PRINCIPLES AND TOOLS OF STATISTICS

Gunther Witzel, 8.4.2022

STATISTICS LECTURE, INTRO

MPIFR, BONN, ROOM 0.02

INTRODUCTION

- Webpage: https://www3.mpifr-bonn.mpg.de/staff/gwitzel/Teaching.html
- **Student questionnaire**
- **Goals of the lecture**
- **Resources**
- Python 3, Jupyter Notebooks, Astroconda, numpy, scipy, matplotlib, numba
- > emcee, corner
- **git**, repositories

WEBPAGE

https://www3.mpifr-bonn.mpg.de/staff/gwitzel/Teaching.html

Personal webpage of Dr. Gunther Witzel

Principles and tools of statist

Textbooks

Textbooks used for this lecture:

Title: Stochastics : Introduction to Statistics

Author: Hans-Otto Georgii E-Book (available for MPIfR-Staff ar institutes)

Two copies in print will be available in the

Title: A Student's Guide to Bayesian Statis Author: Ben Lambert online material:

Student resources (containing links to and "Answer to problem sets" for single k Three copies in print will be available in t

Textbooks available through the Max Library:

MPDL

List of Textbooks on Bayesian statitstics

Home	Research	Outreach	CV	Contact	Teaching	Code	Links
tics							
		Slides					
Probabili	ty and						
nd other	MPG-						
e MPIfR lib	orary.						
stics.							
"Author book-chap the MPIfR	videos" oters) library.						
x Planck	Digital	Code					

Yes

No No

Did you have a course on probability in high school? ³⁷ responses



Did you study combinatorics in high school? 37 responses



Are you familiar with Python 3 and standard Python packages like numpy or matplotlib? ^{37 responses}



Yes

No No

37/40 participants have answered the questionnaire - Thank you!

Did you have a course on probability and/or statistics at university? ^{37 responses}



Are you familiar with Conda environments? ^{37 responses}





Have you learned fundamentals in statistics during lab course experminents? ^{37 responses}



Are you familiar with R? 37 responses



Almost everyone has a laptop.

Yes

🛑 No

Yes

No No



Yes

14





QUESTIONNAIRE



"Random variable is a variable that is not dependent on pre-determined factors. Probability density is the distribution of the probability of certain value."

"A random variable is a set from which a value is randomly drawn. The probability density describes the distribution of the random variable, i.e. it gives the probability for a certain value of the random variable, or (if continuous) the probability inside an interval."



"Given a sufficiently large of a sample, the distribution of the sample will be normally distributed."

"The sum of independent random outcomes tends towards a normal distribution."

Yes

5

Can you explain the difference between frequentist and Bayesian inference?



"Frequentist is about dice and cards, Bayesian about information."

"Frequentist inference has a deterministic approach, while bayesian inference uses the Bayes theorem to measure probability distribution of values."

"A frequentist argues bases on the repetition of an experiment, i.e. if the experiment is repeated infinite times the true probability distribution is obtained. A Bayesian argues based on a degree of belief, i.e. using a prior knowledge about the probability distribution and the experiment, the posterior distribution is estimated.."

Yes 5

Are you using or are you planning to use statistical tools in your PhD work? If yes, please describe them shortly.



"Moments of distributions, censored data (limits), chisquare, uncertainties, correlations, KS-test, linear regression, MCMC, boot strapping, power spectrum, Fisher information, Minkowski functional, student test, bispectrum, Jacknife bias "



Are you familiar with a Bayesian sampler (e.g., MCMC, Multinest, Stan)? If yes, which one?



"MCMC; emcee"

What do you hope to learn from this 3-day course?

"Fundamental understanding of mathematical principles; why use a given tool, robust framework; practical use of python; computational implementation; step-by-step guide"; Bayesian statistics; methods used in astrophysics; goodness of fit; parameter estimation; practical guide to MCMC; examples; model selection; likelihood-free analysis; generative models; ML estimation; corner plots; Fisher information; Jacknife and Bootstrap, tests, error propagation

GOAL AND STRUCTURE OF THE LECTURE

Introduction of concepts (morning sessions):

- Guide to self-study statistics
- Online resources

- Demonstration of concepts
- Familiarity with Python packages
- - the studied concepts and methods

Discussion of fundamental concepts of probability theory and statistics (no mathematical proofs)

• Practical examples and exercises (afternoon sessions): Problem solving and code implementation

Homework (to be submitted 2 weeks after the lecture) : More complicated problems that require several of

RESOURCES

Webpage: https://www3.mpifr-bonn.mpg.de/staff/gwitzel/Teaching.html

Books:

Title: Stochastics : Introduction to Probability and StatisticsAuthor: Hans-Otto GeorgiiE-Book (available for MPIfR-Staff and other MPG-institutes)Two copies in print will be available in the MPIfR library.

Title: A Student's Guide to Bayesian Statistics. Author: Ben Lambert <u>online material</u> Three copies in print will be available in the MPIfR library.

Free Springer books: <u>link</u>

PYTHON ENVIRONMENT

We will install a custom environment for this lecture. This will make sure that we are all on the same page and can execute examples reliably. It also will make sure that the python and package installations are not messing with your current setup.

Conda: <u>webpage</u> <u>command reference</u>

"Package, dependency and environment management for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN, and more."

Anaconda: <u>webpage</u> <u>documentation</u> <u>list of included packages</u>

"A collection of over 7,500+ open-source packages, which includes the package and environment manager Conda, and Jupyter notebook."

Astroconda: get started list of included packages FAQ

"AstroConda is a free Conda channel maintained by the Space Telescope Science Institute (STScI) in Baltimore, Maryland. This channel provides tools and utilities required to process and analyze data from the Hubble Space Telescope (HST), James Webb Space Telescope (JWST), and others."

Jupyter Notebook: webpage documention why are Jupyter notebooks not that bad why do Jupyter notebooks suck

make astroconda kernel available in your Jupyter notebook

PACKAGES FOR BAYESIAN ANALYSIS

Emcee: <u>webpage</u>

"emcee is an MIT licensed pure-Python implementation of Goodman & Weare's Affine Invariant Markov chain Monte Carlo (MCMC) Ensemble sampler"

emcee can be installed via:

conda install –c conda–forge emcee

Corner: <u>webpage</u>

"This Python module uses matplotlib to visualize multidimensional samples using a scatterplot matrix."

corner can be installed via:

conda install -c astropy corner

GIT REPOSITORIES

"Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency."

webpage and documentation

For macOS and Linux, git should be already installed on your system; if not or if you have problems try to install git via:

conda activate astroconda

conda install –c anaconda git

What is git? Git serves three main purposes:

- forth between versions or branches.
- reference.

To develop code in groups of people where everyone contributes while making sure that changes that one contributor made are not overwritten by another contributor; this is achieved by a semi-intelligent "merging" processes.

To version control files on your computer; this enables you to keep images of the progression of changes of code or light data files ("commits") over time; it also enables you to manage several versions at the same time ("branches"), and to go back and

To mirror your project to a git server, a "repository"; this will give you access to your project from everywhere; e.g., it helps to distribute your project to other computers, like the fast server of your research group, by "cloning" your project; it also allows to freeze and publish a version of your code (a "release") for others to download, or to attach a digital object identifier (DOI) for

GIT REPOSITORIES

Git cheat sheet: <u>https://education.github.com/git-cheat-sheet-education.pdf</u>

Example:	mkdir workdi
	cd workdir
	vim new_file
	Type i, then h
	git init
	git branch
	git status
	git add new_ ⁻
	git status
	git commit —
	git log

r

ello world, esc, :wq

file

m 'first commit, created new_file'

GIT REPOSITORIES

Repositories: <u>GitHub</u> <u>Bitbucket</u> **Gitlab (contact your institution)**

Example: git remote rename origin old-origin

git remote add origin url

git push –u origin –all

git push -u origin -tags

git clone url workdir2

Lecture repository (slides, Jupyter Notebooks, homework): git clone https://gitlab.mpcdf.mpg.de/gwitzel/imprs-blackboard-lecture-2022.git

