Tutorial Sheet 1

1 Part A: Censoring and uncertainty calculations.

Task 1

Several methods for dealing with values marked as being at the limit of detection within the paper by Eastoe et al (2006). Read the paper (available below), summarise the different methods compared by the authors and comment on what the conclusions were.

Task 2

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 n_i and total number of individual across all species N.

1. Suppose a fixed number of individuals N are sampled and 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.

Task 3

Waves have a major influence on the marine environment and ultimately on the planet's climate and so are often studied by oceanographers. The period of a particular wave oscillation is measured to be $T = (2 \pm 0.1)s$. What is the uncertainty associated with the frequency, f, of the wave, where $f = 1/T$?

Task 4

10 sets of data (N=13) have been collected. Within each there is a number of (suspected) outliers.

	mean	median	sd	MAD	Nout?
s1	23.82	10.31	33.21	13.98	2
s2	17.72	10.90	24.88	7.64	1
s3	16.58	9.83	25.05	7.52	1
s4	24.13	10.63	33.56	14.13	2
s5	30.68	10.24	39.37	21.07	3
s6	30.82	10.64	39.04	21.08	3
s7	23.79	10.01	33.91	14.47	2
s8	23.95	10.05	34.01	14.39	2
s9	30.96	10.50	39.40	21.33	3
s10	24.31	10.52	33.79	14.32	2

The data for the first sample are shown below:

`9.89, 10.55, 9.67, 10.62, 10.31, 10.21, 10.92, 98.25, 99.03, 9.33, 11.17, 9.78, 9.90`

The R code/output below shows the results of various outlier tests. Comment on this output.

```
# Chauvenet's test for outliers

P <- 1-pnorm(s1[9],mean(s1), sd(s1))
P*length(s1)
[1] 0.152972

# Grubbs test for one outlier

data:  s1
G = 2.2647, U = 0.5370, p-value = 0.06811
alternative hypothesis: highest value 99.03 is an outlier

# Dixon test for outliers

data:  s1
Q = 0.98321, p-value < 2.2e-16
alternative hypothesis: highest value 99.03 is an outlier
```

Task 5

(a) An ecologist wishes to analyse a dataset that contains a variable with around 1% of its data censored at a limit of detection. The ecologist proposes to use a simple substitution method to replace all values at the limit of detection (c_L) with $0.5c_L$. Do you agree with this approach? Why/ why not?

(b) The ecologist wishes to analyse another dataset that contains a variable with around 55% of the data censored at a limit of detection. The ecologist would like to take the same simple substitution approach as in part (a). Do you agree with this approach? What advice would you give?

Task 6

(a) Suppose that chlorophyll data in a freshwater loch are collected twice a week, but the equipment needs to be removed for maintenance once a month, so that there are around 12 missing values per year. Can we assume that these values are missing at random?

(b) Suppose that in another loch, the data are also collected twice a week, but the monitoring device there only needs to be maintained once a year. If this is removed every December, so that there are no values for that month, can we assume that these data are missing at random? Might we need to impute the data?

2 Part B: Sampling and monitoring

> **Note**
>
> This part of the tutorial sheet relates to material in Week 3, which we'll cover in the lectures during the same week as Tutorial 1. You can attempt these questions in Tutorial 1 if you wish, but you may prefer to go through these during Tutorial 2 instead.

Task 7

In the case of a simple random sample $x_1, ..., x_n$ of a random variable, X, assuming the observations are independent,

(a) derive the expected value of X^2
(b) derive the expected value of $\bar{X}^2$
(c) show that the sample variance, s^2, is an unbiased estimator of the population variance, σ^2.

Task 8

Gilbert (1977) reports results of soil sampling at a nuclear weapons test area obtained using stratified random sampling to assess the total amount of Plutonium found in surface soil. Use the information in the table below to:

(a) Estimate the total inventory and derive the estimator for the variance of the totals

(b) Determine the optimal number of population units of measure in each of the 4 strata. Find the total number of units to sample, assuming cost is fixed (where total=£50,000 and cost per unit is £500 for all strata).

strata	Size $\times$ area of the stratum N_l	n_l	Mean for stratum	Variance s_l^2
1	351,000	18	4.1	30.42
2	82,300	12	73	10,800
3	26,200	13	270	127,413
4	11,000	20	260	84,500

Tip

The stratified estimator of the population total is

$$\hat{I} = \sum_{l=1}^{L} N_l \, \bar{y}_l.$$

You may use the fact that the variance for the mean of l-th strata is given by

$$\text{Var}(\bar{y}_l) = \left(1 - \frac{n_l}{N_l}\right) \frac{s_l^2}{n_l}$$

Task 9

Discuss the advantages and disadvantages of the three sampling methods below for mapping a pollutant field:

- Simple random sampling
- Systematic sampling
- Stratified random sampling

Task 10

The Water Framework Directive states:

"Member states must ensure that enough individual water bodies of each water type are monitored and determine how many stations are required to determine the ecological and chemical status of the water body"

Discuss briefly how you would translate this statement into a monitoring programme, given that there are 6 different water body types comprising 10%, 25%, 30%, 20%, 10% and 5% of the total population of 6600 water bodies and that your limited resources only allow you to study a total of 200 water bodies. Knowledge of the within-type variability is not available.

Task 11

Read pages 17–23 of the Analytical Laboratories for the Measurement of Environmental Radioactivity (ALMERA) report on soil sampling (available below). Describe briefly the sampling strategy adopted and also the methods of analysis presented.

Task 12

Read the SEPA survey of business waste (available below) and describe briefly the sampling strategy you would propose. Discuss its advantages and disadvantages.