Week_4_Simulating_sample_data

Cici

2019.12.25

LO:

1. Describe the features of a quality shareable data-set.
2. Explain reasons for using synthetic data-sets, including ethical reasons.
3. Create synthetic data sets in R and Python.
4. Use synthetic data sets to test a data analysis workflow.

Notes:

- What should be delivered to the statistician?

1. raw data. (eg. raw sequencing data from RNA-seq) No software ran on the data. No modification. Do not remove data. Do not summarize data in any way.
2. tidy data set.
3. code book.

- Basic principles for tidy data.

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

- Long vs Wide:

Long data format: Multiple rows constitute a single observation. Hard for computing/statistics. NOT TIDY.

Wide data format: TIDY.

- Code book:

Word file.

Information about variables (units)

Information about the summary choices you made.

Information about the experimental study design you used.

Parameters.

- How to generate synthetic data set?

1. Draw numbers from a distribution. To observe real-world statistic distributions from the original data and reproduce fake data by drawing simple numbers.
2. Agent based-like modeling (ABM). To create a physical model that explains the observed behavior, then reproduce random data using this model.

Functions:

set.seed(n) #to make your code reproducible, remember to set seed when you are random sampling something in R.

t.test(x, y, alternative = c("two.sided","less","greater"),mu = 0, paired = FALSE, var.equal = FALSE)

cbind() #column bind

write.csv() #export a csv file

paste0(collapse = "") #concatenate strings together.