DoE 9: Statistics and Epidemiology

ISES Europe Training Series

Ruairí Weiner, Gerald Bachler

Legal Notice

Fair Use Disclaimer

This presentation may contain copyrighted material used under the fair dealing exceptions in EU copyright law, in accordance with Directive 2001/29/EC and Directive 2019/790. It is intended solely for non-commercial, educational purposes.

Copyright Notice

© ISES Europe 2025

All rights to the original works are retained by their respective copyright holders. Any reproduction, distribution, or use of this presentation, in whole or in part, requires explicit permission from the copyright holder(s) or the author. Unauthorized use of any content is strictly prohibited.

Overview of All Training Videos

DoE 1) Basic Concepts and Principles in Exposure Science

DoE 2) Fundamentals of Environmental Chemistry and (Eco)Toxicology

DoE 3) Exposure Modelling

DoE 4) Exposure Monitoring

DoE 5) Exposure Assessment and Risk Characterisation

DoE 6) Risk Management and Sustainability Assessment

DoE 7) Relevant Legislative Frameworks

DoE 8) Risk Communication and Stakeholder Engagement

DoE 9) Statistics and Epidemiology

Access all videos via: **https://ises-europe.org/**

Domain of Expertise (DoE) 9:
Statistics and Epidemiology

Module 1	Introduction to Statistics
Module 2	Statistics for Exposure Scientists
Module 3	Introduction to Epidemiology

DoE 9: Statistics and Epidemiology

Module 1

Introduction to Statistics

Ruairí Weiner

Presenters

Meet Today's Presenter: **Ruairí Weiner**

- **Current Role:** Teaching Fellow in Research Methods & Statistics, University College Dublin School of Public Health, Physiotherapy and Sports Science
- **Previous Roles:**
  - Teaching data science and statistics, Trinity College Dublin School of Computer Science & Statistics
  - Amyotrophic lateral sclerosis research, Trinity College Dublin School of Medicine
  - Ageing research, The Irish Longitudinal Study on Ageing
  - Programme evaluator for counselling services in third level education
- **Education:**
  - MSc in Applied Social Research

Context and Disclaimers

About This Lecture

This lecture provides an **introductory framework**, with some topics simplified for ease of understanding.

Disclaimer

- The content presented herein does not necessarily reflect the opinions, views, or positions of the presenters' employer or any affiliated organizations.
- References to specific organizations, tools, or entities are for illustrative purposes only and do not imply endorsement or critique.
- While every effort has been made to ensure the accuracy of the information presented, errors or omissions may occur.

Learning Objectives

- Gain an understanding of the statistical methods commonly used in exposure assessment.
- Learn how to explore relationships between variables, summarize data, and handle uncertainties in the context of exposure assessment using these statistical methods.

Content

1. **Introduction to the Statistics used in Exposure Assessment**
2. Normality
3. Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)
4. Describe different Tests and Basic Statistical Analyses
5. Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures
6. Summary

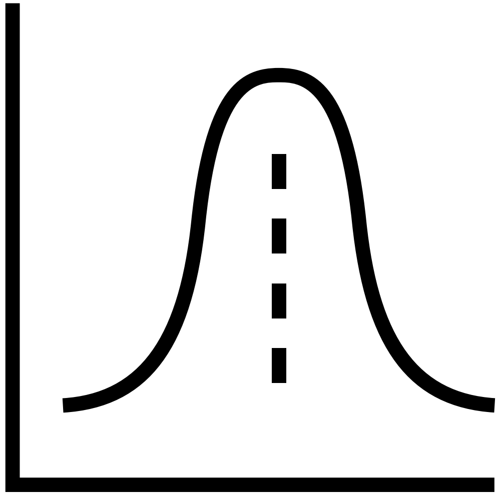
Introduction to the Statistics used in Exposure Assessment

This section will cover descriptive and summary statistics including:

1. **Measures of central tendency:**
   a. Mean
   b. Median
   c. Mode
2. **Measures of variability:**
   a. Range
   b. Standard deviation (SD) and Variance
   c. Interquartile range (IQR)
   d. Proportions

Central Tendency

- **Central tendency** is about describing where most of the data are.
- If I don't know anything else, my **best guess** at a value is using a measure of central tendency.
- The measure of central tendency I use depends on how a **variable** is **measured.**



Type of variable	Measure of central tendency
Numeric	Mean
Ordinal	Median
Nominal	Mode

The Mean

- The mean is what we often think of as 'the **average**.'
- Simply **add** up all the **values** and **divide** by the **number** of values.
- Best for **numeric data** (height, weight, temperature).
- When data are **symmetrical**ly distributed the mean is in the '**middle**.'
- Sensitive to **skew**.
- Sensitive to **outliers** in small samples.

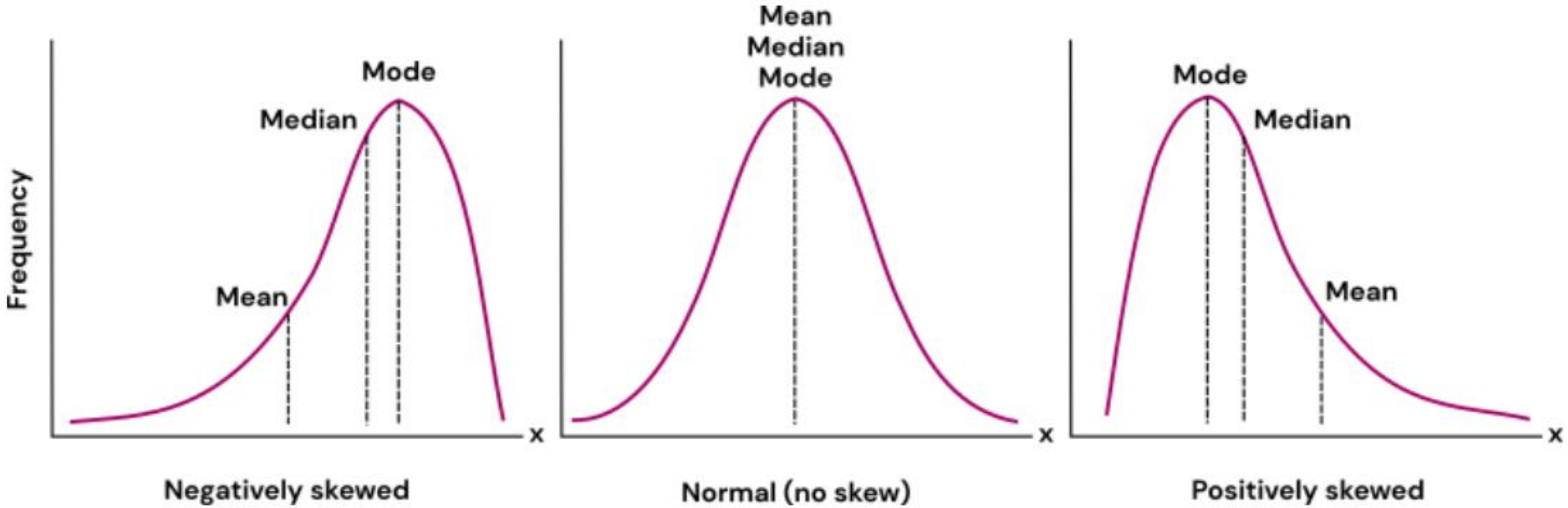
The Median

- If we **rank** all **values** from lowest to highest, the median is the **value** in the **middle**.
- Uses **less information** than the mean – only using the **middle value** where the mean is a little bit of all the values.
- More appropriate when data aren't exactly numeric but can be **ranked in order** e.g., army rank (private, admiral, etc.,)
- Less sensitive to **outliers** and **skew**ed data.

The Mode

- The mode is simply the value which **occurs most often**.
- Doesn't tell us anything about the **other values**.
- Most appropriate for data that **can't be ranked** in order e.g., colour of cars.
- If all car colours appear **equally often** but there is one more red car, we report red as the mode even though all colours are almost equal.
- If both red and green occur 6 times and the others occur less, we have **two modes** (bimodal).
- We don't report information about the **other values** which aren't modal.

Mean, Median and Mode



Source: Van Parys, A. https://ledidi.com/academy/measures-of-central-tendency-mean-median-and-mode

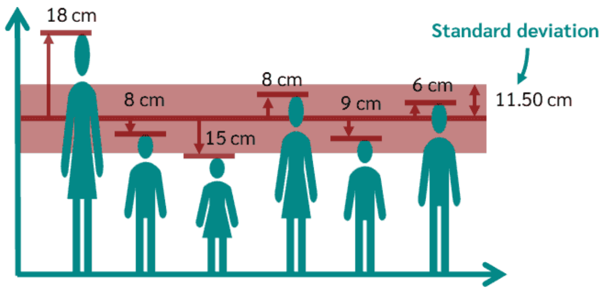
Measures of Variability

Central tendency told us about where the most typical values are.
Measures of variability describe how **spread out** the values are from one another.

Range

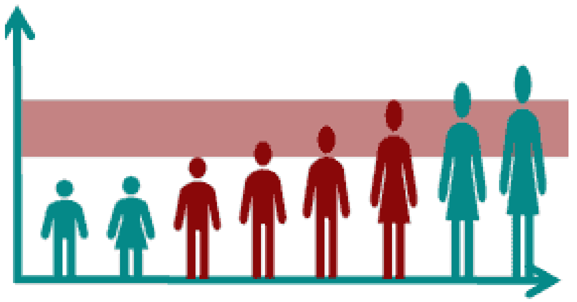
Distance between lowest and highest value of a distribution

Standard Deviation (SD)



Average distance of all measured values from the mean value

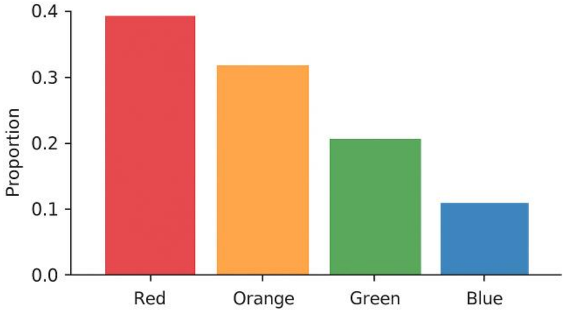
Interquartile Range (IQR)



Spectrum in which the middle 50% of the values lie. Difference between first and third quartile

Source: Volk-Jesussek, H. (2025) https://numiqo.com/tutorial/dispersion-parameter

For data like colours, we report the **proportion** of values in each category.



Content

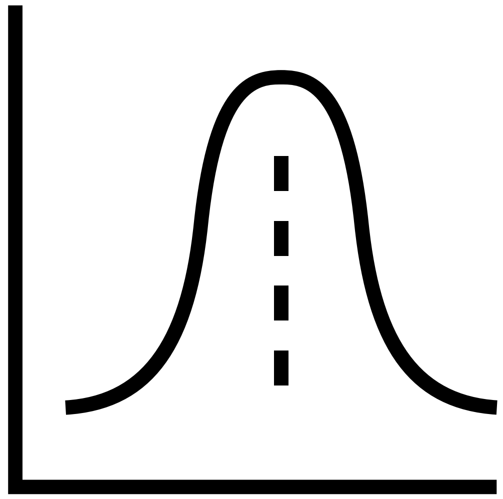
1. Introduction to the Statistics used in Exposure Assessment
2. **Normality**
3. Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)
4. Describe different Tests and Basic Statistical Analyses
5. Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures
6. Summary

Normality I

- Important concept that pops up all around statistics.
- If numeric data **vary randomly** around a **mean**, they will form a **symmetric** shape we call a **normal distribution**.
- Assuming data are **normally distributed** allows lots of powerful statistical techniques.
- If data aren't **normal** different techniques have to be used.

Normality II

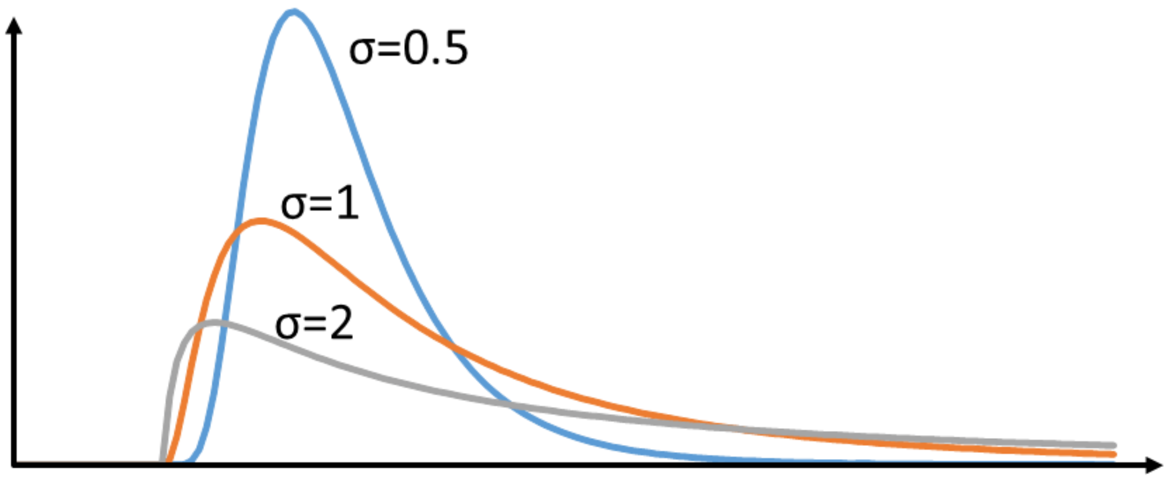
- However, the normal distribution is very **common in nature** because it arises from **random variation**.
- It is very common to see normal data in **epidemiology** – height, weight, biomarkers, etc.



Normality III

- Many exposure studies have log-normal data.
- A distribution is log-normal if its natural logarithm, ln(X), follows a normal (Gaussian) distribution.
- Often seen in these datasets as they are positive real values and positively skewed (tailing off to the right).

Lognormal Distributions (μ=1)



This Photo by Unknown Author is licensed under CC BY-SA

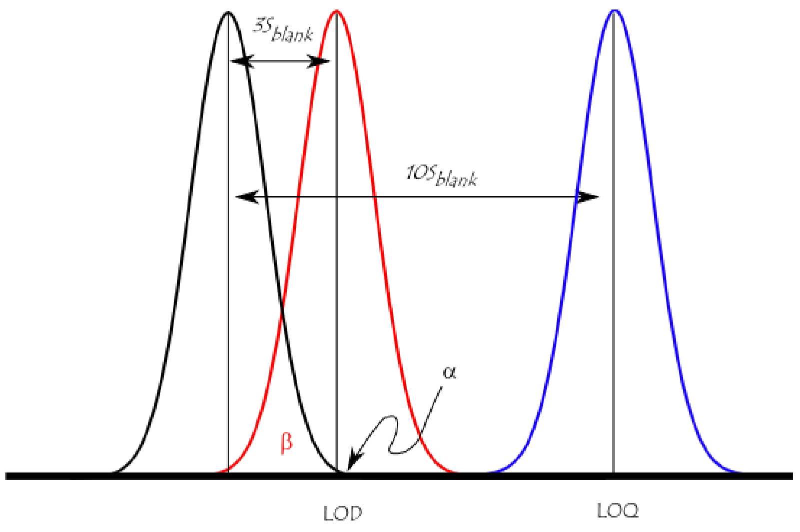
Content

1. Introduction to the Statistics used in Exposure Assessment
2. Normality
3. **Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)**
4. Describe different Tests and Basic Statistical Analyses
5. Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures
6. Summary

Limit of Detection/Limit of Quantitation I

- In statistics we take **measurements** to estimate levels of exposures and to test relationships.
- But sometimes levels are so **low** we **can't detect** them.
- As levels decrease **more sensitive** measures are required.
- This **limit** of measurement is the **LOD**.
- The **LOQ** is the lowest level at which we can reliably quantify the amount of a substance with acceptable precision and accuracy, not just detect its presence.

Definitions



Source: www.wikipedia.org

Limit of Detection (LOD):

- *...is 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.*

Eurachem

- Often calculated as 3 x standard deviation of the blank

Limit of Quantification (LOQ):

- *...is the lowest level at which the performance is acceptable for a typical application.*

Eurachem

- Often calculated as 10 x standard deviation of the blank

Limit of Detection/Limit of Quantitation II

- In cases where we detect something **outside LOQ** we can state whether this was above or below our limit e.g., >1mm.
- We can't compare two things >1mm and say one is longer, but we may say they are long **enough** to meet some criterion.
- We may have both an **upper** and **lower** limit e.g., my thermometer may not go below -30°C or above 80°C so if a reading reaches the max, I could say it was ≤-30°C or ≥80°C.

Limit of Detection/Limit of Quantitation III

- If we have lots of data outside the LOD, we may have lots of **missing data** if the instrument couldn't give a reading.
- Or we may have lots of **false zeros** if the instrument failed to detect anything so indicated there is nothing there.
- If we have lots of data outside the LOD, we might have U-shaped data e.g., loads of data below -30°C and above 80°C.
- If analytical method is not sensitive enough to detect the target substance – left censored data.

Limit of Detection/Limit of Quantitation IV

- These scenarios could give a lot of **nonnormal data**.
- Special techniques are required to leverage the information in data outside the LOQ or handle data that are missing because they are outside the LOD.
- There are special techniques for handling **zero-inflated** data.
- And techniques for handling **skewed** data.
- Methods such as LOD/2, √ LOD/2, LOQ/2, √ LOQ/2, Maximum Likelihood Estimation (MLE) methods.

Content

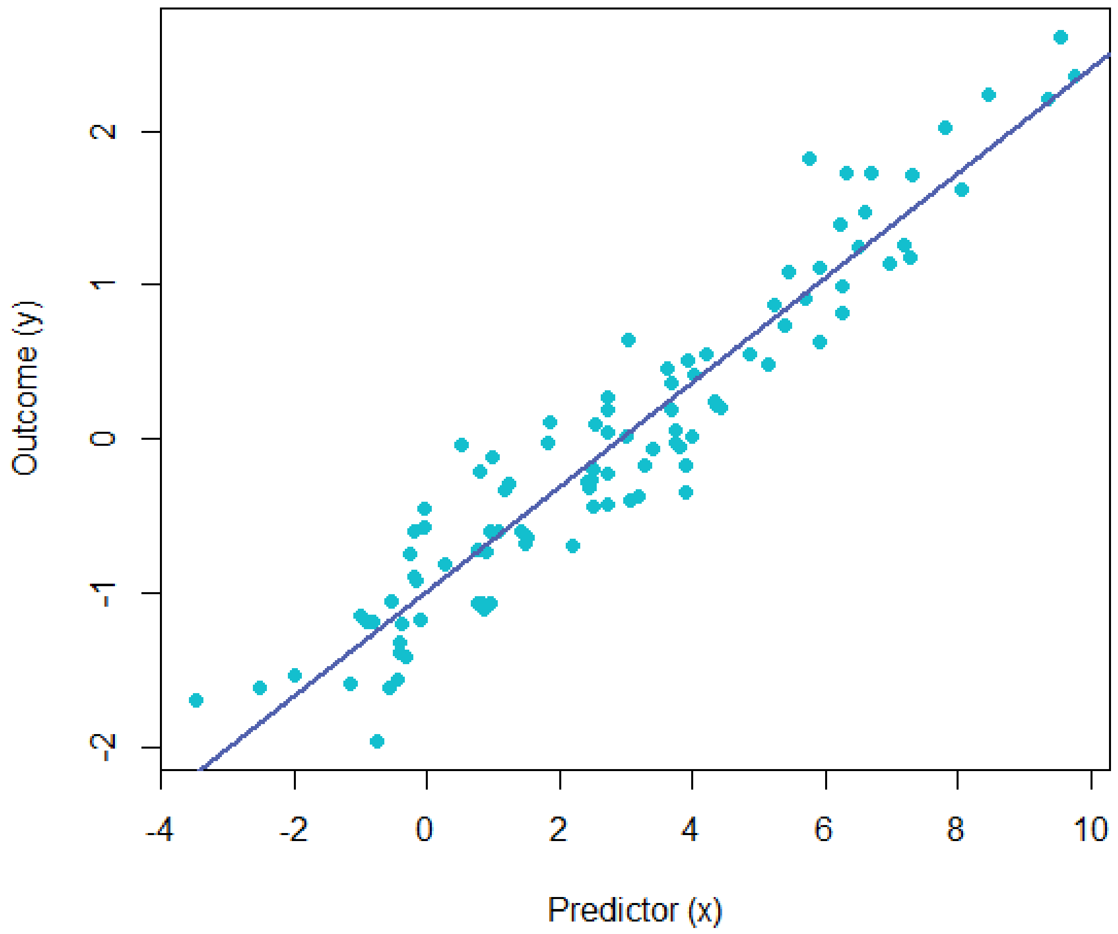
1. Introduction to the Statistics used in Exposure Assessment
2. Normality
3. Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)
4. **Describe different Tests and Basic Statistical Analyses**
5. Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures
6. Summary

Tests and Statistical Analyses - Regression

- Used when there is **one numeric** outcome to predict and **one or more numeric predictors**.

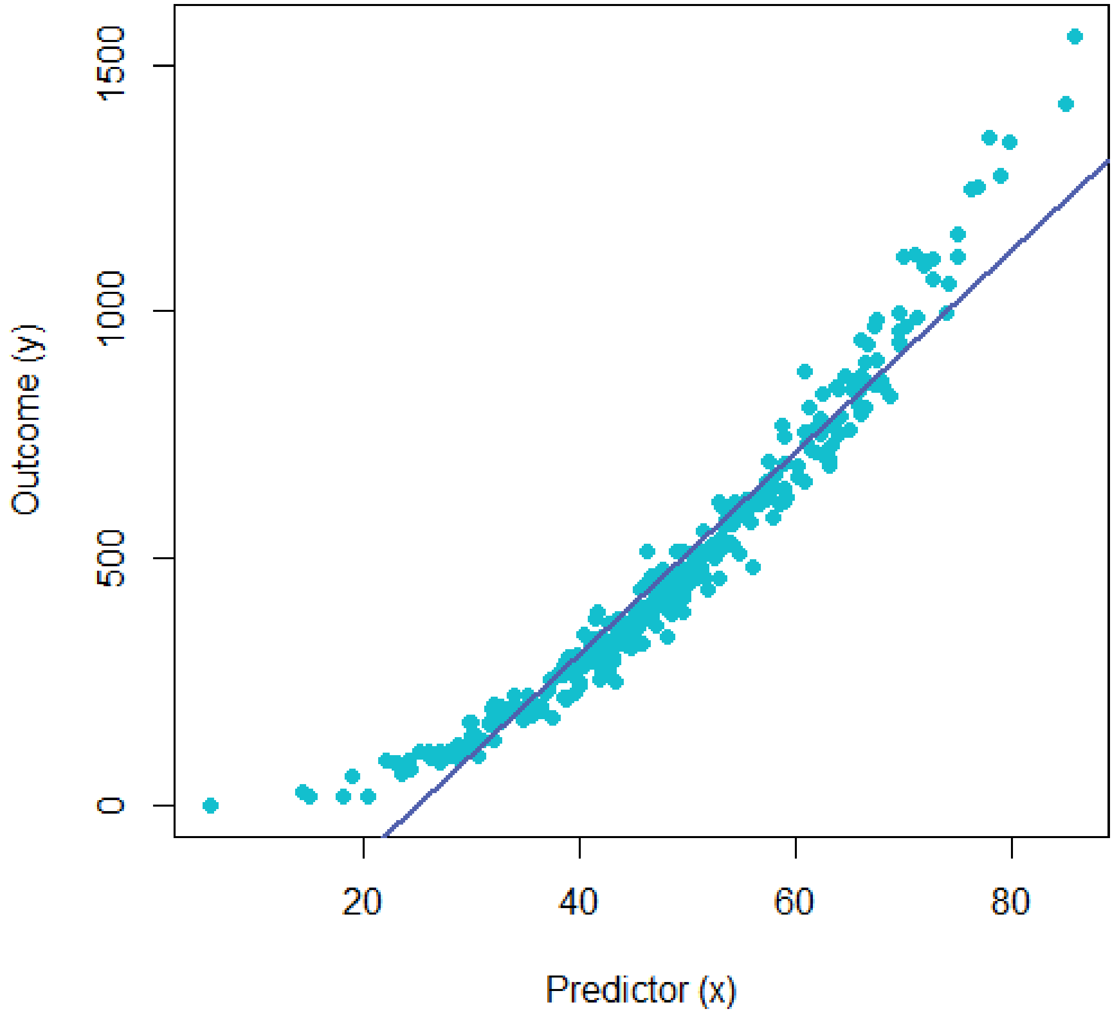
  Asks: for every 1 unit increase in x, how much change do I expect in y.

- Equivalent to plotting x and y, then fitting a line through the data.
- The steeper the line, the more y increases as a function of X.

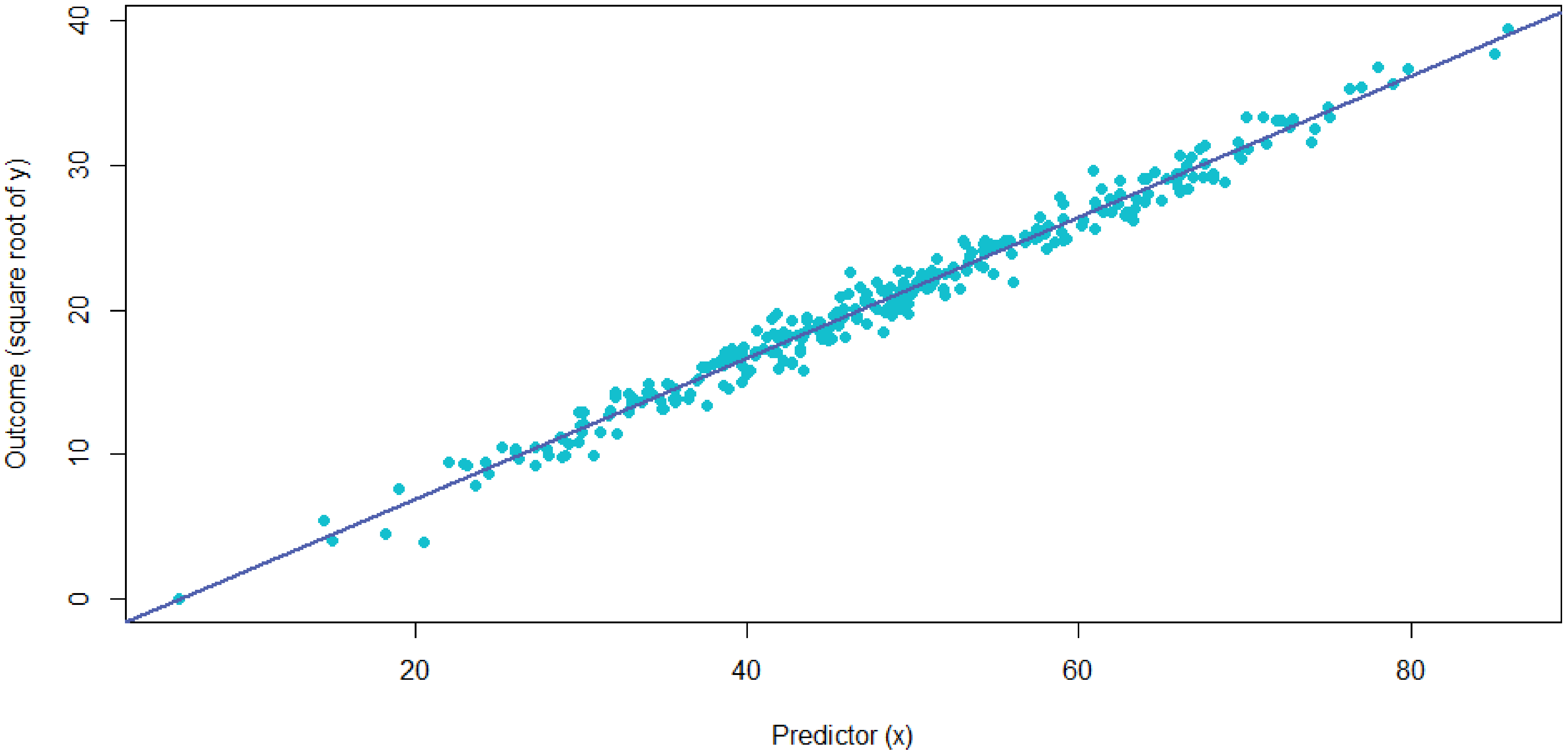


Tests and Statistical Analyses - Data Handling Techniques I

- Sometimes we have to **manipulate** data to make it **suitable** for analysis.
- Here the relationship between x and y does not fit a **straight line**.
- But if I take the square root of y, I resolve this issue.

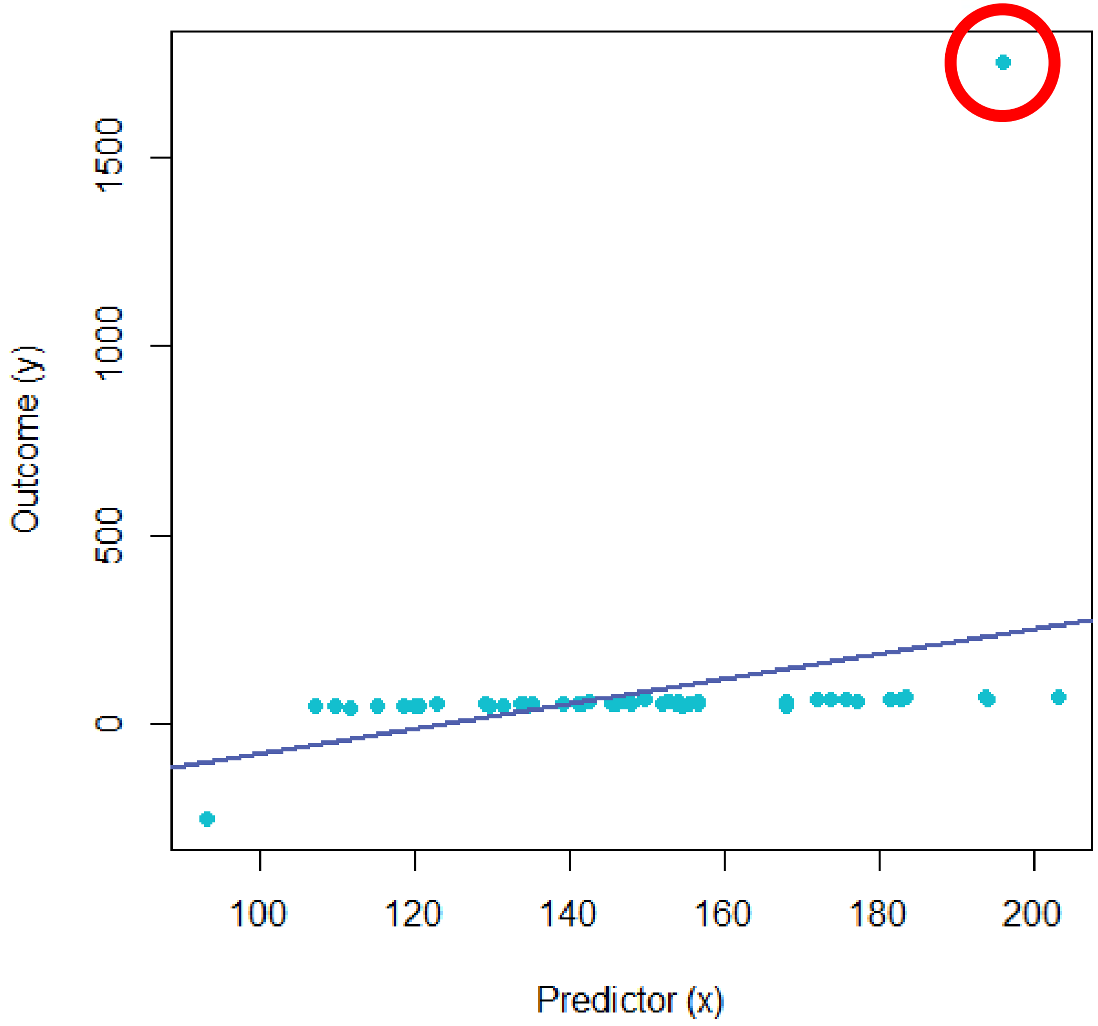


Tests and Statistical Analyses - Data Handling Techniques II



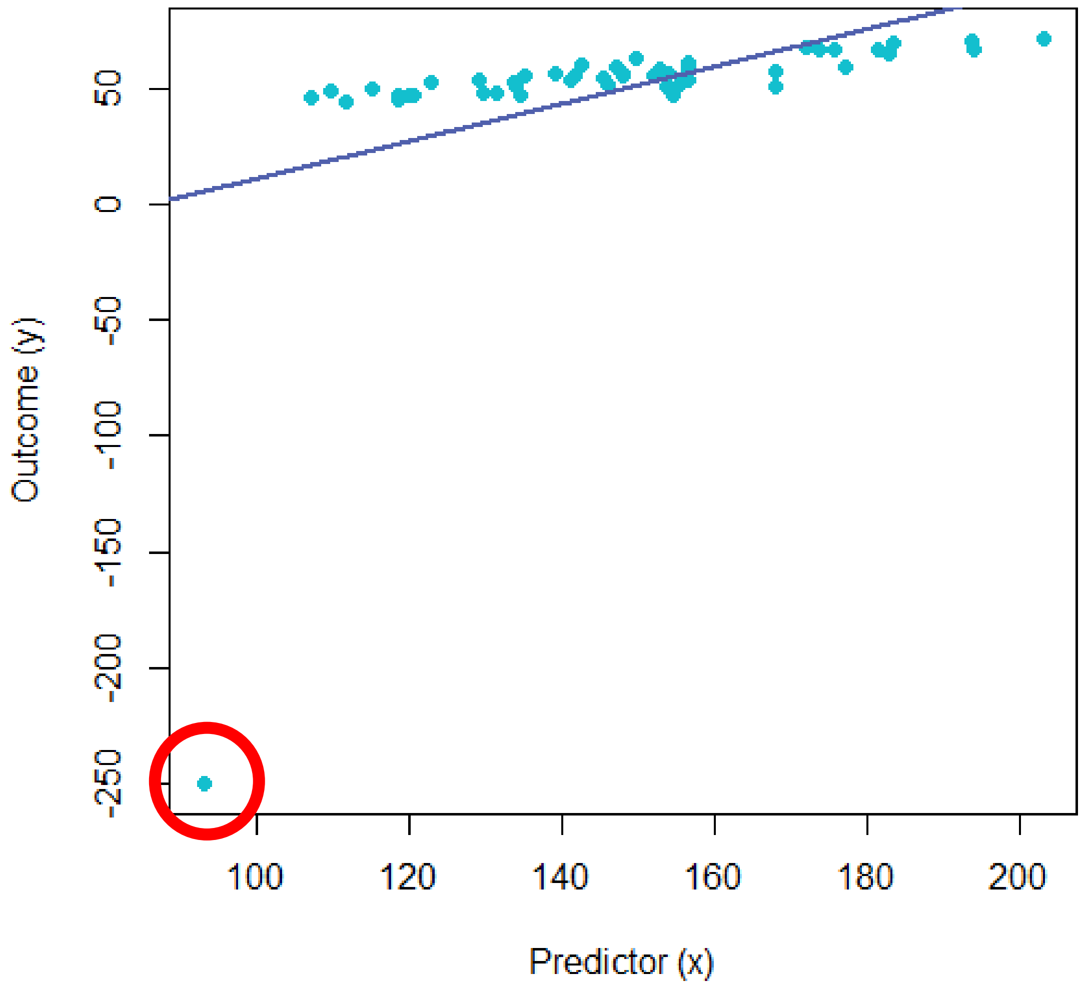
Tests and Statistical Analyses - Data Handling Techniques III

- Sometimes I have one extreme **outlier** messing with the model.
- Here y values shouldn’t be >200 but one value is pulling on the **regression line** a lot.

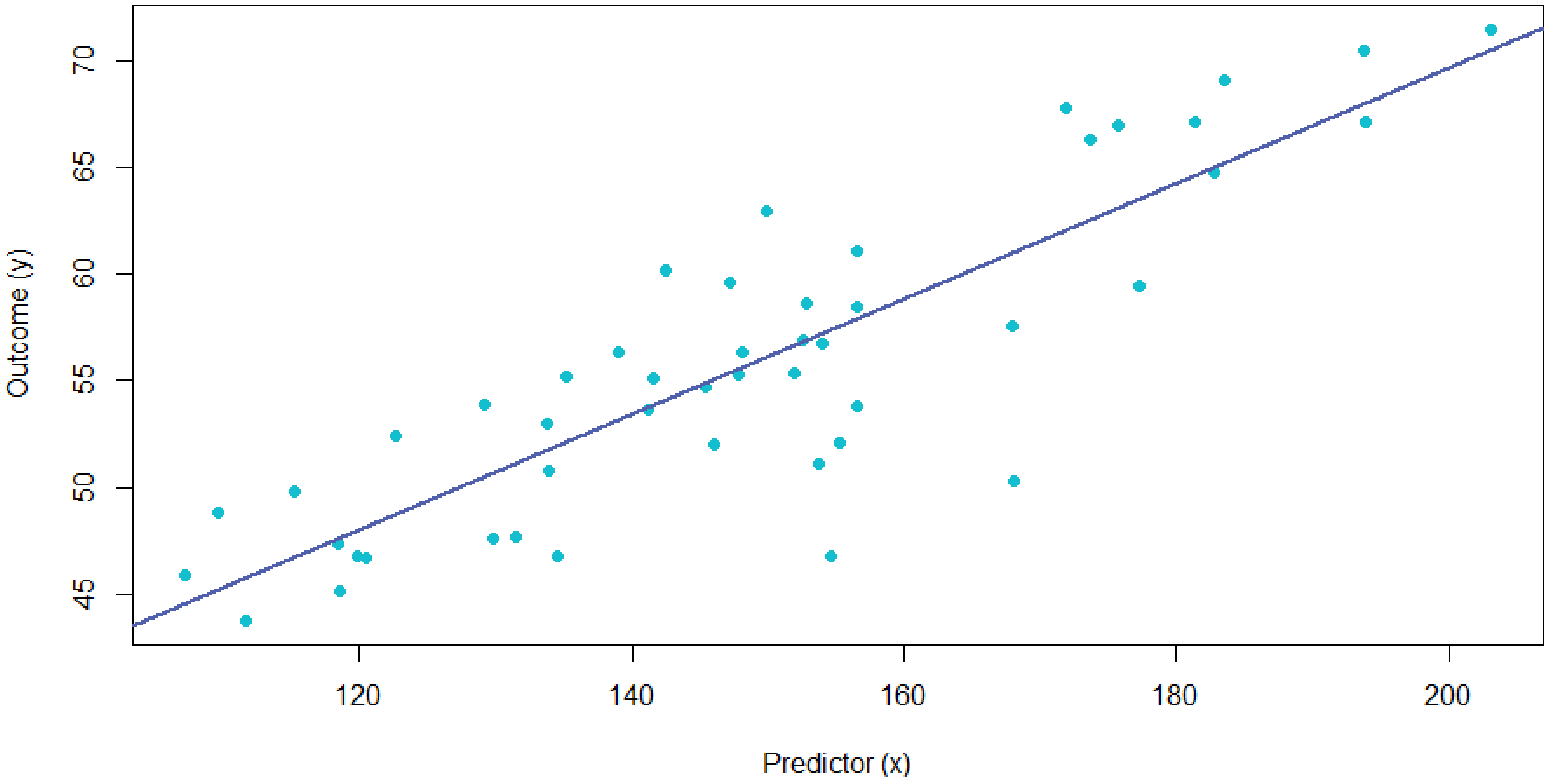


Tests and Statistical Analyses - Data Handling Techniques IV

- After **removing y>**200 there is still a lower outlier exerting **undue influence**.
- I exclude y<0…



Tests and Statistical Analyses - Data Handling Techniques V

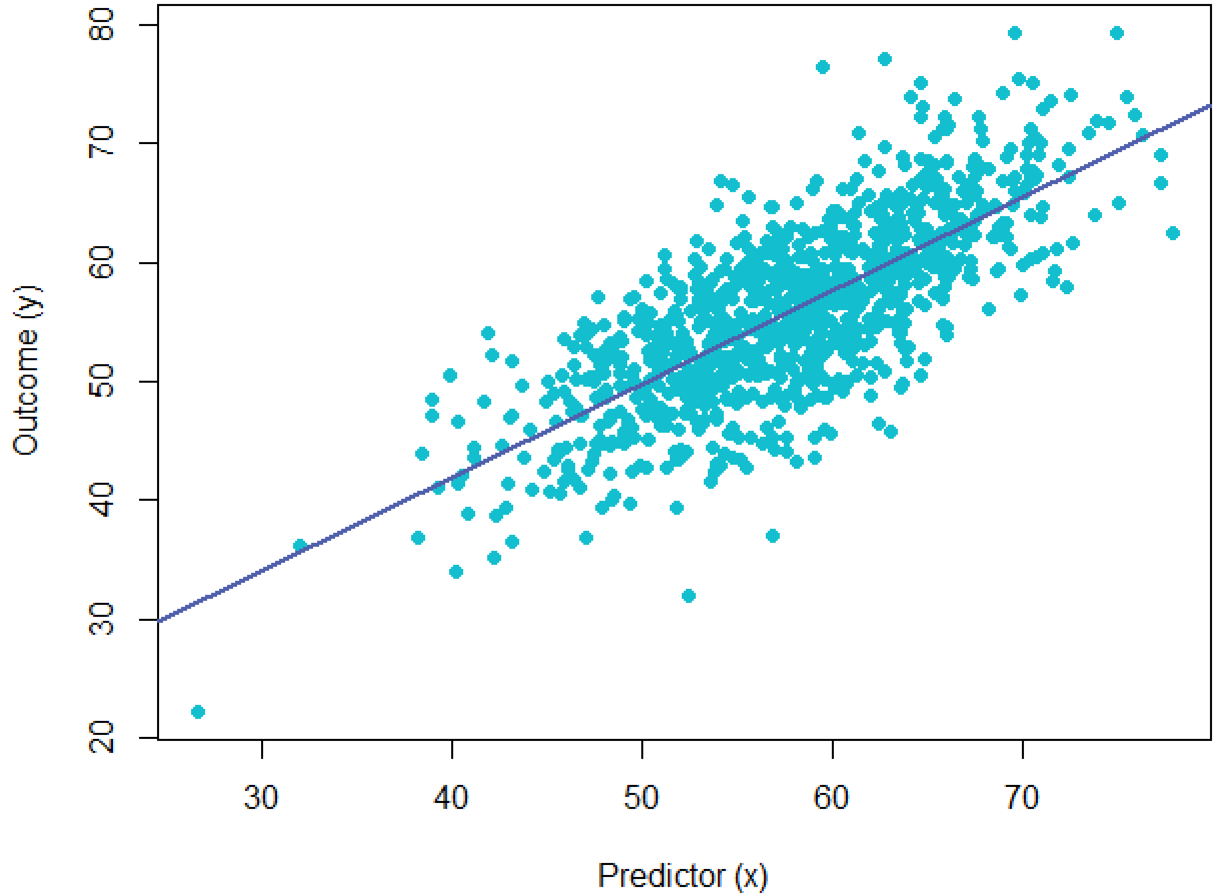


Sensitivity Analysis I

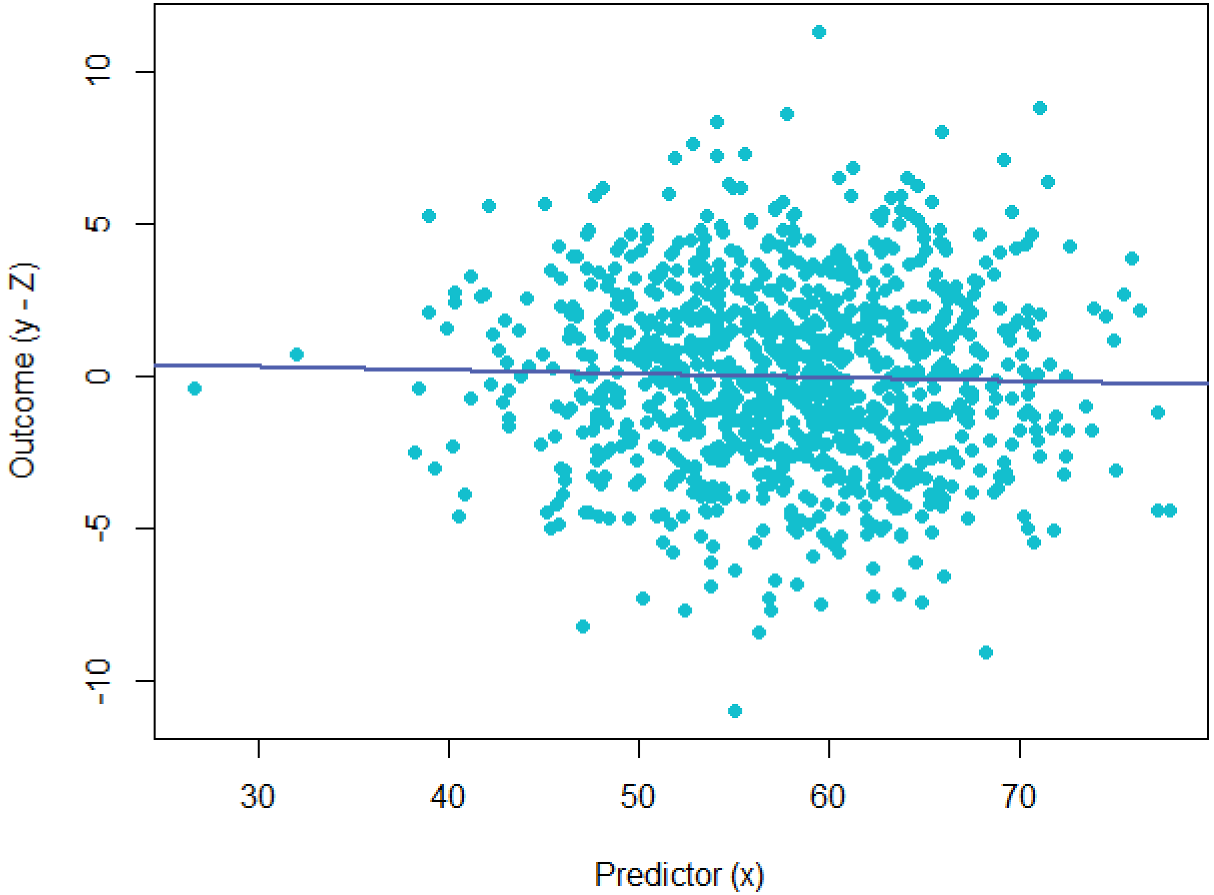
- **Sensitivity analysis** involves testing the robustness of a result.
- How much does our conclusion depend on **influential points**.
- How **unusual** must our sample be for something different to be true of the population.
- How much influence would an **unmeasured confound** need to change our conclusion.

Sensitivity Analysis II

The relationship between x and y



The relationship between x and y after controlling for Z



Identifying Patterns in Exposure Data

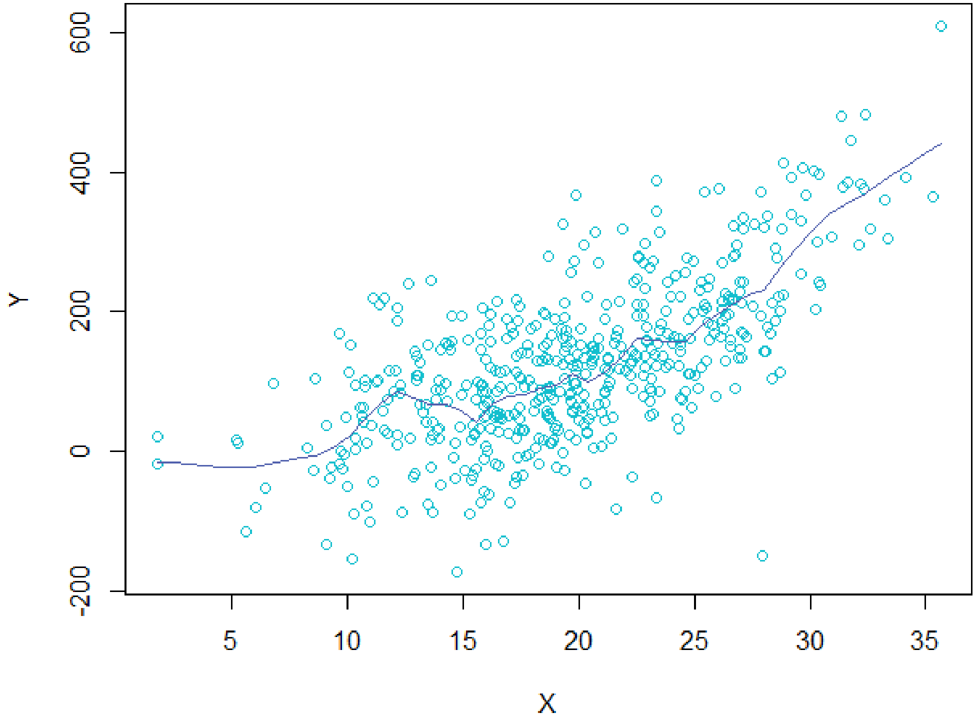
- A fundamental approach to identifying patterns in exposure data is to **compare** the **frequency of an outcome** between individuals who have and have not been **exposed.**

	Disease	No disease
Exposed	A	B
Not exposed	C	D

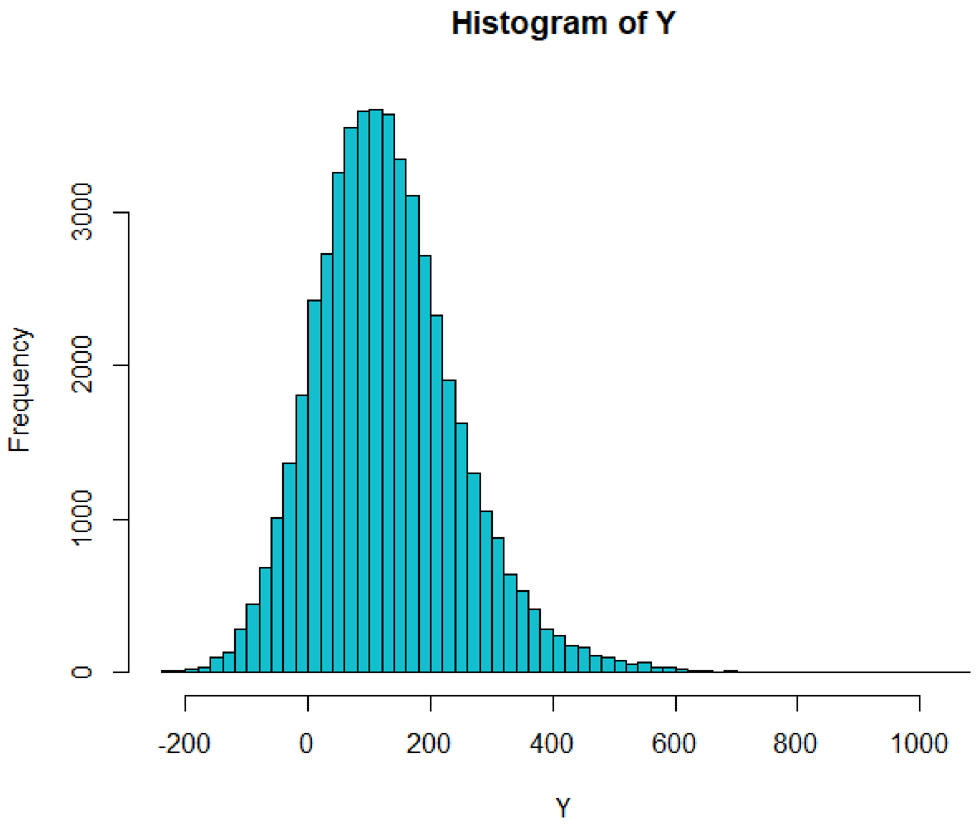
$$relative\ risk\ (RR) = \frac{A/(A+B)}{C/(C+D)}$$

- **Relative risk (RR)** is a **ratio** of the **risk** of a having a disease if you're exposed vs the risk if you're not.
- RR = 1 = no increased risk from exposure; <1 reduced risk; > 1 increased risk.
- If the RR is 1.4 then the exposed group are 1.4 times more likely to contract the disease.

Technical Visualisation Methods



We've already seen scatterplots, but we can add a **smoothed line** to see the relationship better. On the right a smooth line is added using a method called **LOESS**.



A **histogram** allows us to visualise the **distribution** of **one** variable. An example is shown on the left.

Content

1. Introduction to the Statistics used in Exposure Assessment
2. Normality
3. Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)
4. Describe different Tests and Basic Statistical Analyses
5. **Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures**
6. Summary

Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures I

- **Detecting exposure patterns**
  Statistical tools help identify trends, clusters, and distributions in environmental data (e.g., air pollutants, chemicals in water, noise levels).
- **Describing exposure levels**
  Descriptive statistics (mean, median, percentiles) summarise typical and extreme exposure values and help compare populations or time periods.
- **Understanding variability and uncertainty**
  Measures of dispersion (variance, IQR, standard deviation) show how exposure levels fluctuate, while confidence intervals quantify uncertainty.
- **Linking exposures to influencing factors**
  Regression, correlations, and multivariate models reveal relationships between exposures and determinants such as location, behaviour, or season.

Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures II

- **Identifying high-risk groups or hotspots**
  Spatial analysis, thresholds, and exceedance statistics help locate populations or areas with elevated exposure.
- **Supporting risk assessment and standards**
  Percentile-based values and model outputs inform regulatory limits, health-based guidelines, and safe exposure ranges.
- **Evaluating interventions**
  Before–after comparisons and trend analyses assess whether mitigation measures (filters, policy changes, emission controls) reduce exposures.
- **Communicating results clearly**
  Transparent reporting of data, methods, and uncertainty allows policymakers and the public to make informed decisions about environmental health.

Content

1. Introduction to the Statistics used in Exposure Assessment
2. Normality
3. Handling Data below the Limit of Detection (LoD)/Limit of Quantitation (LoQ)
4. Describe different Tests and Basic Statistical Analyses
5. Interpreting Statistics to Identify, Measure, and Mitigate Environmental Exposures
6. **Summary**

Summary: Key Takeaways

- **Descriptive and summary statistics:**
  Describe and summarise data with measures of central tendency (mean, median, mode), and measures of dispersion (variance, interquartile range).

- **Normality:**
  Understand how data in nature tend to follow a normal distribution and how we can leverage the properties of normal distributions in statistical testing.

- **Handling data outside LOD/LOQ:**
  Understand the limits of measurement and how to address these limits.

- **Statistical testing:**
  Understand how outcomes can be explained by other factors with regression and explored with plots.

- **Interpretation:**
  How data and statistics can be used to understand relations between exposures and outcomes.

Consequent Modules

Module 2:

- Statistics for Exposure Scientists

Module 3:

- Introduction to Epidemiology

Future training videos:

- Additional legislations may be covered by ISES Europe in specialised training videos.

Thank You!

We appreciate your participation and attention

Congratulations on completing the ISES Europe training videos.

Access all videos via: https://ises-europe.org/

Further Reading

Introduction to Statistics

Lane, DM

https://onlinestatbook.com/Online_Statistics_Education.pdf

Experimental Design: Statistical Analysis of Data

University of Central Arkansas

https://uca.edu/psychology/files/2013/08/Ch10-Experimental-Design_Statistical-Analysis-of-Data.pdf

Uncertainty and Data Quality in Exposure Assessment

World Health Organization (WHO) - International Programme on Chemical Safety (IPCS)

https://www.who.int/publications/i/item/9789241563765