EcoHealth Alliance

Assessing and ensuring data quality of measurements

Ernest Guevarra

12 October 2021

Outline

- Introduction
- Common data quality issues related to measurements
- Assessing/detecting quality issues with measurement data
- Minimising quality issues with measurement data

EcoHealth Alliance

Introduction

- Collecting measurement data is an important component of most EHA studies
- Ensuring quality of measurement data is critical
- There are many factors that impact on quality of measurement data
- Detecting and/or minimising possible issues with measurement data can help with improving data quality

Common reasons/factors that give rise to issues with measurement data

- Errors in taking measurements (technique/method issues)
- Errors in reading measurements (e.g., ruler-based measurements)
- Errors in recording measurements and/or errors in encoding measurements

Outlier detection

- Compare against a range of possible/plausible measurement values
- Extending the *boxplot* approach (univariate approach)
- Using scatterplot and statistical distance (bivariate approach)

EcoHealth Alliance

Outlier detection examples/demonstration

- For this, we will be using a dataset on bat forearm length available from:

> Penone, C., Kerbiriou, C., Julien, J. F., Marmet, J., & Le Viol, I. (2018). Body size information in large-scale acoustic bat databases. PeerJ, 6, e5370. https://doi.org/10.7717/peerj.5370

- We retrieve the data as follows:

```
exdata1 <- read.csv("data/peerj-06-5370-s003.csv", sep = ";") %>%
  tibble() %>%
  select(Locality_id, Year, Sex, Age, Forearm_length_mm)
```

EcoHealth Alliance

Outlier detection - example dataset

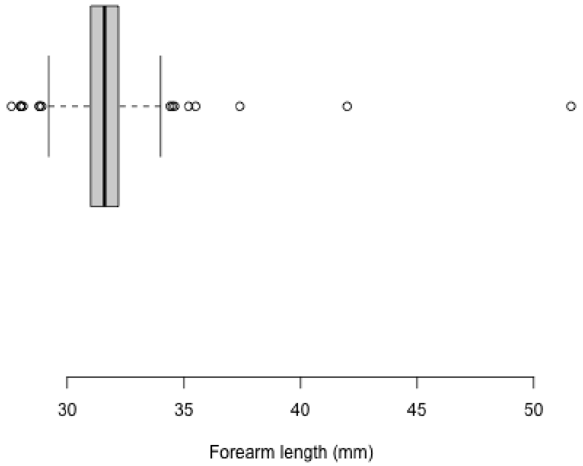
- And the dataset looks as follows:

```
exdata1
```

```
## # A tibble: 1,361 × 5
##    Locality_id               Year Sex   Age   Forearm_length_mm
##    <chr>                    <int> <chr> <chr>             <dbl>
##  1 44_MACHECOUL              2011 M     <NA>               32.1
##  2 44_MACHECOUL              2011 M     <NA>               31  
##  3 44_MACHECOUL              2011 M     <NA>               30.6
##  4 44_MACHECOUL              2011 M     <NA>               32  
##  5 22_SAINT-JACUT-DE-LA-MER  2009 F     <NA>               32.5
##  6 22_SAINT-JACUT-DE-LA-MER  2009 F     <NA>               31.6
##  7 22_SAINT-JACUT-DE-LA-MER  2009 F     <NA>               32.1
##  8 22_SAINT-JACUT-DE-LA-MER  2009 F     <NA>               32.3
##  9 22_SAINT-JACUT-DE-LA-MER  2009 F     <NA>               32.9
## 10 56_SENE                   2009 M     <NA>               30.7
## # … with 1,351 more rows
```

Outlier detection - checking ranges using boxplot()

```
boxplot(
  exdata1$Forearm_length_mm,
  horizontal = TRUE,
  xlab = "Forearm length (mm)",
  frame.plot = FALSE
)
```



Outlier detection - extending the `boxplot()` approach

- We can extend the `boxplot()` approach to checking range using some functions from the `{nipnTK}` package (see https://nutriverse.io/nipnTK)

```
install.packages("nipnTK")
```

- `{nipnTK}` has a function called `outliersUV()` which builds on the techniques used when creating **boxplots**. The function can be used as follows:

```
outliersUV(x = exdata1$Forearm_length_mm)
```

EcoHealth Alliance

Outlier detection - extending the `boxplot()` approach

And produces the following output

```
outliersUV(x = exdata1$Forearm_length_mm)
```

```
##
## Univariate outliers : Lower fence = 29.2, Upper fence = 34

##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [15] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [29] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [43] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [57] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [71] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##   [99] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [113] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALS
##  [127] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [141] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [183] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [197] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALS
##  [211] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRU
```

Outlier detection - extending the `boxplot()` approach

This output can be used to identify the outlier rows of data as follows:

```
exdata1[outliersUV(x = exdata1$Forearm_length_mm), ]
```

```
##
## Univariate outliers : Lower fence = 29.2, Upper fence = 34

## # A tibble: 16 × 5
##    Locality_id                  Year Sex   Age   Forearm_length_mm
##    <chr>                       <int> <chr> <chr>             <dbl>
##  1 44_PORT-SAINT-PERE           2011 M     <NA>               28.8
##  2 44_BONNOEUVRE                2011 F     <NA>               37.4
##  3 35_LIEURON                   2012 M     <NA>               28.9
##  4 35_LIEURON                   2012 M     <NA>               51.6
##  5 44_SAINT-SULPICE-DES-LANDES  2011 F     <NA>               28
##  6 35_MONTAUBAN                 2012 M     <NA>               28
##  7 22_PLUMAUGAT                 2012 F     <NA>               35.2
##  8 35_MARTIGNE-FERCHAUD         2010 M     <NA>               28.8
##  9 72_BEAUMONT-PIED-DE-BOEUF    2011 M     Adult              28
## 10 22_SAINT-LAUNEUC             2009 F     <NA>               34.6
## 11 78_GAMBAISEUIL               2010 F     Adult              42
## 12 22_LOC-ENVEL                 2009 F     Adult              35.5
## 13 22_LOC-ENVEL                 2010 F     <NA>               34.5
```

Outlier detection - extending the `boxplot()` approach

The `outliersUV()` function also allows for adjustments to the "fence" used in the `boxplot()` function to either make the detection of outliers more strict (narrower fence, less than 1.5 times the IQR) or more lax (wider fence more than 1.5 times the IQR). This can be done as follows:

```
exdata1[outliersUV(x = exdata1$Forearm_length_mm, fence = 3), ]
```

```
##
## Univariate outliers : Lower fence = 27.4, Upper fence = 35.8

## # A tibble: 3 × 5
##   Locality_id     Year Sex   Age   Forearm_length_mm
##   <chr>          <int> <chr> <chr>             <dbl>
## 1 44_BONNOEUVRE   2011 F     <NA>               37.4
## 2 35_LIEURON      2012 M     <NA>               51.6
## 3 78_GAMBAISEUIL  2010 F     Adult              42
```

Outlier detection - scatterplot and statistical distance

- For this, we will be using bat morphological dataset taken from:

> Zakaria, N., Tarmizi, A. A., Zuki, M., Ahmad, A. B., Mamat, M. A., & Abdullah, M. T. (2020). Bats data from fragmented forests in Terengganu State, Malaysia. Data in brief, 30, 105567. https://doi.org/10.1016/j.dib.2020.105567

- We retrieve the data as follows:

```
exdata2 <- read.csv("data/bats_malaysia.csv") %>%
  tibble()
```

Outlier detection - scatterplot and statistical distance

- And the dataset looks as follows:

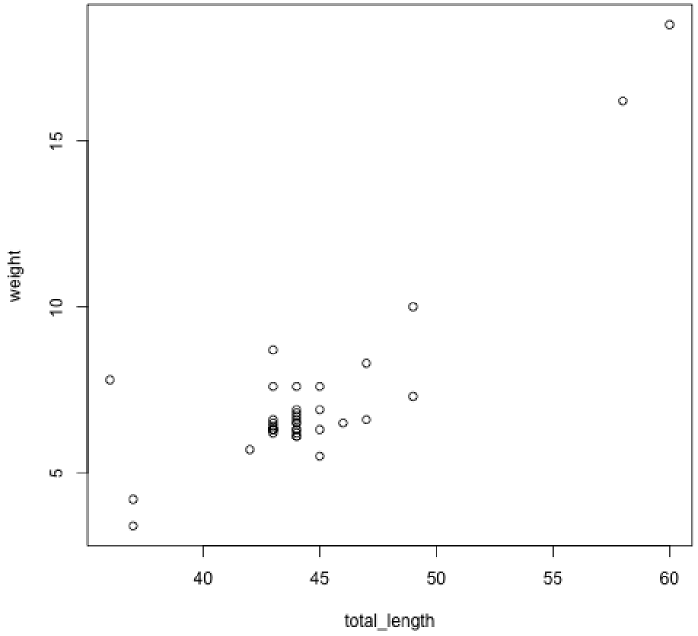
```r
exdata2
```

```
## # A tibble: 78 × 13
##       id site           date   species    trap  sex   stage total_length ear_length tibia_
##    <int> <chr>          <chr>  <chr>      <chr> <chr> <chr>        <int>      <int>
##  1     1 Bukit Kesing… 25-Oc… Balionyct… MN    F     J               47         11
##  2     2 Bukit Kesing… 26-Oc… Cynopteru… MN    M     A               60         14
##  3     3 Bukit Kesing… 26-Oc… Cynopteru… MN    F     A               75         20
##  4     4 Bukit Kesing… 26-Oc… Cynopteru… MN    F     J               63         18
##  5     5 Bukit Kesing… 16-No… Balionyct… MN    F     J               44         10
##  6     6 Bukit Kesing… 16-No… Cynopteru… MN    F     A               71         21
##  7     7 Bukit Kesing… 16-No… Cynopteru… MN    M     A               64         16
##  8     8 Bukit Kesing… 17-No… Cynopteru… MN    F     A               64         19
##  9     1 Ladang Tayor… 27-Oc… Cynopteru… MN    M     A               62         17
## 10     2 Ladang Tayor… 28-Oc… Rhinolopu… HT    F     J               41         16
## # … with 68 more rows, and 3 more variables: hind_foot_length <int>, tail_length <int>,
## #   weight <dbl>
```

Outlier detection - scatterplot

```
exdata2_hipposideros <- exdata2 %>%
  filter(stringr::str_detect(
    string = exdata2$species,
    pattern = "Hipposideros"))

with(exdata2_hipposideros,
  plot(total_length, weight)
)
```



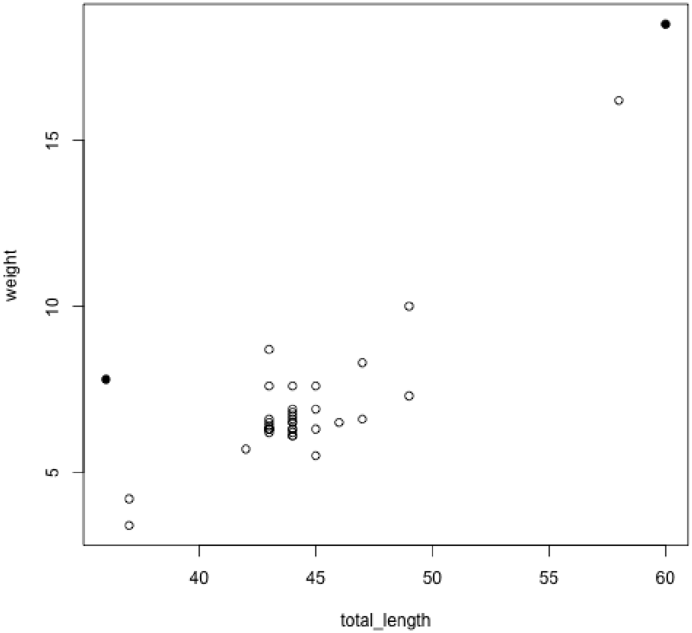
Outlier detection - statistical distance

```
exdata2_hipposideros %>%
  filter(outliersMD(total_length, weight))
```

```
## # A tibble: 2 × 13
##      id site           date   species    trap  sex   stage total_length ear_length tibia_
##   <int> <chr>          <chr>  <chr>      <chr> <chr> <chr>        <int>      <int>
## 1    13 Pengkalan Uta… 05-Ja… Hipposide… HT    F     A               36         13
## 2    14 Sungai Buweh,… 04-Ja… Hipposide… HT    M     A               60         17
## # … with 3 more variables: hind_foot_length <int>, tail_length <int>, weight <dbl>
```

Outlier detection - statistical distance

```r
with(exdata2_hipposideros,
  plot(
    x = total_length,
    y = weight,
    pch = ifelse(
      outliersMD(
        total_length,
        weight), 19, 1)
  )
)
```



Outlier detection - statistical distance

- The `outliersMD()` function has an argument called **alpha** which is set to 0.001 by default
- With `alpha = 0.001` we are looking for records with values so extreme that we would expect to find them with a probability of 0.001 when there are no problems with the data
- Another way of looking at the **alpha** parameter is that it alters the sensitivity of the `outlierMD()` function for detecting outliers by altering the threshold distance that is used to define outliers.
- Larger values of **alpha** will tend to detect more potential outliers

Questsions so far?

EcoHealth Alliance

Digit preference

- observation that the final number in a measurement occurs with a greater frequency than is expected by chance
- can occur because of rounding, the practice of increasing or decreasing the value in a measurement to the nearest whole of half unit, or because data is made up
- common for field staff to round the first value after the decimal point to **0** or **5** or measurements in whole numbers are rounded to the nearest decade or half-decade
- fictitious data often shows digit preference with **2** and **6** appearing as final digits more frequently than expected

Digit preference - simulated example

- We simulate a dataset of finalDigits of a measurement as follows:

```
set.seed(0)
finalDigits <- sample(x = 0:9, size = 1
```

- Examine `finalDigits` as a table

```
table(finalDigits)
```

```
## finalDigits
##   0   1   2   3   4   5   6   7   8   9 
##  95  80  96 102 106  98 109  95 109 110
```

- Examine `finalDigits` graphically

Digit preference - statistical testing

```
chisq.test(table(finalDigits))
```

```
##
##      Chi-squared test for given probabilities
##
## data:  table(finalDigits)
## X-squared = 7.72, df = 9, p-value = 0.5626
```

- No significant difference in the frequency of the last digits of the measurements

- *False-positives* and *false-negatives* can arise when doing this statistical testing approach to detect digit preference. A small number of observations can lead to *false-negatives* while a large number of observations can lead to *false-positives*

- This can be addressed using the **Digit Preference Score** or **DPS** and can be applied using the `digitPreference()` function in the `{nipnTK}` package

Digit preference - digit preference score

- Applying the `digitPreference()` function to the first example data on forearm length, we get:

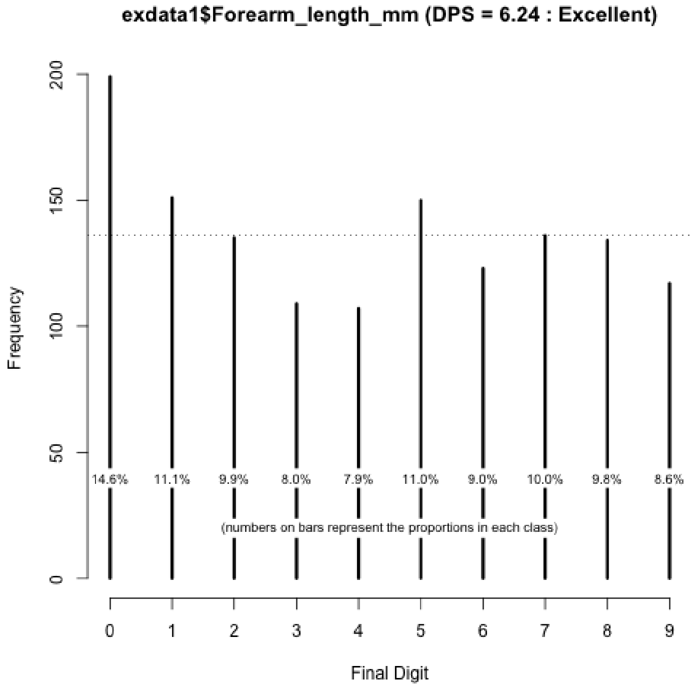
```
digitPreference(x = exdata1$Forearm_length_mm)
```

```
##
##     Digit Preference Score
##
## data:    exdata1$Forearm_length_mm
## Digit Preference Score (DPS) = 6.24 (Excellent)
```

Digit preference - digit preference score

The **DPS** can also be plotted:

```
plot(
  digitPreference(
    x = exdata1$Forearm_length_mm
  )
)
```



Digit preference - digit preference score

- Applying the `digitPreference()` function to the second example dataset on total length, we get:

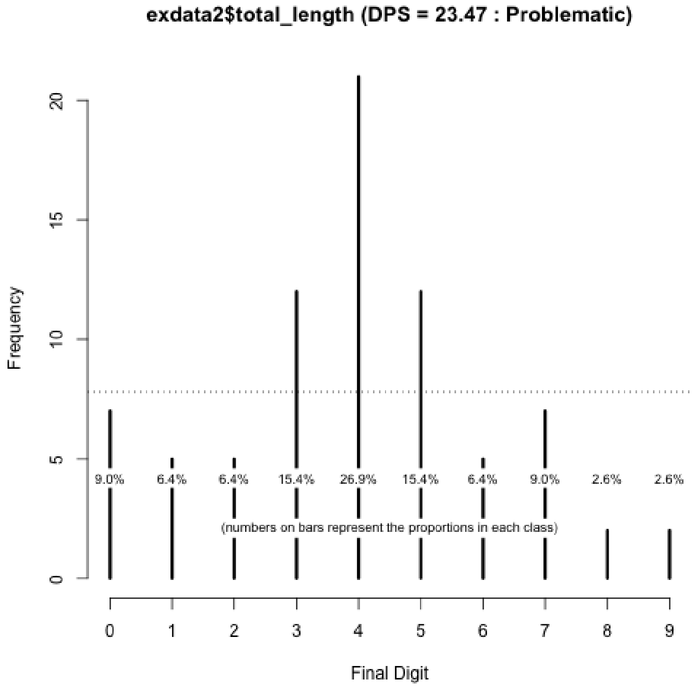
```
digitPreference(x = exdata2$total_length, digits = 0)
```

```
##
##     Digit Preference Score
##
## data:    exdata2$total_length
## Digit Preference Score (DPS) = 23.47 (Problematic)
```

Digit preference - digit preference score

- The **DPS** can also be plotted:

```
plot(
  digitPreference(
    x = exdata2$total_length,
    digits = 0
  )
)
```



exdata2$total_length (DPS = 23.47 : Problematic)

Questsions so far?

Assessing measurers performance during training

- Having measurers undergo standardisation exercises is common practice in health and nutrition sector
- Measurers perform measurements on a minimum of 10 subjects twice and their measurements are assessed for *accuracy* and *precision*

EcoHealth Alliance

Assessing measurers performance - accuracy and precision

- *Accuracy* is calculated against a gold standard which is either the measurements made by an expert or the mean of the measurements made by the entire cohort/group of measuerers being trained

- *Precision* is calculated using the inter-observer technical error of measurement (TEM) metric proposed by Ulijaszek and Kerr in:

  > Ulijaszek, S., & Kerr, D. (1999). Anthropometric measurement error and the assessment of nutritional status. British Journal of Nutrition, 82(3), 165-177. doi:10.1017/S0007114599001348

- The `{anthrocheckr}` package has functions that calculates these metrics. See https://nutriverse.io/anthrocheckr

Assessing measurers performance - limitations as applied to EHA studies

- animal subjects may be more challenging to obtain as compared to human subjects (although recruitment of human subjects for training is also challenging)
- accuracy can be calculated as gold standard is available; however, for precision, setting standard for acceptable TEM will be needed

EcoHealth Alliance

Final questions?

EcoHealth Alliance

Thank you!

Slides can be viewed at https://ecohealthalliance.github.io/dataquality or PDF version downloaded at https://ecohealthalliance.github.io/dataquality/dataquality.pdf

R scripts for slides available at https://github.com/ecohealthalliance/dataquality

EcoHealth Alliance