a small value will only change a bit. If we think of that skewed kind of data, it pulls in that long skewed tail, it pulls in some of those outliers, to make it more of a nice, normal, symmetric-looking shape.

Why does this come up a lot in exposure science?

If you think of exposure to some kind of industrial chemical: if I'm not working in an industry that uses the chemical, my exposure levels will probably be very low. But people who are working in the industry that uses the chemical will have much higher exposure levels, maybe an order of magnitude higher. So most people will have these low values, and then you'll have this long tail with the people who have been exposed, who have the high values. Exposure is often uneven, depending on things like occupation. That's why it comes up very often in exposure science, so it's important to flag. It also comes up a lot with biomarkers. There are a lot of biomarkers—say blood-based biomarkers—where we would have to do a log transform to see a nice normal distribution.

So that's just to be aware of. In the next module, on statistics more specific to exposure science, you'll learn a bit more about those kinds of data.

The next topic we'll cover in this video is, briefly, how you handle data which fall below the limit of detection or the limit of quantitation.

In statistics, we take measurements to estimate levels of exposure and then to test the relationships of, say, my occupation to my exposure levels. I have to measure how much exposure I've had. But sometimes levels of an exposure are so low that our instruments are unable to detect them.

As levels of an exposure decrease, a more sensitive measure is required just to detect the presence of the exposure. The limit of my measurement is the **limit of detection**.

Some exposures are so small that my instrument isn't sensitive enough to detect anything. The **LOQ**, the **limit of quantitation**, is the lowest level where we can reliably quantify the amount of a substance with acceptable precision and accuracy.

I want to do more than just say if the exposure is present or not. I want to do more than detect it. I want to **quantify** it. I want to put a number on your exposure and my exposure and say which one is higher, with a certain level of confidence. If I say your exposure is higher than my exposure, there's

always some measurement error. There's a level of confidence with which I can say your number is higher than my number, because there could be some measurement error. So my level of precision that's acceptable has to be defined, and I can define the limit of quantitation: how precise can I get, or how extreme a level can I measure and still put a number on it?

I may be able to detect that there is an exposure present, but it may be beyond the ability of my measurement to confidently say how much exposure there is, or which person has had more or less exposure.

Here are the proper EuroChem definitions, other than my rough definitions. They define the limit of detection as the lowest concentration of the analyte present in a sample that can be detected using a given measurement procedure with a specified level of confidence. The limit of quantification, or quantitation, they define as the lowest level at which performance is acceptable for a typical application.

In cases where we detect something outside of the limit of quantitation, we could state if it's above or below our limit. I tend to use strange examples, but let's imagine my limit is values above 1 mm. In this case, I won't be able to compare two things that I've measured as greater than 1 mm and say one is longer than another. But I may be able to say that both are long enough to meet some criterion. My measure maxed out at 1 mm, so I have a list of values where I've just said they are greater than 1 mm. I can't tell you exactly how long, but they're greater than 1 mm. Maybe that's enough if being greater than 1 mm is a criterion for some classification.

Take the example of a thermometer. We might have an upper and lower limit. The lowest my thermometer might measure is –30 degrees, and the highest maybe 80 degrees. If a reading is outside those limits, I might be able to record that the temperature was less than –30, or greater than 80. I can't tell you exactly the temperature. I can't say which day was warmer than the other if they're both recorded as less than –30, for example. But it's still giving me some information about the temperature, which is useful. It just limits what I can do with the information.

It also affects how I deal with the data. If I have a lot of data outside the limit of detection, I might have a lot of missing data if the instrument couldn't give a reading, or I might have a lot of **false zeros** if the instrument failed to detect anything. It might give a reading of 0, even though the true value is 0.02, but the instrument wasn't able to pick that up; it just gave a zero.

If we have a lot of data outside the limit of detection, we could have U-shaped data. If you think about it, we have loads of recordings just at that –30 degrees mark, and a lot at that +80 degrees mark, because that was as precise as we could get. So they're equal to or less than –30, but when I graph them, there's a bunch of data stacked up at –30, and a bunch stacked up at 80 degrees, and then the rest of the measurements spread out in between. It forms a funny kind of U-shape—certainly nothing like that normal distribution we saw earlier.

If the method we're using isn't sensitive enough to detect the target substance, we could have what we call **left-censored data**. There was a cutoff: a minimum value we were able to detect, and then

we're missing anything below that value. That could create some bias if there's something meaningful in the measurements we're missing below that cutoff.

These scenarios can give a lot of non-normal data, like a U-shape or all the data cut off at a certain value. Special techniques are required to leverage the information in data that are outside the limit of quantitation, or to handle data that are missing because they fell below the limit of detection. There are special techniques, for example, for handling **zero-inflated data**, where I had a lot of values where I couldn't detect anything, so they just went down as zeros—even though, if we had more precise measurement, maybe it's 0.001, but they're all recorded as zero. And we have special techniques for handling the skewed data that we often see.

It's important to be aware that these methods are out there. Hopefully you'll learn more about them in the future, but I'm just flagging them now.

Now let's look at some different basic statistical analyses that we can do, that are very commonly used in epidemiology.

To start off, I'm going to introduce you to the idea of **regression**.

This is an analytical method that we use when we have a numeric outcome—something that we want to predict, like level of exposure—and it's numeric, so we have a number to represent the value of an exposure. Then we have one or more predictors. These might be numeric predictors, like age, that we use to predict your exposure level.

So your exposure level, represented by a number, is predicted by your age, for example, which is another number.

What regression asks is: for every one-unit increase in X, the predictor (maybe age), how much change do I expect in the outcome Y, like your level of exposure? As age increases, do we expect exposure level to go up?

Regression is equivalent to plotting these values on an X and Y axis and seeing if, as X goes up, Y also goes up. As I get older, does my level of exposure go up, for example?

We fit a line of best fit through these plotted data. In this example, you can see the line is very steep. As the predictor goes up, the outcome goes up very quickly. There's a very strong relationship. The slope is a measure of the strength of the relationship between the outcome and the predictor—the exposure and the predictor, maybe.

Let's think about some data-handling techniques.

If I want to use a method like regression, I'm trying to fit a straight line to data. But as you can see in some examples, the data are curving away from the line, so it doesn't really fit the model very well. In this case, I can manipulate the data to make it suitable for analysis.

For example, if I take the square root of the outcome—so I transform all the values for the outcome Y—now my data might fit a straight-line relationship. That's a data transformation that makes my data suitable for modelling.

The log-normal example we mentioned earlier would also be a data transformation that I might do to make my data easier to analyse.

Sometimes I also have to remove outliers.

In this case, I plotted my data on a scatter plot and tried to fit the line of best fit, but it didn't fit the data very well at all because one point, which is an extreme outlier on Y, was exerting a lot of influence on the regression line.

In this Y variable, there shouldn't be any values greater than 200. So there was some kind of mistake here—a measurement error or data entry error. I decided to exclude any values greater than 200.

Now that outlier is gone, but my data still don't fit the model very well. There's another outlier at the lower end of Y, but in this measure I realise there shouldn't be any values of Y below zero. So I exclude any values below zero.

Then, after excluding my outliers, I have a nice fit. I can fit a straight line to my data.

It's quite common to graph your data and notice an outlier, and sometimes we decide to exclude those outliers or to do a transformation to make the data fit better for analysis.

We also do something called **sensitivity analysis**. This is about testing the robustness of our result.

How much, for example, does the conclusion I've drawn depend on influential points like the outliers we saw? How unusual must our sample be for something different to be true in the population? If there was a very strong relationship between X and Y in our data, we could, by chance, have gotten a very unusual sample from the population.

We can test how sensitive a result is to something like that—to an unusual sample.

We also look at something called an **unmeasured confounder**. An unmeasured confound would be a third variable that, after we account for it, might explain away the relationship between X and Y. Maybe we find a strong relationship between age and an exposure, but maybe it's explained away if we measure a policy change that happened over time. Age is related to exposure, but policies have changed over time to reduce younger people's exposure. So that would be a confounder we could consider.

Here, I've modelled the relationship between an outcome Y and a predictor X, and it's a very strong relationship. Then I controlled for a third variable, Z, which is related to both X and Y. Once I took the effect of Z out of Y, X no longer had any relationship to Y. In fact, it looks like a slight negative relationship, but really there's effectively no relationship.

Sometimes we need to test for the potential for something else to explain what we've observed, and ask: how big would the effect of this third variable Z have to be, on both X and Y, to explain away all of the strong relationship that we see?

Next, we'll talk about how to identify patterns in exposure data.

A fundamental approach to identifying patterns in exposure data is to compare the frequency of an outcome between individuals who have and have not been exposed. In other words, we want to know if an exposure is related to a disease.

We could take people who have been exposed to a chemical and those who haven't, and see if the disease is more frequent, or if the risk of disease is greater, in people with the exposure compared with those without the exposure.

There's a measure we use for this called **relative risk**.

You can see in the formula that we get the risk for those in the exposed group by taking the number in the exposed group with the disease and dividing it by the total number in the exposed group. We do the same for the group who aren't exposed, and then we divide the risk for the exposed group by the risk for the group who aren't exposed to get a relative risk.

If the risk for those who are and aren't exposed is the same, relative risk is equal to 1. There's an even risk, or there's no increased risk from the exposure.

If the relative risk is less than 1, that implies that the disease, or mortality, or whatever the outcome is, is less common in the exposed group, so the exposure would reduce your risk. If relative risk is greater than 1, that tells us that an exposure comes with increased risk of the outcome, like a disease.

For example, if relative risk were 1.5, that means the exposed group are 1.5 times more likely to contract the disease. That helps us measure how much risk comes with an exposure.

Now let's just look briefly at some technical visualisation methods.

We've seen the scatter plot before, where we drew a line of best fit. In the example on the left, I wanted to explore the shape of that relationship without imposing a straight line. I wanted to draw a line that will try to fit all the points as best as possible, to help me see if there's any curve in that relationship or if it looks straight. The line I've fit here is called a **LOESS line**. It tries to smooth out and allow curves to fit the shape of the data and to see if it changes as the value of X increases. I can also change how smooth this line is, so it can be more closely fitted to the data or more smoothed.

On the right, I have a **histogram**, which is a visualisation I use to look at a single numeric variable to see what the shape of that distribution looks like. Thinking back to the curves we had before—the normal distribution—how do I check and see if my data look like they have that shape?

I can't just plot each exact value, because everyone might have a slightly different value. My height might be 173.2, yours might be 173.4. I would just get a flat line if I tried to look at the shape that way.

But if I group people together—for example, if I put everyone between 173 cm and 174 cm tall into one column on the plot—we call that a **bin**. Everyone between 173 and 174 centimetres goes into one bin on the plot.

We create these bins for the full range of our data, then we can see the relative frequency across values. I can see if most people are in a certain range, or if values in a certain range are more common.

That helps me visualise what that smoothed distribution would look like if I'm thinking about what population my sample might have come from.

It's very helpful for me to check and see whether my data look normally distributed. A histogram is very helpful for that.

To finish up, let's summarise how statistics can help us to identify, measure, and mitigate environmental exposures.

Statistics help us to detect exposure patterns, to identify trends, clusters, and distributions in environmental data. They help us to describe exposure levels—tell me the mean of an exposure, tell me what a typical value is, or what a high value looks like—and allow me to make comparisons between populations over time.

Statistics help us to understand variability and uncertainty: what is the spread of an exposure, and how much confidence do we have in our estimate of an exposure?

Statistics, if we think of relative risk, allow us to link exposures to influential factors. Is an exposure higher in one group or another? Is it predicted by age, or something else?

Statistics help us to identify high-risk groups or hotspots, where we can target our focus. They support risk assessments and the development of standards, regulations, and guidelines around safe ranges.

Statistics help us to evaluate interventions. Maybe exposure levels are high, I introduce a control measure, and then I measure exposures later on. I see if exposures are lower following an intervention.

Statistics help us to communicate results clearly and transparently. If I tell you a measure is effective, I do so by showing you how the data have changed.

Now let's summarise our key takeaways.

We looked at descriptive and summary statistics: how we describe what our data look like with the central tendency and the dispersion of the data. We introduced what normality is and what a normal distribution is. We addressed handling data outside the limit of detection or the limit of quantitation. We looked at some statistical modelling—ideas like regression and relative risk—and how to explore data with plots. And we looked at how data and statistics can be used in exposure science.

In the following two modules, we'll follow up on statistics that are useful specifically for exposure scientists. Then in Module 3, we'll give a brief introduction to epidemiology.

Thank you very much for your time in watching this video. I hope it was useful to you. Remember, you can access all the videos on the ISES Europe website. If you're interested in reading more about statistics, we have a few readings you can look at on the final slide.

Thank you again for your time.