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.

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. `mcq(c("random", answer="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 **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 **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.

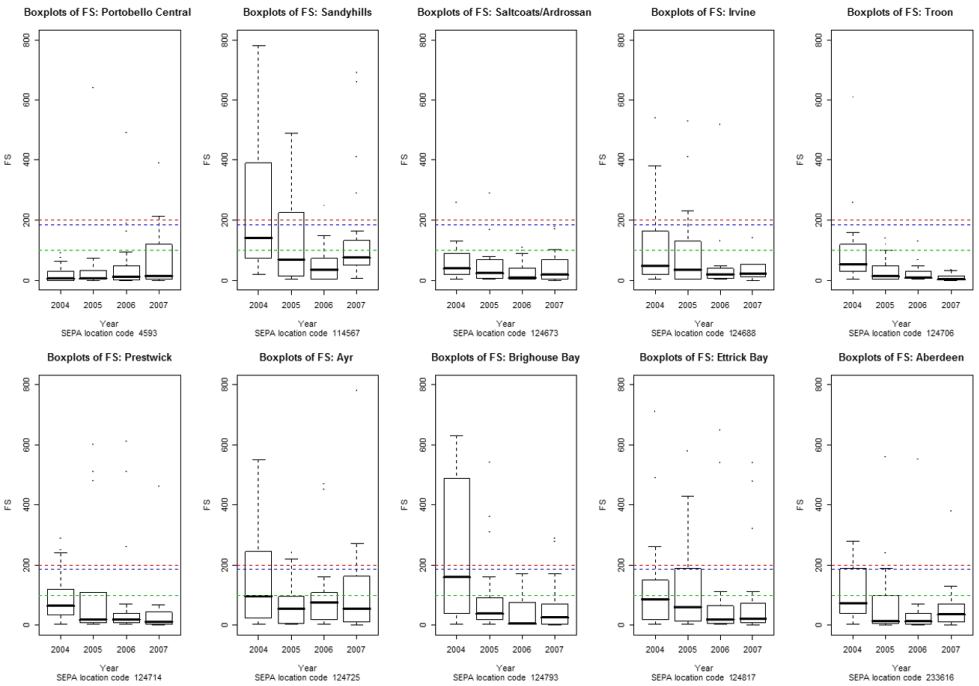
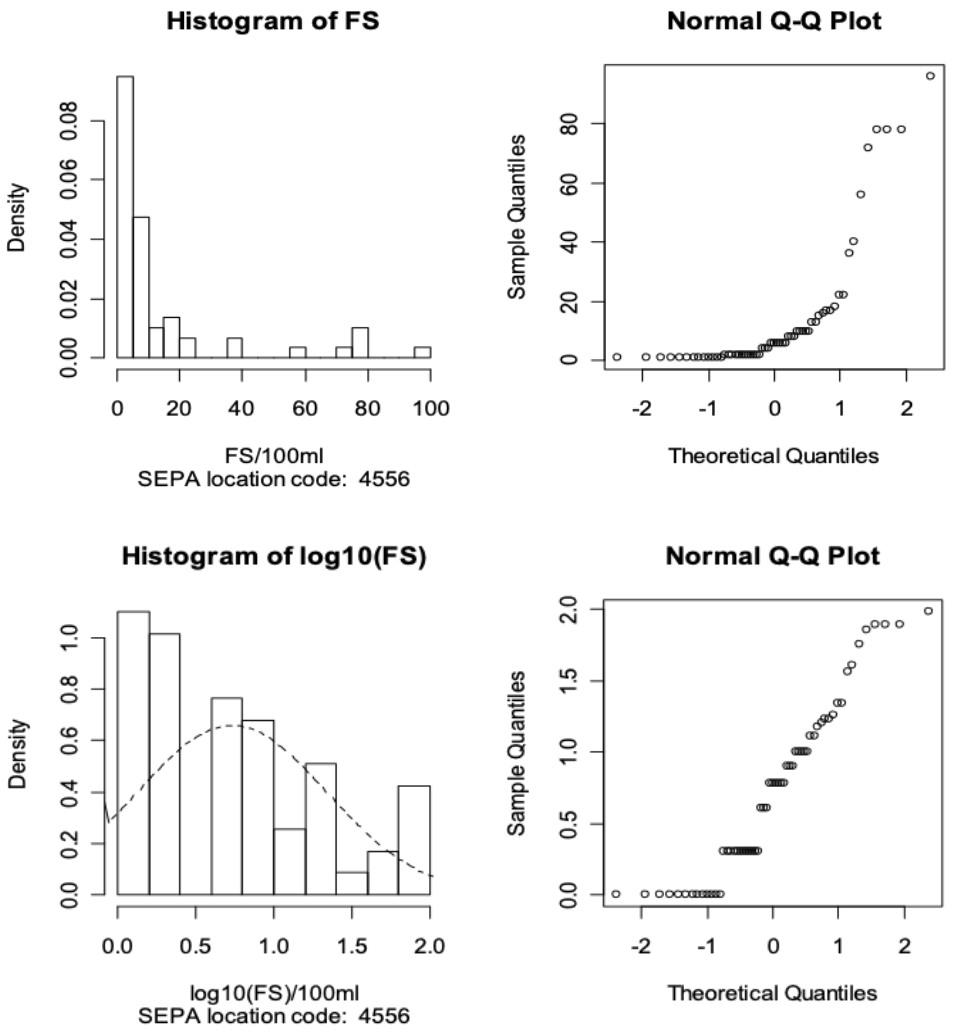


Figure 1: Boxplots of FS by year for 10 sites. The dashed horizontal lines represent "excellent" (green, lowest), "good" (blue, middle) and "sufficient" (red, highest) classification boundaries, respectively.

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:

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

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 $Y = a_0 + a_1X_1 + a_2X_2$ (**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

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

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

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

Exercise 6

The **Shannon index** (or Shannon-Wiener diversity index) is widley 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 $\hat{\theta}_i = \frac{n_i}{N} = p_i$ 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,

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