Alexey Shipunov

Biometry: Biol 240 class
All laboratories

2019

Contents

Laboratory 1

1.1 Background

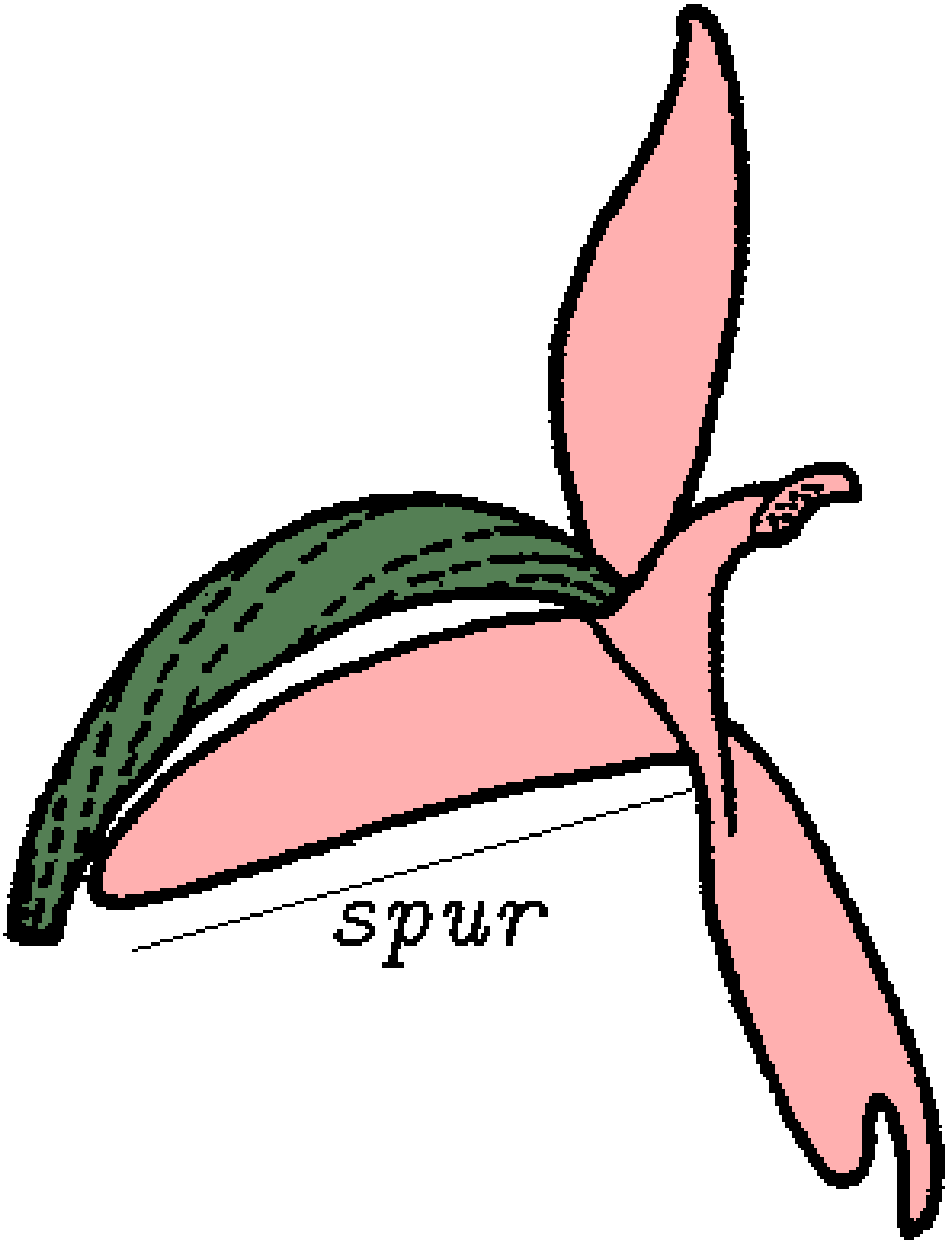
- All generalities like median and mean are taken from sample but should represent the whole statistical population. Therefore, it is possible that these *estimations could be seriously wrong*.

- Newer statistical techniques like *bootstrapping* are designed to minimize the risk of these errors.

- Bootstrap is based only on given sample but try to estimate the population. In fact, bootstrap term originated from Buerger & Raspe's "The Surprising Adventures of Baron Munchausen", where the main character pulls himself out of a swamp by his hair (specifically, his pigtail).
- Statistical bootstrap was actively promoted by Bradley Efron from 1970s but was not used frequently until 2000s because it is computationally intensive.
- We will approach the central tendency of this data with the **mean** and also with the more robust **median**.

1.2 Assignment

The data file is a result of measurements of spur length on 1511 *Dactylorhiza* orchid flowers. The length of spur is extremely important because only pollinators with mouth parts equal to the spur length can pollinate flowers.



Idea is simple: resample the initial sample 100 times, calculate given characteristic each time and finally calculate the average value of all these characteristics. The average will be a *bootstrap estimation*.

1. Open R, check and download the data file from Internet (address is `http://ashipunov.info/data/spur.txt`) into memory object "`spur`".

   Hint 1. Use the command `scan()` (because it is just one column of numbers).

2. Check structure, then calculate mean and median of `spur`.

   Hint 2. Use commands `str()`, `mean()` and `median()`.

3. Resample the whole data **with the replacement**.

   Hint 3. Use the command `sample(spur, length(spur), replace=TRUE)`.

4. Calculate mean and median of the sample. Open spreadsheet and keep results in columns: mean in one column, median in other column.

5. Repeat the previous step 99 times (yes, 99 times!). As a result, you will have two spreadsheet columns, each with 100 numbers.

6. Load these columns into R, preferably using clipboard.

   Hint 4. Use something like: (on Windows and Linux) `medians <- scan("clipboard")` or (on macOS) `medians <- scan(pipe("pbpaste"))`.

   (Another way to do the same is to enter `medians <- scan()`, then paste the clipboard content into R window and press `Enter`.)

7. Calculate the **median of medians** and **mean of means**. These two numbers are **bootstrap estimations** of mean and median of orchid `spur` data.

8. Report:

   (a) Your history of commands from R session. Use file name like to `myname_lab1.r`;

   (b) Your spreadsheet file with two columns. Use similar file name, like `myname_lab1.xls`;

   (c) **Four** numbers:

      i. Mean of `spur` data

      ii. Median of `spur` data

      iii. Mean of means column

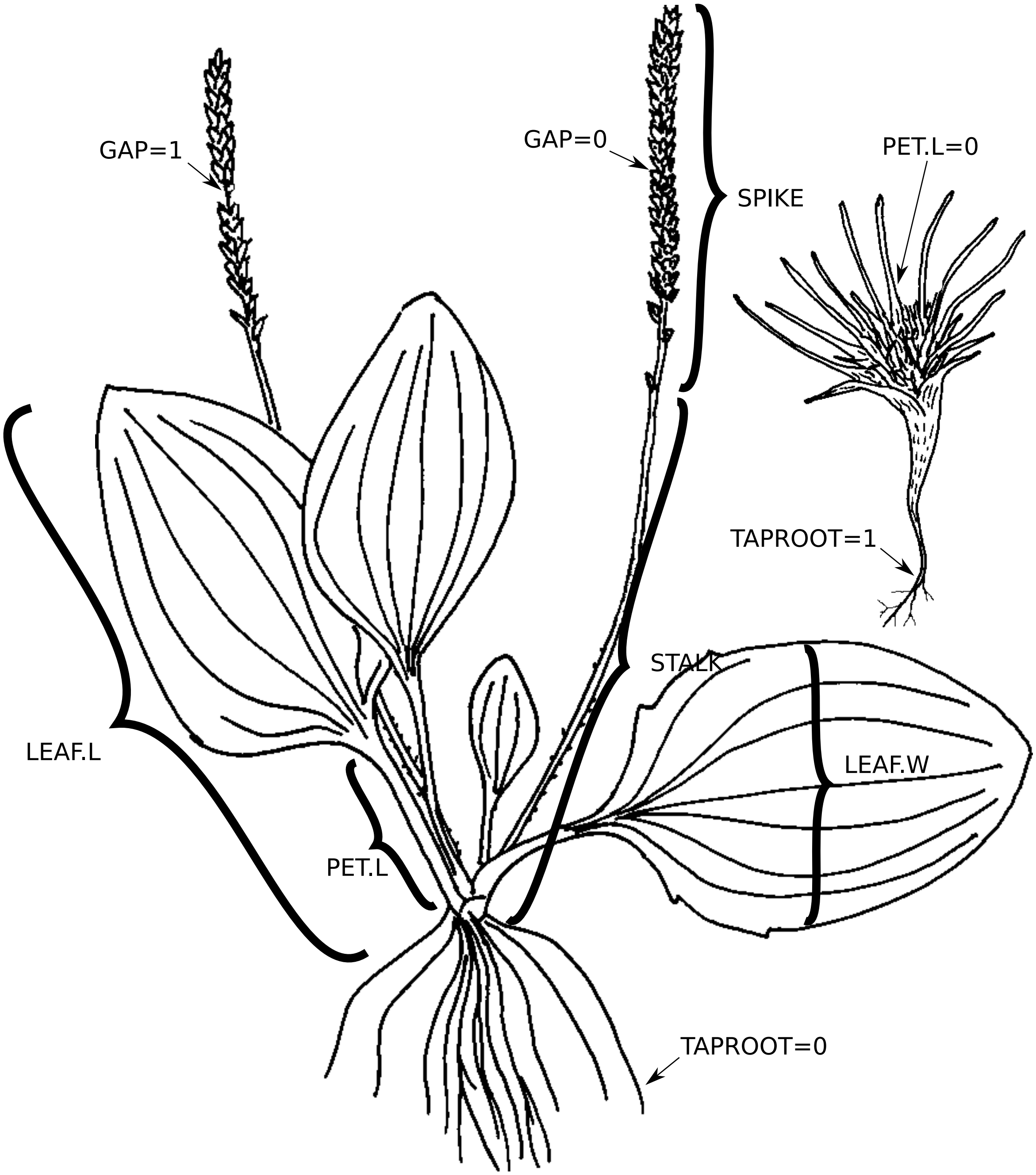
      iv. Median of medians column

   (d) Answer: are (1) **i** *vs.* **iii**, and (2) **ii** *vs.* **iv** different and why?

   Send everything to my e-mail address: alexey.shipunov@minotstateu.edu with the Subject: "Biol 240 Lab 1".

Laboratory 2

2.1 Background



Measurements on Plantago *roots, leaves and spikes*

- We will work today on the morphology of *Plantago* (plaintains), species-rich, worldwide groups of weeds, and transform "real-world data" into computer spreadsheet file for the future work in the class. We also will check how it loads into R, this allows to avoid possible problems.

2.2 Assignment

1. Get a sample (approximately 30 scanned herbarium sheets) and plastic ruler.

2. Enter data into Excel or any other spreadsheet. You might want to download the example of spreadsheet from `http://ashipunov.info/shipunov/school/biol_240/data/plantago.xls` . Your goal is to fill all columns.

3. **Choose** good, non-damaged, non-folded **maximal size** leaf, and the **longest** inflorescence. Then type in the spreadsheet the following:

   PLANT.ID is on the small piece of sticky paper attached to the herbarium sheet and photographed with it.

   Hint 5. IDs are in the form "P-[three_or_four_digits]", e.g. "P-1901" or "P-058". When typing, keep the "P" uppercase.

   MM.20 The length of 20 mm from a photographed, virtual ruler, in real mm from your physical ruler. Use only the millimeter side of rulers!

   Hint 6. Take a plastic ruler. Find a photographed ruler. Put their *millimeter sides* close together. Now measure how long on the real plastic ruler are 20 small tickmarks (20 mm) from photograph. This is your number. Normally, it should be less than 20.

   LEAF.L Length of leaf (from stem base to leaf top), mm

   LEAF.W Maximal width of leaf, mm

   PET.L Length of petiole (from stem base to the place where the leaf *suddenly* widens), mm.

   Hint 7. If there is no obvious petiole, type "0".

   STALK Length of maximal stalk (from stem base to lowest flower), mm

   SPIKE Length of spike on maximal stalk (from lowest flower to the top), mm

   GAP Is there any empty space (stem visible) between flowers near the *middle* of spike? **0** if no, **1** if yes.

   TAPROOT Is taproot present? **0** if no, **1** if yes.

4. Save the file as a *tab-separated text file* in the `data` subdirectory of your working directory. File name should be like `yourname_lab2.txt`

5. Load this file into R object and check its structure.

   Hint 8. Use the `read.table("data/...", head=T)` and `str(...)` commands.

   Hint 9. There should be **nine** columns entitled exactly with names listed above. First column should be in the "factor" mode, others in "num" (numeric) mode.

6. If there are no problems, your data file is ready.

7. Report:

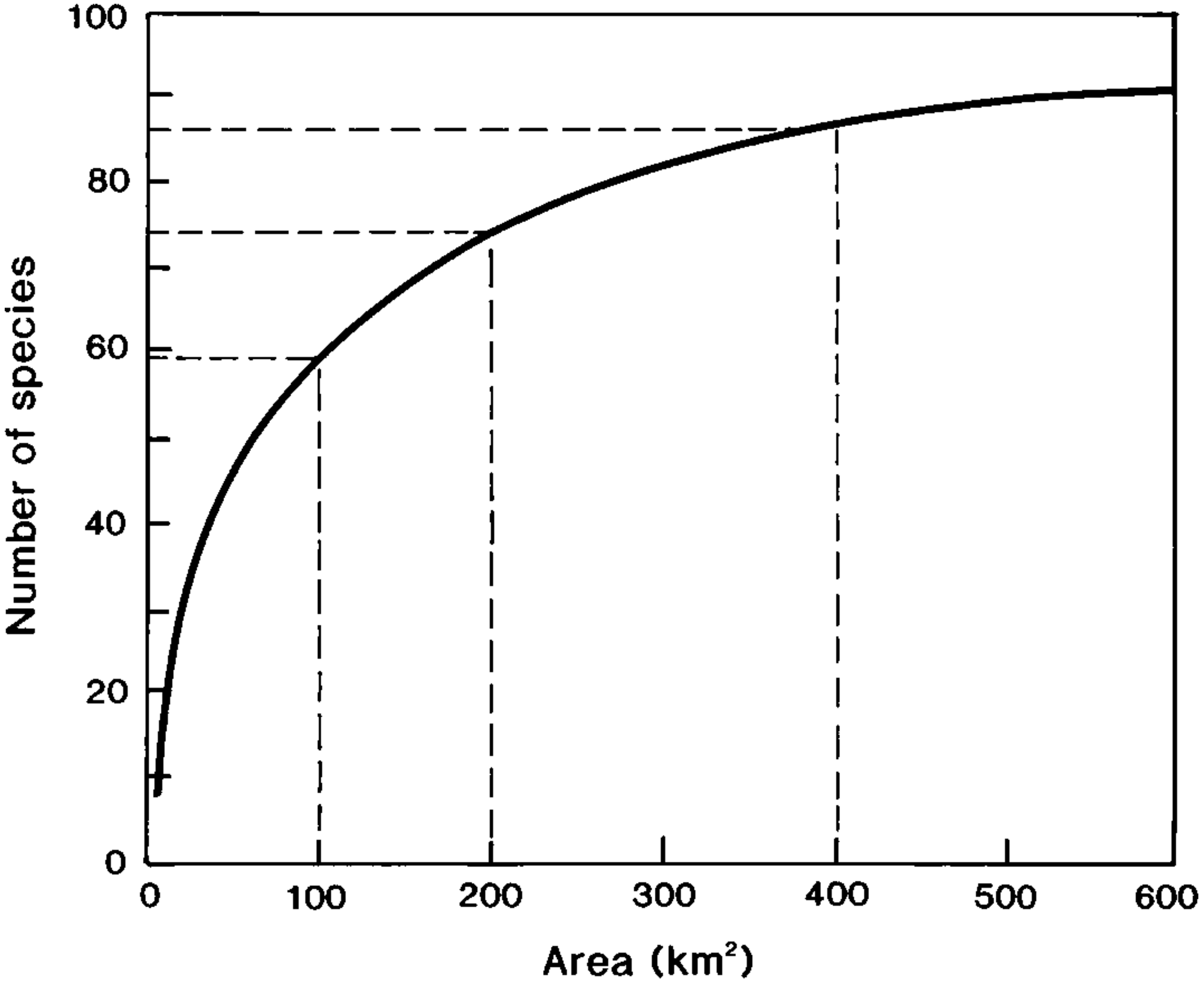
   (a) Attach to the e-mail your text file saved from spreadsheet.

(b) Send it to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: “Biol 240 Lab 2”.

Laboratory 3

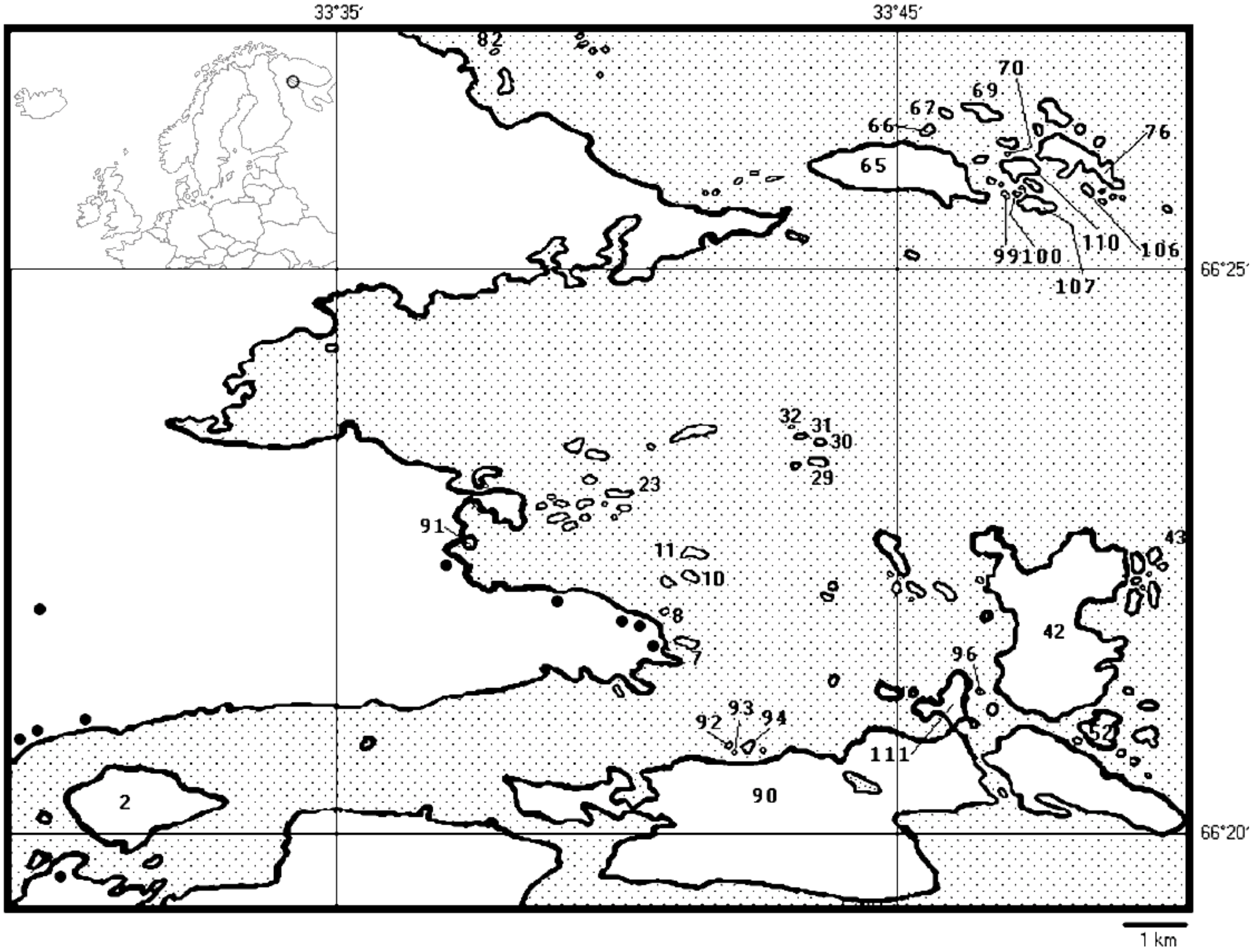
3.1 Background

- Sometimes, graphic analysis is enough to make some conclusions about data. This is the type of work to do today. Of course, these conclusions are only preliminary and still require inferential assessment.
- The island biogeography states that the number of species occurring on island should grow along with the size of island. This is a key point of famous McArthur and Wilson (1967) book.



The theoretical species-area curve.

- From 1964 to 2011, group of botanists and zoologists studied small islands of White Sea nearby Kola peninsula (Russian Arctic). In all, we reviewed more than 100 small islands and islets. On all islands, we registered all plant species.
- The island biogeography theory was based mostly on large tropical islands. Let us check if there is a relation between area of small islands and number of species in Arctic. If there is a relation, then you will see the "virtual curve" (cloud of points) similar to the picture above.



The map of studied islands.

3.2 Assignment

1. Open R, and read the data from Internet: `http://ashipunov.info/shipunov/school/biol_240/data/islets.txt` into some new object. Immediately after reading, check the structure of object.

   Hint 10. Do not forget to check the file with browser or with `url.show()`. To read data, use `read.table()`. To check structure, use `str()` and `head()`.

2. The data file contains a subset of our island data: each of 103 data rows corresponds with some island, header row contains headers of three columns: **length** of island (in steps, step is ≈ 0.6 m), width of **island** (in steps) and number of plant **species**.

3. Calculate the approximate square of each island. To do that, you will need to *multiple* both first columns to 0.6 and then *multiple* one of these columns to another. It will give you the squares of islands in *square meters*. Write the result into a vector.

   Hint 11. If your data frame has a name `data`, then `data[,1]` will give you the first column.

4. Make a scatterplot where your square will be the the horizontal axis and number of species—vertical axis.

   Hint 12. Use `plot()`. If your square has name `x`, and number of species has name `y`, then two variants are possible, either `plot(y ~ x)`, or `plot(x, y)`. Both will give you the scatterplot required. Note the reversed order of arguments.

   Hint 13. You do not have to, but if you "prettify" this plot (that is, make proper axes labels, plot title and so on), you may receive extra points.

5. Save the plot into PDF file.

   Hint 14. Do not forget to use `pdf()` and `dev.off()` commands

6. **Answer the question**:

   Based on the plot, is the theory of island biogeography applicable to small Arctic islands?
   Why or why not?

3.3 Report

For the report I will need today (1) your answer and (2) your **R script**. *This R script should download data, make all calculations and create your PDF plot itself.* This is how to make the script:

1. Save your history of commands, just in case (on macOS, save the whole content of R console).

2. Then copy-paste all necessary commands from R console into the text editor (e.g., open blank file in the R internal text editor on macOS or Windows, or call external text editor on Linux).

   Hint 15. You can use menu or `file.edit()` command.

   Hint 16. Try **not** to use in the script any interactive commands (like `file.show()`); reduce also to the minimum commands which do not create objects (i.e., commands without assignment).

3. Save your script under the name similar to `myname_lab3.r`. Do not forget to change script name in accordance with **your** name.

4. Close R, do **not** save workspace.

5. Make a `test` directory inside your working directory, or (if it already exists) *delete* it (with all contents) and then *make again*.

6. **Copy** (do not move!!!) your script into `test` directory. The master version of your script must stay outside of `test`.

7. Start R, make `test` your working directory.

8. Run your script from within R via `source("myname_lab3.r", echo=TRUE)`.

   If everything is OK, then data will be downloaded, calculations run, and your PDF file with a plot will appear in the test directory. To double check that, open this PDF from within your file manager.

   If anything is wrong, close R and *delete* `test` directory *with all its contents*. Then, open script in text editor and try to find a mistake (this is called "debugging")[1]. Correct your script, close R and repeat. Do it until everything runs well.

Send (1) your answer and (2) attached script to my e-mail address:`alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 3".

[1] You might want to open R temporarily and copy/paste/check your commands one by one.

Laboratory 4

4.1 Background

- This lab has a simple goal, train you with as many R commands as possible. Knowledge-based, we will go far ahead with this training, and run commands which do correlation, linear regression an so on. Please do not worry, we will cover these themes in depth later.
- What is important, is the proper *typing* and *catching* results, both *textual* and *graphical*. These are skills to improve today.

4.2 Assignment

1. Open the textbook on Appendix A "Example of R session". Read the text and type all commands, one by one.

   Hint 17. You can use either `download.file()` approach or URL approach (if the latter, ignore `data` directory creation and replace `data/` with URL, see textbook foreword).

2. Save all plots which you make with `pdf(...)`/`dev.off()` commands.
3. Save your history of commands.
4. There are *two intended mistakes* in the text. One is syntactic (R should react on it), the other is a mere inconsistency between text and R example supplied. Find them.
5. **Report**. As an exception, I do not want R script today. Instead, I want:

   (a) Your command history as attachment.

   (b) All your plots. To speed up your and my work, please pack them in one zipped folder with your file manager, and attach *only one* zip file to the email.

   (c) Description of mistakes found—what are they?

   (d) Send everything to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 4".

Laboratory 5

5.1 Background

- *Aronia* (chokeberry) is a small (different botanists accept from one to five species) North American genus from rose family (Rosaceae). They are well known wild berries in Atlantic states.

- One species, *Aronia mitschurinii*, was somehow originated in Russia in experiments of the famous practitioner of plant selection Ivan Michurin, and it is now re-introduced and started to be widely cultivated as agriculturally promising fruit plant, especially useful in northern regions.

- It is unclear if *Aronia mitschurinii* is a hybrid, and if yes, hybrid of which species. It is also unknown how many species have the genus *Aronia*. DNA and morphological research, taken together, can possibly help in that situation. The goal of today's lab is the exploratory graphical analysis of chokeberry data.

5.2 Assignment

1. The data file contains *15 columns*. To see the meaning of each column, check *description file*.

Hint 18. Either `url.show()` the description file with R, or look on it in the browser window. The URL of description file is `http://ashipunov.info/shipunov/school/biol_240/data/aronia_c.txt`

2. Open R, check the data file (URL is `http://ashipunov.info/shipunov/school/biol_240/data/aronia.txt`) and load it into the R object. Do it directly (with reading URL) or indirectly (download first).

3. Explore the data frame using `str()` and `head()` commands, be sure that separator and header used in a right way.

   Hint 19. Instead of `str()`, you might want to use `shipunov::Str()` which shows numbers of columns and presence of `NA`s. If you loaded `library(shipunov)`, then use it simply as `Str()`.

4. Split the "distance from the base of the smaller branch to the base of biggest leaf" column into three intervals, and plot this new ranked variable. Write the barplot to PDF. Answer:

   Which position of leaves is prevalent: close, distant or intermediate?

   Hint 20. To split, use the `cut()` function.

   Hint 21. Use parameters `main=""`, `xlab=""`, `ylab=""` to make *all* your plots prettier. You do not receive extra points for this anymore. Instead, *points will be taken from you if plot is not prettified.*

5. Make a table and then a dotchart from the variable containing number of secondary veins, save to PDF. Answer:

   What are the most frequent and most rare numbers of secondary veins?

   Hint 22. To make dotchart from the 1-dimensional table, do something like `shipunov::Dotchart(table(aa))`

6. Make the scatterplot from (i) length of petiole and (ii) position of the leaf maximal width, write to PDF. Answer:

   Does the length of petiole relate with the position of maximal leaf width?

   Hint 23. You might want to use something like `na.omit(a[,c("PET.L", "LEAF.L")])` to remove `NA`s from these two variables only. However, this does not affect result.

   Hint 24. Use `scatter.smooth()` instead of `plot()`.

7. Make the similar plot on the other two variables: (iii) length of leaf and (iv) width of leaf. Color the dots by species. Make a legend. Save the scatterplot to the PDF. Answer:

   Is Aronia mitschurinii *different from three other species by length and width of its leaves?*

Hint 25. To avoid `NA`s, you might use something like `na.omit(a[,c("LEAF.L", "LEAF.W", "SPECIES")])` first.

Hint 26. To plot dots with different colors in accordance with species, use something like `plot(bb$LEAF.L, bb$LEAF.W, col=bb$SPECIES)`

Hint 27. To make a legend proper, use something like `legend("topleft", legend=levels(bb$SPECIES), col=1:nlevels(bb$SPECIES), pch=1)`; see also *answer to flowering heads example* in textbook (you do not need to drop factor levels here though).

Hint 28. Finally, if you find how to use `shipunov::PPoints()` function here (again, see in textbook), you will receive extra points.

8. Report:

   (a) Four answers

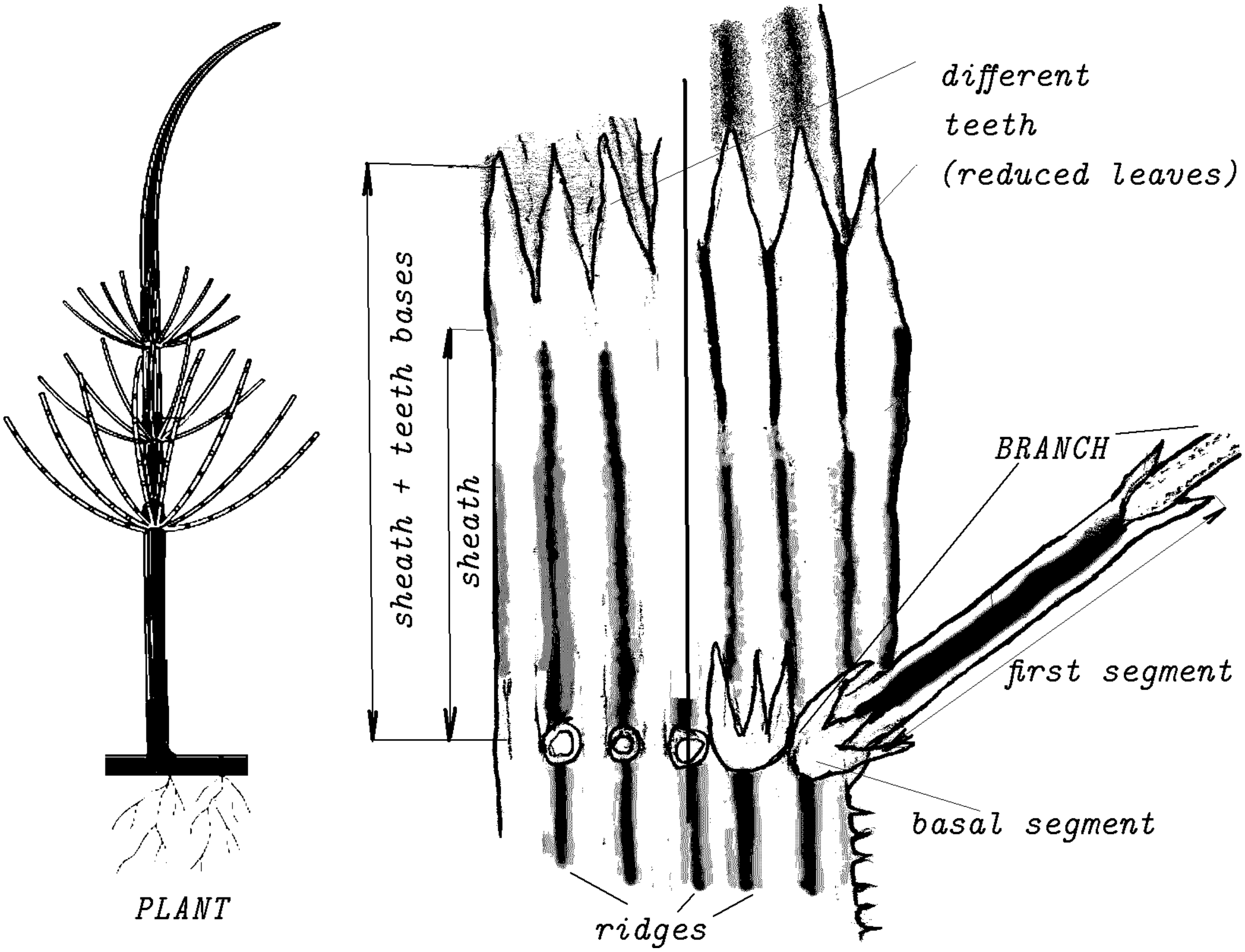
   (b) R script which downloads data and makes plots

9. Send your answers and file attachment to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 5".

Laboratory 6

6.1 Background

- Horsetails (*Equisetum*) are the old, pre-dinosaur plant lineage. Only several dozen species survived, but despite a long evolution the borders between these species are still unclear for researchers.
- In 2005–2006, morphometric analysis was performed of more than 1,000 horsetail plants belong to most widespread Eurasian species. For the analysis, we used 8 morphological characters and also tried to identify species.



Horsetails and their morphological characters.

- **The goal** of today's lab will be to compare two species of horsetails represented in data: *Equisetum* *arvense* and *Equisetum fluviatile.* You will need to find how to distinguish these two species on the basis of studied characters.

6.2 Assignment

1. Open R, download the data file from Internet (URL is `http://ashipunov.info/data/eq.txt`), load it into the R object.

   Hint 29. To see names and meanings of characters, use the companion file `eq_c.txt` (same URL) and the diagram above.

2. Explore the data frame using `Str()` (`shipunov` package) and `head()`.

3. Using `summary()` and `boxplot()` for both species:

   (1) Choose two *characters which you think are* most distinguishable*. How did you decide?*

   Save paired boxplots of these characters into PDFs.

   Hint 30. Refer to the waterlilies exercise from textbook.

   Hint 31. When choosing characters, use `dev.new()` to keep several graphic windows, or `Boxplots()` function from `shipunov` package.

   Hint 32. To make paired boxplots, you might want to use *formula* approach like `boxplot(DL.R ~ SPECIES, data=mydata)`

   Hint 33. Where it is possible, use parameters `main`, `xlab`, `ylab` and colors to improve plots. If you do not, I will take points out.

4. *For each of two* distinguishable characters chosen, decide if they are **normal or not**.

   Hint 34. To decide, use histograms, embedded distributions like `rnorm()` and quantile-quantile plots.

   Hint 35. You may also learn how to use `Normality()` function from `shipunov` package and employ it.

   Hint 36. Do it *per character per species*, similarly to the waterlilies exercise in textbook.

5. Then apply those measures of (1) central tendency and (2) variation *which are applicable most* to these two characters. For central tendencies, **calculate** also 95% confidence intervals. **Save them** to the report.

   (2) Using the most compact way*, report central tendencies (with confidence intervals) and variations* per character per species*. Which measures did you use? Why?*

   Hint 37. At the end, you will have 8 single numbers (center and variation × 2 characters × 2 species) and 4 confidence intervals (4 pairs of numbers).

6. Answer, using your plots, central tendencies and, most importantly, *confidence intervals*:

   (3) Which of two selected characters is the best? Why?

7. Report:

   (a) Your R script which downloads data, runs necessary calculations and tests, and makes two PDF plots.

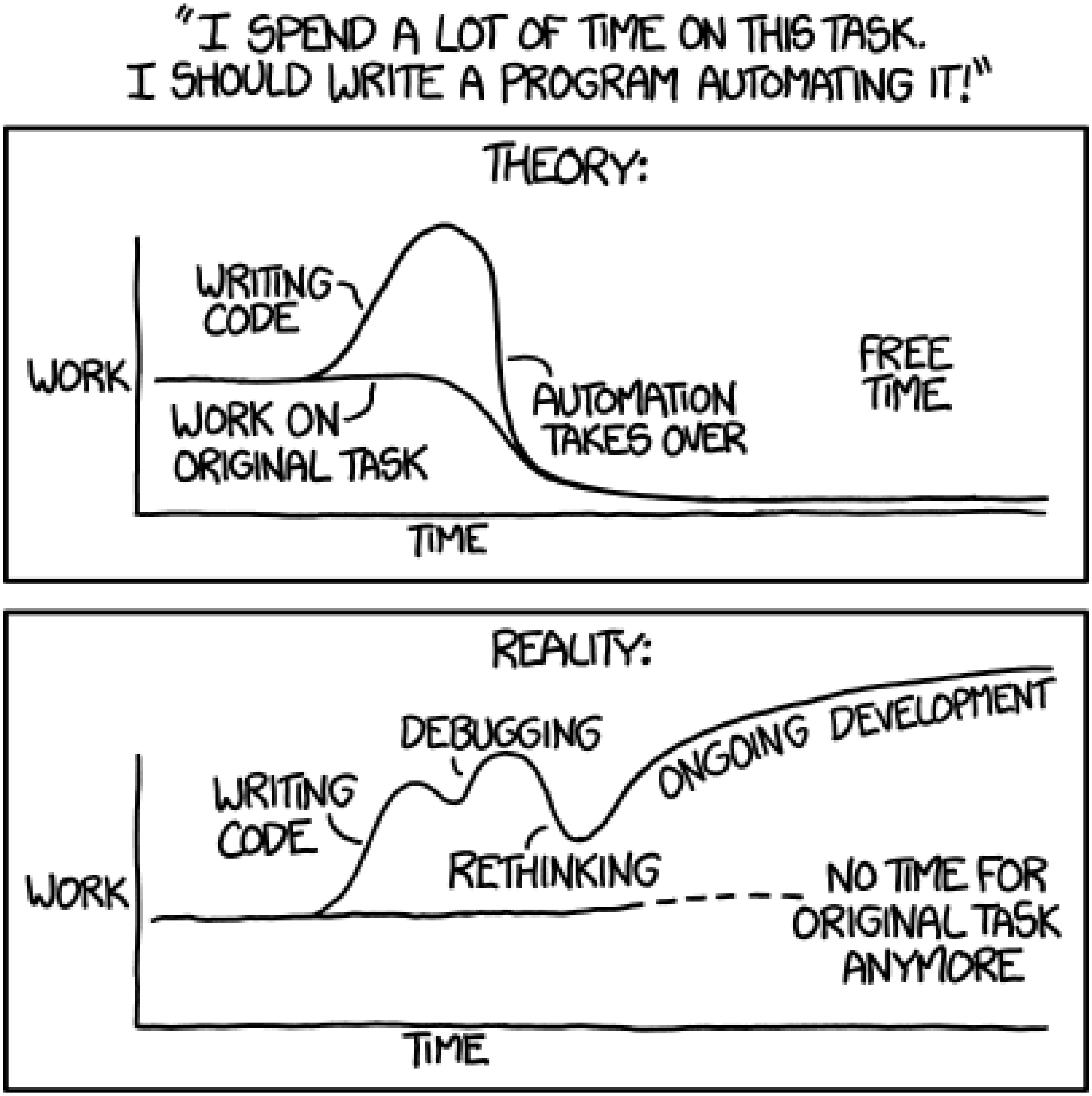
   (b) Your answers.

8. Send your answers and file attachment to my e-mail address:`alexey.shipunov@minotstateu.edu` with the Subject: “Biol 240 Lab 6”.

Laboratory 7

7.1 Background

- Often we need to calculate something complicated many times. Solution is to *automate* it, make it **function**. In fact, most of R consist of such functions.



(From `http://xkcd.com/1319/`)

- Today we will write the function which will automate the calculation of the relative *variability of median*. Relative variability of median might be calculated similarly to the coefficient of variation (CV) which is explained in textbook as exercise, but from median and IQR:

  $$\mathrm{VM} = \frac{\mathrm{IQR}}{\mathrm{median}} \times 100\%$$

  R has no specific function for VM so you need to make it.

- We will use the specific dataset to test if functions are working: data which came from the Ethnobotany lab were two groups of students counted different types pollen grains present inside flax flowers. Both groups worked with the same objects, so their results are directly comparable. Different columns belong to the same flowers, so again, it is easy to use them as *paired* data.

7.2 Assignment

1. Open R, and download `linum` data from `http://ashipunov.info/shipunov/school/biol_240/data/linum.txt`. Companion file with expanations is located on `http://ashipunov.info/shipunov/school/biol_240/data/linum_c.txt`. First column of `linum` data is the group number, three other columns are counts of different types of pollen grains.

2. Write the prototype of `VM()` function.

   Hint 38. Make it similar to the `CV()` function from the textbook. Then, using `fix()` command, modify `VM()` to calculate variability of median instead of CV.

   This is called *debugging* and is a cyclic process: you will modify function, apply it to the data, check how it works, modify again and so on until results become satisfactory.

3. Apply this function to `linum` data to answer:

   Which group of students did count pollen less accurately?

   (Or, in other words, where the variability of median is higher,
   in the data from the *first* group, or in the data from the *second* group?)

   Hint 39. To apply function simultaneously for several columns and two subsets of data, you might want to use something like `aggregate(linum, by=list(linum$GROUP), MYFUNCTION)` which splits one of your columns (first) or the whole data frame (second) on the fly.

   BONUS (extra points available): If you have extra time, load the other data file, `linum_short.txt` which is almost the same data but in *short format* (i.e. each column represents the smallest piece of data: one group and one type of observation). **Apply** your function to this data (find out how, this is simpler then above).

4. As a report I will need today:

   (a) Your **R script** which contains definition of your function and all necessary calculations to support your answer.

   Hint 40. When you make the script, you need to insert there the whole function definition. The best way to obtain it is just (1) **type** function name without parentheses, (2) **copy-paste** output into the script and (3) in the beginning, **add** function name and assignment operator.

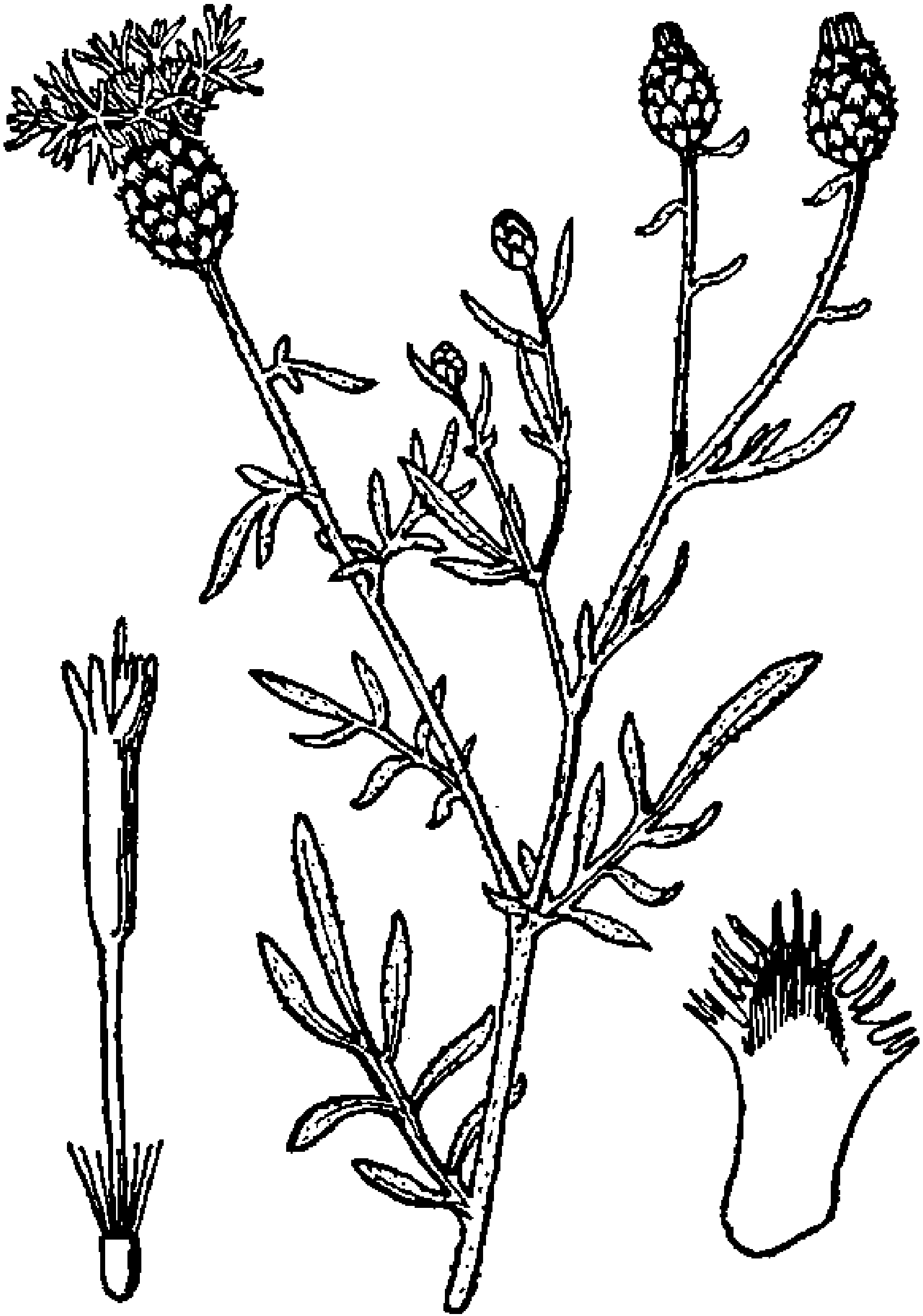
(b) Answer. Please do not forget to provide numerical arguments.

5. Send your report to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 7".

Laboratory 8

8.1 Background

- Knapweeds (*Centaurea stoebe*) are important invasive plants in North America. It is still unclear what make knapweeds so successful invaders. One of hypotheses is that knapweed extract chemicals suppressing the growth of native American plants ("novel weapons" hypothesis). It is also possible that knapweeds employ symbiotic endophyte fungi to produce these chemicals.



Spotted knapweed

- In 2007, the large experiment has been performed in order to receive a support for the idea that sterile knapweed plants inoculated by its own fungi will suppress the growth of native American grass, fescue *Festuca idahoensis*. We inoculated knapweed seedlings with given fungus and transplanted them to pots with five fescue plants as neighbors. After about 18 weeks, we harvested plants and measured the dry biomass of fescues and knapweeds.

8.2 Assignment

1. Today, you will need to analyze the part of this data. Only two variants are present in data file `knapweed1.txt`, sterile control plants (marked as CID 0) and plants inoculated with fungus *Cladosporium herbarum* (CID 63). Variables in the data file are described in the companion file, URL is `http://ashipunov.info/shipunov/school/biol_240/data/knapweed_c.txt`

2. **The goal** of today's lab is to provide inferential answers for the following six questions:

   (a) Are dry weights of knapweed different between inoculated and sterile plants?

   (b) Are dry weights of fescue different between pots contained inoculated and sterile knapweed plants?

   (c) Are stem lengths on tenth week different between inoculated and sterile plants?

   (d) Are stem lengths on eighteenth week different between inoculated and sterile plants?

   (e) Are numbers of flowering heads different between inoculated and sterile plants?

   (f) Are stem lengths on eighteenth week differ from stem lengths on harvest time?

3. Open R, download the data file from Internet (address is `http://ashipunov.info/shipunov/school/biol_240/data/knapweed.txt`), load it into the R object.

4. Explore the data frame using `str()` and `head()`.

5. Using plots, `Normality()` function and common sense, *decide* which kind of two-sample test to use in each of six cases.

   Hint 41. You have a choice between `t.test()` and `wilcox.test()`.

   Hint 42. Also, consider `paired=TRUE` or `paired=FALSE`

6. Run six tests and answer questions from the first part. To answer each question, use p-values.

   Hint 43. Since the data file is already in *long form,* the easiest way is to use *model formula* (like `t.test(a ~ CID, data=x)` where `a` is a name of tested variable and `x` is a name of your data object).

7. For every test, make a boxplot with proper labels and save it into PDF file

   Hint 44. If you are using the model formulae, you simply need to change name of command to `boxplot()`.

8. Report:

   (a) Your answers for six questions with supporting data (like p-values, degrees of freedom, statistic values—all with interpretation)

   (b) The R **script** which will download data, check normality, make 5 boxplots and 5 tests.

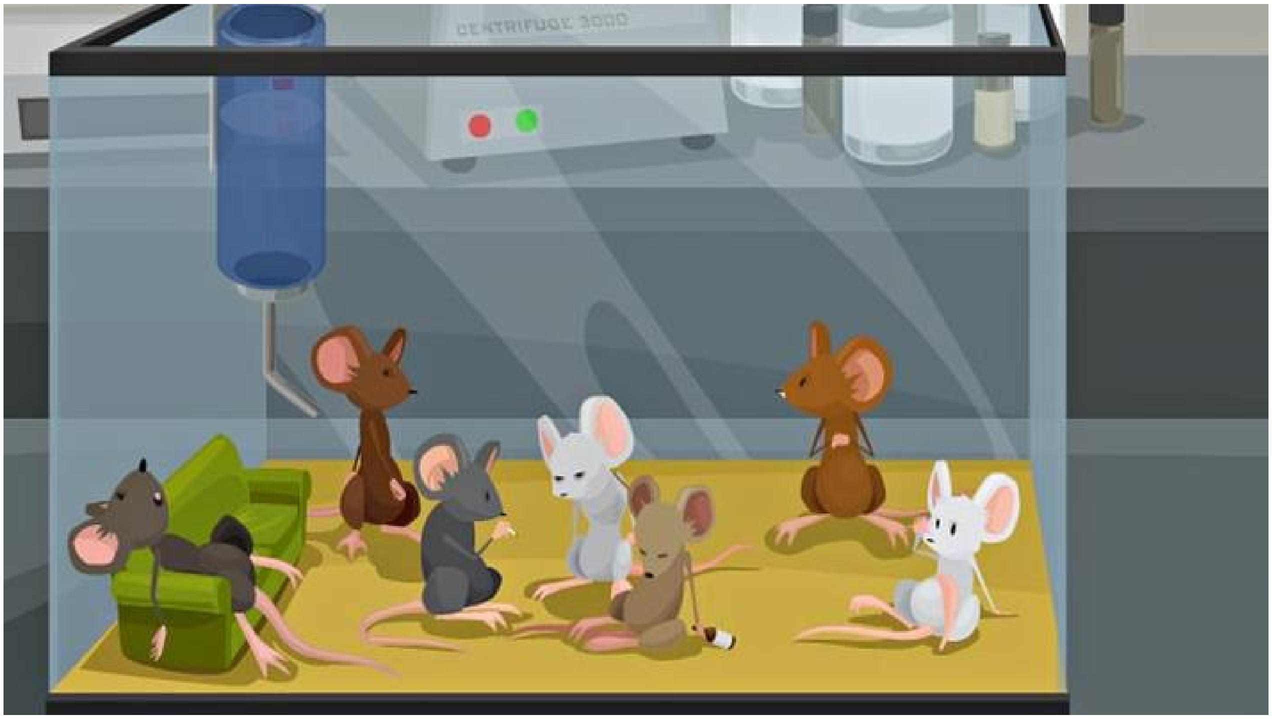
   Hint 45. Be sure to **test your script** properly! If you do not do that, you may loose your points.

9. Send your answers and R script as an attachment to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 8".

Laboratory 9

9.1 Background

- Drug addiction is a complicated phenomenon. It could be (for example) facilitated by multiple external factors which include *social interactions*. In turn, addiction-related behavior would influence social interactions back again.



Addicted mice (from: "Mouse Party",
`http://learn.genetics.utah.edu/content/addiction/mouse/`)

- Dr. Shabani (former MSU faculty, now at Grand Valley State University, MI) studied the relation between sexual pairing and drug addiction. There were three groups of mice:

  1. Females paired with males
  2. Females paired with females
  3. Not paired, single females

  These mice were given methamphetamine in order to understand if there is an effect of pairing on the drug addiction. Consequently, the response was how much methamphetamine solution is consumed on the particular day of the experiment.

- Our goal today is to perform the analysis of variation (ANOVA) to understand if there is any effect, and then run the appropriate *post hoc* tests to check *which* group(s) is divergent.

9.2 Assignment

1. Open R, download the data file from Internet (address is `http://ashipunov.info/shipunov/school/biol_240/data/pairing.txt`), read it into R, explore with `str()` and/or `head()`.

   Hint 46. There are `GROUP` and `UPTAKE` columns (see above for the explanation).

2. Check assumptions and make conclusions.

3. With ANOVA or its robust/non-parametric analogs, answer **if drug solution uptake is different between groups**. Make necessary plot(s).

4. With the appropriate *post hoc* test, answer **which group has a significant difference(s) from which**.

5. Report:

   (a) Your R script which downloads everything and makes all calculations and plots (if you decide to supply them).

   (b) Answers on the two questions.

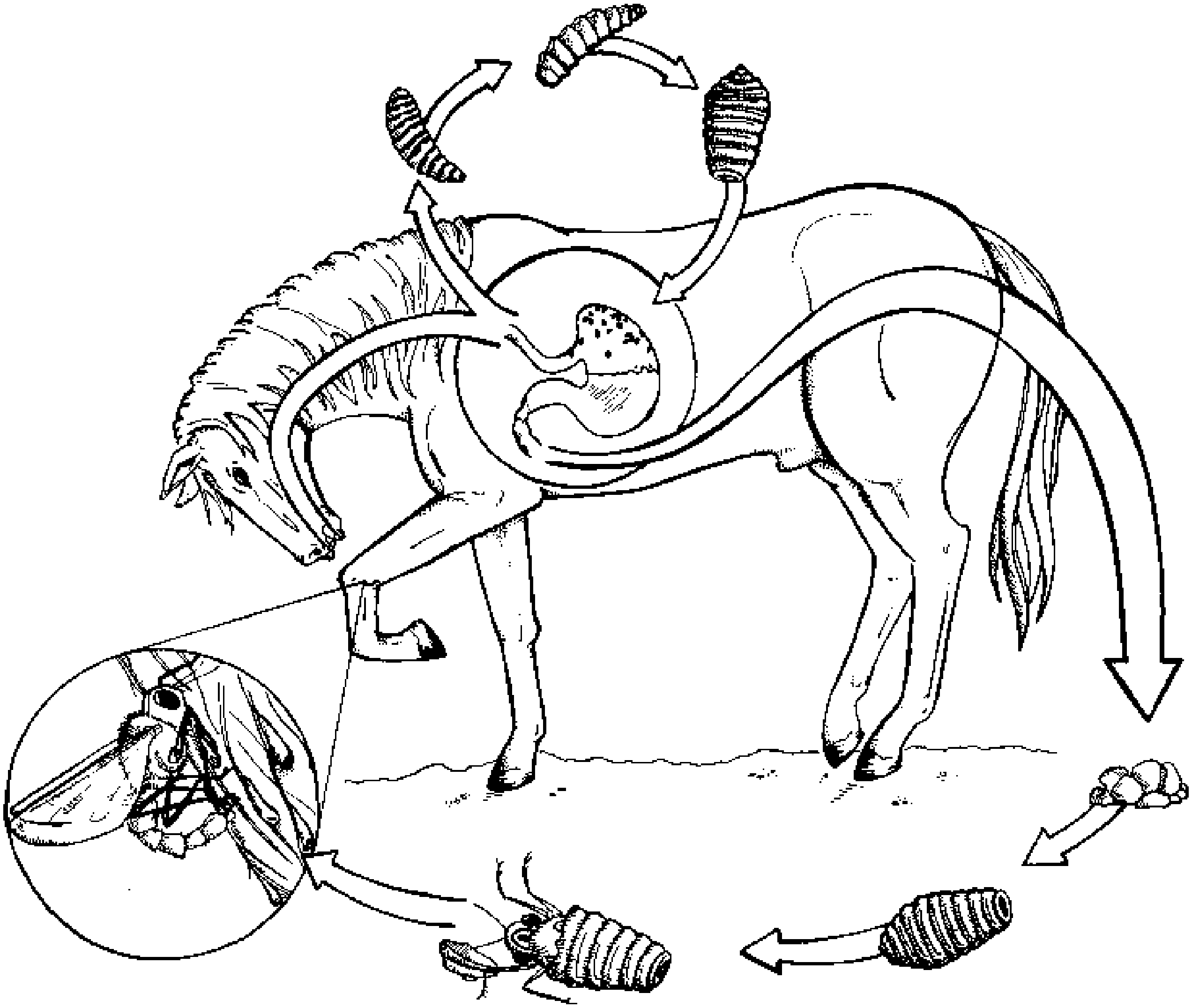
6. Send your answers and file attachments to my e-mail address:`alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 9".

Laboratory 10

10.1 Background

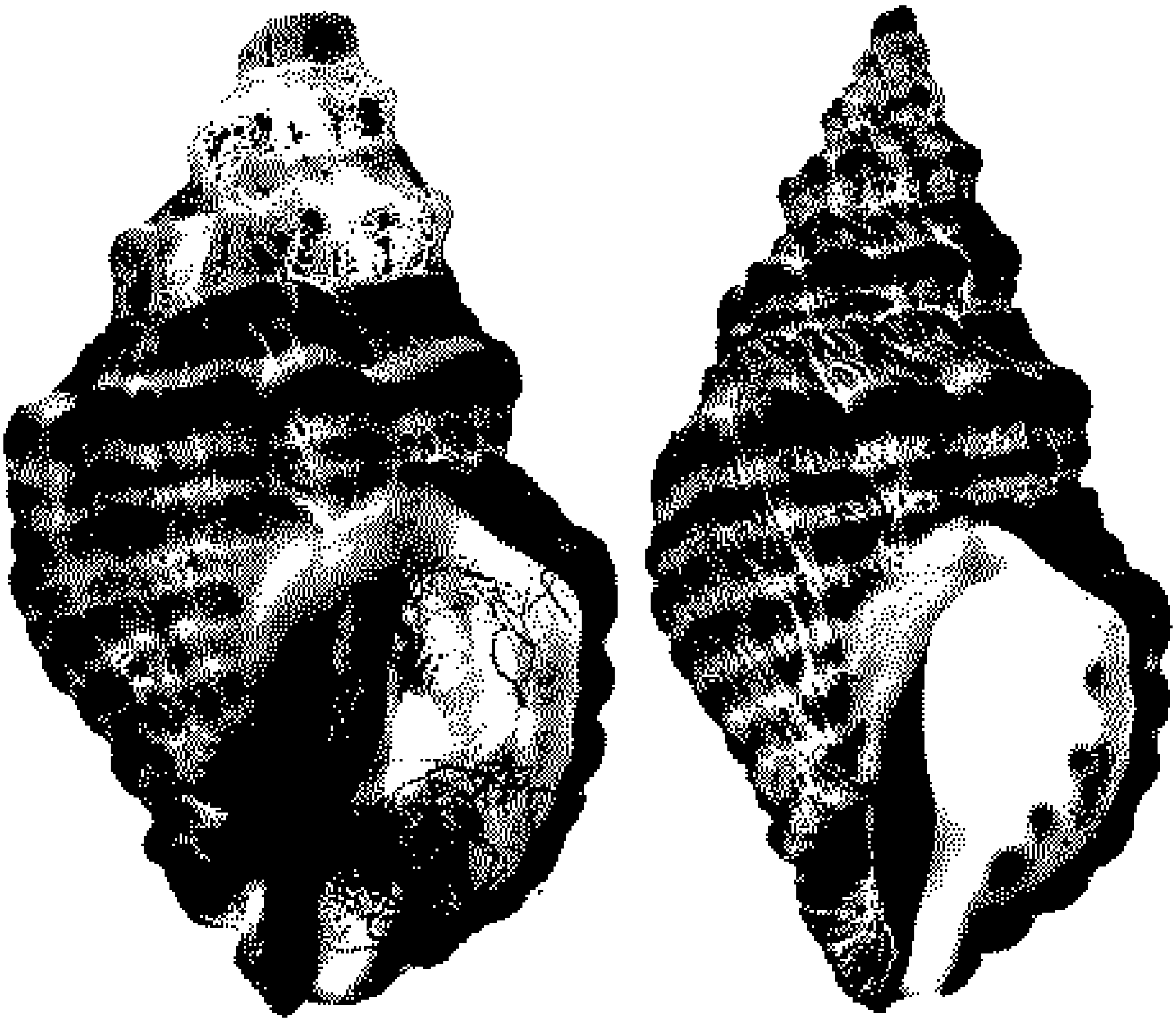
Today's data is split in two parts:

- Bot fly larvae is a parasite of mammals. Their larvae develop under the skin leaving an easily identifiable wound. This dataset contains records on small mammal captures at six research sites (first variable) in New York state. The dataset includes research site, species of mammal (`PL` = *Peromyscus leucopus*, the white-footed mouse and `TS` = *Tamias striatus*, the eastern chipmunk), and whether there was any sign of bot fly infection (`Y` or `N`).



Life cycle of bot fly (please do not mind a horse `;-)`)

- Egg capsules of the predatory gastropod *Lepsiella vinosa* where collected from the littorinid (more shallow) and mussel (more deep) zones on a rocky tidal shore of Australia. Egg capsules were dissected and the number of eggs counted. The variables in the dataset are `ZONE`—where the egg capsule was collected from (`Littor` or `Mussel`) and `EGGS`—the number of eggs found in the egg capsule.



The shell of southern Australian whelk, *Lepsiella vinosa*

10.2 Assignment

1. Open R, download data (`http://ashipunov.info/shipunov/school/biol_240/data/bots1.csv` and `http://ashipunov.info/shipunov/school/biol_240/data/gastropod.csv`), load them into R objects, explore your data.

   Hint 47. Data files could be in unusual format, please find out how to load them. Since data came from external source, there are no companion files.

2. Answer the following questions about bot fly data:

(a) **Is there the evidence that two studied mammal species differ in their infestation rates? How big is the difference, if any?**

(b) **Is there the evidence that research sites differ in their infestation rates? How big is the difference, if any?**

Hint 48. Make two tables of two variable pairs (function `table()`), test the equivalence (independence) of factor distribution with `chisq.test()`

3. **Ask table-based questions about mollusk data and answer them using statistical approach(es). These questions should concern (c) association, (d) overall effect and (e) *post hoc* (pairwise) test**.

   Hint 49. You might split egg numbers into *three* intervals with `cut()` function, give them names "`small`", "`medium`" and "`large`"; then make a table of this new variable and zone variable and use `chisq.test()` and friends to answer your questions.

4. For every pair of proportion or chi-tested variables, make a PDF plot of the appropriate kind (mosaic plot, spineplot, association plot, fourfold plot).

5. Report:

   (a) Your answers based on p-values, effect sizes and *post hoc* results.

   (b) Your R script.

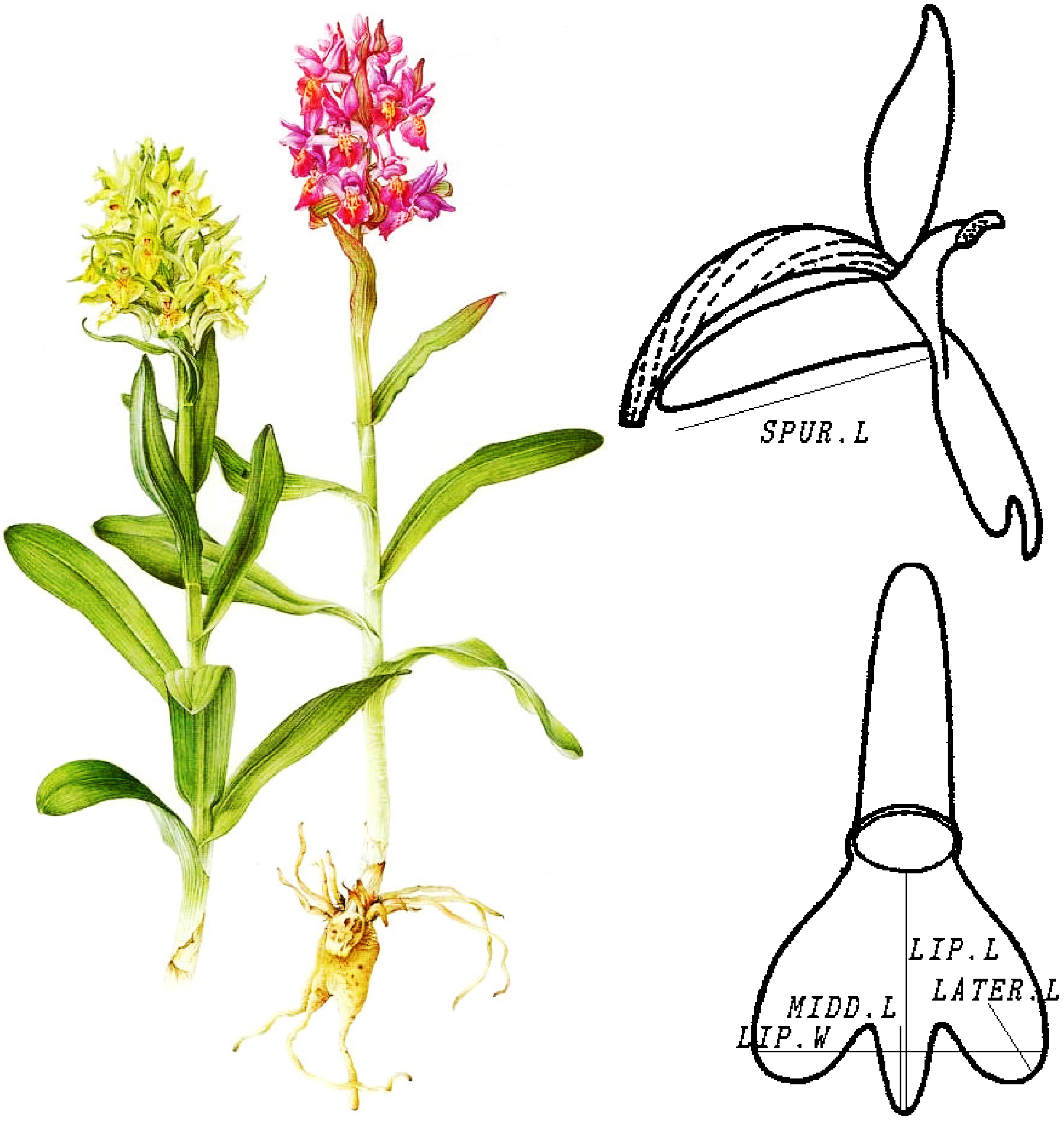
6. Send your answers and script attachment to my e-mail address:`alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 10".

Laboratory 11

11.1 Background

1. The data file today came from the morphometric observations on Eurasian dactylorhids (*Dactylorhiza*), terrestrial orchids which normally grow in forests and wet meadows. Hundreds of populations from England to Siberia were observed in 2002–2006. Two species were selected for today's analysis.

2. Our data has 11 variables which are described in the companion file and on the picture below.



Dactylorhids (*Dactylorhiza* spp.) and their morphological characters

11.2 Assignment

1. Open R, download the data (`http://ashipunov.info/shipunov/school/biol_240/data/dact2.txt`), load it into the R object. Check the companion file `dact2_c.txt`.

2. Explore the data frame, check normality for every measurement character.

3. Answer the following questions (do not forget to supply numerical arguments based on p-values):

   (a) **Is leaf width different between *Dactylorhiza incarnata* and *D. maculata*?**
   Hint 50. Here you should supply p-value and effect size.

   (b) **Does the association exist between lip color and presence of leaf spots?**
   Hint 51. Please also supply p-value and effect size.

   (c) **Which pair of measurement characters is most correlated? Is this correlation significant?**
   Hint 52. Do not forget p-value!

   (d) **Make the linear model for those two most correlated characters. How good is the model, is it significant? If yes, what is the shape of dependence?**
   Hint 53. To answer, use both model summary and diagnostic plots.

4. Please supply one plot per question, 4 plots in total. Do not include diagnostic plots in the script.

5. Report:

   (a) R script

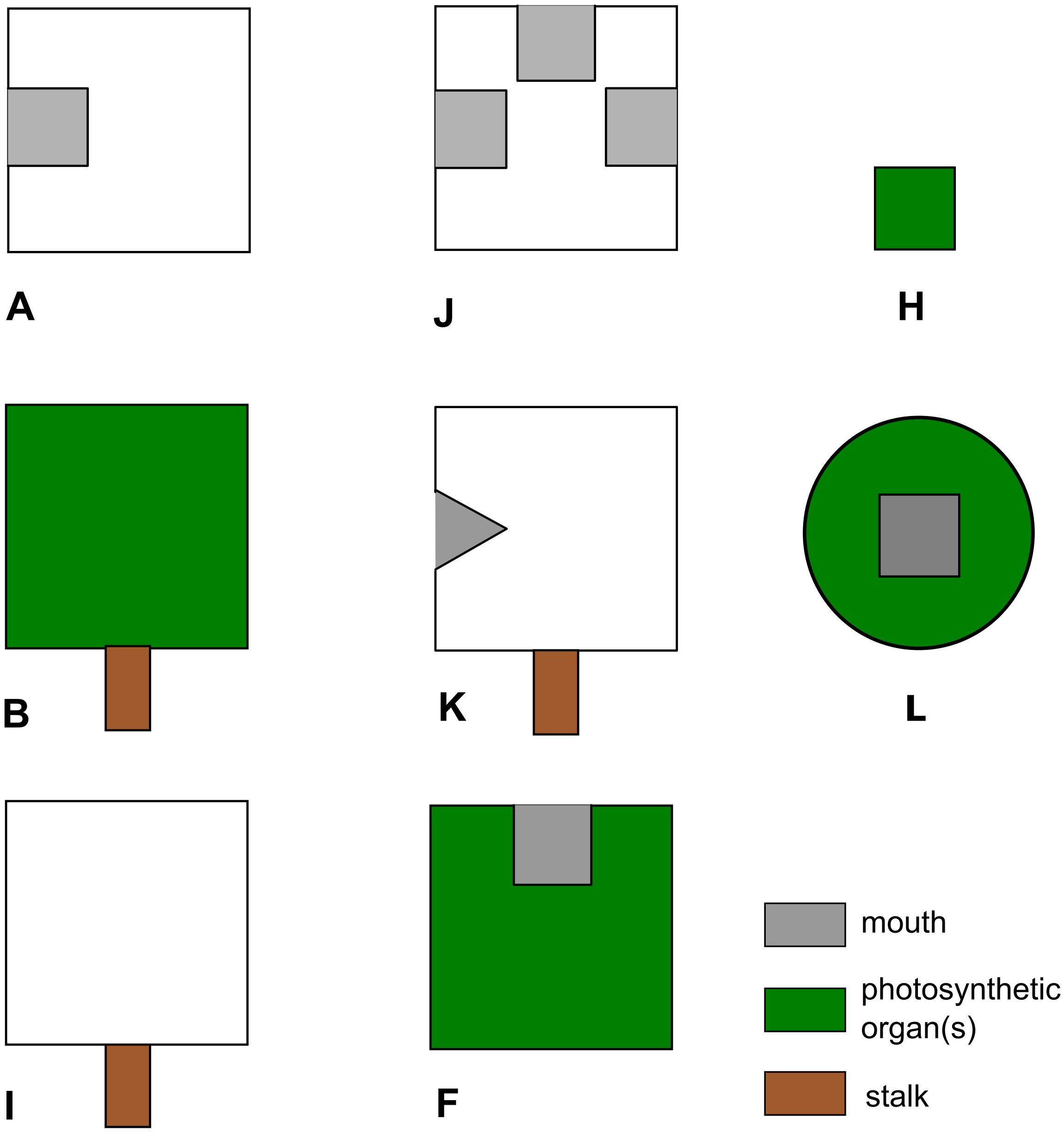
   (b) Four answers

6. Send your textual answers and file attachments to my e-mail address: `alexey.shipunov@minotstateu.edu` with the Subject: “Biol 240 Lab 11”.

Laboratory 12

12.1 Background

- Planet Aqua is entirely covered with shallow water. The ocean is inhabited with various flat organisms (see the figure). These creatures (let us call them "kubricks") can photosynthesize and/or eat other organisms or their parts (which match with their mouths), and move (only if they have no stalks).



Kubricks. Note that common scale is used, so you can measure them.

- Your first goal is to create data (character table) yourself. Then, you will to categorize kubricks into two or three groups and provide support for your classification with **two** different multivariate classification methods.

12.2 Assignment

1. Open R, open Excel or any spreadsheet software and create the data file. This data file should be a table where kubricks are rows and characters are columns (variables). Every row should have a name (kubrick's letter), and every column should also be names (short uppercased names are preferable). In cells, there will be either 1 (character present) or 0 (character absent). Topmost / leftmost cell might be left empty.

   Hint 54. Every character (column) is some feature of kubricks. As an example, you might want to use characters like "presence of stalk" or "presence of three mouths", or "ability to make photosynthesis". You can even use measurement characters and measure features on picture with the millimeter rule.

Hint 55. Try to make at least 8 characters.

2. Save your table as a **text** file (it will be part of your report).

   Hint 56. On macOS, the best way is probably to save text file from your spreadsheet editor as tab-delimited text file.

   Hint 57. On Linux and Windows, copy your spreadsheet cells into any text editor like Geany, then save as a text file.

3. Read your kubricks file into R, explore with `str()` and/or `head()`.

   Hint 58. Use something like `read.table(..., h=T, row.names=1)`.

4. Choose and apply **two** methods which classify your data without learning. PCA, MDS or cluster analysis will all work. Different distances metrics and different hierarchical clustering algorithms do **not** count as different methods. Answer:

   What are groups of kubricks? How exactly do your results support this classification?

   Hint 59. If you like, you can name your groups. Optimal is 2 or 3 groups. Ideally, your groups should have some *biological meaning*.

   Hint 60. You will need to supply plots, at least one per method.

5. Report:

   (a) First attachment: your kubricks textual data file with names and characters. *Rename* this file with your name.

   (b) In the e-mail body: explanation of every character (i.e., what names of columns mean).

   (c) Second attachment: your R script which makes all calculations and plots. Name it as you typically name your script files for the lab.

   (d) In the e-mail body: your answers.

6. Send your two answers and two file attachments to my e-mail address:`alexey.shipunov@minotstateu.edu` with the Subject: "Biol 240 Lab 12".

7. Do not forget to answer the survey!