Thống kê và vẽ đồ thị trong R

Lời mở đầu

Tác giả

Duc Nguyen | tuhocr.com

Nội dung cuốn sách này điểm qua hầu hết các chủ đề thống kê và vẽ đồ thị thường gặp, bao gồm các trích dẫn đến tài liệu toàn văn để thuận tiện cho người đọc dễ tra cứu.

Cách tiếp cận đi từ làm rõ định nghĩa, thuật ngữ, kế đến là công thức, thuật toán, bài tập ví dụ và lời giải, sau cùng là tình huống cụ thể.

Trích dẫn

```
Duc Nguyen (2025). "Thống kê và vẽ đồ thị trong R". TUHOCR. https://thongkevavedothi.com

@Book{Nguyen2025,
  author    = {Duc Nguyen},
  publisher = {TUHOCR},
  title     = {Thống kê và vẽ đồ thị trong {R}},
  year      = {2025},
  url       = {https://thongkevavedothi.com},
}
```

1

Các chủ đề thường gặp

1.1 Person

1.1.1 Statistician

Samiran Sinha

`https://samiransinha.github.io/teaching/`

Laurent Smeets

`https://www.rensvandeschoot.com/colleagues/laurent-smeets/`

1.1.2 Psycholinguist

Luca Campanelli

`https://www.lcampanelli.org/`

1.2 Dataset

Vanderbilt Biostatistics

`https://hbiostat.org/data/`

Datasets for the survival data modelling on engineering applications

`https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data#overviewHardDriveData`

Clinical proteomic datasets from NCI

`http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`

Kaggle, a platform for different kinds of data used for data science competitions.

`https://www.kaggle.com/data`

It is a repository of shared datasets available through AWS resources. https://registry.opendata.aws/‘

1.3 Mixed effects model

Mixed effects model analysis using R

http://samiransinha.github.io/files/teaching/685part1.html

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using Lme4." Journal of Statistical Software 67, no. 1 (2015).

https://doi.org/10.18637/jss.v067.i01

http://book.thuviencanhan.com:8033/results?query=%22Bates+et+al.+-+2015+-+Fitting+Linear+Mixed-Eff

Bates, Douglas M. Lme4: Mixed-Effects Modeling with R. 2022.

https://people.math.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf

Luca Campanelli. Introduction to mixed-effects modeling using the `lme4` package.

https://web.archive.org/web/20230313184038/https://www.lcampanelli.org/mixed-effects-modeling-lme4

LME4 Tutorial: Popularity Data

https://www.rensvandeschoot.com/tutorials/lme4/

Fixed vs Random vs Mixed Effects Models – Examples

https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/

What is a difference between random effects-, fixed effects- and marginal model?

https://stats.stackexchange.com/questions/21760/what-is-a-difference-between-random-effects-fixed-

Concepts behind fixed/random effects models

https://stats.stackexchange.com/questions/33984/concepts-behind-fixed-random-effects-models

A brief introduction to mixed effects modelling and multi-model inference in ecology

https://pmc.ncbi.nlm.nih.gov/articles/PMC5970551/

1.4 Survival analysis

Hosmer, David W., Stanley Lemeshow, and Susanne May. Applied Survival Analysis: Regression Modeling of Time-to-Event Data. John Wiley & Sons, Ltd, 2008.

https://doi.org/10.1002/9780470258019.fmatter

http://book.thuviencanhan.com:8033/results?query=%22Hosmer+et+al.+-+2008+-+Applied+Survival+Analys

1.5 B-splines

A short note on B-splines, and two related files for computing spline basis functions R script, Fortran subroutines

`http://samiransinha.github.io/files/teaching/note1.pdf`

`http://samiransinha.github.io/files/teaching/code4Splines.R`

`http://samiransinha.github.io/files/teaching/spline.f`

`https://samiransinha.github.io/teaching/`

1.6 Epidemiology

1.6.1 Case-control study

Case-control studies in epidemiological research

`http://samiransinha.github.io/files/presentation/TAMU_Vet_School_Nov2021.pdf`

1.7 Single cell RNAseq

Benchmarking of a Bayesian single cell RNAseq differential gene expression test for dose-response study designs

`https://samiransinha.github.io/files/presentation/WNAR2023_presentation.pdf`

1.8 Multilevel analysis

Multilevel analysis: Techniques and applications

`https://multilevel-analysis.sites.uu.nl/`

1.9 Bayesian

Bürkner, (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1–28.

`https://doi.org/10.18637/jss.v080.i01`

Magnusson et al. (2019). Bayesian leave-one-out cross-validation for large data (2019)

https://proceedings.mlr.press/v97/magnusson19a/magnusson19a.pdf

Vehtari et al (2013). Understanding predictive information criteria for Bayesian models.

https://sites.stat.columbia.edu/gelman/research/published/waic_understand3.pdf

Vehtari et al. (2018). R-squared for Bayesian regression models

http://www.stat.columbia.edu/~gelman/research/unpublished/bayes_R2.pdf

Vehtari et al. (2019). Bayesian R2 and LOO-R2

https://avehtari.github.io/bayes_R2/bayes_R2.html

Vehtari et al. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). Bayesian Data Analysis.

https://projecteuclid.org/journals/bayesian-analysis/advance-publication/Rank-Normalization-Foldin

1.10 Randomness

https://en.wikipedia.org/wiki/Randomness#cite_note-5

1.11 Normal distribution

https://en.wikipedia.org/wiki/Normal_distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1.12 Sample size

How to calculate sample size in randomized controlled trial?

https://pubmed.ncbi.nlm.nih.gov/22263004/

2

Sách tham khảo

(Crainiceanu, Caffo, and Muschelli 2018)

Part I

Kỹ thuật vẽ đồ thị

3

Hướng dẫn vẽ đồ thị trong R bằng package `ggplot2`

Cú pháp `ggplot2` (Wickham 2010)

4

Tài liệu tham khảo

Crainiceanu, Ciprian, Brian Caffo, and John Muschelli. 2018. *Methods in Biostatistics with R*. Leanpub. https://leanpub.com/biostatmethods.

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28. http://book.thuviencanhan.com:8033/results?query=&dir=tuhocr/R+programming/ggplot2/wickham2010&after=&before=&sort=relevancyrating&ascending=0&page=1.