Introduction

PRINCIPLES AND TOOLS OF STATISTICS

Gunther Witzel, 8.4.2022

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

Principles and tools of statistics

Textbooks

Textbooks used for this lecture:

Title: Stochastics : Introduction to Probability and Statistics
Author: Hans-Otto Georgii
E-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
online material:
Student resources (containing links to "Author videos" and "Answer to problem sets" for single book-chapters)
Three copies in print will be available in the MPIfR library.

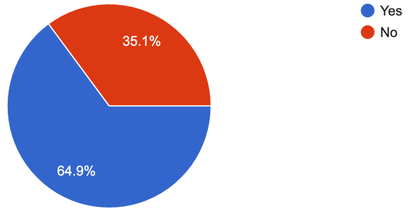
Textbooks available through the Max Planck Digital Library:
MPDL
List of Textbooks on Bayesian statitstics

Slides

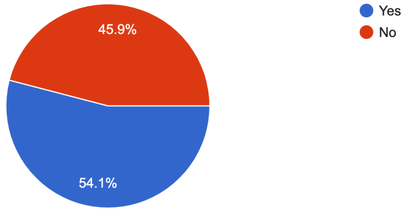
Code

QUESTIONNAIRE

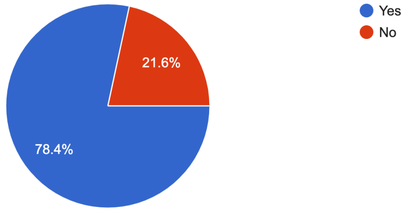
Did you have a course on probability in high school?
37 responses



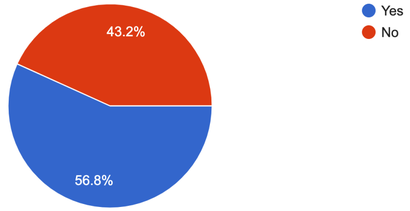
Did you study combinatorics in high school?
37 responses



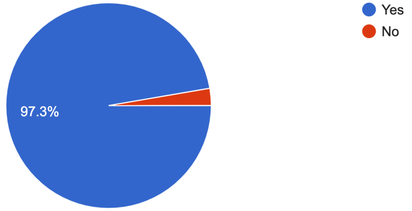
Did you have a course on probability and/or statistics at university?
37 responses



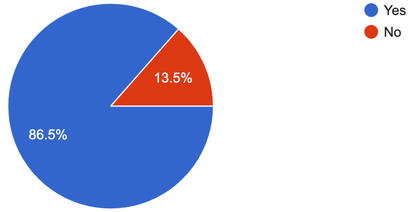
Have you learned fundamentals in statistics during lab course experminents?
37 responses



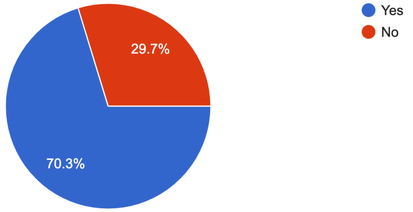
Are you familiar with Python 3 and standard Python packages like numpy or matplotlib?
37 responses



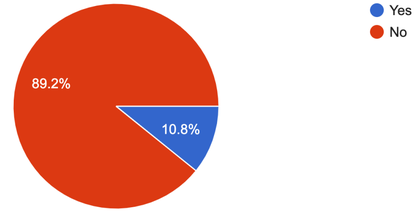
Are you familiar with Jupyter notebooks?
37 responses



Are you familiar with Conda environments?
37 responses



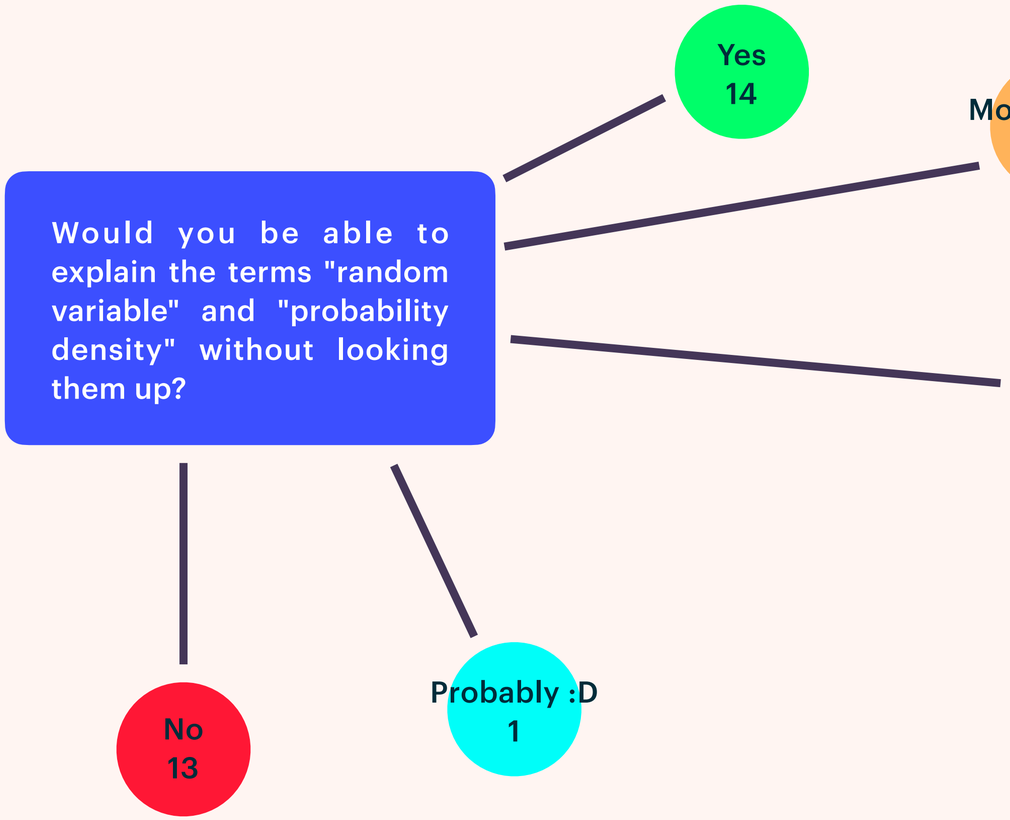
Are you familiar with R?
37 responses



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

Almost everyone has a laptop.

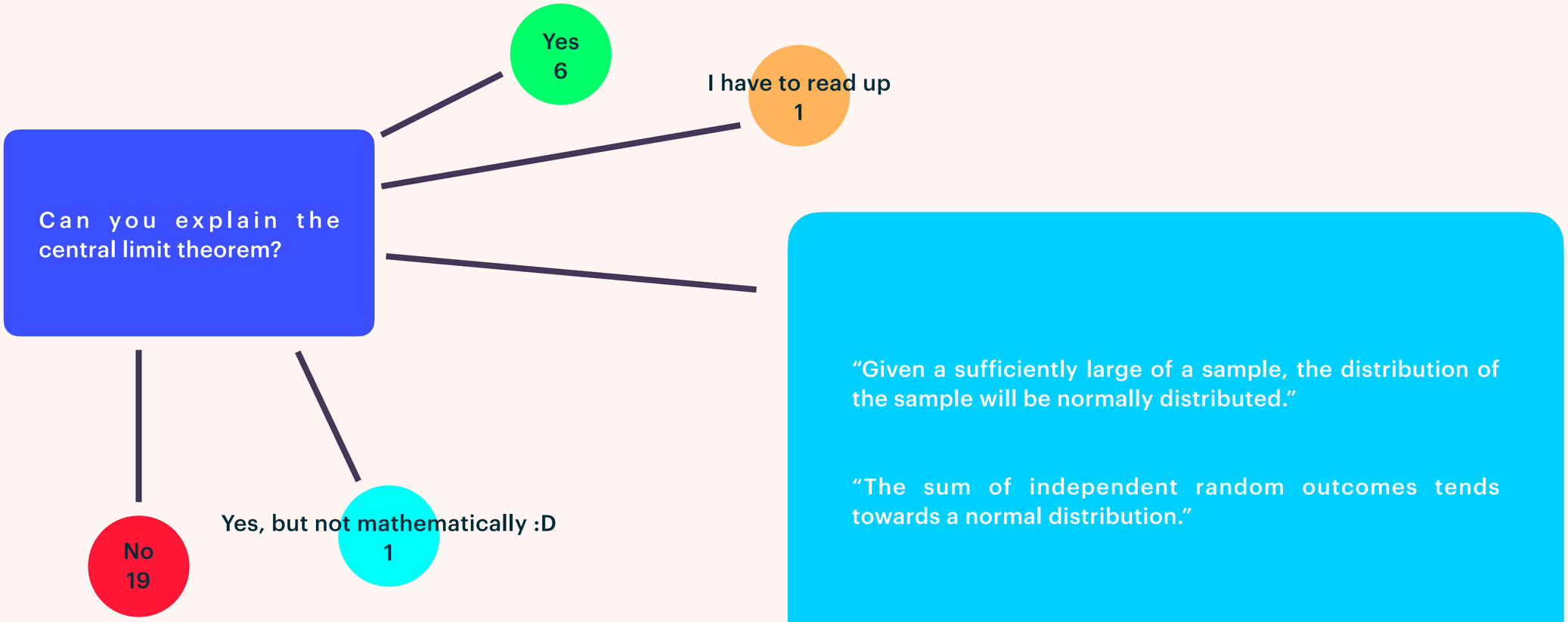
QUESTIONNAIRE



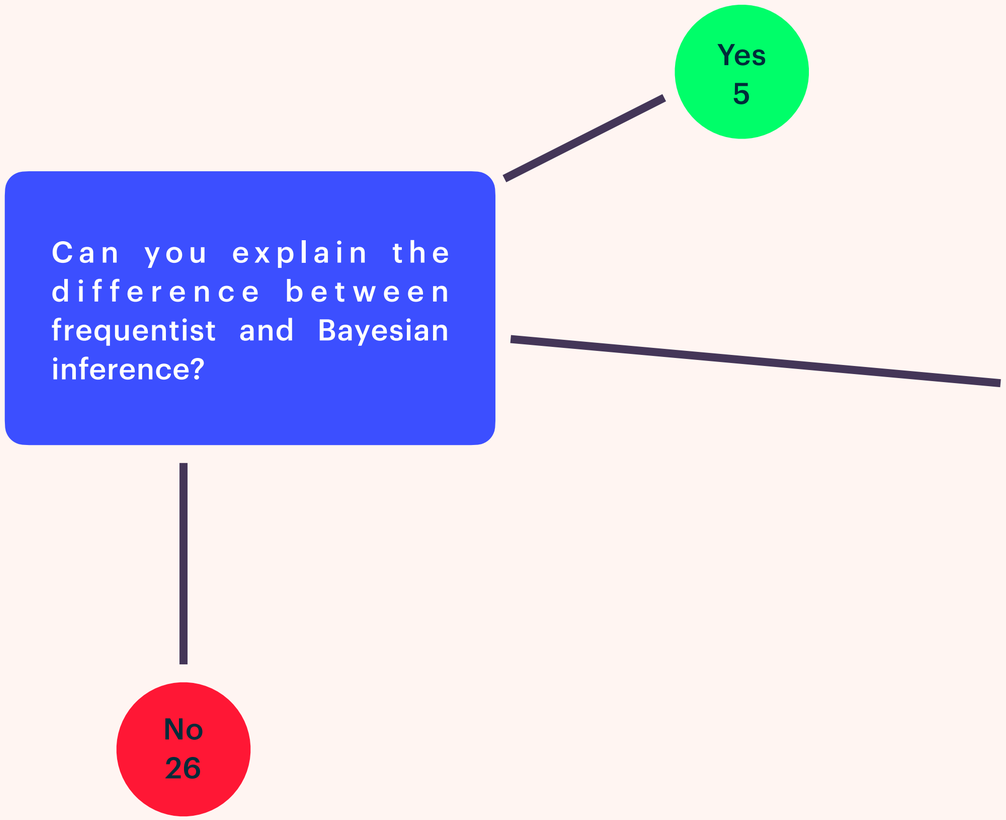
"Random variable is a variable that is not dependent on pre-determined factors. Probability density is the distribution of the probability of certain value."

"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."

QUESTIONNAIRE



QUESTIONNAIRE

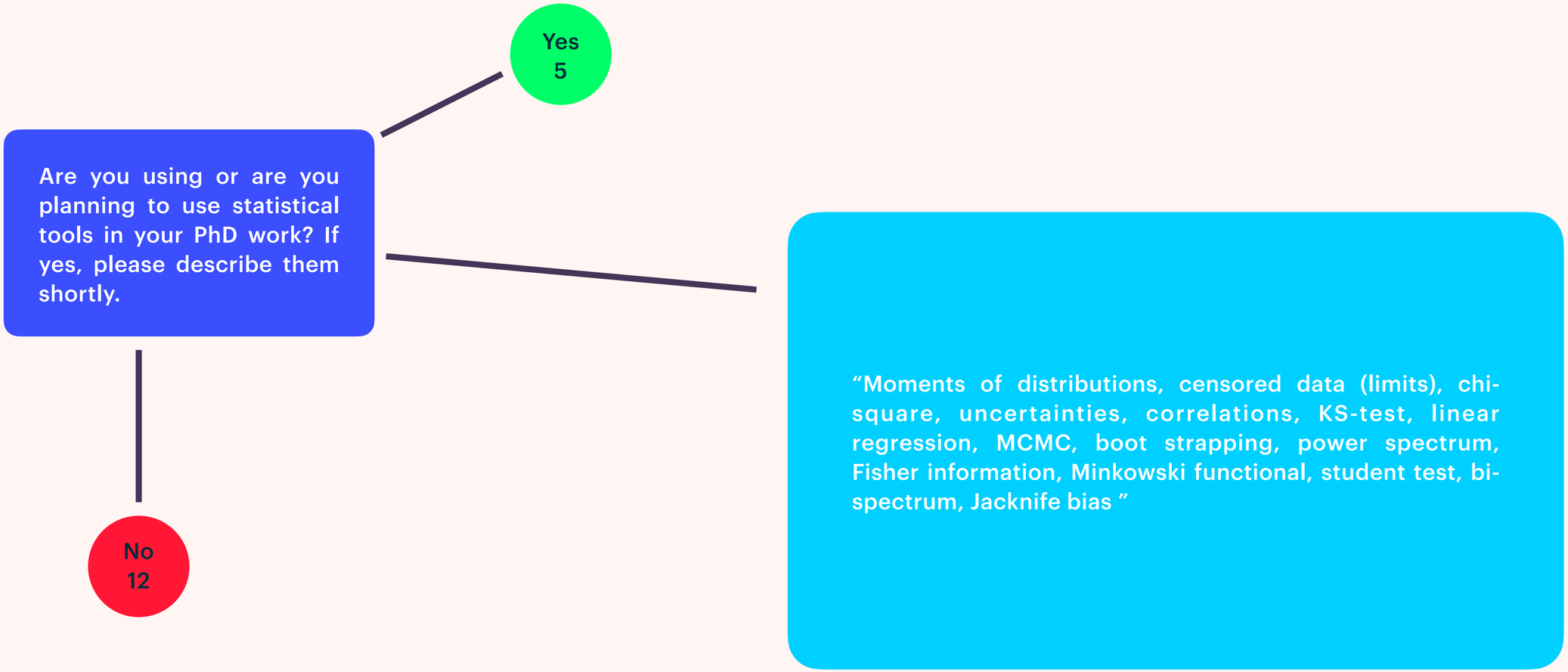


"Frequentist is about dice and cards, Bayesian about information."

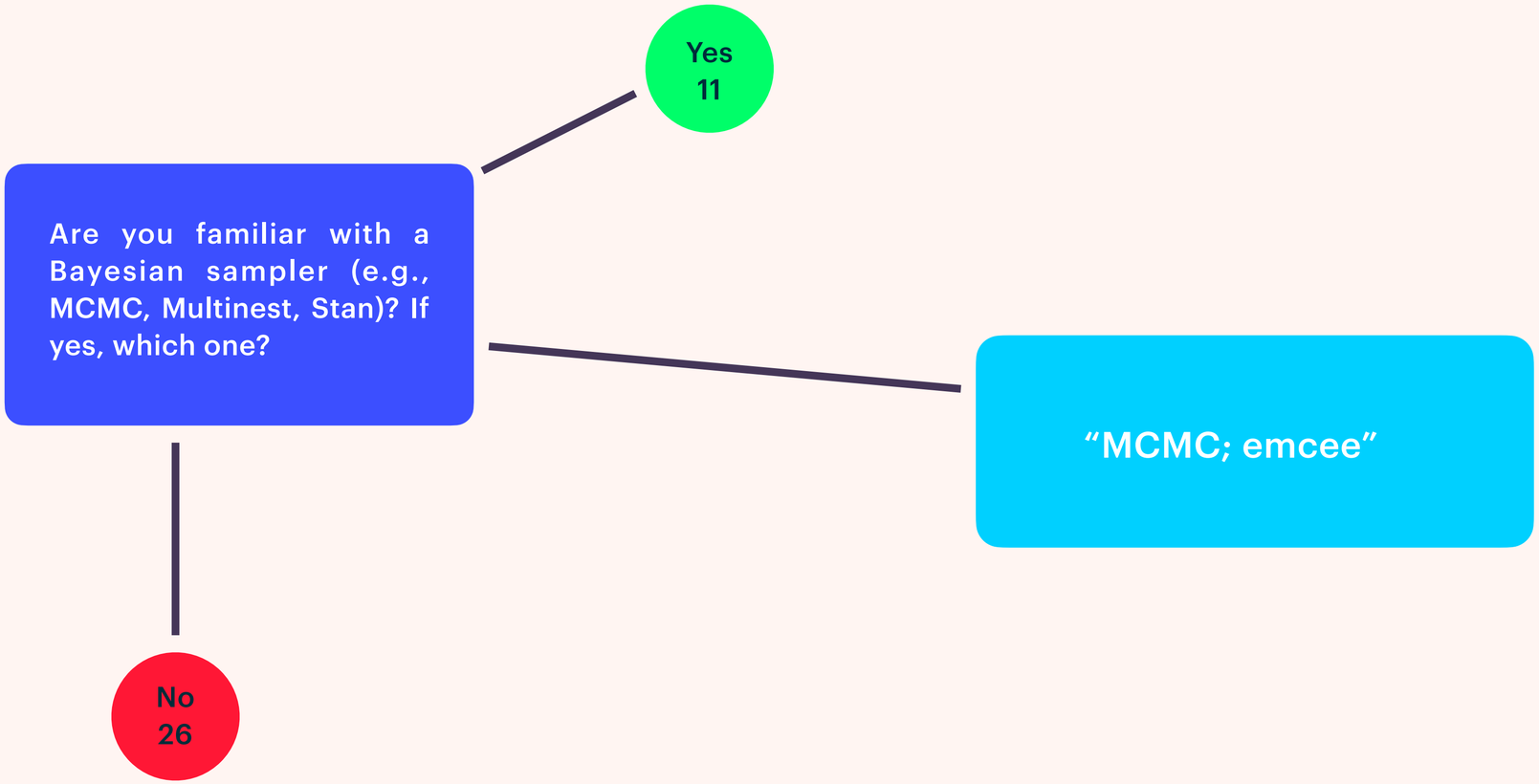
''Frequentist inference has a deterministic approach, while bayesian inference uses the Bayes theorem to measure probability distribution of values."

"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.."

QUESTIONNAIRE



QUESTIONNAIRE



QUESTIONNAIRE

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):
  - Discussion of fundamental concepts of probability theory and statistics (no mathematical proofs)
  - Guide to self-study statistics
  - Online resources

- **Practical examples and exercises** (afternoon sessions):
  - Demonstration of concepts
  - Problem solving and code implementation
  - Familiarity with Python packages

- **Homework** (to be submitted 2 weeks after the lecture) :
  - More complicated problems that require several of the studied concepts and methods

RESOURCES

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

- Books:

  Title: Stochastics : Introduction to Probability and Statistics
  Author: Hans-Otto Georgii
  E-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
  online material
  Three copies in print will be available in the MPIfR library.

SYLLABUS

Preperatory meeting

+ Installation of Python and Astroconda
+ git version control
+ cloning the repository with programming exercises and examples

Day 1

+ Probability Spaces and random variables
+ Conditional Probability and Independence
+ The law of large numbers and the central limit theorem
+ Markov chains

Day 2

+ Parameter estimation and confidence intervals
+ Tests and regression
+ The Bayesian Formula
+ Priors
+ Likelihoods
+ The evidence
+ Posteriors

Day 3

+ Practical Bayesian sampling methods
+ Markov Chain Monte Carlo with STAN
+ Approximate Bayesian Computation
+ Hierarchical models and linear regression models

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: webpage command reference

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

Anaconda: **webpage** documentation list of included packages

"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: webpage

"**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: webpage

"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:

- 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 forth between versions or branches.

- 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 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.

GIT REPOSITORIES

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

Example:

```
mkdir workdir
cd workdir
vim new_file
```

Type `i`, then `hello world`, `esc`, `:wq`

```
git init -b main
git branch --show-current
git status
git add new_file
git status
git commit -m 'first commit, created new_file'
git branch
git log
```

GIT REPOSITORIES

Repositories: **GitHub** **Bitbucket** **Gitlab (contact your institution)**

Example:

```
git remote add origin https://github.com/guwitzel/lecture_test2.git
git branch -M main
git push -u origin main
cd ..
git clone url workdir2
```

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