Tutorial Sheet 1 Solutions

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.
Solution

(a) Replacement analysis; replace values marked at the LOD by 2/3 of that value. Rationale is that true value will be somewhere between zero and the LOD. After replacing the censored values, the distribution of the data are assumed to be log normally distributed and the likelihood function is maximized in order to obtain parameter estimates. Replacement analysis is a standard approach that is commonly used in practice. We will call the model fitted with parameters estimated with this approach Model R.

(b) Survival analysis; In this case a parametric distribution is specified – in this case – again log normal distribution was assumed. The likelihood function is then made up of the product of the density functions for the observed values and the probabilities of getting a censored value. The likelihood is then maximized to estimate the parameters of the distribution - all the data were used within this (both observations marked at the LOD and those without). We will call the model fitted with parameters estimated with this approach Model S.

For each of the approaches a Kaplan Meier estimator was computed in order to explore the survival function of the standardized residuals from models R and S. This was done to assess goodness of fit (and to see if the assumed distribution was normal).
Conclusions: LOD values cannot be ignored! The estimated survival functions showed the models fitted the data well and the distributional assumptions were satisfied – particularly for model S. When the proportion of LOD values was low (<10%) there was little difference between the approaches. With a greater proportion of censored observations survival analysis models were more robust in terms of estimating a trend.

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 $n_i \sim$ Binomial(N, θ_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

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

Let

$$f(p_i) = -p_i \log p_i$$
$$f'(p_i) = -\left(\frac{d}{dp_i}(p_i \log p_i)\right)$$

Applying product rule (i.e., $[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$

$$f'(p_i) = -(\log(p_i) + 1)$$

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

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

Assuming $n_i \sim$ 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,

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

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 $f = 1/T$?

Solution

The estimated frequency is:

$$\hat{f} = \frac{1}{\hat{T}} = \frac{1}{2s} = 0.5s^{-1}$$

The estimated uncertainty $u(f)$ is:

$$u(f) = \sqrt{\left(\frac{\delta f}{\delta T}\right)^2\Bigg|_{T=2} u(T)^2}$$

$$\left(\frac{\delta f}{\delta T}\right) = -T^{-2} = \frac{-1}{T^2}$$

$$\left(\frac{\delta f}{\delta T}\right)^2 = \left(\frac{-1}{T^2}\right)^2 = \left(\frac{1}{T^4}\right)$$

$$u(T) = 0.1s^{-1}$$

$$\therefore u(f) = \sqrt{\left(\frac{1}{2^4}\right)(0.1)^2} = \sqrt{0.000625} = 0.025$$

So, for the frequency of the wave, we have:

$$f = (0.5 \pm 0.025)s^{-1}$$

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
```

Solution

There is a lot of variability in the means of each of the samples than there is in the median. The medians are all close to 10, indicating that this is likely close of the mean of the data if it was not influenced by large/outlier values. We could not confirm this without observing all data, however.

The median absolute difference provides a more robust estimate of the spread of the data if outliers are present. The MAD values are all much lower than the variance estimates provided. The greater the number of suspected outliers, the higher the MAD.

In terms of the sample data, there are two values which are much larger than the others (observations 8 and 9). We would need to go back to how the data has been

collected and recorded before we assess whether or not these are truly 'outliers' but they may be identified as such using the statistical test results presented.

Chauvenet's Criterion

This is assessing if observation 9 can be termed an 'outlier'. As the value of P multiplied by n (=13) is less than 0.5 (0.15) this indicates there is evidence that observation 9 is an outlier by Chauvenet's criterion.

Grubbs Test

Here we are testing the null hypothesis that the maximum value is not an outlier against the alternative that it is an outlier. As the p value is 0.06 it is just above the significance level of 0.05 (5%) and so the maximum is not deemed to be an outlier using this test. The output for the test with the second highest value is presented below – in this case the highest value is deemed to be an outlier. This indicates that Grubbs test may be better suited to cases where there is only a single suspected outlier.

```
grubbs.test(s1[-8])
```

```
  Grubbs test for one outlier

data:  s1[-8]
G = 3.17470000, U = 0.00043801, p-value < 2.2e-16
alternative hypothesis: highest value 99.03 is an outlier
```

Dixon's Criterion

Again we are testing the null hypothesis that the maximum value is not an outlier against the alternative that it is an outlier. As the p value is <0.05 it is much less than the significance level of 0.05 (5%). Hence we can reject the null hypothesis and accept that the maximum value (observation 9) is an outlier using this test.

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?

Solution

(a) Since we have a very small proportion of data at the limit of detection, using a simple substitution approach here seems to be appropriate.

(b) With more than 50% of the data points being censored at a limit of detection, a simple substitution approach will not be appropriate. Taking one of the distribution-based approaches would be more appropriate.

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?

Solution

(a) Since the monitoring device is removed every month, there is unlikely to be a seasonal pattern that we are missing here, so it would be possible to assume that the data are missing at random.

(b) Since the monitoring device is removed at the same time every year (and we miss a whole month of data), it would not be appropriate to assume that the data are missing at random. It is likely that we are missing a peak or trough of a seasonal pattern/ cycle in the data. We should consider an imputation approach. (We do not have enough information in the question to come up with an approach here, but we would have to consider the seasonal patterns (e.g. through exploring the data via plots) and consider whether we could use other variables to help with imputing these values (e.g. by looking at relationships between the variable with missing values and other relates variables with non-missing values).)

2 Part B: Sampling and monitoring

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.

Solution

(a)

$$Var(X_i) = E(X_i^2) - \{E(X_i)\}^2$$
$$E(X_i^2) = \mathsf{Var}(X_i) + \{E(X_i)\}^2$$
$$E(X_i^2) = \sigma^2 + \mu^2$$

(b)

$$E(\bar{X}^2) = \mathsf{Var}(\bar{X}) + \{E(\bar{X})\}^2$$

$$= \text{Var}\left(\frac{\sum X_i}{n}\right) + \left\{E\left(\frac{\sum X_i}{n}\right)\right\}^2$$

$$= \frac{1}{n^2}\text{Var}\left(\sum X_i\right) + \frac{1}{n^2}\left\{E\left(\sum X_i\right)\right\}^2$$

$$= \frac{1}{n^2}(n\sigma^2 + n^2\mu^2)$$

$$= \frac{\sigma^2}{n} + \mu^2$$

(c) Show that $E(s^2) = \sigma^2$.

$$E(s^2) = E\left(\frac{1}{n-1}\sum(X_i - \bar{X})^2\right)$$

$$= \frac{1}{n-1}E\left(\sum(X_i - \bar{X})^2\right)$$

$$= \frac{1}{n-1}E\left(\sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2\right)$$

$$= \frac{1}{n-1}E\left(\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right)$$

Remember that $2\bar{x}\sum x_i = 2\bar{x}\cdot n\bar{x} = 2n\bar{x}$

$$= \frac{1}{n-1}\left(E\left(\sum X_i^2\right) - E\left(n\bar{X}^2\right)\right)$$

$$= \frac{1}{n-1}\left(\sum E\left(X_i^2\right) - nE\left(\bar{X}^2\right)\right)$$

$$= \frac{1}{n-1}\left(n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right)$$

$$= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

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

Solution

The total inventory in all strata is (the original units are in microcuries (μCi))

$$\hat{I} = \sum_{l=1}^{4} N_l \hat{\mu}_l = \sum_{l=1}^{4} N_l \bar{y}_l$$

(a)

$$\begin{aligned}\hat{I} &= (351000 \times 4.1) + (82300 \times 73) + (26200 \times 270) + (11000 \times 260) \\ &= 17381000 \mu\text{Ci} \approx 17.4\text{Ci}\end{aligned}$$

An estimator for the variance of the totals is given by:

$$\begin{aligned}\text{Var}(\hat{I}) &= \sum_{l} \frac{1}{N_l^2} \text{Var}(\bar{y}_l) \\ &= \sum_{l=1}^{L} N_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_l^2}{n_l} \\ &= (351000)^2 \left(1 - \frac{18}{351000}\right) \left(\frac{30.42}{18}\right) + \\ &(82300)^2 \left(1 - \frac{12}{82300}\right) \left(\frac{10800}{12}\right) + \\ &(26200)^2 \left(1 - \frac{13}{26200}\right) \left(\frac{127413}{13}\right) + \\ &(11000)^2 \left(1 - \frac{20}{11000}\right) \left(\frac{84500}{20}\right) \\ &= (1.354 \times 10^{13})\end{aligned}$$

$$s(\hat{I}) = \sqrt{1.354 \times 10^{13} = 3679674 \mu\text{Ci} \approx 3.7\text{Ci}}$$

(b) Total size of population (N) = $351k + 82.3k + 26.2k + 11k = 470.5k$

What sample size can we afford?

$£50,000/£500 = 100 = n$

How do we split the 100 samples across each of the strata?

As the costs are the same for each stratum then we can use Neyman Allocation.

$$n_l = n \frac{W_l \sigma_l}{\sum_{i=1}^{L} W_l \sigma_l}$$

Alternatively, we could use proportional allocation

Proportion of total population $\frac{N_l}{N} = W_l$	$W_l s_l$	Proportional allocation nW_l	Neyman allocation $n_l = n\frac{W_l s_l}{\sum_{l=1}^{L} W_l s_l}$
$351/470.5 = 0.746$	$0.746 \times \sqrt{30.42}$	74	$\frac{100 \times 0.746 \times \sqrt{30.42}}{48.98} = 8$
$82.3/470.5 = 0.175$	$0.175 \times \sqrt{10800}$	18	37

$26.2/470.5 = 0.056$	$0.056 \times \sqrt{127413}$	6	41
$11/470.5 = 0.023$	$0.023 \times \sqrt{84500}$	2	14
	Total: 48.98		

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

Solution

Advantages and disadvantages of sampling schemes:

Simple random sampling and stratified sampling should provide a representative sample, thus ensuring that we can quantify the sampling variability and hence precision. Therefore, estimation should be unbiased. The stratified random sample will be more efficient (have lower variance) than simple random sampling and so may be preferred. However, stratified random sampling requires us to know about the different strata and the relative proportions in the population.

Systematic is much more practical for field work, although it is not often random except for identifying a random starting point. The other potential drawback of systematic sampling is that we may miss hidden periodicities in the data. The analysis is more complex than the other two methods.

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.

Solution

In this case it may make sense to adopt a stratified random sampling scheme. We have the strata identified as well as their relative proportions within the population. We have been told that we can 'afford' to have a sample size of $n = 200$. As there is no information regarding within-type variability available we cannot use Neyman allocation. The simplest approach is to use proportional allocation where the allocation of samples to the stratum is pro-rata to the size of the stratum within the population. So, for population discussed we have:

- Population size (N) 6600
- Sample size $n = 200$

W_l	$N_l = W_l N$	Proportional allocation nW_l
0.10	$0.10 \times 6600 = 660$	$200 \times 0.10 = 20$
0.25	1650	50
0.30	1980	60
0.20	1320	40
0.10	660	20
0.05	330	10

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.

Solution

ALMERA Sampling strategy

The aim was to compare the protocols of different institutes by having them each determine the mean value of several radionuclides in an agricultural area of about 10000 square meters.

The area selected was a reference site – meaning it was an area where the spatial and temporal variability of one or more of the element concentrations of interest are well understood.

Soil sampling over a site was carried out by dividing the site into 100 sub-areas (cells), each measuring 10m $\times$ 10m. From each sub areas a systematic sample was collected, with samples 2m apart (25 per sub area). The 25 samples per sub-area were pooled to give 100 composite samples. These composite samples were used to look at long range variability.

For two sub areas (cells) each of the 25 samples/cell were analysed separately in order to verify within cell variability. There is no explanation given as to how these two sub areas were selected.

The sampling approach used is fundamentally systematic. The sample preparation and analytical activity were performed by a single laboratory in order to rule out the variabilities due to different analytical laboratories.

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.

Solution

SEPA Business waste survey

The key thing here is to recognise that the population of businesses in Scotland is divided into strata by both the type of business and also by the size of the business. Therefore, a natural approach may be to apply a stratified sampling plan.

The advantage of this approach is that the estimates we get would be more precise than if a simple random sampling approach was used and hence this is a more efficient approach.

However, there are potential disadvantages. One of which is to use a proportional sampling scheme we need to have knowledge about the relative sizes of the different strata in the population. An additional potential disadvantage is that there may be some strata where there are very small numbers of firms, and so there are possible issues surrounding information disclosure or identifiability and the entire population of the strata is sampled.