DESCRIPTIVE STATISTICS

Erik Kusch

erik.kusch@i-solution.de

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

1. Background
    - Introduction
    - Methods Of Descriptive Statistics
2. Parameters In Descriptive Statistics
    - What Are Parameters?
    - Getting Practical
    - Location Parameters/Measures Of Central Tendency
    - Dispersion Parameters/Measures Of Spread
    - Making Life Easier in `R`
3. Exercise
    - Excel data
    - Parameters

Introduction

Descriptive statistics are used to **summarize data**.

The aim: To describe a given set of data records n in regard to a certain **variable** p_j or set of variables p.

The procedure: Using an adequately chosen set of methods to **summarize or visualize** the data at hand.

Characteristics of **variables** are often expressed via **parameters**.

Methods & Quirks

Information is usually handed to descriptive statistics as $n \times p$ (row count $\times$ column count) data frames.

This information is used to calculate informative **parameters**:

Location Parameters (Measures Of Central Tendency):

- Arithmetic Mean
- Mode
- Median
- Minimum, Maximum, Range
- ...

Dispersion Parameters (Measures Of Spread):

- Variance
- Standard Deviation
- Quantile Range
- ...

Descriptive statistics **do not allow generalisation** beyond the data!

Parameters And Their Meaning

What is a parameter?

In the case of descriptive statistics, a *parameter* presents some information on the shape of the distribution of the values of a certain variable.

What's the fuss?

Parameters can be used to summarise data properties and make large data sets with a multitude of values per variable more accessible.

So?

To know which parameters to use one must know which ones there are and how to calculate them.

Parameters are, more or less, **digested data**.

Creating Some Data

For the following computation of descriptive statistics parameters, we will need the following data:

```r
set.seed(42) # making the code reproducible
data_vec <- rnorm(mean = 20, sd = 2, n = 54)
matrix(sort(data_vec), nrow = 6)
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,] 14.69 17.22 18.55 19.28 19.73 20.64 21.01 21.52 22.89
## [2,] 15.12 17.26 18.72 19.39 19.79 20.73 21.27 22.07 23.02
## [3,] 15.17 18.30 18.78 19.43 19.81 20.81 21.27 22.43 23.15
## [4,] 16.44 18.38 18.87 19.44 19.87 20.87 21.29 22.61 23.79
## [5,] 16.47 18.43 19.14 19.49 20.07 20.91 21.31 22.64 24.04
## [6,] 16.57 18.43 19.14 19.66 20.41 20.92 21.41 22.74 24.57
```

→ Calculation of parameters of descriptive statistics is reserved almost exclusively for `numeric` data records

Arithmetic Mean (Theory)

Definition: Also called average, this metric is the mathematical average of the given data values.

Non-resistant to outliers and asymmetric distributions.

Calculation: $\overline{x} = \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$

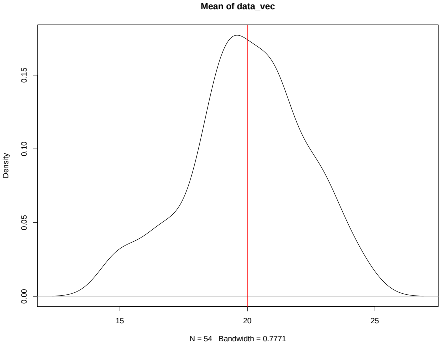
$\overline{x} = \mu$	Arithmetic mean
n	Number of samples ($=$ number of values for the variable in question)
i	Index of variable values ($i = 1, 2, .., n$)
x_i	i^{th} value of variable x

Arithmetic Mean (Calculation in R)

The arithmetic mean is calculated using the `mean()` function contained within base R.

```
# calculation
mean(data_vec)
```

```
## [1] 20
```



Median (Theory)

Definition: The median is the value separating the higher half of the data values from the lower half.

Resistant to outliers and asymmetric distributions.

Calculation:

$median(x) = (\frac{n+1}{2})^{th}$ odd numbers of data values

$median(x) = \frac{\left((\frac{n}{2})^{th}+(\frac{n}{2}+1)^{th}\right)}{2}$ even numbers of data values

$median(x)$ Median of the values available for variable x

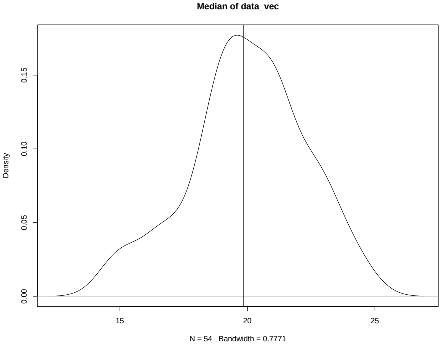
n Number of observations available for x

Median (Calculation in R)

The median is calculated using the median() function contained within base R.

```
# calculation
median(data_vec)
```

```
## [1] 19.84
```



Mode (Theory)

Definition: The mode of a set of data values is the value that is the most common.

Resistant to outliers but the shape of the distribution might be crucial.

Calculation: $mode(x) = max_{k=1}\big(I_{i=1}(x_i = x_k)\big)$

$mode(x)$ Mode of the values available for variable x

$max_{k=1}()$ Maximising argument for k in 1 to p

$I_i()$ Identifier that returns 1 if the internal statement is true with i in 1 to n

Mode (Calculation in R)

One may wish to use the `max()` and `table()` function contained within the base R or through `mlv(..., method="mfv")` within the `modeest` package:

```r
# counts of values in rounded vector
table <- table(round(data_vec))
table # counts
```

```
##
## 15 16 17 18 19 20 21 22 23 24 25
##  3  2  3  4 11  7 12  3  6  2  1
```

```r
# most common appearance
max <- max(table)
max # maximum appearances
```

```
## [1] 12
```

```r
# position of maximum in table
pos <- which(table == max)
pos # mode position
```

```
## 21
##  7
```

```r
# value at maximum position
mode <- names(table)[pos]
as.numeric(mode) # mode
```

```
## [1] 21
```

Minimum, Maximum, Range (Theory)

Sometimes, one may want to use the following, simple information on data values:

Maximum: The highest value available for a given variable.

Minimum: The lowest value available for a given variable.

Range: The span of values that the data distribution defined by minimum and maximum extends over.

Minimum, Maximum, Range (Calculation in R)

```
# calculation
min(data_vec)
```

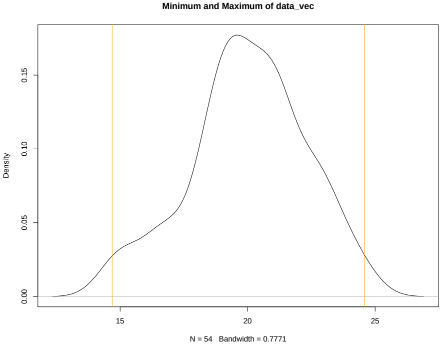
```
## [1] 14.69
```

```
max(data_vec)
```

```
## [1] 24.57
```

```
range(data_vec)
```

```
## [1] 14.69 24.57
```



Which Location Parameter Do I Use?

All measures of central tendency describe the central position of a frequency distribution of values of a given variable in the data set at hand.

The *arithmetic mean* is only really useful when concerned with symmetric distributions of data values.

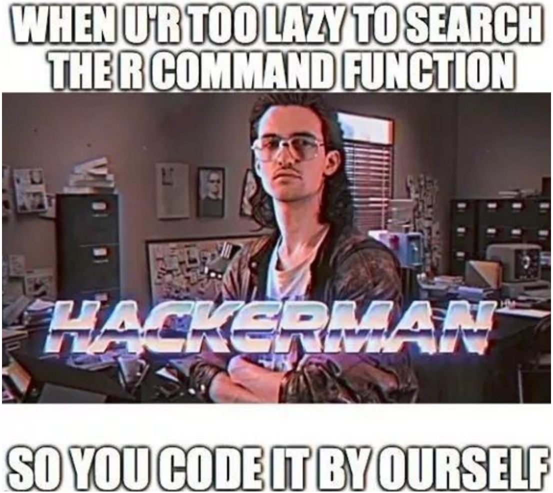
The *median* exhibits robust behaviour when faced with asymmetric distributions of data values.

The *mode* is most applicable to the classification setting and rarely used.

$\rightarrow$ The **median** will usually do.

You Can Do It Yourself

Remember: You can code most of these basic parameter calculations yourself.


WHEN U'R TOO LAZY TO SEARCH THE R COMMAND FUNCTION
HACKERMAN
SO YOU CODE IT BY OURSELF

Variance (Theory)

Definition: Variance measures how much data values are spread out from their average value.

Non-resistant to outliers and asymmetric distributions.

Calculation: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$

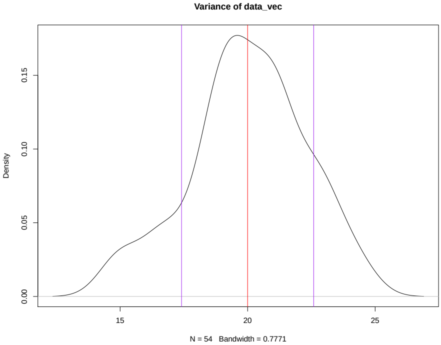
s^2	Variance
n	Number of samples ($=$ number of values for the variable in question)
i	Index of variable values ($i = 1, 2, .., n$)
x_i	i^{th} value of variable x
$\overline{x}$	Arithmetic mean

Variance (Calculation in R)

The variance is calculated using the `var()` function contained within base R.

```
# calculation
var(data_vec)
```

```
## [1] 5.181
```



Note that his plot shows the span of the variation around the mean.

Standard Deviation (Theory)

Definition: The standard deviation quantifies the amount of variation or dispersion of a set of data values.
Non-resistant to outliers and asymmetric distributions.

Calculation: $SD = s = \sqrt{s^2}$

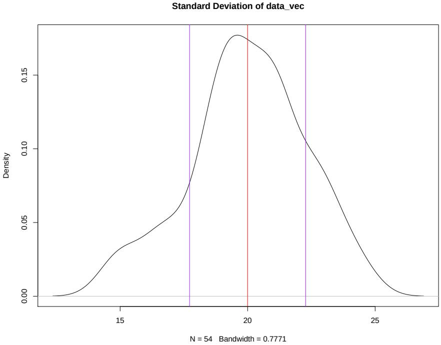
$SD = s$	Standard Deviation
s^2	Variance

Standard Deviation (Calculation in R)

The standard deviation is calculated using the `sd()` function contained within base R.

```
# calculation
sd(data_vec)
```

```
## [1] 2.276
```



Note that his plot shows the span of one standard deviation above and below the mean.

Quantile Range (Theory)

Definition: Quantiles are cut points dividing the range of a distribution of data values into adjacent intervals with equal probabilities. You will always receive one cut-point less than quantiles are produced.

Resistant to outliers and asymmetric distributions.

Most often, one uses the following quantiles:

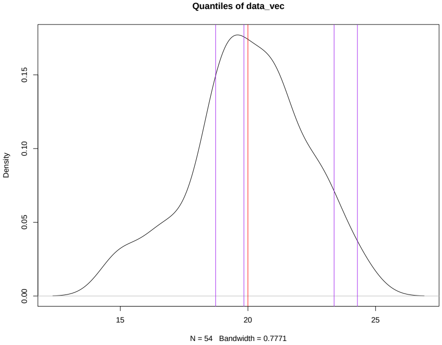
- *Quantile 50:* This is basically the median.
- *Quantile 25 and 75:* These are also known as *quartiles*.

Quantile Range (Calculation in R)

Quantiles are calculated using the `quantile()` function contained within base R. A second argument, within this function can be specified to call certain quantiles.

```r
# quantiles we want
q <- c(0.25, 0.5, 0.95, 0.99)
# calculation
quantile(data_vec, q)
```

```
##   25%   50%   95%   99%
## 18.74 19.84 23.38 24.29
```



Which Dispersion Parameter Do I Use?

All measures of spread describe the spread of a frequency distribution of values of a given variable in the data set at hand.

The *variance* is only really useful when concerned with symmetric distributions of data values.

The *standard deviation* is only really useful when concerned with symmetric distributions of data values.

The *quantiles* exhibit robust behaviour when faced with asymmetric distributions of data values.

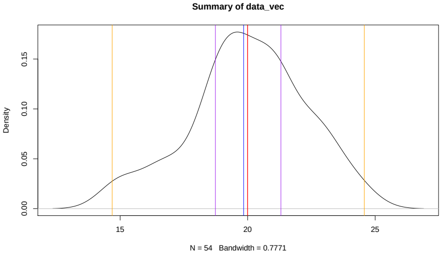
$\rightarrow$ The **quantiles** will usually do.

The summary() Function

```
# calculation
summary(data_vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.7    18.7    19.8    20.0    21.3    24.6
```

The summary() function can be called on a vector object in R to return some of the most useful information on measures of central tendency and measures of spread.



Loading Excel data into `R`

Excel is a **valuable tool** for **data accquisition** but almost **useless** when it comes to **statistical analyses** or **data visualisation** in biological sciences.

So **how do you get your excel data into `R`**?

Loading procedure depends on **file format**:

.csv - I recommend using this format as it allows for less alteration and is compressed.
Functions: `read.csv()` and `read.table()` (also works on .txt files)

.xls, .xlsx, etc. - Go for this if you need to alter your data by hand (which you shouldn't. EVER!).
Functions: `read.xlsx()` (included in `xlsx` package)

You can also **use `R`** to **save data in excel format**.

Inspecting Data

The **most common** form of data is the **`data frame`** which you can:

Inspect using functions such as:

- `dim()` to access the dimensions
- `str()` to access types and modes
- `colnames()`/`rownames()` to asses column and row names
- `head()`/`tail()` to show only the top or bottom five rows of the data set
- `table()` to show a count of items in a vector

Subset using the different sub-setting methods:

- `[r,c]` can be used to index rows (r) and columns (c)
- `$` can be used to index column names

→ This is also how you **extract data from a `data frame`** (and most objects within `R`).

Calculating parameters of descriptive statistics

Your **ToDo-List** for this exercise:

- Load the file DescriptiveData.csv into R
- Identify what kind of information it contains
- Calculate the location parameters and parameters of spread for any of the variables contained within the data set that catch your interest.
- Question the validity of your findings and the data

The solution file will deal with all of the variables contained within the data set so don’t worry about which one to pick and just **have fun**.