Understanding our Data

1 Overview

In this session, we will be looking at uncertainty and variability, and how we can measure these and incorporate them into our conclusions. Next, we will examine various environmental and ecological data sources, highlighting critical pre-processing steps such as handling censored data, outliers, and missing values.

We often talk about uncertainty and error as though they are interchangeable, but this is not quite correct.

- **Error** is the difference between the measured value and the 'true value' of the thing being measured.
- **Uncertainty** is a quantification of the variability of the measurement result.

1.1 Statistical distributions

Practically speaking, we make use of common statistical distributions to account for uncertainty. These include both continuous and discrete distributions.

1.1.1 Continuous distributions

- *Normal*: perhaps the most commonly used distribution in statistics. $X \sim N(\mu, \sigma^2)$.
- *Exponential*: distribution of the time (λ) between events. $X \sim Exp(\lambda)$.

1.1.2 Discrete distributions

- *Poisson*: distribution of the probability of observing a specific count (θ) within a particular time period. $X \sim Po(\theta)$.
- *Binomial*: distribution of the number of successes in n independent trials where θ is the probability of success. $X \sim Bi(n, \theta)$.
- *Negative binomial*: distribution of the number of trials until the kth success is observed. $X \sim NeBi(k, \theta)$.

1.2 Observational Error

The observational **error** in a measurement is a single result, namely the difference between the measured and the true value. The error may include both a random and a systematic component.

Random error is variation that is observed randomly over a set of repeat measurements. As you make more measurements, these errors tend to average out and your estimates will improve in accuracy.

Systematic error is variation that remains constant over repeated measures. This is typically due to some feature of the measurement process. Making more measurements will not improve accuracy, since all new measurements will be affected in the same way. Systematic error can only be eliminated by identifying the cause of the error.

Figure 1: Representation of systematic vs random error and its relationship with bias/accuracy and precision.

Exercise 1

For each of the examples below, consider whether the error is **random** or **systematic**.

- A meter reads 0.01 even when measuring no sample.
- (A) random
- (B) systematic
- An old thermometer can only measure the temperature to the nearest 0.5 degrees. (e.g., 23.5 °C becomes 23 °C or 24 °C)
- (A) random
- (B) systematic
- A poorly designed rainfall monitor often leaks water on windy days.
- (A) random
- (B) systematic
- You are asked to measure the volume of an ice cube in a warm laboratory.
- (A) random
- (B) systematic
- To estimate the abundance of a fish species in a lake, scientists use a net with a mesh size equal to the average fish length.
- (A) random
- (B) systematic

2 Quantifying uncertainty

2.1 Standard uncertainty and expanded uncertainty

When presenting our results, it is important that we are clear the uncertainty associated with them. A common approach is to use a **standard uncertainty**, which is just the standard deviation, reported as:

$$\text{estimated value} \pm \text{standard uncertainty}$$

The standard uncertainty ($u(\bar{\mathbf{x}})$) for a vector $\mathbf{x}$ of length n is computed as follows:

$$u(\bar{\mathbf{x}}) = \frac{sd(\mathbf{x})}{\sqrt{(n)}}$$

More generally we can use an **expanded uncertainty**, which is obtained by multiplying the standard uncertainty by a factor k. You have already seen this in statistics as the key building block of a confidence interval. The value of k is chosen based on the quantiles of a standard normal distribution, with a value of $k = 1.96$ (or $k = 2$) giving a 95% confidence interval. The 95% CI for the mean of $\mathbf{x}$ is given as $\bar{\mathbf{x}} \pm 1.96 \times u(\bar{\mathbf{x}})$.

Example: Bathing water quality

All bathing water sites in Scotland are classified by SEPA as "Excellent", "Good", "Sufficient" or "Poor" in terms of how much fecal bacteria (from sewage) they contain. The minimum standard all beaches or bathing water must meet is "Sufficient". The sites are classified based on the 90th and 95th percentiles of samples taken over the four most recent bathing seasons.
The figure below shows the data from some selected sites.

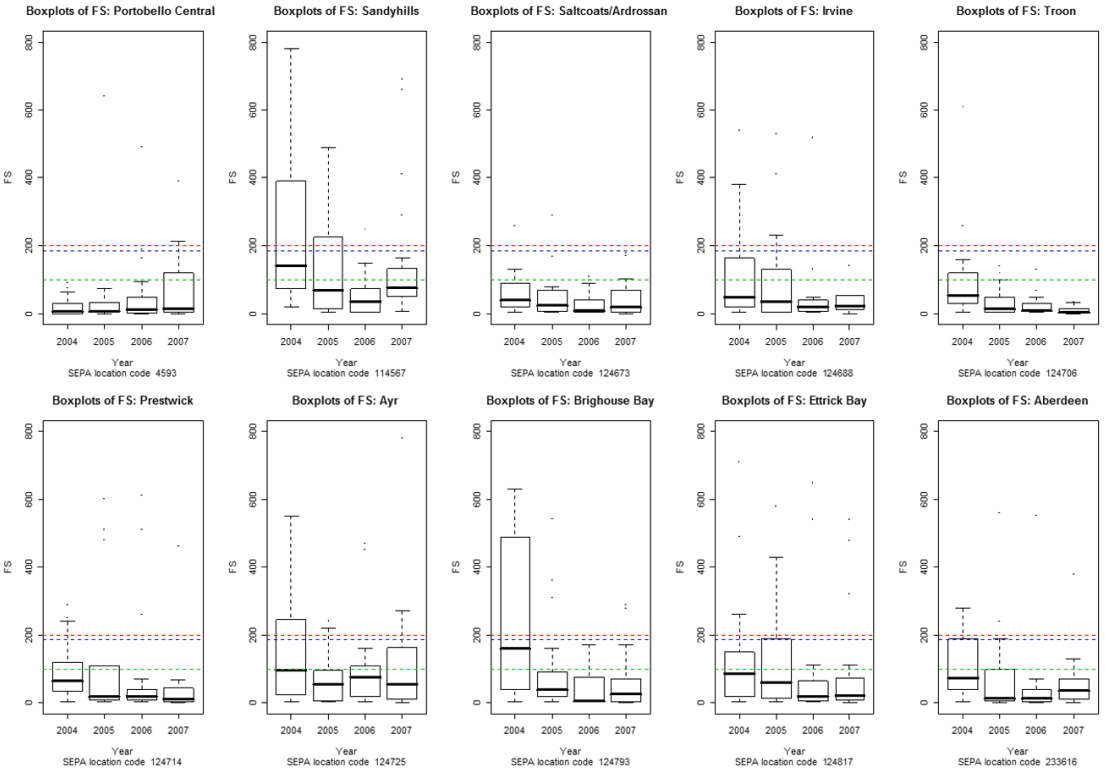
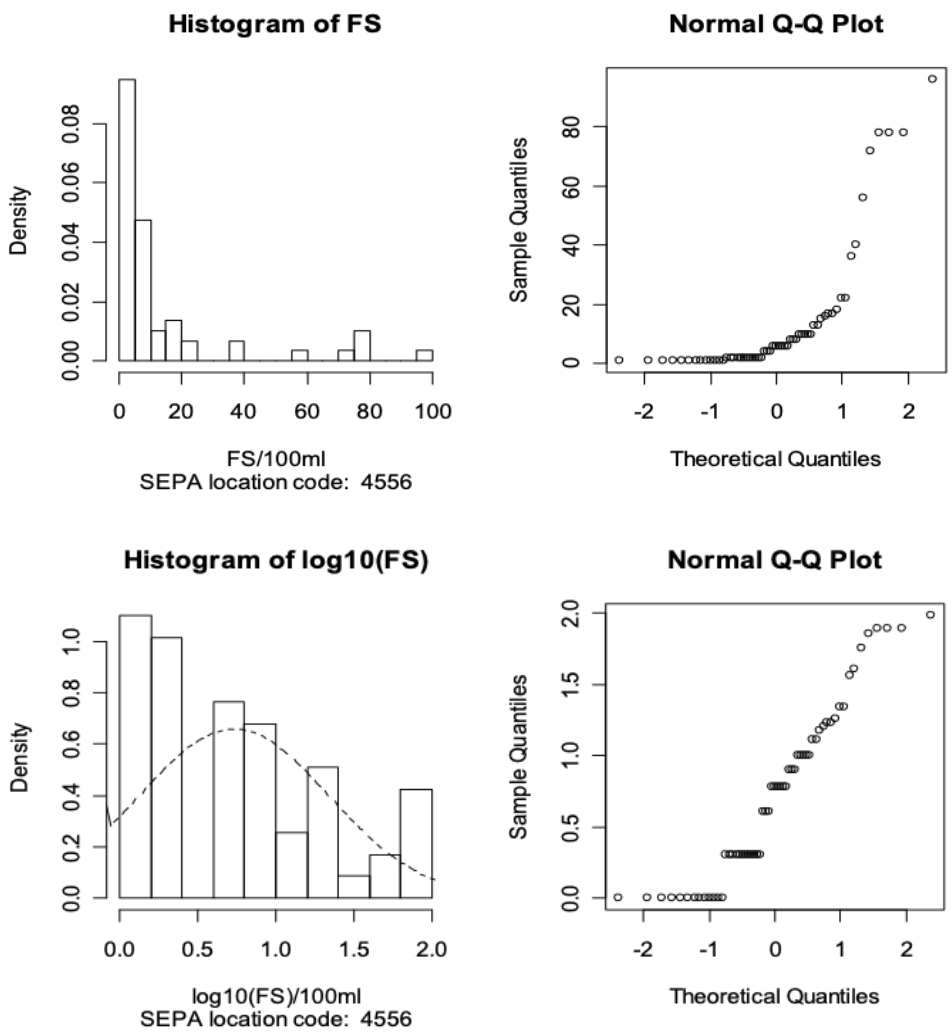


Figure 2: Boxplots of FS by year for 10 sites. The dashed horizontal lines represent "excellent" (green, lowest), "good" (blue, middle) and "sufficient" (red, highest) classification boundaries, respectively.

The classification is based on a belief that the samples at each site follow a log-normal distribution. If this assumption does not hold, then our classifications would not be accurate. Therefore, it is crucial that we regularly assess this assumption to ensure the safety of our bathing water. We can use our standard plots to assess log-normality. In the figure below, the top plots are produced using the untransformed data and the bottom plots are produced after taking a logarithmic transformation of the data (FS).



Exercise 2

Can we assume that the samples at each site follow a log-normal distribution?

- (A) Yes
- (B) No

Solution

Yes, we can assume that the samples at each site follow a log-normal distribution. From the plots, there is no strong evidence to suggest we have breached our assumptions.

Specifically, the histogram of $\log_{10}$(FS) shows that the distribution is not far from a bell shape, and the points on the Normal Q-Q plot lie close to the line of equality.

Example: Bathing water quality (continued)

In the following exercises, we will calculate the standard uncertainty and a 95% confidence interval for the mean of log(FS).

Exercise 3

(a) We have 80 measurements of log(FS), with a mean of **3.861** and a standard deviation of **1.427**. Use these to calculate the standard uncertainty of the population mean log(FS) using our vector **x**.

Answer (to 3 decimal places): _____
Solution

$$u = \frac{sd(\mathbf{x})}{\sqrt{(n)}} = \frac{1.427}{\sqrt{80}} = 0.160$$

(b) Given the standard uncertainty that calculated in part (a), calculate a 95% confidence interval for the population mean of log(FS).

Answer (to 3 decimal places): (_____,_____)
Solution
A 95% confidence interval for $\bar{x}$ is:

$$\bar{x} \pm 1.96 \times u = 3.861 \pm 1.96 \times 0.160 = (3.574, 4.175)$$

2.2 Uncertainty propagation

For a measure Y that is a linear combination of n quantities $X_1, ..., X_n$ (i.e. $Y = a_1X_1 + ... + a_nX_n$, with $\mathbf{a} = (a_1, ..., a_n)$ being a row vector of coefficients), the **combined uncertainty** $u(Y)$ is calculated as follows:

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\left(\sum_{j=1}^{n} a_j X_j\right) \\
&= \sum_{i=1}^{k} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j) \\
&= \sum_i \sum_j a_i a_j \underbrace{\mathrm{Cov}(X_i, X_j)}_{\rho_{ij}\sigma_i\sigma_j} \\
&\Rightarrow \\
u(Y) &= \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(u(X_i) \times u(X_j) \times a_i \times a_j \times \rho_{ij}\right)}
\end{aligned}
$$

where $u(X_i) = \sigma_i$ and $u(X_j) = \sigma_j$ are the standard uncertainties of X_i and X_j, respectively, and ρ_{ij} is the correlation between X_i and X_j.

If $X_1, ..., X_n$ are independent, this reduces to:

$$
u(Y) = \sqrt{\sum_{i=1}^{n} \left(u(X_i)^2 \times a_i^2\right)}
$$

Exercise 4

Show that the combined uncertainty $u(Y)$ for $Y = a_0 + a_1 X_1 + a_2 X_2$ (**not** assuming that $X_1, ..., X_n$ are independent) reduces to:

$$
u(Y) = \sqrt{a_1^2 u(X_1)^2 + a_2^2 u(X_2)^2 + 2\rho_{12} u(X_1) u(X_2) a_1 a_2}
$$

Solution
We have:

$$
u(Y) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \left(u(X_i) \times u(X_j) \times a_i \times a_j \times \rho_{ij}\right)}
$$

where $i = 1, 2$ and $j = 1, 2$, i.e.

$$
u(Y) = \sqrt{u(X_1)u(X_1)a_1a_1\rho_{11} + u(X_1)u(X_2)a_1a_2\rho_{12} + u(X_2)u(X_1)a_2a_1\rho_{21} + u(X_2)u(X_2)a_2a_2\rho_{22}}
$$

$$
\therefore u(Y) = \sqrt{u(X_1)^2 a_1^2 + 2u(X_1)u(X_2)a_1a_2\rho_{12} + u(X_2)^2 a_2^2}
$$

since $\rho_{11} = \rho_{22} = 1$ and $\rho_{12} = \rho_{21}$.

The **general uncertainty propagation formula** is as follows. The standard uncertainty of $Y = f(X_1, ..., X_n)$ is:

$$u(Y) = \sqrt{\sum_{i=1}^{n} f'(\mu_i)^2 u(X_i)^2}$$

where $f'(\mu_i)$ is the partial derivative of Y with respect to X_i evaluated at its mean μ_i.

Exercise 5

We wish to calculate the area A of a rectangle, with height h and width w. ($A = h \times w$.) Height and width are measured with uncertainty $u(h)$ and $u(w)$, respectively. Evaluate the uncertainty on the area A.
Solution

$$u(A) = f(h, w) = u(h \times w)$$

$$\frac{df}{dh} = w \quad \text{and} \quad \frac{df}{dw} = h$$

$$\therefore u(A) = \sqrt{w^2 u(h)^2 + h^2 u(w)^2}$$

Exercise 6

The **Shannon index** (or Shannon-Wiener diversity index) is widely used in Ecology to quantify the diversity of a biological community by considering both species richness and evenness. It is calculated as:

$$H = -\sum_{i=1}^{S} p_i \log(p_i)$$

where p_i is the proportion of species i in the community computed as the ratio between the num. of individuals of a given species and total number of individual across all species.

1. Suppose that the proportion of each species is estimated with some uncertainty $u(p_i)$. Provide the general form for the uncertainty propagation of these proportions on the calculation of H.

2. Imagine you go to your garden an find out there are $S = 3$ different species of arthropods living there. Then you go out one day and sample $N = 100$ individuals and end up collecting $n_1 = 50$ ants, $n_2 = 30$ beetles and $n_3 = 20$ spiders. Assuming that number of individuals of a given species follows $n_i \sim \text{Binomial}(N, \theta_i)$, and let

$\widehat{\theta}_i = \frac{n_i}{N} = p_i$ be the estimator of θ_i show that $u(p_i)^2 = p_i(1-p_i)/N$ and then compute the uncertainty propagation for the Shannon Index.

Solution

$$u(H) = \sqrt{\sum_i^S \left(\frac{\partial H}{\partial p_i}\right)^2 u(p_i)^2}$$

where

$$\frac{\partial H}{\partial p_i} = -(\log(p_i) + 1)$$

$$\therefore u(H) = \sqrt{\sum_i^S \left(-\log(p_i) - 1\right)^2 u(p_i)^2}$$

Assuming $n_i \sim \text{Binomial}(100, p_i)$ for $i = 1, 2, 3$ where $p_1 = 50/100; p_2 = 0.3; p_3 = 20/100$. The partial derivatives are given by

$$\frac{\partial H}{\partial p_1} = -(\log 0.5 + 1) \approx -0.307$$
$$\frac{\partial H}{\partial p_2} = -(\log 0.3 + 1) \approx 0.204$$
$$\frac{\partial H}{\partial p_3} = -(\log 0.2 + 1) \approx 0.609$$

First, the variance of p_i (i.e., $u(p_i)^2$) is given by:

$$\begin{aligned}\text{Var}(p_i) &= \text{Var}\left(\frac{n_i}{N}\right) \\ &= \frac{1}{N^2}\text{Var}(n_i) \\ &= \frac{Np_i(1-p_i)}{N^2} = \frac{p_i(1-p_i)}{N}\end{aligned}$$

Thus,

$$u(p_1)^2 = \frac{0.5 \times 0.5}{100} = 0.0025$$
$$u(p_2)^2 = \frac{0.3 \times 0.7}{100} = 0.0021$$
$$u(p_3)^3 = \frac{0.2 \times 0.8}{100} = 0.0016$$

Then,

$$u(H) = \sqrt{0.0025 \times (-0.307)^2 + 0.0021 \times (0.204)^2 + 0.0016 \times (0.609)^2} \approx 0.03$$

3 Data Sources

3.1 Ecological and Environmental Data sources

Over the last decade, the information available for surveying and monitoring ecological and environmental resources has changed radically. The rise of new technologies, novel collection methods, and modern data-submission platforms have facilitated the access to large volumes of environmental and ecological data.

Today's ecological and environmental data landscape is overwhelmingly vast - far too extensive to cover comprehensively in one session! Instead, we'll focus on key data sources and digital technologies that are currently shaping policy decisions, enabling scientific breakthroughs, and driving innovation in research.

Figure 3: Environmental monitoring over the year (UKCEH)

3.2 Institutional Monitoring Programmes

Institutional Monitoring programmes have long been a primary source of information for long-term environmental assessment, producing **structured datasets** essential for detecting ecological change and informing evidence-based policy.

These initiatives rely *field surveys* conducted on established *monitoring networks* to track trends in species populations, habitat quality, and ecosystem processes - a topic we will explore in detail in the following session. Their strength lies in rigorous implementation of *standardized sampling protocols*, which reduces the observational errors associated with data collection methods. However, these are typically constrained by other factors. For example, large-scale programmes are inherently resource-intensive to maintain and often limited in taxonomic scope (typically focusing on key species), spatial/geographic coverage, and temporal resolution. Some popular monitoring schemes are shown below:

Monitoring Scheme	Description
United Kingdom Butterfly Monitoring Scheme (UKBMS)	Protocolized sampling scheme run by butterfly conservation that has monitored changes in the abundance of butterflies throughout the United Kingdom since 1976.
UK Environmental Change Network (ECN)	UK's long-term ecosystem monitoring and research programme that has produced a large collection of publicly available data sets including meteorological, biogeochemistry and biological data for different taxonomic groups (Rennie et al. 2020).
National Hydrological Monitoring Programme (NHMP)	The NHMP, particulalry the National River Flow Archive conveys a national scale management of hydrological data within the UK hosted by the UKCEH since 1982 collating hydrometric data from gauging station networks operated by multiple agencies.
Natural Capital and Ecosystem Assessment (NCEA)	Long-term environmental monitoring of natural capital including data from freshwater Surveillance Networks, ecosystem condition & soil health, forest inventory, estuary and coast surveillance, etc.
Breeding Bird Survey (BBS)	Main scheme for monitoring the population changes of the UK's common breeding birds. It covers all habitat types and monitors 110 common and widespread breeding birds using a randomised site selection.

3.3 Citizen Science Programmes & Platforms

Citizen science (CS) monitoring involve public participation to collect large volumes of ecological & environmental data at a low cost across broad spatiotemporal scales.

Data submission platforms like iNaturalist and eBird have become an important groundwork for citizen scientist to submit records and generate vast, real-time datasets, enabling researchers to track species distributions, phenology, and ecosystem responses to environmental change in ways that were previously logistically and financially unfeasible.

Despite this, the analysis of such data remains challenging as there is little o no design involved in the sampling protocols of most CS data recording schemes. A major issue with the lack of a standardized sampling protocol is that sampling efforts tend to be uneven over time and space and biased towards human activity centers, locations that are easy to access, or sites where species are more likely to be found such as protected area. If unaccounted for, such sources of bias can mislead Figure 4 (left) shows the sampling effort based on CS records submitted through the Pl@anet Net App in the French mediterranean region. If we compare the spatial effort against the elevation of the region we can clearly see that a *sampling bias* towards lower elevation values. This would imply that small populations at lower elevation could be over-sampled and if we had worngly assumed sampling was evenly distributed, then species distributions at higher elevation would be under-estimated.

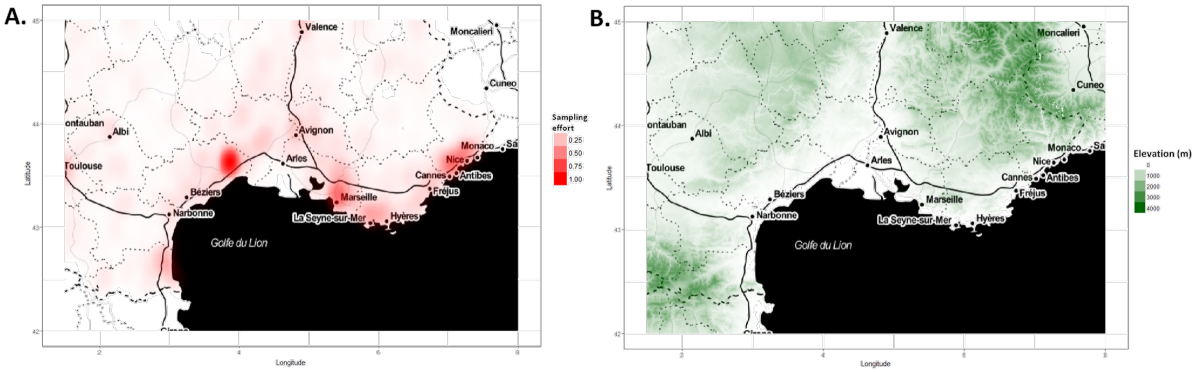


Figure 4: Elevation versus sampling effort (obtained through the Pl@net Net App) in the French mediterranean region (Figure taken from Botella et al. (2020))

Harnessing the power of CS data is not an easy task.

Advantages	Disadvantages
Extensive taxonomic, spatial and temporal coverage.	Under-reporting of rare and inconspicuous species.
Eye-catching species that are easily identifiable by participants.	Varying recording skills and uneven sampling effort.

3.4 Biological Collections

Biological collections constitute probably the oldest form of historical data reservoirs. For over 300 years, naturalists have been collecting and preserving biological specimens, initially for personal curiosity and public display. Today, their value has expanded far beyond their original purpose; they are now recognized as critical sources of information for addressing modern global challenges like biodiversity loss and climate change. Now housed mostly in **museums** and **herbaria** throughout the world, these biological collections, and their associated systematic research, provided the basis for much biological research.

For instance, the Natural History Museum in London safeguards a collection of over 80 million specimens, spanning 4.5 billion years of Earth's history to the present. This unparalleled archive, along with many others, is increasingly accessible through digital data portals, enabling researchers worldwide to analyse historical trends and understand the distribution of biodiversity and geodiversity through time.

Despite their immense value, biological collections data are subject to significant limitations and biases that need bo considered.

- Most historic collection were obtained in an opportunistic manner without following any particular sampling protocol (largely dependent on the particular interests of the collector).
- Often there is no information about the collection methods or effort employed.
- Limited in the geographic coverage and typically biased near centres of human activity and along the roads or during specific seasons.
- The information associated with each collection or specimen (e.g., species, sex, collection date, location, collector's name, morphological measurements, habitat description) may vary widely which limits the environmental context and ecological questions that can be addressed.
- Strongly biased towards specific taxonomic groups, especially birds and mammals

Exercise 7

Read the paper by Pyke and Ehrlich (2010) and discuss three scenarios where biological collections have been used to address different environmental issues and ecological questions.

3.5 Data Repositories & Portals

Data repositories have become a major sources of information for modern environmental and ecological research, serving as centralized, curated platforms that aggregate, preserve, and disseminate vast quantities of data from diverse sources.

These digital archives - ranging from global biodiversity databases like the Global Biodiversity Information Facility (GBIF) to thematic collections such as the National Biodiversity Network (NBN) Atlas - standardize and harmonize heterogeneous datasets, enabling researchers to access, share, and reuse data across disciplines and geographic boundaries. Often, these repositories are integrated into comprehensive **data portals** that host interactive visualisaion tools, web-based applications, programming interfaces (APIs) and data catalogues, transforming static archives into dynamic platforms for exploration and discovery (see e.g. UK-SCAPE plant diversity trends and Grasshoppers and Allied Species Recording Scheme).

Exercise 8

Select two or three of the following data repositories. For each, examine some of the available datasets and list the types of uncertainty or error that might affect the data quality and reliability.

Data Repository	Description
Move Bank	Movebank is a global, open-data repository and research platform that specializes in managing, sharing, and analyzing **animal tracking** and bio-logging data.
Global Biodiversity Information Facility (GBIF)	GBIF is an international open-data infrastructure that provides free and universal access to over two billion **species occurrence records** from a vast network of museums, research institutions, and citizen science platforms worldwide.
National Biodiversity Network (NBN)	The NBN Atlas is the UK's largest repository of publicly available biodiversity data, aggregating and providing open access to millions of **species records** from a wide range of recording societies, conservation NGOs, and research institutions across the country.

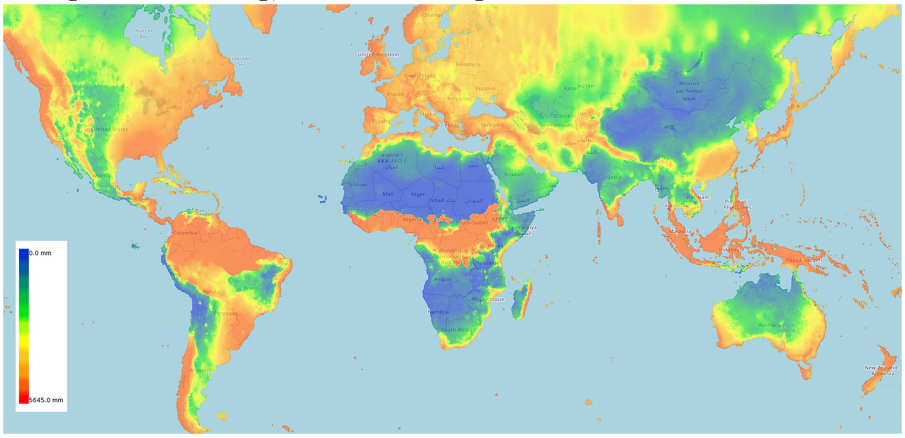
Biological Records Centre (BRC)	The Biological Records Centre (BRC) is a national UK facility that supports and coordinates a network of volunteer recording schemes and societies to collect, manage, and disseminate high-quality data on terrestrial and freshwater **species distributions**.
National River Flow Archive (NRFA)	The National River Flow Archive (NRFA) is ta hydrometric data repository hosted by the CEH, curating and providing open access to river flow, groundwater level, and rainfall time-series from a national network of monitoring stations.
UK Lakes Portal	The UK Lakes Portal is a comprehensive data hub that provides access to physical, chemical, and biological monitoring data for lakes and reservoirs across the United Kingdom
World Ocean Database (WOD)	The World Ocean Database (WOD) is the world's largest publicly available, quality-controlled repository of uniformly formatted oceanographic profile and plankton data, spanning centuries of global marine observations.
UK Water Resources Portal	The UK Water Resources Portal is an interactive online platform that provides access to current and historical data on water availability, including river flows, groundwater levels, rainfall, and reservoir stocks.
Water quality data archive (WIMS)	The Water Quality Archive provides data on water quality measurements. Samples are taken at sampling points around England and can be from coastal or estuarine waters, rivers, lakes, ponds, canals or groundwaters.
Ecology & Fish Data Explorer	This is an online data portal that provides access to ecological monitoring data for English rivers and lakes, including fish populations, invertebrate surveys, and plant communities.
Knowledge Network for Biocomplexity (KNB)	The Knowledge Network for Biocomplexity (KNB) is an open-source data repository that enables the discovery, management, sharing, and synthesis of complex, heterogeneous ecological and environmental datasets.

3.6 Processed information products

Processed information products transform raw measurements into refined, analysis-ready resources tailored for decision-makers and researchers. Unlike primary data repositories, these products undergo rigorous calibration, integration, and modelling to generate authoritative maps, indicators, and synthesized datasets.

Example: WorldClim

WorldClim is a widely used set of global, high-resolution climate surfaces (raster maps) that provide interpolated estimates of historical and future projections (using global climate models CMIP) of temperature, precipitation, and other bioclimatic variables. The historical data layers are generated by applying advanced spatial interpolation algorithms to an extensive global network of weather station records, creating continuous, gap-free rasters. These surfaces serve as the foundational data for species distribution modeling, ecological forecasting, and a vast range of other environmental research applications.



Nowadays, it is common that contemporary data products are synthesized based on a combination of multiple data sources, including field surveys, citizen science and advanced **remote sensing** data from satellite and aerial platforms.

3.6.1 Remote sensing

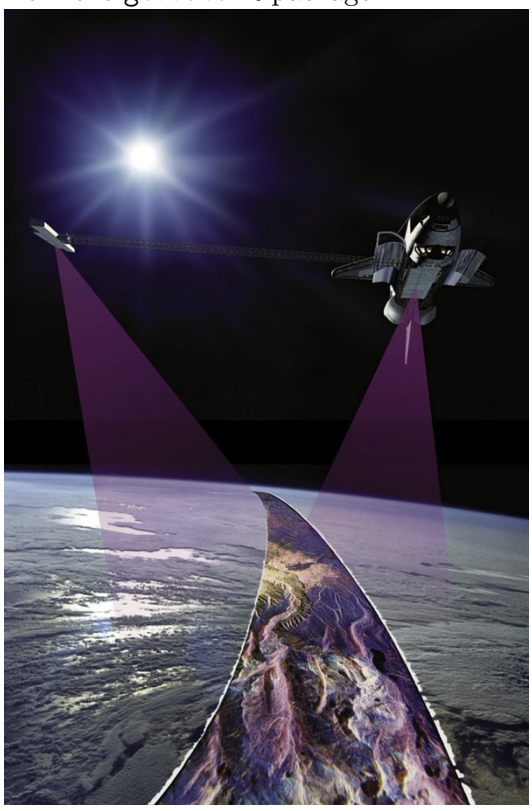
Remote sensing refer the process of obtaining information of an object from a distance, typically from aircraft or satellites. Recent advances in bioinformatics, GIS technologies and remote sensing techniques have changed radically how we monitor the Earth's environment at multiple spatial and temporal scales. These technologies enables the systematic, non-invasive, and often near-real-time collection of data across vast and inaccessible regions. The resulting data are then calibrated, classified, and modeled using specialized algorithms to generate

diverse information products, such as land cover classifications, vegetation indices, and digital elevation models.

While remote sensing-based products enable the quantification of ecological and environmental parameters across extensive geographic scales, they are often subject to substantial uncertainties. These include systematic errors from sensor calibration, spatial and temporal resolution constraints, and generally lower accuracy compared to direct *in-situ* field measurements. Consequently, remote sensing data are often validated using data collected *in-situ* to assess and ensure their accuracy.

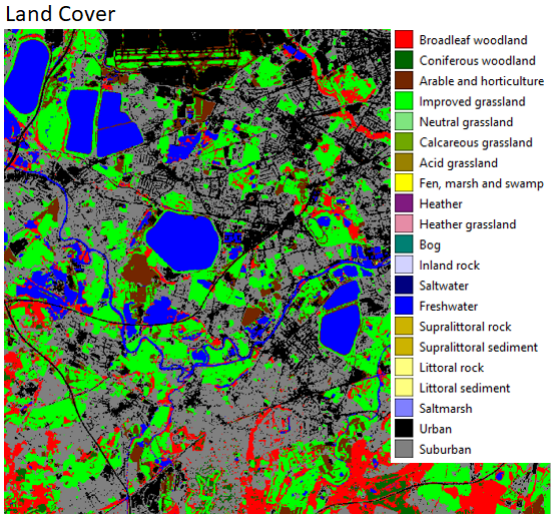
Example: Digital Elevation Models

Digital Elevation Models (DEMs) are digital representations of the earth's topographic surface. DEMs providing a continuous and quantitative model of terrain morphology and are typically stored as a raster grid where each cell (pixel) contains an elevation value. The accuracy of DEMs is determined primarily by the resolution of the model (the size of the area represented by each individual grid cell). For example, the Shuttle RaDAR Topography Mission (SRTM), aquired by NASA using a Synthetic Aperture Radar (SAR) instrument, provide elevation data for any country and is available from the `geodata` R package.



Example: Land Cover Maps

Land cover maps describe the physical material on the Earth's surface. They are created by applying automated algorithms to satellite or aerial imagery to identify features such as grassland, woodland, rivers & lakes or man-made structures such as roads and buildings. For example, UK CEH has produced a series of Land Cover Maps which are a series of spatially continuous, raster-based classification products, derived from the automated analysis of Earth observation data (primarily from the Sentinel-2 satellite constellation), which provide consistent, national-scale representations of surface vegetation and land use classes.



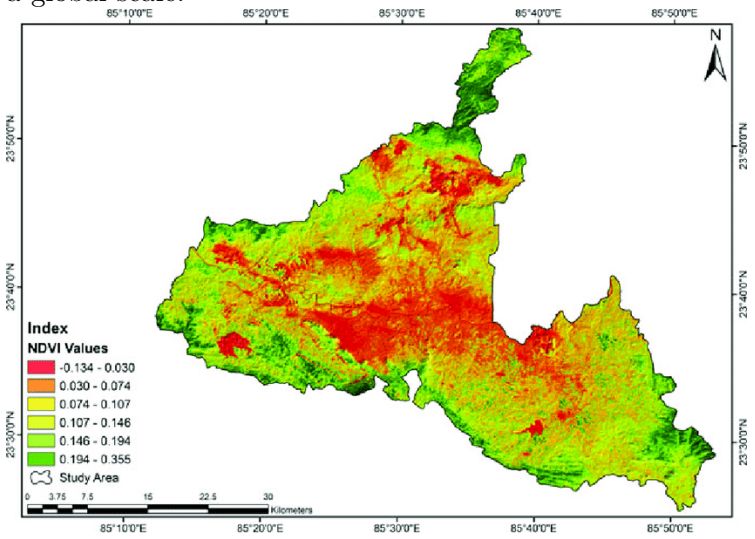
Other widely used global products include MODIS Land Cover, which offers a long-term, coarse-resolution record of global change since 2001, and ESA WorldCover, which provides a high-resolution (10m) global snapshot designed for detailed thematic mapping (the later is available on the `geodata` R package).

Example: NDVI Vegetation Index

Vegetation indeces derived from remote sensing utilize spectral data from satellite or aerial sensors to quantify and monitor plant health, structure, and function across landscapes. These indeces are founded on the principle that vegetation absorbs red light (around 660 nm) for photosynthesis while highly reflecting near-infrared (NIR) light (around 800 nm) due to its internal leaf structure.
This contrast is captured by the Normalized Difference Vegetation Index (NDVI), which is calculated as

$$NDVI = \frac{(NIR - Red)}{(NIR - Red)}$$

The resulting value, which ranges from -1 to +1, provides a standardized measure of greenness; values close to +1 indicate dense, healthy vegetation, values near 0 represent bare soil, and negative values typically correspond to water. By translating raw spectral data into this simple yet robust index, remote sensing enables the tracking of phenological cycles, the assessment of drought stress, and the estimation of primary productivity on a global scale.



3.7 Research-Generated Data

Research-generated data repositories, such as Dryad and Zenodo, are cornerstone platforms in the modern scientific workflow, explicitly designed to uphold the principles of transparency, reproducibility, and open data access. Unlike passive archives, these repositories require researchers to actively deposit the precise datasets, code, and scripts used to generate the results published in peer-reviewed journals. By assigning persistent digital object identifiers (DOIs) to these materials, they create a permanent, citable record that allows other scientists to independently verify, replicate, and build upon published findings. This process is fundamental to detecting errors, reducing redundancy, and accelerating scientific discovery, effectively transforming a single study's output into a reusable resource for the entire research community and safeguarding the integrity of the scientific record.

Exercise 9

1. **Choose a Repository:** Select either Dryad or Zenodo.
2. **Find a Dataset:** Browse or search for a dataset related to a topic in environmental science or ecology that interests you (e.g., "pollination," "microplastics," "forest fragmentation," "climate change adaptation").

3. **Select and Record:** Choose one specific dataset and note down:
    - The full citation for the dataset (including its DOI).
    - The title of the associated publication (if provided).

How did you find this dataset? Was the search intuitive? Is the dataset openly available? Could you download the files without restrictions?

4 Data Preprocessing

Environmental and Ecological systems are inherently complex due to the large number of interrelated biological, physical, and social components. This complexity arises stochastic processes that operate across vastly different spatial and temporal scales. Adding to this complexity, analyzing these systems becomes a challenging task due to the heterogeneity of available data and the different sources of uncertainty that impact the quality of the data. Data collection methods vary widely and spatial and temporal sampling schemes may be too sparse to fully capture overall system behavior. Consequently, we often have to deal with issues such as outliers, missing values, and highly uncertain information.

Fortunately, many of these data quality issues can be addressed. This is typically done through a rigorous data *pre-processing* phase before formal analysis, and through statistical models that explicitly account for the observational process of how the data was collected.

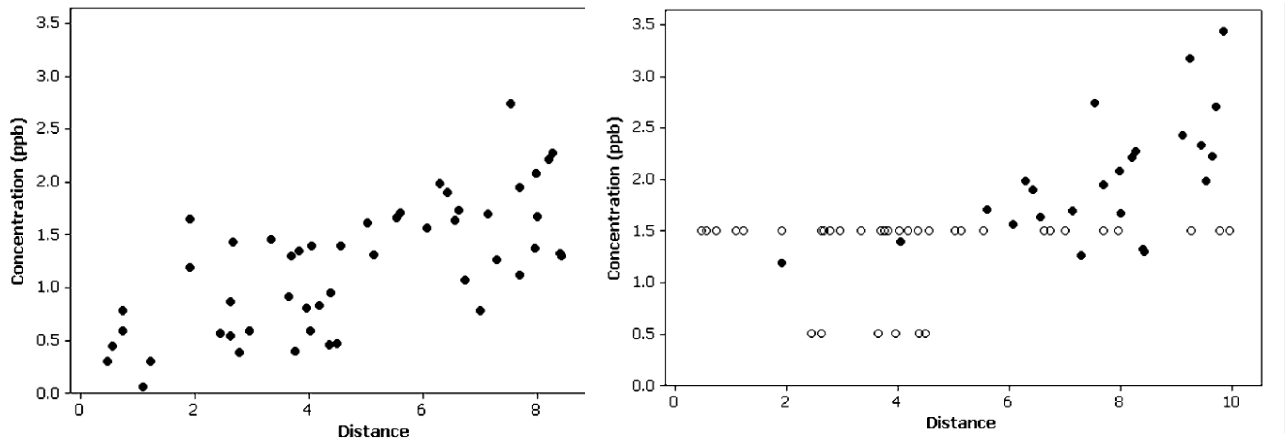
Data pre-processing is crucial stage in any sort of ecological or environmental data analysis and it includes data cleaning, outlier detection, missing value treatment, handling censored data, transformation, and the creation of new derived variables. The goal is to create a robust, consistent dataset ready for analysis while carefully documenting all changes to preserve the integrity of the original information.

4.1 Censored Data

Censored data are data where we are restricted in our knowledge about them in some way or other. Often this will be because we only know that the data value lies below a certain minimum value (or above a certain maximum). For example, if we had scales which only weighed up to 10kg, we would not know the exact weight of any object greater than 10kg.

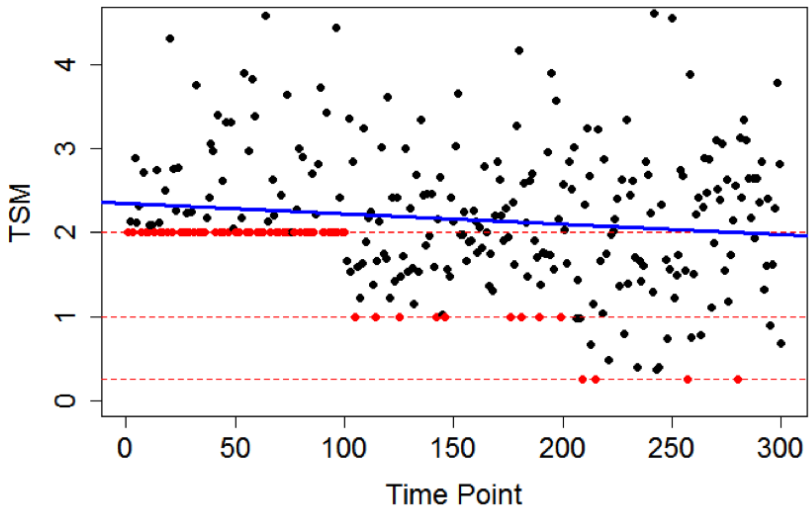
For environmental data, it is more common to have data which are censored at some minimum value. This is because many pieces of measuring equipment will have an analytical **limit of detection**. A limit of detection is *the lowest concentration that can be distinguished with reasonable confidence from a blank* (a hypothetical sample with a value of zero). The limit of detection is often denoted c_L.

Censoring has a huge impact on how we interpret our data. The two plots below show the same data, but the right panel is 'censored' with two different limits of detection (some with an LOD of 0.5, others with an LOD of 1.5).



Censored observations are not completely without information. We still know they are equal to or more extreme than the limit. For an LOD, we might therefore report the data point as either "not detected" or '$< c_L$'. Removing them from our study would not be sensible, since this would lead to us *overestimating* the mean and probably also *underestimating* the variance. We therefore need to find a way to incorporate these censored data points into our analysis.

We can't simply use the minimum value of the LOD. This would ignore the fact that the values are often *below* this. In the plot below, the LOD reduces after every 100 observations (e.g. because of better quality equipment), and this leads to an artificial trend.



4.1.1 Simple Substitution

The simplest approach for dealing with LODs is via **simple substitution**. This involves taking the LOD value and multiplying it by a fixed constant, e.g. by replacing all $< c_L$ values with $0.5c_L$.

This approach is fairly popular because it is simple and easy to implement. However, it only works if there is a small proportion of censored data (maximum 10–15%). If there is a higher proportion, it tends to overestimate the mean.

4.1.2 Distribution-based approaches

It is generally preferable to use a more statistics-based approach which accounts for the data distribution. The basic idea is that we estimate the statistical distribution of the data in a way that takes into account the censoring. We can then use this estimated distribution to simulate values for our censored points.

Commonly used distribution-based approaches are **Maximum Likelihood**, **Kaplan-Meier**, and **Regression on Order Statistics**.

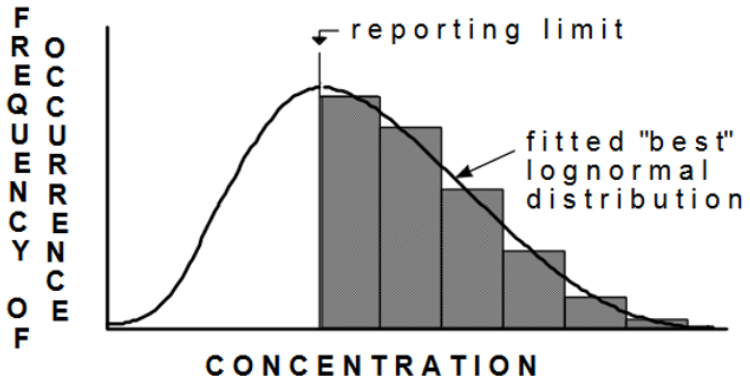
4.1.2.1 Maximum Likelihood

The maximum likelihood approach is a *parametric* approach. It requires us to specify a statistical **distribution** which is a close fit to the data. We then identify the **parameters** of this distribution that maximize the likelihood of obtaining a dataset like ours.

This ML approach has to take into account the likelihood of obtaining:

- the observed values in our dataset
- the correct proportion of data being censored, i.e. the proportion falling below our detection limit(s)

Maximum Likelihood (MLE) -- fits 'best' lognormal distribution to the data, and then



determines summary statistics of the fitted distribution to represent the data.

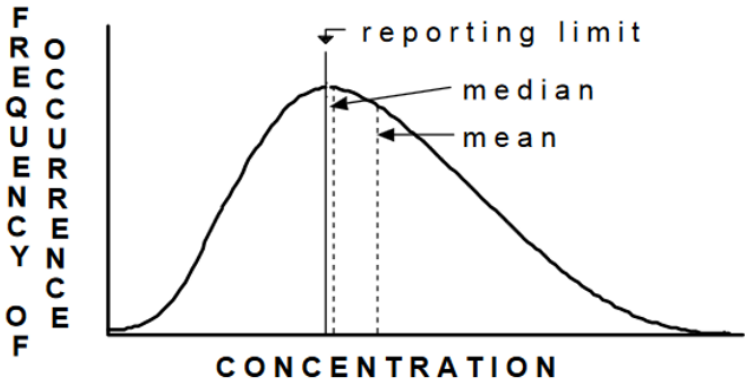


Figure 13.2. Distributional (MLE) method for computing summary statistics.

Advantages

- Able to handle multiple limits of detection.
- Good for estimating summary statistics with a suitably large sample size.
- MLE explicitly accounts for the underlying distribution of the data (if known).

Disadvantages

- More applicable to larger datasets (n > 50).
- Reliant on specifying the correct distribution, otherwise estimates can be incorrect.
- Transforming data to fit a distribution can potentially cause biased estimators.

4.1.2.2 Kaplan-Meier Approach

The Kaplan-Meier approach is a *nonparametric* approach. I.e., it doesn't require a distributional assumption. It's often used in survival analysis for estimating summary statistics for right-censored data. However, it can be applied to left-censored data by 'flipping' the data and subtracting from a fixed constant.

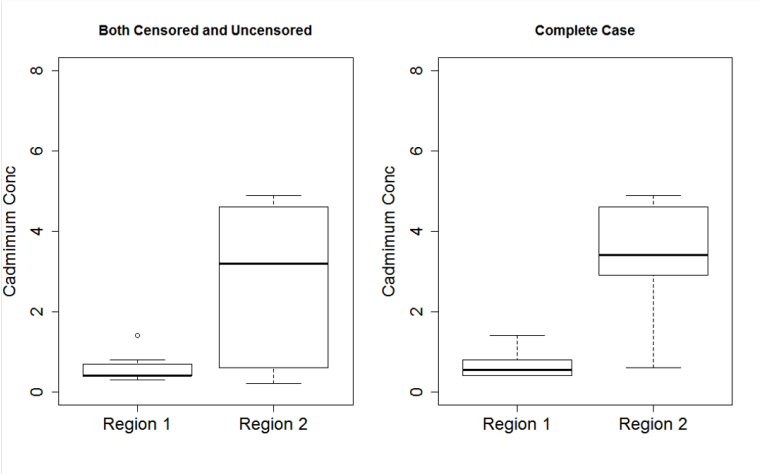
In survival analysis, Kaplan-Meier estimates the probability that an observation will survive past a certain time. In our 'flipped' context, it gives the probability that an observation will fall below the limit of detection.

Example: Cadmium in fish

We can illustrate this using an example of cadmium levels in fish. Cadmium is a heavy metal identified as having potential health risks. We observe cadmium levels in fish livers in two different regions of the Rocky Mountains.
Due to variation in data collection, there are four different LODs (0.2, 0.3, 0.4 and 0.6 μg per litre).

Cd	Region	CdCen
81.3	2	FALSE
3.5	2	FALSE
4.6	2	FALSE
0.6	2	FALSE
2.9	2	FALSE
3	2	FALSE
4.9	2	FALSE
0.6	2	FALSE
3.4	2	FALSE
0.4	1	FALSE
0.8	1	FALSE
0.3	1	TRUE
0.4	1	FALSE
0.4	1	FALSE
0.4	1	TRUE
1.4	1	FALSE
0.6	1	TRUE
0.7	1	FALSE
0.2	2	TRUE

Plotting the data shows the potential impact of censoring. The left panel shows all the data (plotting censored values as equal to the LOD), while the right panel excludes those which have been censored.



We can use the NADA (Nondetects and Data Analysis) package in R. The `cenfit` function applies the Kaplan-Meier method. This package automatically 'flips' the data, since it is designed for environmental data.

```
blinky <- cenfit(obs, censored, groups)
```

```
          n  n.cen  median    mean      sd
groups=1  9      3     0.4   0.589   0.352
groups=2 10      1     3.0  10.540  25.069
```

There are clear differences between the locations in terms of both median and standard deviation.
The `cendiff` function tests for significant differences between the groups. This uses a chi-squared hypothesis test:

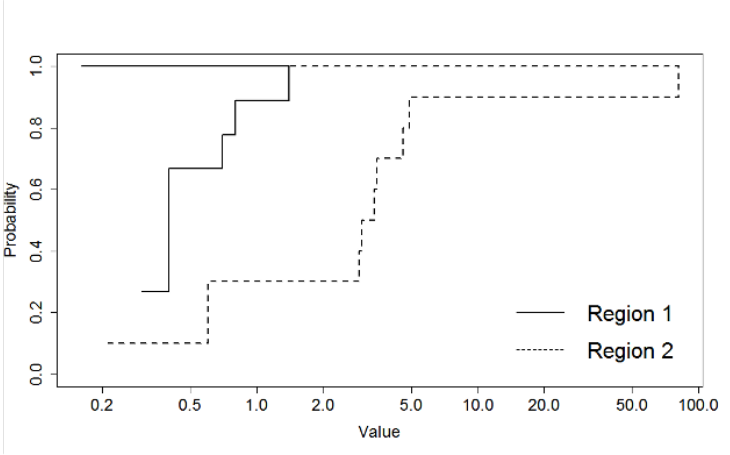
- H0: Median cadmium levels are the same in Region 1 and Region 2
- H1: Median cadmium levels are different in Region 1 and Region 2

```
cendiff(obs, censored, groups)
```

```
               N  Observed Expected (O-E)^2/E (O-E)^2/V
groups=1       9      2.84     6.13      1.76      7.02
groups=2      10      6.84     3.55      3.05      7.02

Chisq=7 on 1 degrees of freedom, p= 0.00808
```

We can also plot the empirical cumulative distribution function (ECDF), taking into account the LODs. Note that this works in the opposite direction from regular survival plots due to the 'flipping' of the data.



Advantages

- Nonparametric, so no need to assume underlying distribution.
- Can easily account for multiple LODs.
- Works for large numbers of censored datapoints ($> 50\%$).

Disadvantages

- Quite simplistic — identical to simple substitution if we only have one LOD.
- Less reliable for values near and below the LOD.
- The mean tends to be overestimated — need to rely on median.

4.1.2.3 Regression on Order Statistics (ROS)

Regression on Order Statistics is a *semi-parametric* approach. I.e., it combines elements of parametric and nonparametric models. It follows a two-step approach:

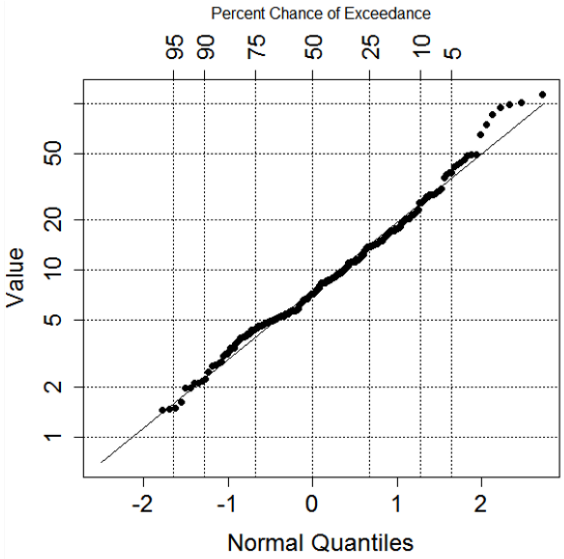
1. Plot the uncensored values on a probability plot (QQ plot) and use linear regression to approximate the parameters of the underlying data distribution.
2. Use this fitted distribution to impute estimates for the censored values.

There is an assumption that the censored measures are normally (or lognormally) distributed.

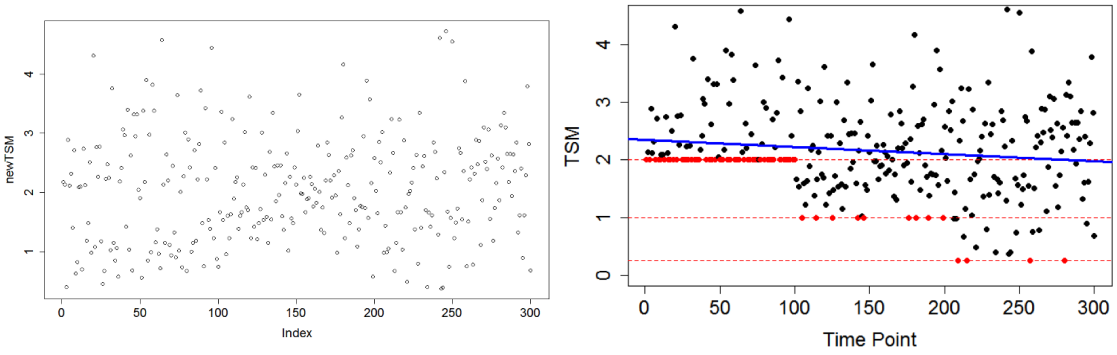
Example: Bathing water quality (continued)

The plot shows the uncensored points and their probability plot regression model. The `NADA` package in R uses lognormal as default. The plot suggests this is sensible. We then use this fitted model to estimate the values of the censored observations based on their

normal quantiles.



We can compare our ROS approach to simple substitution for the bathing water example used earlier. The left panel (ROS) shows no trend present, the right panel (simple) has an artificial trend.



Advantages

- Can be applied to a wide variety of environmental datasets.
- Works with multiple LODs, but still not too simplistic with a single LOD.
- Can be used with up to 80% censored datapoints.

Disadvantages

- Semiparametric approach — requires a distributional model to be assumed.
- Specifically requires normality (or lognormality) for estimation of parameters.
- Two-stage model introduces extra source of variability.

4.2 Outliers

An outlier is an extreme or unusual observation in our dataset. These will often (but not always) have a large influence on the outcomes of our analysis. We have to find ways to identify and deal with outliers.

There are two main categories of outlier: (1) genuine but extreme values, and (2) data errors. If we have a genuinely extreme value, we should try to accommodate these in our analysis. Not doing so would mean that we are ignoring a real feature of our data. There are robust modelling techniques that allow us to incorporate outliers. On the other hand, if we have an outlier due to data error, we can either try to correct it (where possible) or remove it, since this does not reflect a real observation.

4.2.1 Identifying outliers

An outlier is an observation that does not follow the general trend of the rest of the data. It is often helpful to plot your data, since outliers are sometimes very obvious in boxplots or scatterplots or even maps! E.g., Figure 5 shows an Elk's animal track with two unusual observations. To assess whether these are errors from the tracking device or genuine unusual movement patterns (e.g., escaping a predator), we could compute the velocity between points. This is done using the time stamp and the step length (the straight-line distance traveled). We can then assess how likely it is for the animal to have moved at that calculated speed.

Graphical techniques are a commonly used for detecting outliers, but sometimes identifying outliers visually is not straightforward and can become difficult with larger data sets. This has led to the development of a variety of outlier detection techniques. There are also several statistical approaches that can be applied to identify datapoints that are significantly different from the rest. These include (but are not limited to) *tests of discordancy*, *Chauvenet's criterion*, *Grubb's test* and *Dixon's test*.

4.2.1.1 Test of discordancy

This is a hypothesis test, where the null hypothesis is that each datapoint comes from a given data distribution F. We look at the maximum (or minimum) value of our sample, $x_{(n)}$ and test whether it is a reasonable sample from F. If the maximum (or minimum) value could reasonably come from this distribution, we have no outliers. If the maximum (or minimum) value is an outlier, we check the second highest (or lowest) and so on.

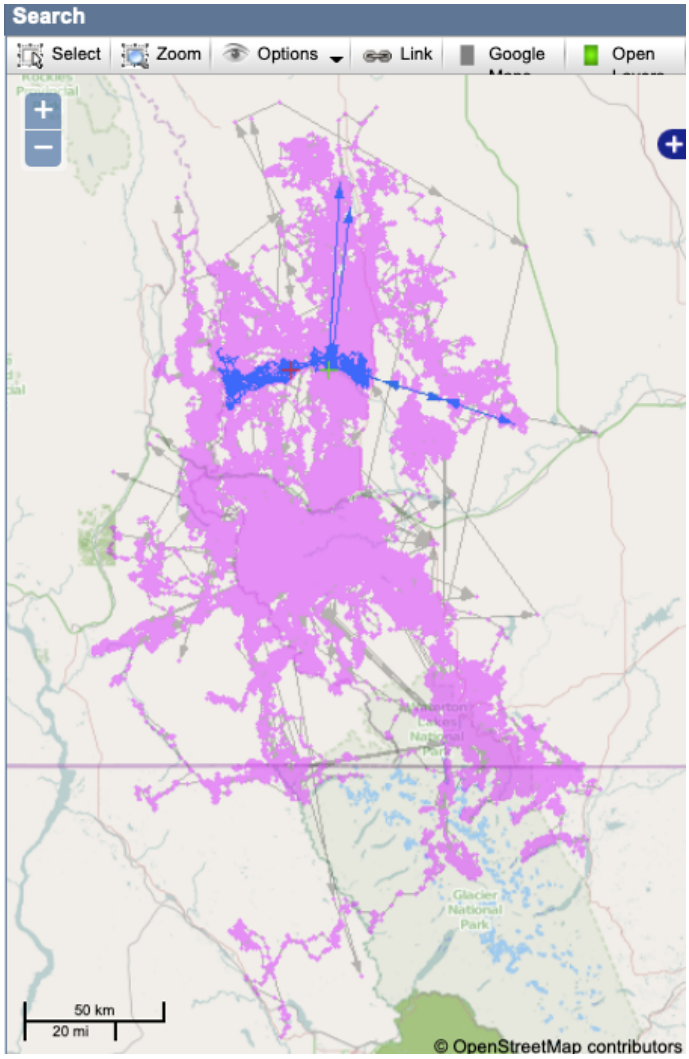


Figure 5: Elk tracking data in southwestern Alberta. The Blue line indicates the tracking for one individual with blue and red crosses showing the start and end point of the track respectively.

4.2.1.2 Chauvenet's criterion

This test assumes that our data are from a Normal distribution. For our dataset of size n, we calculate the mean (μ) and standard deviation (σ). We then use the Normal probability density function to estimate the probability (p) of obtaining a value as extreme our more extreme than or suspected outlier. If $p \times n < 0.5$, then our value is considered to be an outlier.

4.2.1.3 Grubbs' test

Again, this test assumes that our data are from an $\mathrm{N}(\mu, \sigma^2)$ distribution. The Grubbs' test statistic is the largest absolute deviation from this mean in terms of units of the standard deviation:

$$G = \frac{\max\limits_{i=1,\ldots,n} |y_i - \mu|}{\sigma}.$$

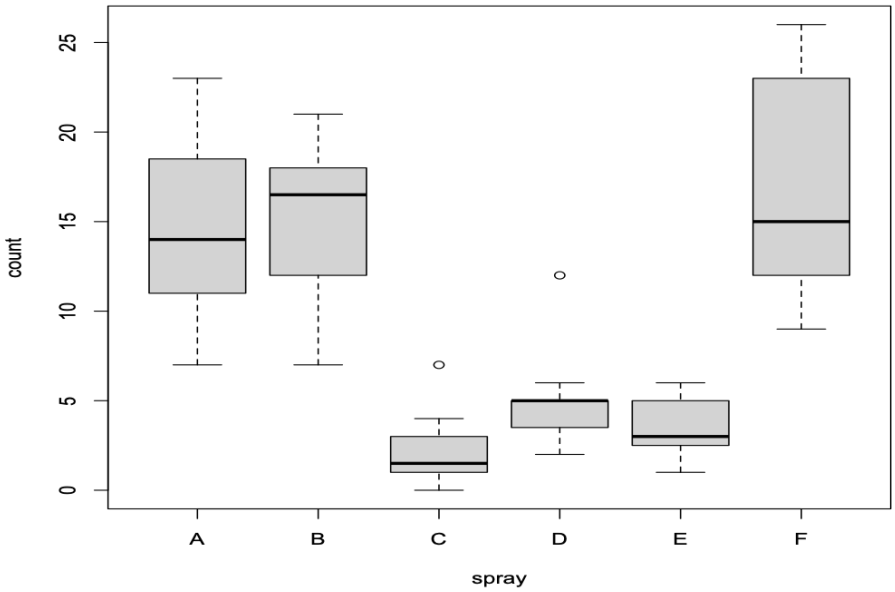
This is compared to the $t(N-2)$ distribution to obtain a p-value. If we identify an outlier, we repeat the process for the next most extreme value.

4.2.1.4 Dixon's test

Rather than looking at the dataset as a whole, Dixon's test compares the outlier to the next most extreme value. Let the *gap* be the distance between our suspected outlier and the closest value, and the *range* be the full range of the dataset. Then we compare $Q = \frac{\text{gap}}{\text{range}}$ to a specifically designed reference table.

This test is only suitable if there is a single distinct suspected outlier. If there were two similar outliers, then the gap would not be large.

> Example: Effectiveness of mosquito sprays
>
> The example dataset below shows the effectiveness of several mosquito sprays.



The `outliers` package in R contains functions for Grubbs' and Dixon's tests. Here we apply the Grubbs' test to the insect spray data.

```
grubbs.test(InsectSprays$count)
```

```
Grubbs test for one outlier

data:  InsectSprays$count
G = 2.29062, U = 0.92506, p-value = 0.719
alternative hypothesis: highest value 26 is an outlier
```

Exercise 10

What conclusion can we draw from the results of the above test?

- (A) There is no statistically significant evidence that the point with the highest value is an outlier
- (B) There is statistically significant evidence that the point with the highest value is an outlier

Solution

Since the p-value for the Grubbs' test is greater than 0.05, we can conclude that there is no statistically significant evidence that the point with the highest value is an outlier.

Example: Effectiveness of mosquito sprays (continued)

We can also apply the Grubbs' test to the data for just one spray type (spray C).

```
grubbs.test(InsectSprays$count[InsectSprays$spray=="C"])
```

```
G = 2.48917, U = 0.38553, p-value = 0.0153
alternative hypothesis: highest value 7 is an outlier
```

Exercise 11

What conclusion can we draw from the results of the above test?

- (A) There is no statistically significant evidence that the point with the highest value is an outlier
- (B) There is statistically significant evidence that the point with the highest value is an outlier

Solution
Since the p-value for the Grubbs' test is less than 0.05, we can conclude that there is statistically significant evidence that the point with the highest value is an outlier.
This is unsurprising from the plot, as since the point is drawn as a circle above the upper whisker of the boxplot, we see that R has identified this point as being more than 1.5 times the interquartile range above the upper quartile. Note, however, that this does not necessarily mean that this point is an outlier, so that a test was still required.

4.2.2 Dealing with outliers

We generally do not want to discard outliers. Sometimes, we can fit a model with and without them to assess their impact on the results. Additionally, we can use robust alternatives to summary statistics, for example median instead of mean, and median absolute deviation (MAD) instead of standard deviation.

The median absolute deviation is defined as:

$$\text{MAD} = \text{median}|y_i - \tilde{y}|$$

where $\tilde{y}$ is the median of our dataset. In other words, we find all the distances between our points and the median, and then take the median of those distances.

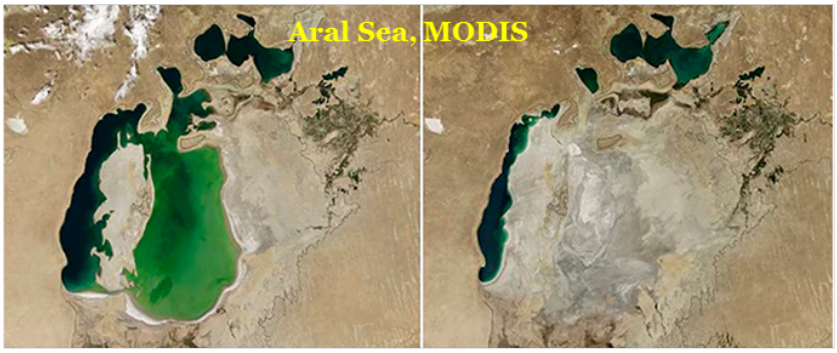
4.3 Missing Data

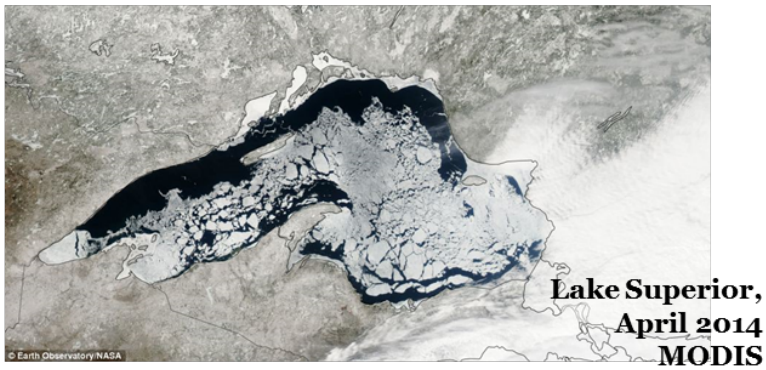
Environmental data are very prone to missing values. There is a whole discipline of statistics related to this, and we will just touch on the topic.

Data can be missing for any number of reasons. Adverse weather (e.g., rainfall, snow, drought or wind) can affect measuring equipment or prevent access to the location. We can have missing data due to the failure of scientific equipment or samples being lost or damaged. Monitoring networks also often change in size over time, with data considered "missing" at a certain site before that site was introduced or after it was removed.

Exercise 12

What causes of missing data do each of the three examples in the images below illustrate? Are these data missing at random?





Images from ESA (top) / NASA (middle/ lower)

- Top image: MERIS data over Ireland.
  Solution
  Cloud cover means that the satellite cannot observe the lakes, oceans or land. These data are not missing at random, since cloud cover is likely to change over the seasons.
- Middle images: MODIS data over the Aral Sea, for two timepoints.
  Solution
  The Aral Sea has decreased in size over the years, due to human impacts. Suppose

that we wish to measure chlorophyll levels in a certain location in the lake. The changing size of the lake means that some locations that had data in previous years will have missing data (since there is no water present at that location) in more recent years. The data are therefore not missing at random.

- Bottom image: MODIS data over Lake Superior.
  Solution
  Ice cover means that the satellite cannot observe the lake water. These data are not missing at random, since ice cover occurs during the coldest times of the year. This may be problematic, if it occurs during peaks or troughs of patterns of the variable that we wish to measure (e.g., chlorophyll may reach its lowest values at the same time that ice cover appears and prevents the satellite from recording measurements).

The technique we use to deal with missing data depends on the type of missingness. If there are a handful of datapoints missing at random, we can essentially ignore this and carry out our analysis as usual. However, if they are missing in some sort of systematic way (e.g., a whole month missing due to bad weather), we may instead look at some form of **imputation**.

4.3.1 Imputation

Imputation is a process which involves predicting the missing values via some form of statistical method. There are two main forms of imputation:

- **Single imputation** involves generating one value in place of each missing value.
- **Multiple imputation** involves generating several values in place of each missing value.

Single imputation has the advantage of being simpler, allowing for a straightforward analysis once the missing values have been estimated. Multiple imputation does a better job of accounting for the uncertainty of the imputation process, but it makes the final analysis more complex.

Our approach for generating the imputed value will vary depending on the context. In the simplest case, we may replace missing values with the overall mean — usually only if we have very limited information. More commonly, we may use neighbouring values, or some form of seasonal mean. These will usually work reasonably well as long as we do not have too many missing data. A more complex approach is to fit a more general statistical model, perhaps taking account of other variables and/or using random components.

To handle missing data, models that incorporate it should be chosen over simply ignoring the missing values. If the missing data can be predicted based on the observed data, imputation models can effectively estimate them.

Botella, Christophe, Alexis Joly, Pascal Monestiez, Pierre Bonnet, and François Munoz. 2020. "Bias in Presence-Only Niche Models Related to Sampling Effort and Species Niches: Lessons for Background Point Selection." Edited by Mirko Di Febbraro. *PLOS ONE* 15 (5): e0232078. https://doi.org/10.1371/journal.pone.0232078.

Pyke, Graham H., and Paul R. Ehrlich. 2010. "Biological Collections and Ecological/Environmental Research: A Review, Some Observations and a Look to the Future." *Biological Reviews* 85 (2): 247–66. https://doi.org/10.1111/j.1469-185x.2009.00098.x.

Rennie, Susannah, Chris Andrews, Sarah Atkinson, Deborah Beaumont, Sue Benham, Vic Bowmaker, Jan Dick, et al. 2020. "The UK Environmental Change Network Datasets – Integrated and Co-Located Data for Long-Term Environmental Research (1993–2015)." *Earth System Science Data* 12 (1): 87–107. https://doi.org/10.5194/essd-12-87-2020.