Environmental Statistics

Chapter 2: Modelling variability and handling uncertainty

Session 2020/2021

University of Glasgow

VIA VERITAS VITA

What we will cover

- Distributions (revision!)
- Uncertainties
- Dealing with censored observations
- **Outlier detection**
- **Missing data**

Dealing with Outliers

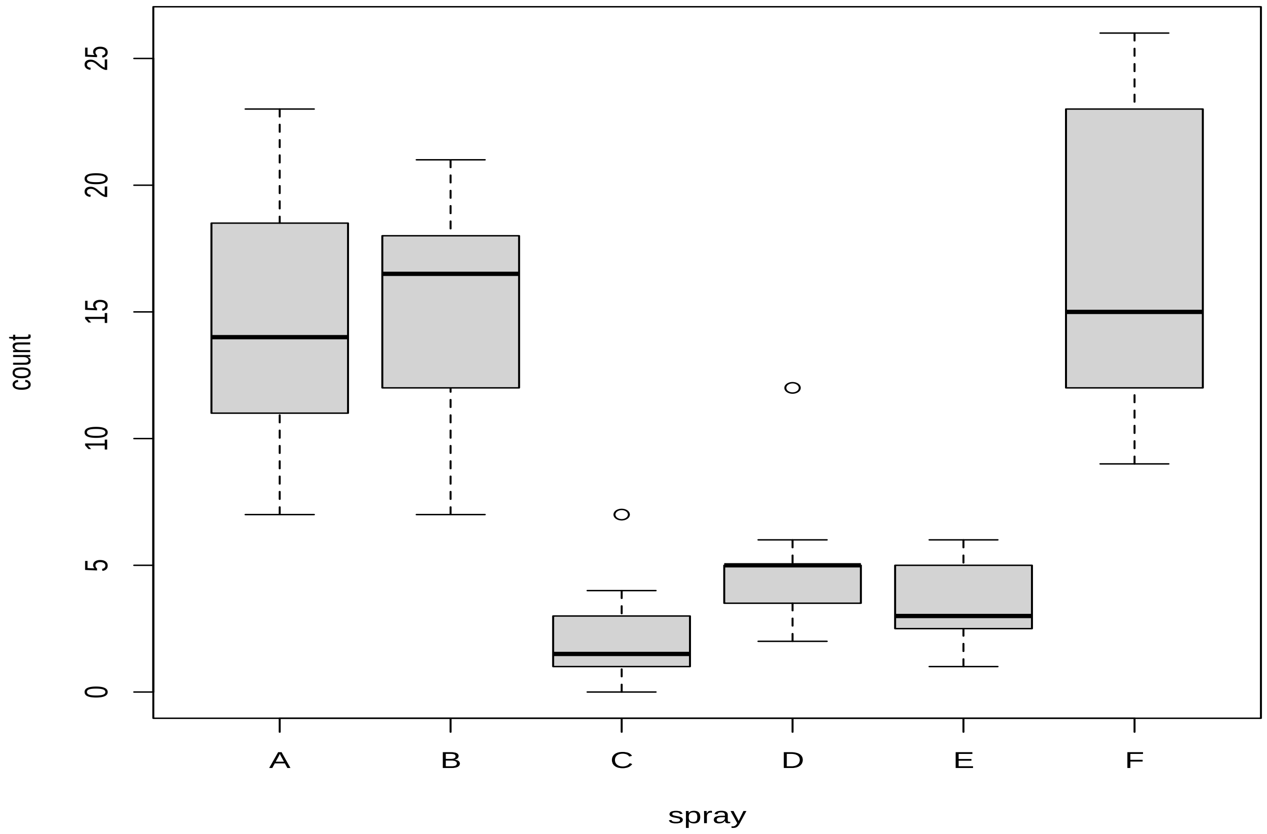
Recall- last lecture?

- Dealing with censored observations
- What are censored observations?

- Which approaches may be used to deal with these?
  - remove
  - replace by a single value
  - ML approach
  - Kaplan Meier approach
  - regression of order statistics

What is an outlier?

- How many outliers do we have?



What is an outlier?

- An extreme observation which may or may not be having an undue influence on our analysis

- What do we do?
  - ▫ Reject them
  - ▫ Identify them for special consideration
  - ▫ Accommodate them using robust statistical modelling techniques

- …but outliers are surely just extreme values?

Test of Discordancy

- Null Hypothesis

  H_0: Sample arises from distribution F

- Find the maximum of the sample, $x_{(n)}$
- Is this a reasonable sample from F?
- Suppose $X_{(n)}$ has distribution $F_n\,(x) = \{F(X)\}^{\,n}$
- If $F_n(x)$ is known we can work out the probability of $F_n\left(x_{(n)}\right)$

Test of Discordancy
Exponential Distribution

- Null Hypothesis

$$H_0\text{: Sample arises from distribution F}$$

- Maximum of the sample, $x_{(n)}$
- Is this a reasonable sample from F?

- Suppose $X_{(n)}$ has distribution

$$F_n\,(x) = \{F(X)\}^{\,n} = (1 - e^{-\lambda x})^n$$

What is the probability that $X_{(n)}$ is at least as great as $x_{(n)}$?

Other common tests for outliers

Chauvenet’s criterion

(Assumes data are from a normal distribution)

- Calculate mean and sd of observed data (including outliers)
- Use normal pdf:
  - Estimate the probability of a value as (or more) extreme as (than) the suspected outlier take from a normal population with this mean and sd
  - Call this P
- Multiply P by the number of observations: Pxn

- **Criterion:**
  If Pxn < 0.5 we have an outlier!

Other common tests for outliers

Grubb’s test

(Assumes data are from a normal distribution)

- Looks for outliers one (or two) at a time - downside ….

- Can be used to check if either min or max is an outlier

- Hypothesis test:
  Test statistic
  - reduced sample (obtained by omitting the possible outlier)
  - ratio of the maximum deviation from the mean and the sample standard deviation

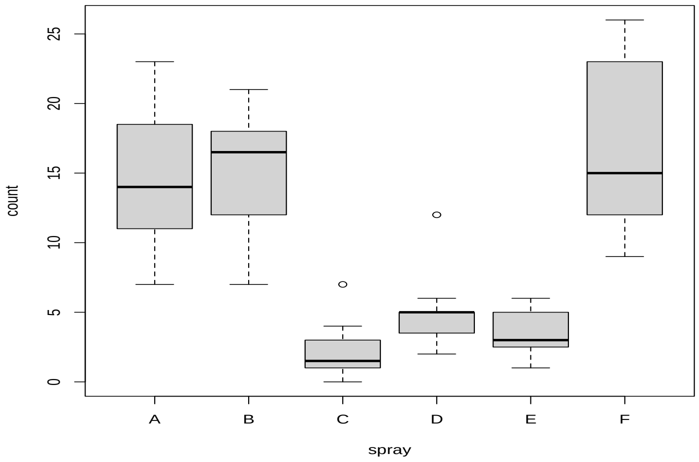
Other common tests for outliers

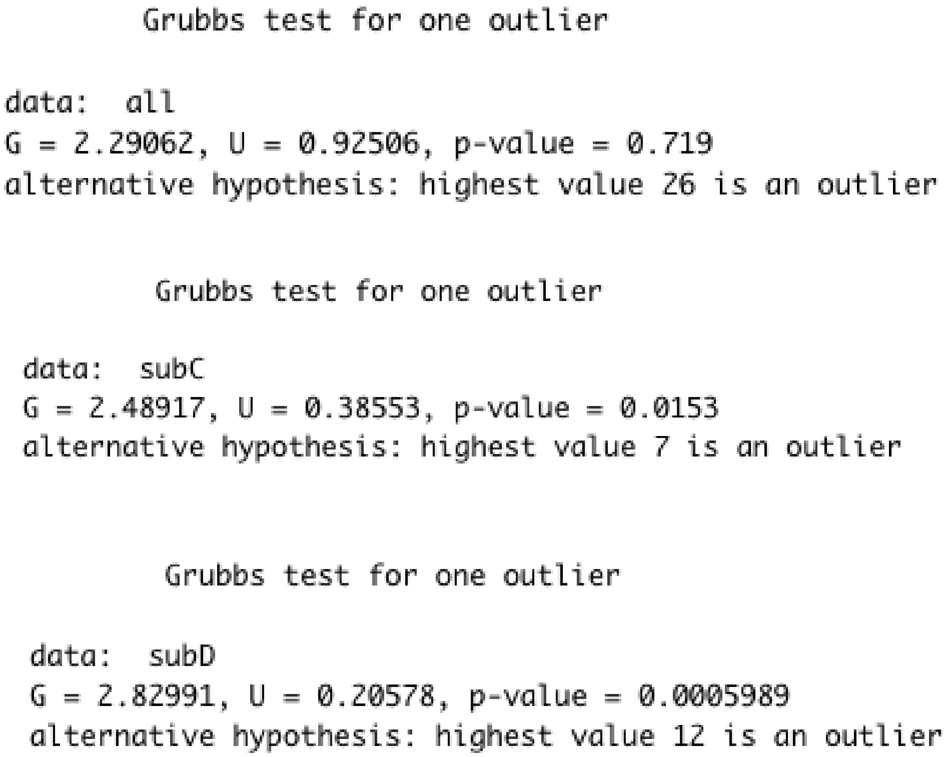
- Dixon’s test
  (Assumes data are from a normal distribution)

- Hypothesis test:
  Test statistic:
- ratio of the distance between a suspected observation and its nearest or next-nearest (assumed unaffected) neighbour to the range of the sample (with zero, one, or two observations omitted)

Let’s see…

- Grubb’s test for insect spray data




```
        Grubbs test for one outlier

data:  all
G = 2.29062, U = 0.92506, p-value = 0.719
alternative hypothesis: highest value 26 is an outlier

        Grubbs test for one outlier

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

        Grubbs test for one outlier

 data:  subD
 G = 2.82991, U = 0.20578, p-value = 0.0005989
 alternative hypothesis: highest value 12 is an outlier
```

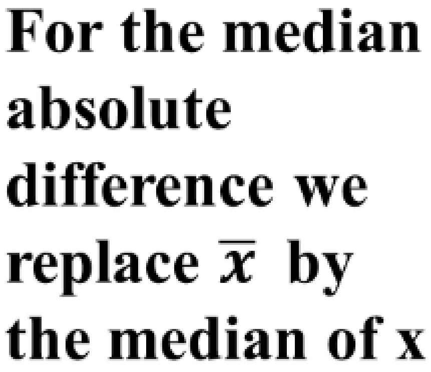
University of Glasgow

Robust Statistics

- Don’t need to identify or discard outliers

- Use robust alternatives for the simple statistics such as mean and standard deviation:
  - the median (a robust statistic)
  - median absolute difference (MAD)

- $MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}_{\text{median}}|$



For the median absolute difference we replace $\bar{x}$ by the median of x

Dealing with Missing Data

Missing Data

- Environmental data are very prone to missing values
- Dealing with missing values is a whole area of statistics in itself – we'll just touch on this
- Lots of reasons why data might be missing

Lake Superior,
April 2014
MODIS

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

Causes of Missing Data

- If the underlying ecological or environmental system is complex the reason for every missing observation often cannot be determined.

- **Examples:**
  - ▫ Adverse weather events:
    - sites are inaccessible which prevents samples being collected
    - particularly in the winter months.
  - ▫ Failure of scientific equipment used to analyse samples
  - ▫ Samples becoming lost or damaged in transit.
  - ▫ Monitoring networks change in size throughout time
  - ▫ Additional stations entering a network can cause problems due to differences in the quantity of data available at different locations.

Dealing with Missing Data

- In situations where e.g. we have a time series of monthly observations (e.g. water temperature in Loch Lomond), we see that frequently the month of January is missing.
- This may have an impact on our analysis (e.g. what is the trend in temperature?)
- **Imputation methods** are used in this context

Data Imputation Methods

2 broad methods

- **Single imputation**; one value generated in place of each missing value
- **Multiple imputation**, several values are generated for each missing value (aims to reflect uncertainty associated with the missing values)

- Advantage of single imputation methods:
  - only generate one value per missing observation
  - complete data analysis can be applied directly after the missing data values are in place
- More complex for multiple imputation

Data Imputation Methods

Examples of single imputation:

- Replace the missing data by the series mean
- Use neighbouring values
- Replace the missing data by the seasonal mean

These usually work well in practise provided level of missing is not too great

Alternatively…

- Fit a more general model (including a random component)

What we have learned …

- Outliers and missing data can cause problems
- Again, as with LOD values we cannot simply ignore or remove these values
- There are often straightforward methods which can be applied in order to assess outliers and impute missing values
- Note: There are lots of other methods available for dealing with these problems.

References

- Statistical evaluation of measurement errors. G Dunn, Arnold
- An Introduction to Uncertainty in Measurement. L Kirkup, B Frenkel, Cambridge University Press.
- Non-detects and data analysis. D R Helsel, Wiley

Accompanying Materials
THINGS TO READ

- LOD: Papers by Eastoe **(tutorial question 1)** and Lee and Helsel on limit of detection
- RSC Statistics notes (on precision and accuracy and error propagation)
- Extract from IPCC reports on reporting on uncertainty
- Look at the **additional example slides** on Moodle (Chapter 2 additional examples)

Tutorial Sheet 1 (Part A)

- This will be available on Moodle after the lecture today.
- Some of the questions have the solutions provided, some of the questions are left for you to try and discuss the solutions in the first tutorials (week of 8th Feb).

Past Exam Questions

1. “Historically, if the data set contained **non-detect** results, the **substitution method** was used to replace non-detect results with a **set value**, **typically one-half the LoD**. Currently, the best practice is to use statistical methods to handle the non-detect results such as **Regression on Order Statistics** methods for known distributions and the **Kaplan-Meier method** for non-parametric data sets.”

Briefly describe three widely used non–statistical techniques for dealing with such LoD observations in the environmental science literature and contrast them with the two statistical approaches, highlighted above. Comment critically on the different approaches. 5 Marks

Past Exam Questions Solutions

1. In environmental sciences, ignore the < value, and simply take the numerical values, else take the "<value" and replace it by a fixed constant (eg $0.5C_L$), or treat the value as zero. Another approach is to replace the "<value" from a probability distribution (uniform over 0, C_L)
Statistical approaches include the Kaplan Meier approach, based on non-parametric estimate of the distribution function or the robust regression on order statistics- where a distribution will be assumed. Note that KM approach common in survival with right censored- in the environmental context data are left censored (so we change the signs of the values).
These methods are based on replacing non-detect results with values generated to match the distribution of the rest of the data set

Past Exam Questions

2. A frequent problem in environmental data concerns the presence of outliers. Describe some formal hypothesis tests for the identification of outliers. Devise a test of discordancy for the maximum value in a sample, where the population density is known to be Exponential (λ) and where data $x1,\ldots xn$ assumed independent and identically distributed.

Past Exam Questions Solutions

2. outlier detection by various means- graphical such as in boxplot- observation is flagged if a certain distance (usually 1.5* IQR) from median, also more formal tests- Grubbs, Dixon or Chauvenet criterion- all similarly based on whether an observation is greater than a given distance from the mean.

Having detected them options are a) ignore/remove them from the analysis b) consider whether a transformation might help or c) undertake some form of robust analysis.

Past Exam Questions Solutions

2. cont….

null hypothesis: H_0: sample arises from distribution F
find $x_{(n)-}$ the maximum. Is this a reasonable value from F?
suppose $X_{(n)}$ has distribution function $F_n(x)$
$F_n(x) =\{F(x)\}^n$
so if F(x) is known we can work out the probability of $F_n(x_{(n)})$.
If X~Ex (λ) then $X_{(n)}$ has distribution function $F_n(x)$
$F_n(x) =\{F(x)\}^{n\,=}\,(1-e^{-\lambda x})^n$
Then we can work out the probability that $X_{(n)}$ is at least as great as $x_{(n)}$ to allow us to identify whether we believe the observation to be an outlier.