Biometry. Lecture 8

Alexey Shipunov

Minot State University

February 8, 2012

Outline

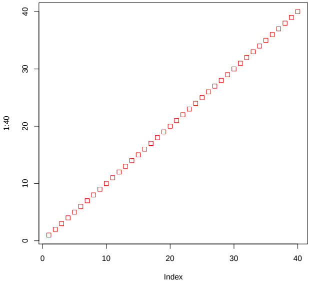
1. Questions and answers
2. Types of data
   - Measurement data
   - Ranked data
   - Categorical data

Outline

1. Questions and answers
2. Types of data
   - Measurement data
   - Ranked data
   - Categorical data

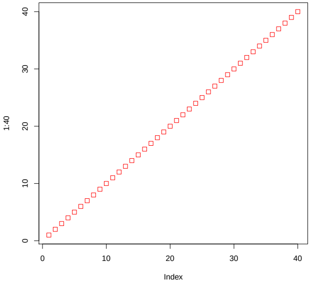
Previous final question: the answer

Which command will produce this plot?



Previous final question: the answer

Which command will produce this plot?



```
plot(1:40, pch=0, col=2)
```

Starting...

```
> setwd("<working folder>")
```

or

“`Change dir`”

in menu!

Commands to look around

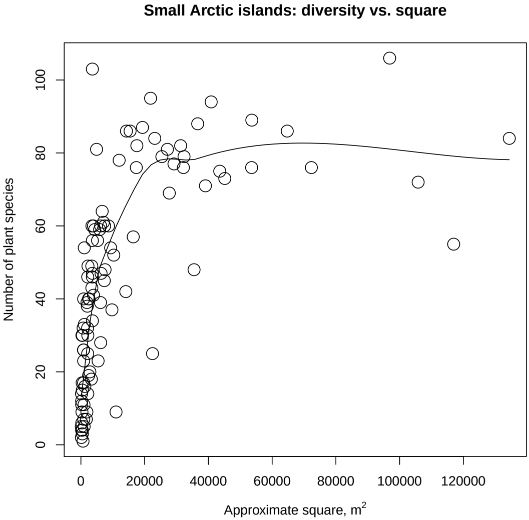
```
> dir() # shows files in working folder
> file.show() # shows content of file
> ls() # lists all objects
> str() # shows the structure of object
> head() # shows first rows of table object (data frame)
```

Enhancing islets

```
> download.file("http://ashipunov.info/data/islets.txt",
+ "data/islets.txt")
> i <- read.table("data/islets.txt", h=T)
> sq <- (i[,1]*.6) * (i[,2]*.6)
> pdf("pics/islets.pdf")
> plot(sq, i[,3], main="Small Arctic islands:
+ diversity vs. square",
+ xlab=expression("Square, m"^2),
+ ylab="Number of plant species")
> lines(loess.smooth(sq, i[,3]), lty=1)
> dev.off()
```

I added the estimated curve (command `loess.smooth()` — from LOESS, *locally weighted scatterplot smoothing*) plus better axes labels (note the `expression()` function for superscript) and title.

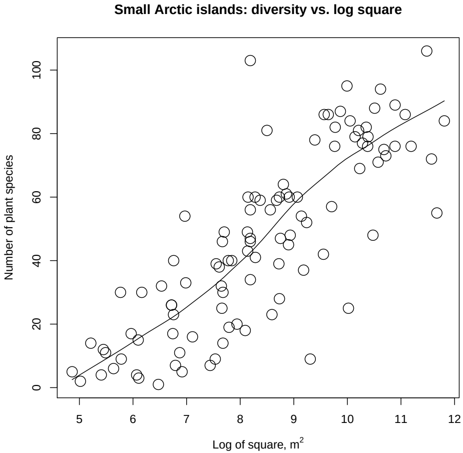
Enhanced islets



Minot State University

What if we use logarithm of square?

With `plot(log(sq), ...)` curve becomes almost linear!



Types of data

Measurement data

Measurement data

- For any two measurements, the third between them also has sense
- Best example: location on the ruler. Continuous, could be zero, positive and negative.
- Temperature has a restriction: there is a minimal temperature
- Angle is worse: there are both minimal and maximal angles

Discrete measurement data: counts

- Tis is the other kind of measurement data
- Number of items is always a whole number so there is the third between 2 and 4
- But the third number between 2 and 3 is a nonsense

“Parametric” and “non-parametric” data

- (a) Only *continuous measurement* data may be parametric
- In addition, parametric methods require: (b) suspected *normal distribution* of data and (c) sample ≥ 30
- Everything else should be studied with non-parametric methods

Measurement data in R

```
> x <- c(174, 162, 188, 192, 165, 168, 172)
> str(x)
 num [1:7] 174 162 188 192 165 168 172
> is.numeric(x)
[1] TRUE
> is.vector(x)
[1] TRUE
```

Some rules about vectors

- For every type of R object, there are functions `is.<something>()` and `as.<something>()` (e.g., `as.vector()` and `as.numeric()` will convert to vector and to numeric vector, respectively).
- Object names must not start with a number
- R is case-sensitive
- Please avoid to use names of popular functions (like `c()`) and keywords: `T` (TRUE), `F` (FALSE), `NA` (missing data), `NaN` (not a number), `Inf` (result of dividing by zero), `pi`

Minot State
UNIVERSITY

Types of data

Ranked data

What if we cannot measure?

- In this case, we cal use scale-like representation
- E.g., we can rank the student success from 1 to 5 (“very bad” to “excellent”)
- Or softness of mattress from 0 to 10 (“hard as a plank” to “soft as a cloud”)

Ranked and measurement data

- Similarity: for every two ranks, the third between them has sense
- E.g., it is possible to imagine mattress with softness between 2 and 3
- However, ranks are not represent intervals correctly!
- Ranked data should be studied with non-parametric methods

How to create ranked data

In R, ranked data is normally represented by the same numerical vector or *ordered factor*. Command `cut()` will break continuous data into ranks:

```
> height <- trees[,2]
> cut(height, 3, labels=c(1:3), ordered=T)
> cut(height, 3, ordered=T)
```

Types of data

Categorical data

Just observations

- Some data cannot be ordered at all
- Sex, color, absence/presence are good examples
- If even we label red color as “1” and green color as “2” the “1.5” is a nonsense.
- Therefore, if we use numbers for categorical data, they are only *labels*.

Binary data

- Absence/presence is a specific subset of categorical data which only two possible values
- One of the easiest representation is with numbers 0 and 1
- Computers normally prefer binary data over non-binary

Categorical data in R

Character and logical vectors may be used for categorical data:

```
> sex <- c("male", "female", "male", "male",
+ "female", "male", "male")
> is.character(sex)
> is.vector(sex)
> str(sex)
> presence <- c(F, T, T, F, F)
> presence
> str(presence)
> presence * 1 # convert to 1/0
> (presence * 1) == 1 # convert back
```

“==” is a logical test: “Is equal?”. In R, “=” has a different meaning, it is a replacement for “<–”.

Finishing...

```
> savehistory("20120208.r")
```

Final question (2 points)

Final question (2 points)

Which command will convert measurement data to ranked data?

Summary: most important commands

- `str()`—shows the structure of object

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access: `http://ashipunov.info/shipunov/school/biol_299`

P. Dalgaard
Introductory Statistics with R. 2nd edition.
Springer, 2008.
Section 1.2.1–1.2.8.