ISES Europe Training Series

DoE 9: Statistics and Epidemiology

Module 2: Statistics for Exposure Scientists

Transcript Notice:

This is the transcript of the presentation. Please note that the actual spoken text may differ slightly from what is written here.

Slide 1 – Welcome

Hello and welcome to today's session in the ISES training series.

This presentation falls under Domain of Expertise 9: Statistics and Epidemiology. I'm delighted that you've joined us for this part of the training, and I look forward to guiding you through today's topic.

Slide 2 – Legal notice

Before we begin, just a quick legal note. All rights to the materials used here remain with the original copyright holders.

If you would like to reuse any part of this presentation, please make sure to seek explicit permission first. Thank you for respecting these rules.

Slide 3 – Overview of All Training Videos

This training is part of a broader ISES Europe initiative. In total, there are nine Domains of Expertise, together covering the breadth of exposure science.

They range from the fundamentals of exposure science and environmental chemistry, through topics like exposure modelling, risk communication, and sustainability, all the way to today's focus: Statistics for Exposure Scientists.

If you'd like to revisit any session, all training videos are available at ises-europe.org.

Slide 4 – Domain of Expertise (DoE) 9

Our focus today is Domain of Expertise 9: Statistics and Epidemiology.
This domain is divided into three submodules:

- Module one, which we covered earlier, introduces the principles of statistics.
- Module two, which we will cover today, builds directly on Module 1 and aims to translate the basic concepts of statistics into the key requirements for exposure scientists.
- Module three will provide an introduction to epidemiology.

Together, these modules offer a comprehensive overview of the fundamentals in statistics and epidemiology, and show how these disciplines intersect with Exposure Science.

Slide 5 – Module two

The title of today's session is *Statistics for Exposure Scientists.*
We will explore the main statistical approaches and tools that an exposure scientist typically needs to analyse data, understand variability and uncertainty, and support well-founded risk assessments.

Please note that this is a short introduction and only covers basic statistical concepts.

Slide 6 – Presenter

A few words about me before we begin. My name is Gerald Bachler. I currently work as a Regulatory Officer at the European Chemicals Agency, where I am a member of the Industrial Emission, Chesar and Exposure Group. Before joining ECHA, I worked on product safety in various companies within their Product Stewardship, Regulatory Affairs, and Toxicology/Risk Assessment teams.

I received a PhD in Nanotoxicology from ETH Zurich and hold an MSc in Health Care Engineering and in Health and Environmental Science.

Slide 7 – Context and Disclaimers

Please note that this lecture is designed as an introductory-level session. Some aspects have been simplified in order to provide a clear overview.

Slide 8 – Learning objectives

By the end of this lecture, you should be able to:

- Understand why selecting an appropriate exposure value is crucial in risk assessment
- Recognize exposure distributions and the role of SEGs
- Interpret confidence intervals for exposure values
- Apply methods to handle left-censored (non-detect) data
- Identify software tools for exposure data analysis

Slide 9 – Content

Here is today's roadmap:

1. We will start with a short introduction.
2. Then, we will look at percentiles and distributions commonly used in risk assessments.
3. After that, we will discuss how variability influences exposure estimates and what this means for confidence intervals and data requirements.
4. Next, we will briefly consider statistical methods for handling left-censored data.
5. Before the summary, we will take a quick look at software tools that support the application of these statistical approaches.
6. Finally, we will wrap up with a concise summary and key takeaways.

Let's begin with the introduction.

Slide 10 – Refresher: How to Calculate Risk

In Domain of Expertise 5, we discussed how to characterise risk.

In principle, risk is evaluated by comparing an exposure estimate with a threshold limit value. The two main approaches are shown on the slide:

- Dividing the exposure estimate by the threshold limit value to calculate the Risk Characterisation Ratio (RCR), or
- Dividing the threshold limit value by the exposure estimate to calculate the Margin of Safety (MoS).

Regardless of the method, the decision is based on *one exposure value*.

However, only in very rare circumstances does an exposure assessment truly conclude with a single number – for example, when using a simple screening tool that generates only one worst-case estimate.

So how is this typically handled in practice?

Slide 11 – Exposure Estimates in Practice

As mentioned, exposure assessments rarely result in a single estimate.

Using a higher-tier tool or conducting measurements usually produces far more complex outputs – such as full exposure distributions or large datasets that require additional processing.

For large datasets, or when many receptors are assessed at once, it is not productive to calculate risk individually for each data point. Instead, exposure distributions allow us to describe the entire population and capture variability in a structured way.

At the bottom of the slide you see two real-life examples:

- **Left**: An inhalation exposure output from the Advanced REACH Tool (ART), a higher-tier model. It provides a percentile selection and a confidence interval around that percentile.
- **Right**: A small extract from the NHANES dietary intake database. This dataset contains physiological information, detailed dietary intake across many food categories, and extensive biomonitoring data for thousands of individuals. Exposure scientists must distill such information into distributions that describe the population as a whole.

With that context, let's move into the key statistical concepts needed to process such exposure data.

Slide 12 – Content

We begin with an introduction to exposure percentiles and distributions.

This provides the foundation needed to understand how to select appropriate exposure values for risk assessment.

Slide 13 – What is a Percentile

Let's start with the basics: *what is a percentile?*
According to Encyclopædia Britannica, a percentile is "*a number denoting the position of a data point within a numeric dataset by indicating the percentage of the dataset with a lesser value*".

For example, at the bottom of the slide you see ten individuals with different heights, ranked from smallest to tallest.

To find the 90th percentile height, we identify the ninth person – 180 cm in this example. This means that 90% of the group is 180 cm or shorter, and 10% is taller.

Now let's see how the same concept applies to exposure distributions.

Slide 14 – Percentiles and Distributions I

Here we see two types of probability distributions frequently encountered in exposure science: the normal distribution and the log-normal distribution. Many exposure datasets follow one of these patterns, and they are also commonly assumed by exposure modelling tools.

Percentiles can be determined in the same way for both distributions. The slide shows examples of the median – which is the 50^{th} percentile – as well as the 75^{th} and 90^{th} percentiles. The 75^{th} percentile, for instance, represents the value below which 75% of the measured exposures fall.

You can also see that the log-normal distribution is typically the default assumption in exposure science. It does not produce negative values and features a longer tail at higher exposures, reflecting the reality that some individuals or scenarios tend to experience much higher exposures.

Slide 15 – Percentiles and Distributions II

This slide shows four examples to illustrate that real-world datasets can follow many different shapes. They may be normal or log-normal, as shown at the top. But sometimes we see distributions like the one on the bottom left, where two log-normal groups overlap. This usually signals that the dataset was not prepared correctly – two populations with distinct exposure profiles were merged. We will discuss this challenge further in the next slide.

In contrast, the bottom right shows a random-looking distribution, which can for example occur for some dietary exposure categories where intake varies widely across the population. In such distributions, a small subgroup with very high exposure could be overlooked if only the 90th percentile is considered, which you can also see in the depicted example. At the very right you have a group with very high exposure.

Maybe worth mentioning, when many different distributions are combined – for example, when dozens of food categories are aggregated into a total dietary exposure estimate – the resulting distribution often tends to approximate a log-normal shape again.

In any case, it is the responsibility of the exposure scientist to understand the dataset and determine the most appropriate distribution to use in risk assessment.

Slide 16 – Similar Exposure Group (SEG)

Let's now look into a specific type of distribution: Similar Exposure Group, or SEG. This is common tool in exposure science. A Similar Exposure Group is a group of receptors who share a comparable exposure profile for the agent being assessed.

Grouping is based on similarities in tasks, materials, processes, work practices, or consumer-product use. This approach allows us to take representative measurements and draw conclusions about the group without needing to assess each individual separately, which is often not feasible.

Establishing Similar Exposure Groups relies heavily on expert judgement and typically combines observational insights with representative sampling. Most Similar Exposure Groups are assumed to follow a log-normal distribution, which enables statistical tests to check whether the group is homogeneous, or whether what appears to be one group is actually two distinct exposure groups — as seen earlier in the overlapping distributions example.

Slide 17 – Selecting an Appropriate Exposure Value

After identifying the appropriate distribution, an appropriate exposure value has to be selected. This requires expert judgement, and unfortunately, there is no universal rule-set. Several factors must be considered.

First, the population type matters – whether it is workers, consumers, or environmental receptors. Vulnerable sub-groups may require special attention, especially if they cluster in the higher exposure tail.

Next, the distribution type influences the selected percentile. When the distribution is unknown or highly variable, more conservative percentiles may be justified. When a log-normal distribution or well-defined Similar Exposure Group exists, the percentile selection can be more data-driven.

The toxicological endpoint also plays a key role. Chronic effects may be less sensitive to occasional high exposures, whereas acute endpoints may depend heavily on single high exposure events. For example, for carcinogenic exposures, some jurisdictions consider lifetime cumulative exposure, and selecting an overly high daily percentile would artificially inflate lifetime exposure. A different example are substances that can sensitise. Here brief exposure above the threshold limit value may be significant and has to be avoided.

The source and quality of data are equally important. Peak-based measurements differ from typical measurements, and high variability may support selecting a higher percentile.

In some cases, regulatory guidance may exist, such as REACH guidance for occupational exposure assessments, although even this does not provide a one-size-fits-all solution.

Whatever value is selected, the rationale must be clearly documented in the risk assessment report for future reviewers, including auditors or authorities.

Slide 18 – Content

Now, let's look at how we can determine the reliability of the selected value. For this, we will look at confidence intervals and the associated data requirements.

Slide 19 – Confidence Interval (CI)

Lets start with the definition. A confidence interval "*shows the range of values you expect the true estimate to fall between if you repeat the study many times*".

Let's consider a simple example. Imagine that the true population median exposure is 15. A measurement study might estimate that the median lies between 14 and 16 with 95% confidence. If you repeat the study, the interval might shift to 12–18 with a 95% confidence. If you repeat the study another 18 times, then 19 out of those 20 times– that is, 95% – the confidence intervals should contain the true median. You are thus 95% sure that the true median is within the confidence intervals.

A related concept that is sometime used is **reliability.** It "*refers to the consistency of the measurement, or the ability to repeat the measurement and obtain the same result"*.

With this understanding in mind, let's examine the properties of confidence intervals.

Slide 20 – Properties of Confidence Intervals (CI)

When only a small number of data points is available, the confidence interval becomes wide. This reflects uncertainty and low reliability due to the limited dataset.

However, even large datasets can yield wide intervals. In such cases, the cause is often high variability or poor data quality, both of which should prompt the exposure scientist to investigate further.

Improving the confidence interval may require collecting more data – particularly if the dataset is small. If the dataset is already large, it may be necessary to re-examine the Similar Exposure Group definition to ensure that the group is homogeneous.

When the confidence interval cannot be improved – for example, because a higher-tier tool produces it as part of its output – this may justify selecting a more conservative exposure estimate for risk characterisation.

In general, it is advisable to use the upper bound of the confidence interval, such as the 70% or 90% upper confidence limit (UCL), as an added margin of safety.

If the exposure estimate is close to the toxicological threshold, refinements may include collecting additional data, narrowing the confidence interval, or redefining Similar Exposure Groups.

Again, all decisions and their justifications must be documented in the risk assessment report.

Slide 21 – Content

Next, let's briefly discuss how to handle left-censored, or non-detect, data using statistical approaches.

Slide 22 – How to Handle Values below LOQ

In many exposure assessments, a portion of the data falls below the limit of quantification – the LOQ. A common approach is the substitution method, where non-detects are replaced with a value such as zero, LOQ divided by two, LOQ divided by the square root of two, or the LOQ itself.

While this method is simple, it can introduce bias, especially when many values fall below the LOQ, which is quite common in practice.

A more accurate approach involves applying statistical estimation methods. Since this is only an introductory session, we cannot cover these techniques in full detail here. However, methods such as Maximum Likelihood Estimation (MLE) or Regression on Order Statistics (ROS) allow us to fit a distribution while properly accounting for censored values.

The idea – shown at the bottom of the slide – is presented here in a simplified, high-level way. Even when values fall below the LOQ, they are still treated as part of the underlying distribution, for example a log-normal distribution. Although this depiction is not technically complete, the general concept is that these methods estimate the distribution's shape more reliably, which in turn leads to better approximations of percentiles and confidence intervals.

These approaches can become statistically complex, so using specialised software is generally recommended. And this brings us to the final topic of the presentation.

Slide 23 – Content

Software tools that can support exposure scientists in applying the methods discussed today.

Slide 24 – Tools for Statistical Exposure Assessment

Several tools are available to support exposure scientists in analysing exposure distributions, defining Similar Exposure Groups, handling censored data, and selecting appropriate percentiles.

However, it is crucial to understand the statistical principles behind the tools to use them correctly. Before applying any of these tools, users should consult the manual or ideally receive training. And, as always, the assumptions and outputs should be documented in the risk assessment report.

The slide shows three examples of freely available tools – each with increasing capability and complexity.

- First, **BWStat**, provided by the Belgian Society for Occupational Hygiene (BSOH). It focuses on occupational inhalation exposures following the European Norm 689. It supports Similar Exposure Group development, handles left-censored data, and calculates percentiles and confidence intervals in line with the standard.
- Second, **ProUCL** from the US Environmental Protection Agency. It focuses on environmental datasets and offers numerous options for handling censored data. It allows user-defined percentiles, multiple distribution fits, and supports Similar Exposure Group development.
- Finally, the **R** software from the R Foundation provides a general statistical computing platform capable of applying nearly any method an exposure scientist may need.

Although BWStat and ProUCL are designed mainly for occupational and environmental data, respectively, both can also be used for other receptor types within their limitations.

Slide 25 – Content

Now that we have covered all sections of the presentation, let's conclude with a short summary.

Slide 26 – Summary: Key Takeaways

Statistical approaches play a central role in risk assessment and must be well-understood by assessors. Percentiles help describe population exposure and support risk management decisions. Choosing an appropriate value from an exposure distribution is not always straightforward – it depends on factors such as the distribution shape, the characteristics of the exposed population, the toxicological endpoint of concern, and the overall quality of the underlying data.

Data reliability is equally important. Sample size, variability, and the presence of left-censored values all influence the final exposure estimate and the confidence we can place in it. Software tools can support distribution fitting, percentile estimation, and other statistical analyses, but only when they are applied correctly, transparently, and with a sound understanding of their limitations.

One final comment: since this was only a short introduction, many relevant topics could not be addressed. For example, input values for exposure estimation models often require the same type of statistical consideration. They are frequently not single point estimates but complex datasets that follow the same kinds of distributions we discussed today. Although we could not explore these in detail here, it is important to note that many of the principles covered in this session apply equally to model inputs.
And to briefly follow up on this: An important point is that when an exposure estimation model requires several input parameters, and conservative values are selected for each of these parameters, the resulting exposure estimate can become **unrealistically high**. This is a common pitfall. It is therefore essential that exposure scientists apply their expertise to make balanced, practical decisions and clearly document the reasoning behind those decisions as part of the overall risk assessment report.

Slide 27 – Consequent Modules

Looking ahead, **Module three** of this training will provide an introduction to epidemiology.

And further down the line, ISES Europe may release additional specialised materials, allowing us to explore individual topics in greater depth.

Slide 28 – Closing thanks

That brings us to the end of today's session.

Thank you very much for your attention and participation. I hope the material has provided a clear understanding of why statistical approaches are essential for robust risk assessment, and what level of statistical literacy an exposure scientist should have.

Please continue exploring ISES's training series for more insights into exposure science and risk assessment.

As a reminder, all training materials are available on **ises-europe.org.**

Slide 29 – Further Reading

On this final slide, you'll find suggested reading materials for further study. We won't go through them here, but I highly recommend you review them afterwards.
Thank you very much, and goodbye.
