Biometry. Lecture 9

Alexey Shipunov

Minot State University

February 10, 2012

Outline

Outline

Starting...

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

or

“Change dir”

in menu!

Previous final question: the answer

Which command will convert measurement data to ranked data?

Previous final question: the answer

Which command will convert measurement data to ranked data?

- `cut(..., ordered=TRUE)`

Types of data

Categorical data

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 subsetting.

Factors to numbers

```
> as.numeric(sex.f)
> w <- c(69, 68, 93, 87, 59, 82, 72)
> x <- c(174, 162, 188, 192, 165, 168, 172)
> plot(x, w, pch=as.numeric(sex.f), col=as.numeric(sex.f))
> legend("topleft", pch=1:2, col=1:2, legend=levels(sex.f))
```

Objects `x`, `sex` and `w` could be height, gender and weight of seven people in small office, respectively.

Factors to ranks

```
> m <- c("L", "S", "XL", "XXL", "S", "M", "L") # t-shirts
> m.f <- factor(m)
> levels(m) # Wrong order, alphabetical
> m.o <- ordered(m.f, levels=c("S", "M", "L", "XL", "XXL"))
> levels(m.o)
```

The danger of factors

```
> a <- factor(3:5)
> a
> as.numeric(a) # wrong!!!
> as.numeric(as.character(a)) # correct
```

Types of data

Secondary data

Types of secondary data

- Fractions (and percents)
- Counts and ranks

Dotcharts for fractions and percents

These kinds of plots are often much better to read than bar plots and pie charts:

```
> height.2 <- cut(height, 3,
+ labels=c("low", "middle", "high"))
> dotchart(table(height.2)) # ignore the warning
```

Ranks and ties

```
> a1 <- c(1,2,3,4,4,5,7,7,7,9,15,17)
> a2 <- c(1,2,3,4,5,7,7,7,9,15,17)
> names(a1) <- rank(a1)
> a1
> names(a2) <- rank(a2)
> a2
```

Ranks may be decimal, same numbers will have same ranks (ties) and R will tell about that: try, for example

```
> wilcox.test(a2)
```

Types of data

Missing data

No silver bullet

- All real world data have problems
- Numbers may be corrupted, missed, unknown, duplicated etc.
- “Missing data” is often used to cover part of these cases

Entering missing data

We asked seven office co-workers about their average time of sleep. One denied to answer, other answered “I don’t know”, the third was not in the office at the moment:

```
> h <- c(8, 10, NA, NA, 8, NA, 8)
> h
```

How to operate with NA

```
> mean(h) # error!
> mean(h, na.rm=TRUE)
> mean(na.omit(h))
```

Missing data imputation

```
> h.old <- h
> h[is.na(h)] <- mean(h, na.rm=TRUE)
> h
```

Types of data

Outliers

Too big or too small

- Outliers are results of mistypes, misprints and other random events
- It is possible to recognize them using `table()` and `summary()` functions

Types of data

Data conversion and normalization

Logarithmic conversion

- Very often the bell-shape curve or linear relation may me achieved by logarithmic conversion
- In R, it is possible to apply `log()` (natural logarithm) function to any vector. The only problem is that the data should not contain zeroes.
- Other mathematical conversions also exist, e.g., square root conversion.

Scaling data

Function `scale()` will convert all columns of data frame to the same scale:

```
> a <- 1:10
> b <- seq(100, 1000, 100)
> d <- data.frame(a, b)
> d
> scale(d)
```

Columns `a` and `b` became identical!

Finishing...

```
> savehistory("20120210.r")
```

Final question (3 points)

Final question (3 points)

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

Summary: most important commands

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

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.

Minot State UNIVERSITY