Biometry. Lecture 7

Alexey Shipunov

Minot State University

February 19, 2014

Outline

Outline

Starting...

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

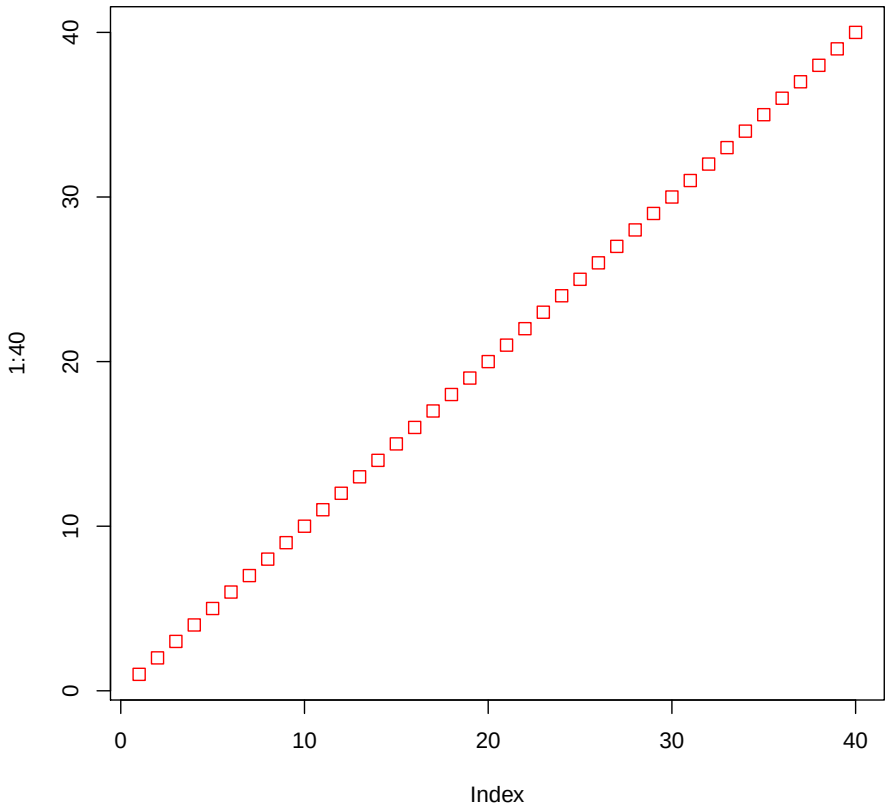
or

“Change dir”

in menu!

Previous last question

Which command will produce this plot?




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

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

Logical data

Practically, it is another kind of binary data:

```
> height < 72
> height >= 72
> height == 72 # not "="!
> presence <- c(F, T, T, F, F)
> 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 "<−".

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)
> str(presence)
```

Squeezing the categorical data

```
> sex <- c("male", "female", "male", "male",
+ "female", "male", "male")
> presence <- c(F, T, T, F, F)
> table(sex)
> table(presence)
```

The `table()` command will let us to have some numbers even from categorical data!

Character to factor

```
> plot(sex) # error!
> sex.f <- as.factor(sex)
> plot(sex.f) # makes bar plot
```

Features of factors

```
> is.factor(sex.f)
> is.character(sex.f)
> str(sex.f)
> levels(sex.f)
> sex.f[6:7] # two levels!
> sex.f[6:7, drop=TRUE] # one level
```

Factor has levels which will not automatically drop with a sub-setting.

Finishing...

```
>savehistory("20140219.r")
```

Final question (2 points)

Final question (2 points)

What is a difference between factor and character vector in R?

Summary: most important commands

- `as.<something>()`—converts objects
- `table()`—summarizes categorical data

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access:
http://ashipunov.info/shipunov/school/biol_240

A. Shipunov, and others.
Visual statistics. Use R!
DMK Press, 2012. [Translating from Russian.]