Ornis Hungarica 2016. 24(1): 172–181.
DOI: 10.1515/orhu-2016-0010

Hitchhikers' guide to analysing bird ringing data

Part 2: distributions, summary statistics and outliers

Andrea Harnos[1*], Tibor Csörgő[2] & Péter Fehérvári[1,3]

Received: May 31, 2016 – Accepted: June 12, 2016

Andrea Harnos, Tibor Csörgő & Péter Fehérvári 2016. Hitchhikers' guide to analysing bird ringing data – Part 2. – Ornis Hungarica 24(1): 172–181.

Abstract This paper is the second part of our bird ringing data analyses series (Harnos *et al.* 2015a) in which we continue to focus on exploring data using the R software. We give a short description of data distributions and the measures of data spread and explain how to obtain basic descriptive statistics. We show how to detect and select one and two dimensional outliers and explain how to treat these in case of avian ringing data.

Keywords: distribution types, outlier, standard deviation, coefficient of variation, descriptive statistics

Összefoglalás A sorozat második részében folytatjuk a madárgyűrűzési adatok kezelését, elemzését az R statisztikai program használatával (Harnos *et al.* 2015a). A különböző eloszlás típusok ismertetése, a szórás fogamának bevezetése után bemutatjuk, hogyan számolhatunk egyszerű módon leíró statisztikákat az adattáblázatunkra. Módszereket mutatunk a lehetséges egy- és kétdimenziós kiugró értékek felimerésére, és tanácsokat adunk a kezelésükre. Megmutatjuk, hogyan lehet egyszerűen leválogatni az olyan eseteket, amelyeknél gyanús értéket találunk valamely változóban.

Kulcsszavak: eloszlás típusok, kiugró érték, szórás, relatív szórás, leíró statisztika

[1] Department of Biomathematics and Informatics, Szent István University, Faculty of Veterinary Science, 1078 Budapest, István utca 2., Hungary, e-mail: harnos.andrea@univet.hu
[2] Department of Anatomy, Cell- and Developmental Biology, Eötvös Loránd University, 1117 Budapest, Pázmány Péter sétány 1/C, Hungary
[3] Department of Zoology, Hungarian Natural History Museum, 1088 Budapest, Baross utca 13., Hungary
**corresponding author*

Introduction

This paper is the second part of our bird ringing data analyses series (Harnos *et al.* 2015a) in which we continue to focus on exploring data. We will give a short description of data distributions, and explain how to obtain basic descriptive statistics. In general our template dataset is of Pied Flycatchers trapped and ringed at the Ócsa Bird Ringing Station (Central Hungary) between 1984–2014 (for details see Harnos *et al*. 2015b). The dataset is available through our `ringR` package or from the online appendix of this paper along with the code used in this part (OH_2016_24(1)_172-181_appendix.zip). We used R 3.3.0 for the analysis (R Core Team 2016) on the Ubuntu 14.04 platform. The codes were written with RStudio 0.98.1103.

Distribution types

Under a given variable's distribution we mean the pattern of observed values on the number line (Reiczigel *et al.* 2014). We can distinguish uniform, unimodal, multimodal and skewed distribution patterns *(Figure 1)*. In case of uniform distributions the values are spread more or less evenly, without observable clutter or aggregation along the number line. In practice, this pattern occurs rarely and typically is a product of artificial data generation. The other pattern types all show that the frequency of observed values is higher at given locations. If this aggregation occurs around a single value, the pattern is unimodal, as opposed to multiple clutters along the number line in which case, we classify the shape as multimodal (bimodal for two modality centres). Distribution shape and symmetry are also to be considered; distributions can be right-skewed (i.e. more extreme observations are present at larger values) or left-skewed (i.e. more extreme observations are present at smaller values) and can be symmetrical. Distributions can be typically illustrated with histograms and smoothed histograms (see Harnos *et al.* 2015a for details), however boxplots may also show symmetry and give a hint on the shape of distribution. We illustrate distribution types by plotting observed values *(Figure 1)* and we show, how these values shape distributions and how they are represented by boxplots and histograms *(Figure 2a-e)*.

One of the most common and most frequently referred unimodal, bell-shaped, symmetric distribution is the normal or Gaussian distribution. In avian ringing datasets, most of the variables that are measured on a continuous scale (typically biometric measurements) can be expected to have normal distributions (McDonald 2014). This often makes data analyses convenient as most of the commonly used statistical procedures rely on normally distributed

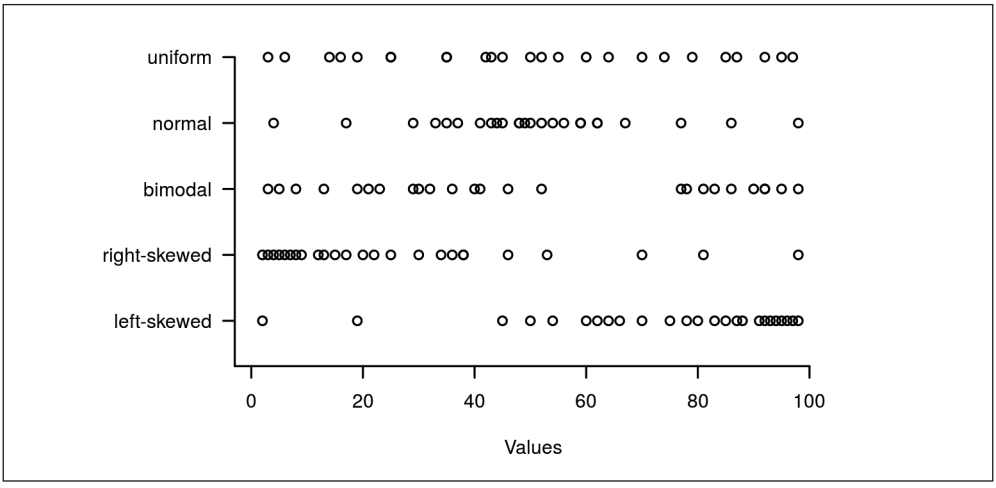

Figure 1. Different distribution patterns described with 25-25 points: evenly distributed on the whole data range (a), symetrically aggregating in the middle (b), aggregating around two values(c), right-skewed (d) and left-skewed (e)

1. ábra Különböző eloszlás-mintázatok 25-25 ponttal ábrázolva: az egész tartományon egyenletesen sűrű (a), középen szimmetrikusan sűrűsödő (b), két értéknél sűrűsödő (c), jobbra ferde (d), balra ferde (e)

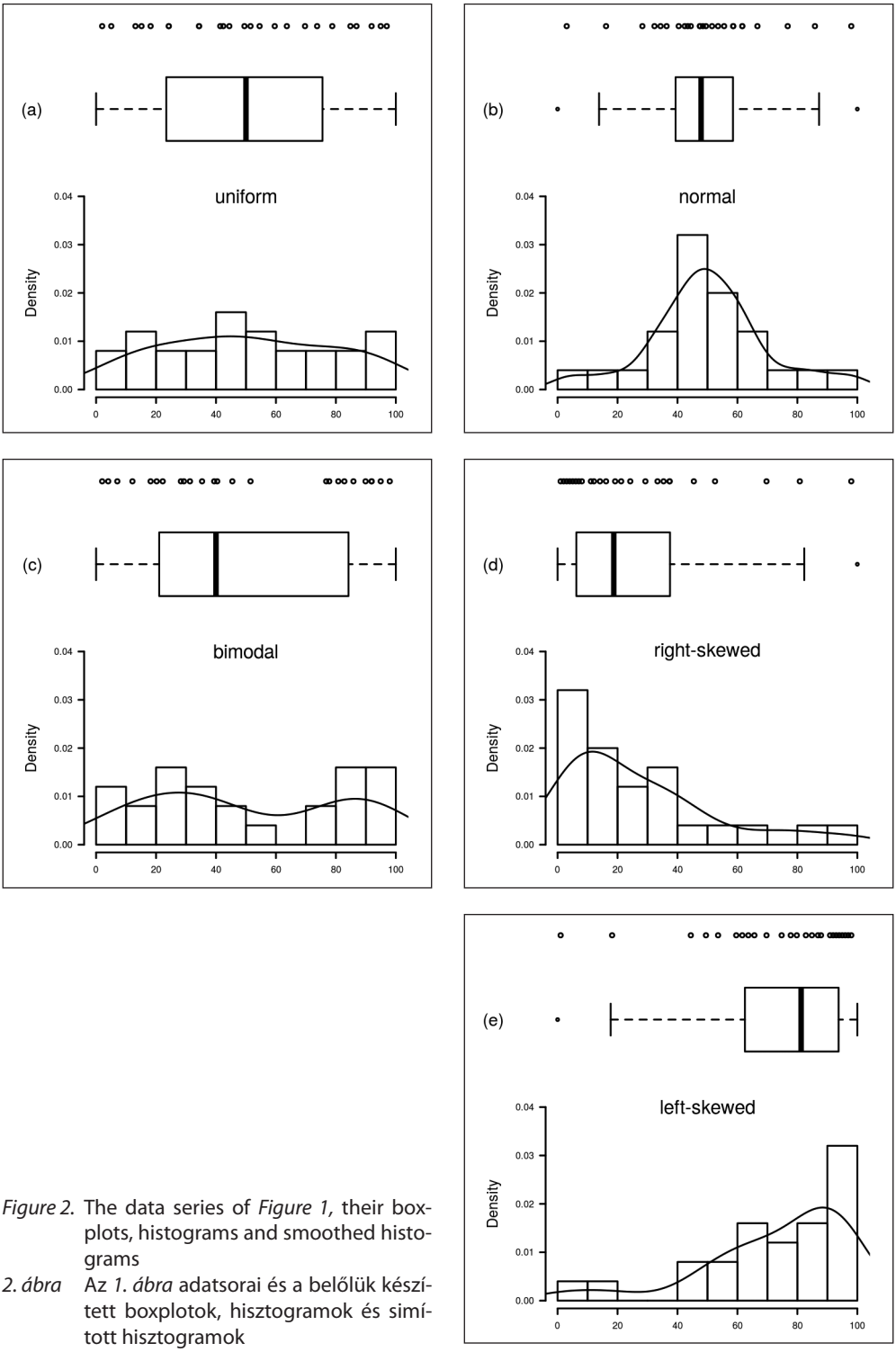

Figure 2. The data series of *Figure 1*, their boxplots, histograms and smoothed histograms

2. ábra Az *1. ábra* adatsorai és a belőlük készített boxplotok, hisztogramok és simított hisztogramok

data. However, in cases when this does not apply, quite often analysts utilise data transformation to artificially create a normal distribution. While transforming data may produce distributions that fulfil the requirements of statistical procedures, one may loose the biological meaning of the data (Ieno & Zuur 2015) making inference on obtained results difficult. Therefore we advise to use common sense and utmost care when applying transformations.

The spread of the data

The distribution of values of a variable can be summarized by two essential statistics, one depicting the central tendency of the data, the other the spread of the data. While central tendency is usually represented with the mean or the median, occasionally with the mode, the spread of the data with the sample variance and sample standard deviation. The variance and the standard deviation measure the average deviation of observed values from the mean as the centre of the distribution. More precisely, the variance is the sum of the squared deviations over all observations divided by n-1, where n is the number of observations. The standard deviation is the square root of the variance, and it is measured on the same scale as the variable. It may be useful to keep in mind, that in case of normally distributed data, approximately 68% values are within one standard deviation away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This quality of normal distributions is often referred to as the 1-2-3 standard deviation rule *(Figure 3)*.

Using variance (or standard deviation) to describe the spread of data has its limits. For instance, we are interested to know whether Sparrowhawks *(Accipiter nisus)* or Goshawks *(Accipiter gentilis)* have more variable wing length measurements. In our hypothetical

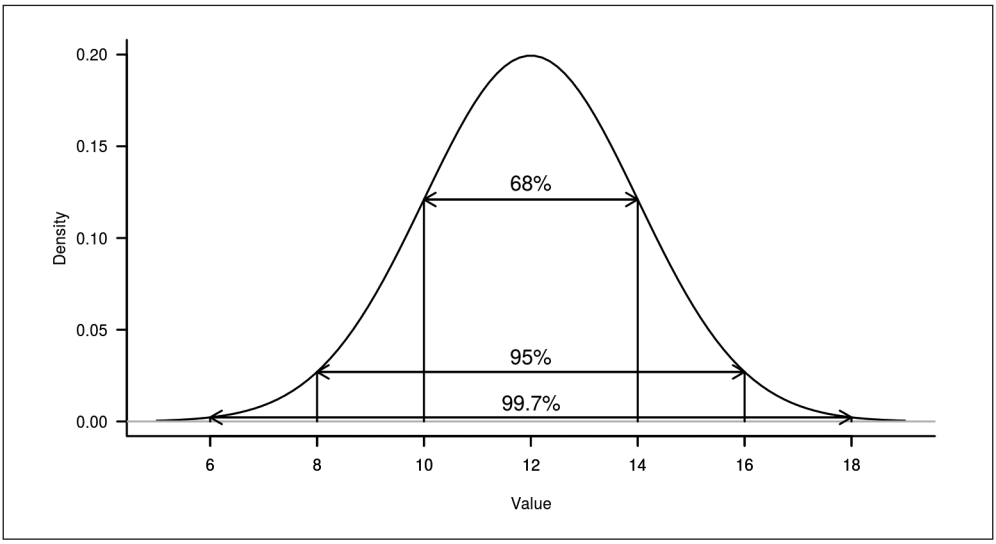

Figure 3. Illustrating the 1-2-3 standard deviation rule in case of a normal distribution if the mean is 12 and the standard deviation is 2

3. ábra Az 1-2-3 szórás szabály bemutatása 12 átlagú és 2 szórású normális eloszlás esetén

example mean Sparrowhawk wing length is 249 mm with a standard deviation of 51 mm while for Goshawks the same statistics are 533 mm and 51 mm. One might intuitively think that both species have similar variation in wing length, as the standard deviations are equal. However, we have to consider the value of the mean when accounting for spread. In order to do this we can calculate the coefficient of variation (CV) as the standard deviation divided by the mean. It can be given in percentages (CV%). This statistics gives a mean-independent assessment of data spread, and as such can be used to evaluate our example, the CV% for Sparrowhawks is 20.5% while for Goshawks it is 9.6%.

Table of the basic descriptive statistics

Let us observe the basic descriptive statistics *(Table 1)* of the distributions shown in *Figure 1*. The table was created with the `RcmdrMisc` package's `numSummary()` command (Fox 2014). The parameters we used were the name of the dataframe and the descriptive statistics to be printed (see. `?numSummary`). Notice that for the first three symmetrical distributions the mean is close to the centre of the value range, while for the left-skewed distribution the mean shifted towards larger, for right-skewed distribution towards smaller values. For symmetrical distributions the median (50% quantile) is similar to the mean, however for skewed distributions it is larger (left-skewed) or smaller (right-skewed) than the mean. The bimodal distribution has the largest standard deviation (`sd`) while the unimodal distribution has the smallest, but the difference is not substantial. However, the coefficients of variation (`cv`) show a marked difference among distributions. Probably the best approach for skewed distributions is to examine the different quantiles.

```
>options(digits = 3)   # setting the number of digits
>library(RcmdrMisc)
>distr_data = read.table("distr_data.csv", sep=";",header=T)
>numSummary(distr_data, statistics = c("mean", "sd", "cv", "quantiles"))
```

Type of distribution	mean	sd	cv	0%	25%	50%	75%	100%
uniform	54.1	27.9	0.516	3	42	52	74	97
normal	50.2	20.1	0.399	4	41	49	59	98
bimodal	50.8	32.8	0.645	3	23	41	83	98
right-skewed	27.7	25.6	0.924	2	8	20	38	98
left-skewed	73.0	24.6	0.337	2	62	80	92	98

Table 1. Basic descriptive statistics of the data used for *Figure 1* and *2*
1. táblázat Az *1.* és *2. ábrák*hoz használt adatok leíró statisztikái

Summary statistics for groups of data

With the `numSummary()` function we can calculate the basic numerical summaries for groups defined by factors or factor combinations of numeric variables. For example, if we are interested in the basic descriptive statistics of the wing length in the age and sex groups of Pied Flycatchers, we should type (note that we use the data cleaned as shown in Part 1. of this series):

```
>setwd("D:/mydirectory")
>mydata = read.table("FICHYP1.csv", sep=";",header=T)
>numSummary(mydata$WING, statistics = c("mean", "sd", "cv", quantiles"),
groups = mydata$age:mydata$SEX)

          Statistic
   Group     mean      sd       cv      0%     25%    50%    75%   100%
   adult:F   79.5    1.96    0.0246    76    78.0     79     81     85
   adult:M   80.8    1.87    0.0232    77    79.5     81     82     85
   juv:F     79.3    1.77    0.0224    70    78.0     79     80     85
   juv:M     80.3    1.92    0.0239    73    79.0     80     81     87
```

Outliers

In some cases observed values may be out of bounds, or simply impossible (we have already handled the latter in Harnos *et al.* 2015a). These odd values or outliers are values that are somehow out of proportion, too small or too large yet not small or large enough to be able to a priori exclude from the observations. It is important to avoid subjectivity when identifying and treating outliers and here we demonstrate a few methods that may help the evaluation. For normally distributed variables, values that are 3 standard deviations larger or smaller than the mean are considered outliers. For non-normally distributed variables a frequently used rule of thumb is to consider values below or above 1.5 (`const`) interquartile range (`IQR`) from the lower (`Q1`) or upper quartile (`Q3`) to be outliers, and this latter rule is used also by R. These observations can be retrieved using the `boxplot.stats()` command. In the example below we show the outliers of juvenile female Pied Flycatcher wing lengths stored in the `WING` variable. Note that it is important to calculate these values only within homogeneous subgroups of the observations. Calculating the outliers for all ringed Pied Flycatchers regardless of age and sex would produce erroneous results, hence the necessity of subsetting the data. It is useful to evaluate outliers together with the other variables of the individual. To retrieve all observations related to the individuals (i) we can use the `subset()` command as shown below.

```
>mydata_1 = subset(mydata,(mydata$age == "juv" & mydata$SEX == "F" &
>mydata$season == "autumn"))

>boxplot.stats(mydata_1$WING)$out

$out
[1] 70 84 85 84

>Q1 = quantile(mydata_1$WING, na.rm = T)[2]
>Q3 = quantile(mydata_1$WING, na.rm = T)[4]
>IQR = IQR(mydata_1$WING, na.rm = T)
>const = 1.5
>outliers = subset(mydata_1, (mydata_1$WING < (Q1 - const * IQR) |
mydata_1$WING > (Q3 + const * IQR)))

>outliers
          ID       DATE   RECAP     RING AGE  SEX FAT  MUSCLE  MASS WING  THIRD
 258    71051   9/9/1990       0   L52435  1Y    F   4    <NA>  11.7   70     53
 1436  429635   9/2/2006       1  T468844  1Y    F   2       2  13.7   84     60
 1585  503787  9/21/2007      NA   8E3738  1Y    F   0       2  11.5   85     65
 1616  534644  8/26/2008       0   W55664  1Y    F   0       3  11.7   84     64

       TAIL   year   yearday  period.recap   season   age
 258     43   1990       251             0   autumn   juv
 1436    56   2006       244             1   autumn   juv
 1585    56   2007       263             0   autumn   juv
 1616    58   2008       238             0   autumn   juv
```

The code stores the observations in the `outliers` object. Our output shows an interesting pattern; all four outliers in the `WING` variable have correspondingly small or large other measurements indicating that probably the observations are valid and not artefacts due to erroneous data entry. This example demonstrates that outliers are not values that are to be excluded automatically. In fact if we automatically exclude outliers, we alter the distribution pattern, thus new outliers may appear. In extreme cases we may exclude the whole dataset before we realize that there is something wrong. Bear in mind that having outliers may be an indication of skewed distribution and instead of leaving the values out of the analyses we may be more successful in using methods that are applicable to skewed distributions. Quite often we only notice the potential outliers in later phases of the analysis. For example, the outlier indicated with an arrow on *Figure 4* (see *Box 1* for scatterplot) can only be noticed when we examine the two variables together, but separately it is impossible. Such bivariate outliers are useful for checking data entry, e.g. we expect that flycatchers with longer wings have longer tarsi. Such an outlier may help finding misspelled data.

To explore the relationship of two numerical variables and the outliers, we can use a more advanced scatterplot function from the `car` package. We show its usage with the `WING` and `THIRD` variables in the case of juvenile female birds measured during the autumn migration season *(Figure 5)*. The variables for the plot are given in a so called formula, as `y ~ x`, where `y` is the dependent and `x` is the independent variable. Since the variables are measured

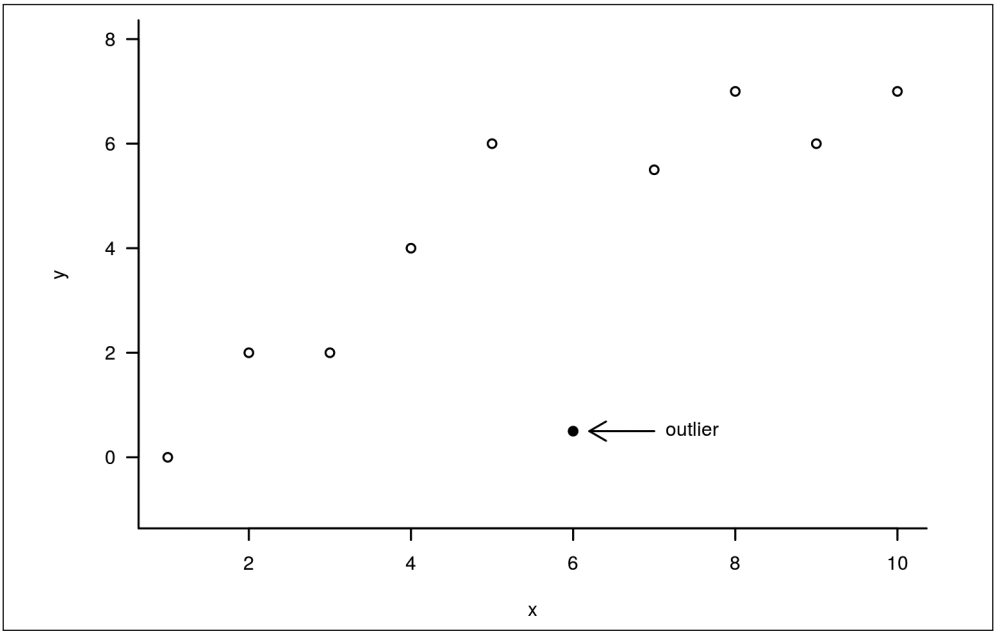

Figure 4. A scatterplot with an outlier
4. ábra Szórásdiagram kiugró értékkel

Scatterplot

To describe the two dimensional distribution of two variables we use the scatterplots, where data is displayed as a set of points. One variable determines the position of the points on the horizontal axis, while the other variable determines the position on the vertical axis. We can inspect the plot visually and notice if there are points that lie distant from the main data distribution. These points are potential outliers in a two dimensional sense.

Box 1. Scatterplot
1. doboz Szórásdiagram

with 1 mm precision, both of them are discrete, therefore, the scatterplot is typically uninformative because the data points are overplotted. Jittering the data (i.e. adding a small random quantity to each coordinate (Cleveland 1994)) can be useful in these cases. We used the `jitter()` function in the formula.

The `scatterplot()` function of the `car` package (Fox & Weisberg 2011) makes a usual scatterplot with several additional options (see `?scatterplot`).

```
>library(car)
>scatterplot(jitter(WING) ~ jitter(THIRD), data = mydata, smooth = F,
subset = (season == "autumn" & SEX == "M" & age == "juv"), reg.line = F,
ellipse = T, levels = c(0.95, 0.99), id.method = "identify")
```

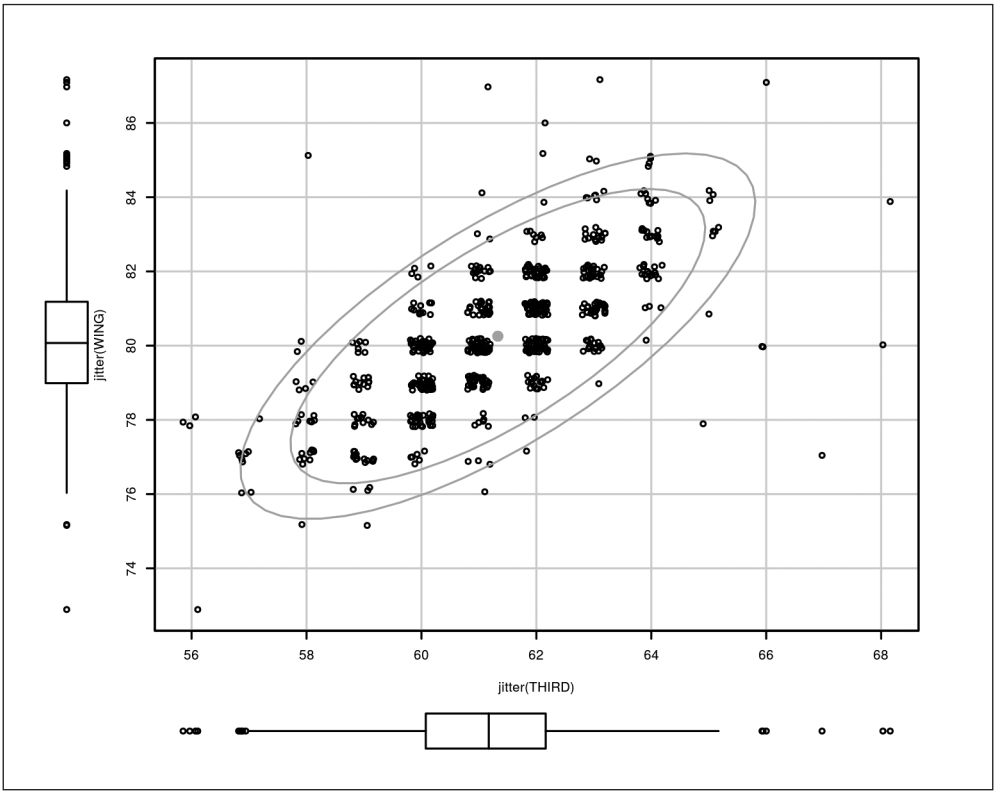

Figure 5. Scatterplot of the jittered `WING` and `THIRD` variables with boxplots on the sides and 95% and 99% concentration ellipses

5. ábra A `WING` és `THIRD` változók szórásdiagramja boxplotokkal a tengelyek mellett, valamint az adatok 95 és 99%-át tartalmazó ellipszisekkel. Az átfedések miatt a pontokhoz hozzáadtunk egy kis véletlen értéket (jitter) a jobb áttekinthetőség kedvéért

This adds marginal boxplots for both variables *(Figure 5)* and plots a nonparametric-regression curve (`smooth`) and a regression line (`reg.line`) to the plot by default. Since we do not need the regression line now, we set the `reg.line = F`. We call this function with a subset of the data as specified using a logical expression. If we can assume, that the two variables are linearly related and their two dimensional distribution is normal, than the `ellipse` parameter is also useful to find possible outliers (Fox & Weisberg 2011). If it is set to `TRUE`, data-concentration ellipses are plotted. With the `levels` parameter, the levels of concentration ellipses can be set, which means that in case of a two dimensional distribution, we expect that a specified proportion (set by this parameter) of the data fall inside the ellipses. In this case we used two levels: 95 and 99%. With the parameter `id.method = "identify"`, we can get the position of the graphics pointer when the (first) mouse button is pressed. It then searches the coordinates given in x and y for the point closest to the pointer. If this point is close enough to the pointer, it's index will be returned as part of the value of the call (Becker *et al*. 1988). Usually the identification process can be terminated by pressing the `ESC` key or by closing the graphics device (for more details see `?identify`).

Two dimensional outliers can be caused by erroneous data entry, thus the evaluation procedure should be similar to outliers of a single variable. If we decided that the value is implausible we may exclude the observation. However, if we are unsure, the best approach is to run the statistical analyses with and without the observation in question and evaluate the results. If we plan to publish our results, it is necessary to mention and advisable to report analyses with and without the outliers. Maybe the outliers are our most interesting observations that may help further our knowledge in our field.

Acknowledgments

The authors express their gratitude for the work of all the volunteers that collected data at the Ócsa Bird Ringing Station throughout the years. We are grateful for our colleagues – especially for János Kis and Zsolt Lang – who helped us improve this manuscript. This work was supported by OTKA under Grant No. 108571.

References

Becker, R. A., Chambers, J. M. & Wilks, A. R. 1988. The New S Language. – Wadsworth & Brooks/Cole

Cleveland, W. S. 1993. Visualizing data. – Hobart Press, Summit, NJ.

Fox, J. 2014. RcmdrMisc: R Commander Miscellaneous Functions. R package version 1.0-2. https://CRAN.R-project.org/package=RcmdrMisc

Fox, J. & Weisberg, S. 2011. An {R} Companion to Applied Regression, 2nd ed. – Thousand Oaks CA, Sage

Harnos, A., Fehérvári, P. & Csörgő, T. 2015a Hitchhikers' guide to analysing bird ringing data Introduction, Part 1: data cleaning, preparation and exploratory analyses. – Ornis Hungarica 23(2): 163–188. DOI: 10.1515/orhu-2015-0018

Harnos, A., Lang, Zs., Fehérvári, P. & Csörgő, T. 2015b Sex and age dependent migratory phenology of the Pied Flycatcher in a stopover site in the Carpathian Basin. – Ornis Hungarica 23(2): 10–19. DOI: 10.1515/orhu-2015-0010

Ieno, E. N. & Zuur, A. F. 2015. A Beginner's Guide to Data Exploration and Visualisation with R. – Highland Statistics Ltd. Newburgh

McDonald, J. H. 2014. Handbook of Biological Statistics, 3rd ed. – Sparky House Publishing, Baltimore, Maryland. This web page contains the content of pages 133–136 in the printed version. http://www.biostathandbook.com/normality.html

R Core Team 2016. R: A language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reiczigel, J., Harnos, A. & Solymosi, N. 2014. Biostatisztika nem statisztikusoknak [Biostatistics for non statisticians]. 3rd ed. – Pars Kft., Budapest (in Hungarian)