

## How to organize a data set for QPweb – essentials before upload

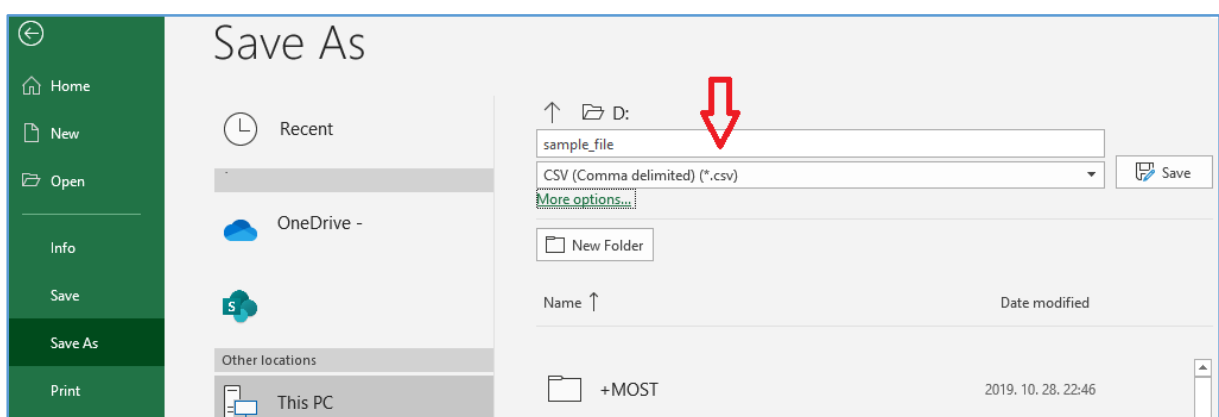
(version 1.0)

### 1. A single sheet with one column of data

**This is useful when we have one sample of hosts infected by one species (or one type) of parasites.**

Arrange data into a column; this column may or may not have a header (a title in the first cell). Each other cell represents a host individual. Parasite abundance data (the number of parasite individuals, including the zero values of non-infected hosts) are arranged into a single column. By definition, only zero and positive values are accepted. This is how it looks like in Excel. Choose the 'save as' command, then choose the extension '.csv' ('comma-separated values', also called 'comma delimited'). This file extension is safe to use in QPweb, but also see other options at point 5.

	A	
	number_ of_ticks_ on_each_ dog	it is facultative to have a title (reminder) here
1		
2	1	
3	1	
4	0	an uninfected dog
5	0	
6	0	
7	144	a heavily infected dog
8	13	



With this one-column data format, we can calculate prevalence, mean and median intensity, mean abundance, mean crowding and create the 95% confidence intervals of all these measures. Moreover, we can calculate aggregation indices (variance/mean, index of discrepancy, and the negative binomial distribution's exponent 'k') to characterize the distribution of parasite individuals across the host sample.

## 2. A single sheet with two or more columns

This format does not allow comparing infection indices between the two or more parasite species. For that purpose, see points 3 and 4.

### 2.1. One sample of hosts containing two or more species of parasites

Each row represents a host individual. Within a row, the cells represent the abundance of parasite1, parasite2, etc., **co-occurring in the same host individual**. Each column represents one particular type (most often: a species) of parasites. Other host characters (say, body mass) can also be represented. Thus infections by different parasites appear as linked (paired) values. It is advisable to put headers (titles) in the first row. Using two columns together, we can apply Spearman's rank correlation coefficient (with p-values based on Monte Carlo replications) to explore whether infections by two different parasites co-vary or not.

	A	B	
	ticks_on_cats	lice_on_cats	
1			
2	0	1	
3	0	NA	← this indicates a missing value
4	3	0	
5	0	0	
6	14	8	
7	3	1	
8	2	5	
9	4	0	
10			

these values are linked to each other because they characterize the same host individual

### 2.2. A single sheet with several columns to estimate the true species richness of parasites

By choosing all (3 or more) columns describing parasite quantities, we can estimate the true species richness of a parasite assemblage (a multispecies parasite community inhabiting a host population). This means that a particular sample of hosts (with limited sample size) is used to extrapolate to the host population as a whole in order to estimate the true species richness of the parasite community. For example, this sample contains 16 host individuals infected by 4 species of parasites. Applying the Chao2 species richness estimator indicates that the true species richness of parasites is likely to be about 6. A 95% Confidence Interval for the estimate is also presented (here: 4.2–24.9). In reality, we need a much larger sample size (say, hundreds) to provide a more reliable estimation (with a narrower Confidence Interval).

	A	B	C	D
1	par_sp_1	par_sp_2	par_sp_3	par_sp_4
2	0	0	0	0
3	1	0	0	0
4	13	0	1	0
5	0	0	1	0
6	0	0	0	0
7	0	0	4	0
8	0	0	0	0
9	0	0	0	0
10	0	0	23	0
11	0	0	11	0
12	0	0	0	0
13	0	1	0	1
14	0	0	1	0
15	0	0	1	0
16	0	0	2	0
17	0	0	1	0
18				

### 2.3. A two-column sheet to analyze parasite sex-ratio

The same data format can be used for parasite sex-ratio studies. Let the two 'types' of parasites (in the same row) represent conspecific male versus female parasites co-occurring in the same host individual. Here, we can calculate the parasite sex-ratio and its confidence interval. Apply Spearman's rank correlation coefficient (p-values are based on Monte Carlo replications) to explore whether or not parasite sex-ratio covaries with the intensity of infection.

	A	B
1	males	females
2	0	1
3	5	22
4	1	0
5	0	1
6	4	12
7	2	3
8	2	5
9	1	0
10	11	13

a single female parasite occurred on this particular host individual

### 3. Using several (2-6) data sheets (as separate files) to compare infection indices between two or more samples

In QPweb, we can upload two or more sheets (as separate files) in parallel, each with one or more columns of data to compare infection indices across them. Naturally, the infection values are not paired between the files (samples). In this way, comparisons of prevalence, mean abundance, mean intensity, median intensity, mean crowding, index of discrepancy (aggregation), and intensity distributions (Neuhauser's test and stochastic equality) are available. In addition, other host characters, such as body mass, can also be compared. This is how it looks like as Excel files:

File name: **donkeys.csv**

	A	B
1	worms	lice
2	2	0
3	0	0
4	0	1
5	23	0
6	12	0
7	0	7
8	4	0
9	0	11
10	3	2
11		

File name: **horses.csv**

	A	B
1	worms	lice
2	1	0
3	24	0
4	2	0
5	0	2
6	0	99
7	1	2
8	2	17
9	1	1
10	4	0
11	45	0
12	1	0
13	5	0
14		



comparisons of prevalence, mean intensity, etc.

these values are linked to each other because they characterize the same host individual

these values are not linked

#### 4. An alternative way of representing several (2-6) samples of hosts to test whether levels of infection depend on group identity


File name: *crocodiles.csv*

Let one of the columns (say, the 1<sup>st</sup> column) act as a group identifier. Any number can identify a group (say, 1, 2, ..., 6). These identifiers may refer to host groups by seasons, locations, host sex or age classes, etc. The other column(s) represent parasite abundance values for one or more species (types) of parasites or other host characters like body size. This is how it looks like in Excel:

By choosing two columns (one of which must be the Group\_ID), we can test whether or not prevalence, mean intensity, or mean abundance differs across groups. The latter test can also be used for other host characters, such as host body size, etc.

These tests are called '*group comparisons*' in the menu. Many other comparisons that are accessible by the former way (point 3) are not available here.

this column identifies 3 groups of hosts each with 4 individuals



	A	B	C	D
1	Group_ID	par_sp_1	par_sp_2	mass_kg
2	1	1	0	1.25
3	1	8	0	1.33
4	1	0	0	NA
5	1	1	1	2.22
6	2	NA	4	2.12
7	2	0	2	0.92
8	2	2	11	1.45
9	2	0	6	1.34
10	3	1	1	1.87
11	3	0	0	3.01
12	3	11	0	2.67
13	3	0	1	0.99

#### 5. Using other software

Spreadsheet programs like Excel (MS Office), Numbers (Mac) or Calc (LibreOffice) are also suitable to organize data. Save data matrix as a simple text (extension ".txt") or as a "comma-separated values" (extension ".csv") file. In .txt files the values are separated by blanks or tabs, in a .csv files by commas or semicolons (in those countries where the decimal symbol is comma).

Alternatively, text editors like Notepad (Windows) or TextEdit (Mac) can also be applied. In this case, choose a comma or semicolon as the delimiter between the numbers because they are safer than using space or tab (that we cannot see on the screen). Two delimiters with no number between them (such as ; ;), or a delimiter at the beginning or the end of the line means missing data. This data matrix entered as a .txt file looks like this:

File name: *crocodiles.txt*

```
Group_ID; par_sp_1; par_sp_2; mass_kg
1; 1; 0; 1.25
1; 8; 0; 1.33
1; 0; 0;
1; 1; 1; 2.22
2; ; 4; 2.12
2; 0; 2; 0.92
2; 2; 11; 1.45
2; 0; 6; 1.34
3; 1; 1; 1.87
3; 0; 0; 3.01
3; 11; 0; 2.67
3; 0; 1; 0.99
```

← missing data

← missing data

---

Choose these options on the 'Import Data' screen:

- \* file type: text file
- \* field separator: semicolon
- \* variables in the first row: yes
- \* decimal point character: period

## **6. Using subsets of large data sheets**

When we have uploaded a large and complex data set (say, several sheets each with several columns), we can still choose to analyze a smaller subset of that (say, one or two columns of one sheet). In this case, we will get the statistical functions suited for that smaller data set.

## **7. Some technical details**

The 'Import data' screen allows for choosing the appropriate separator character (blank, tab, comma, semicolon, or other). If you use decimal numbers, also be sure to specify the decimal symbol on the same screen.

File size is limited to 200 kB. Would you need larger files to work with, contact us by e-mail ([parasite.ecology@gmail.com](mailto:parasite.ecology@gmail.com)). Avoid special characters (like ; ) and accented characters (like ä or ô) in file names and column titles. Dots and underscores work well to structure these names (e.g., 'viper\_females', 'killer.whale.winter').

## **8. Citation**

This is not a citable document. When using QPweb in scientific research, please, refer to it by citing the following paper:

Reiczigel J, Marozzi M, Fábíán I, Rózsa L 2019. Biostatistics for parasitologists – a primer to Quantitative Parasitology. *TRENDS IN PARASITOLOGY* 35 (4): 277-281.