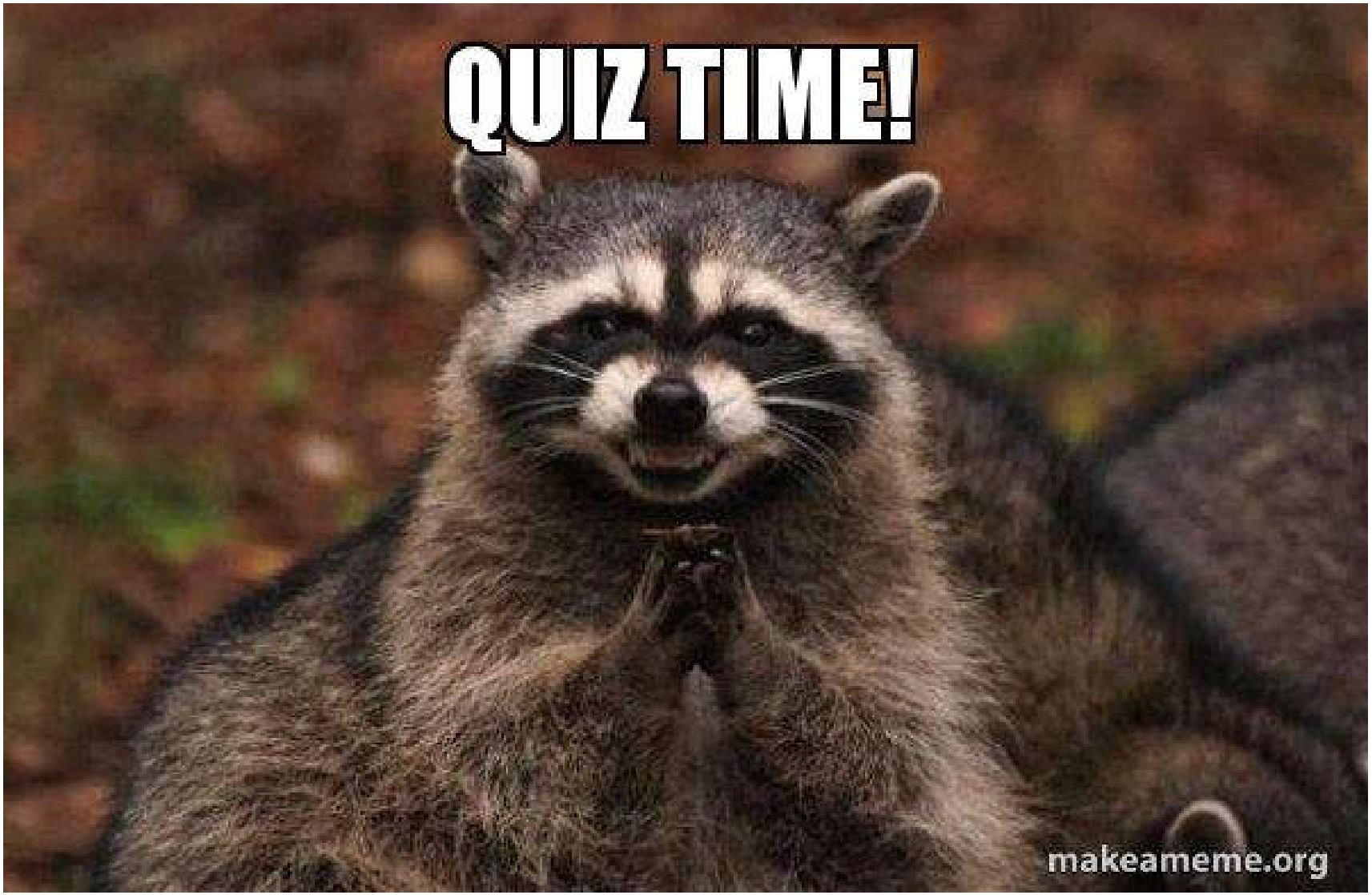
Introduction to statistics using R

Seminar series - session 2

Session 2 – Learning objectives

- Understand the concept of probability distribution
- A very short intro to R and Rstudio

But first,



Distribution

The concept of distribution is the soul of data analysis

If you understand distribution, then you'll understand stats

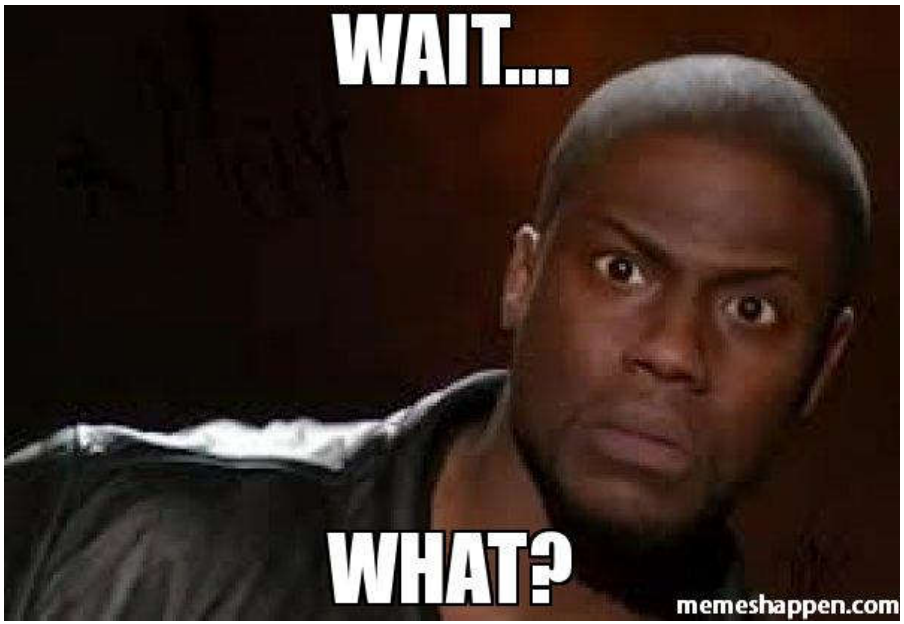
Statistical tests assume data follow specific distributions

What is distribution?

- **Probability** distribution
- Mathematical **function** defining all the possible values of a **random variable**, and how often they occur
- Within a given range
- Describes the long-run behavior of random variables
- How the possible values are plotted depends on:
  - Central tendency
  - Spread
  - Skewness
  - Curtosis (among others)

Do not confuse

Frequency vs probability distribution



Frequency and probability distribution

Frequency:

- Empirical
- Summarizes observed data

Probability:

- Theoretical
- Calculates the observation probabilities

An example

- We captured 5 sandgrouses, 7 ravens, and 3 falcons
- Frequency distribution of species is 5, 7 and 3
- Probability distribution of species is 5/15, 7/15 and 3/15

Why use probability?

- Data come from experiment or complex systems
- Variability and randomness
- Measure uncertainty in the data
- Probability measures uncertainty

Some definitions

- **random variable:** variable which outcome varies from measurement to measurement
- **Probability distribution:** possible values a random variable can take, and how likely they are
- Behavior of random variables must be defined using **probability**.

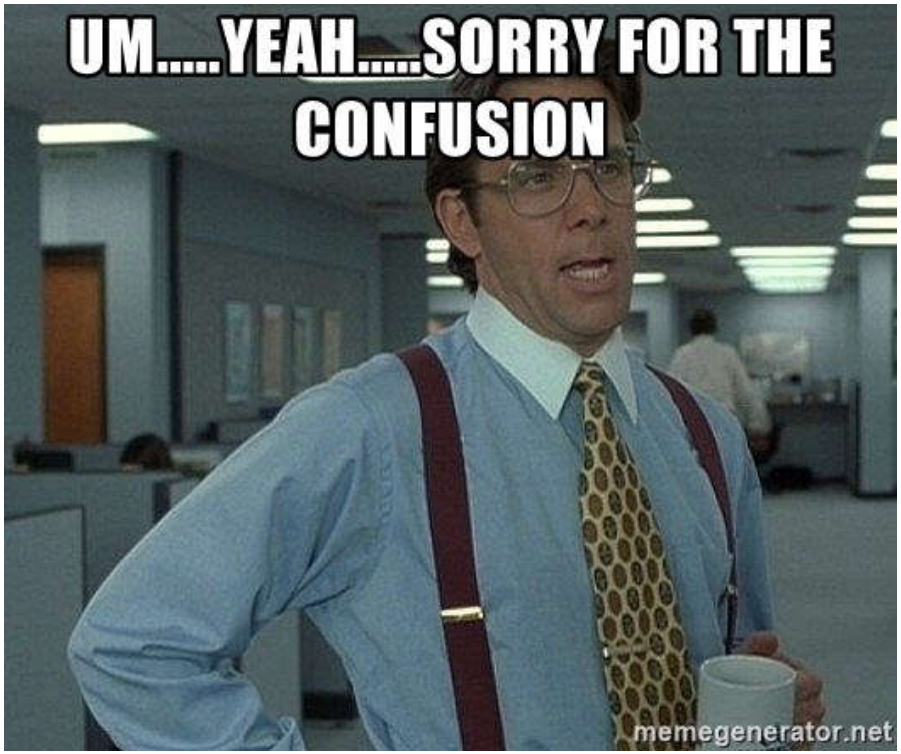
Do not confuse random variable and variable

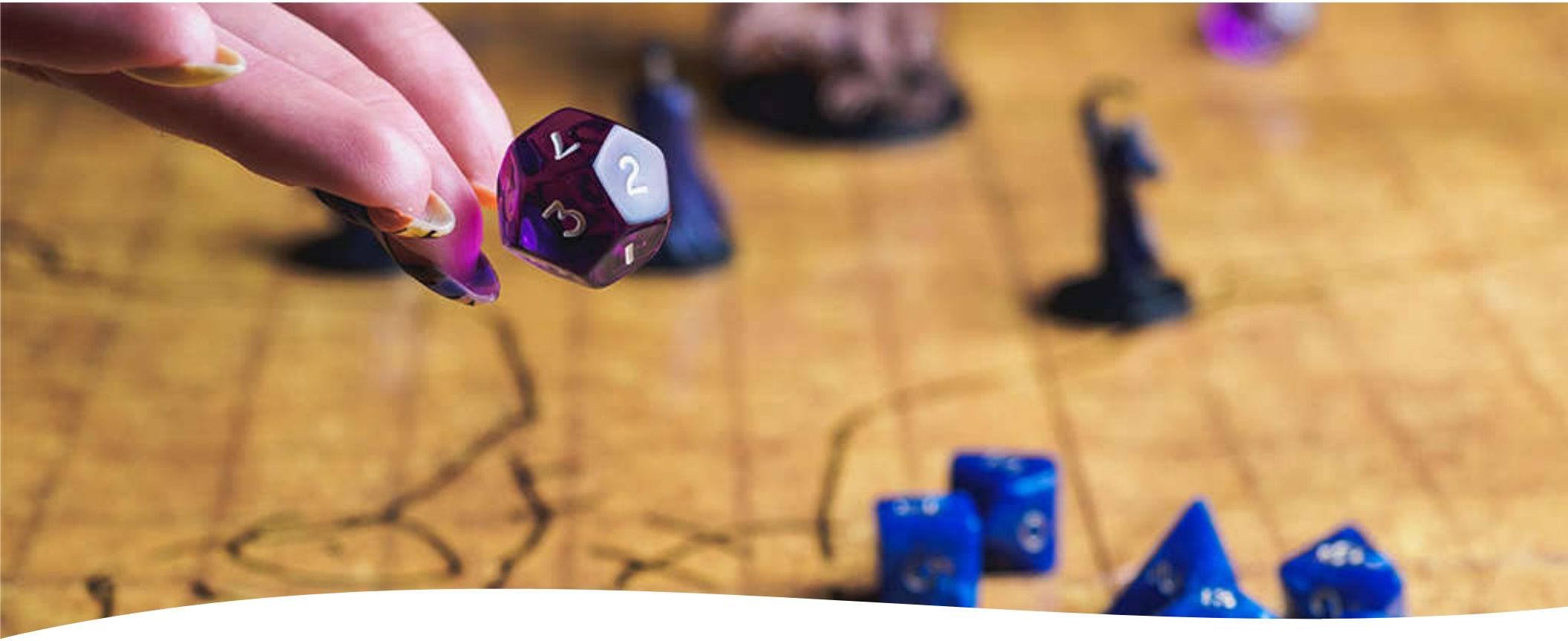
A **random variable** is always numerical and depends on the outcome of a chance experiment.

A **variable** is any property that you can measure or control

Unfortunate choice of words

- Random variables: not random & not variables. They are functions mapping from possible outcomes of sample space to measurable space
- Categorical variables can also describe random outcomes

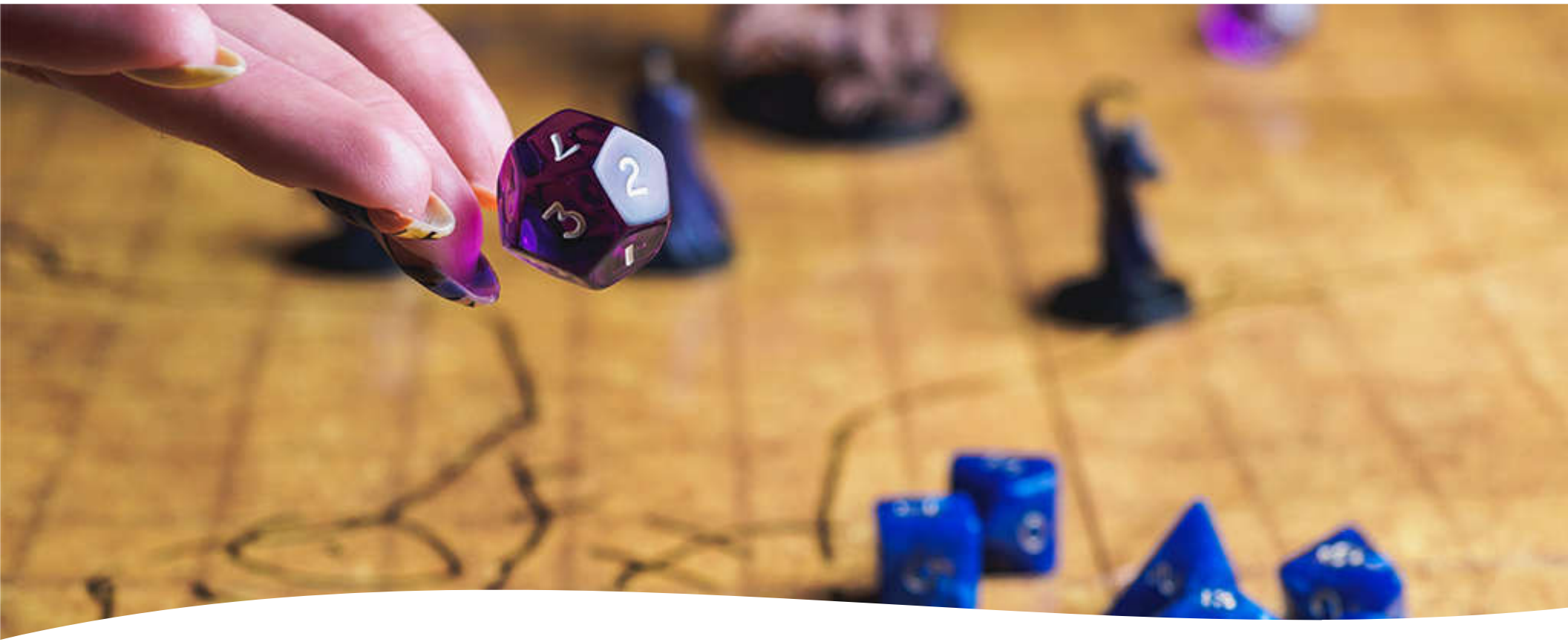




Some more definitions

- **Chance experiment**

Uncertain situation with 2 or more possible outcomes



- **Chance experiment**

Uncertain situation with 2 or more possible outcomes

Your dragonborn barbarian wants to open a closed door in the dungeon using her Sword of Vengeance.

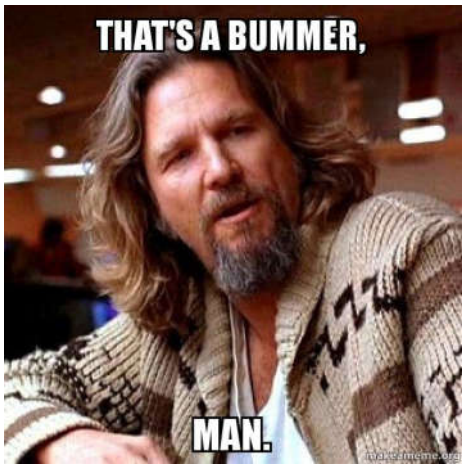
When you roll your d20 to perform the deed, you'll have 20 possible outcomes from nat 1 to nat 20.

Some more definitions

- **Outcome**

Result of a single trial

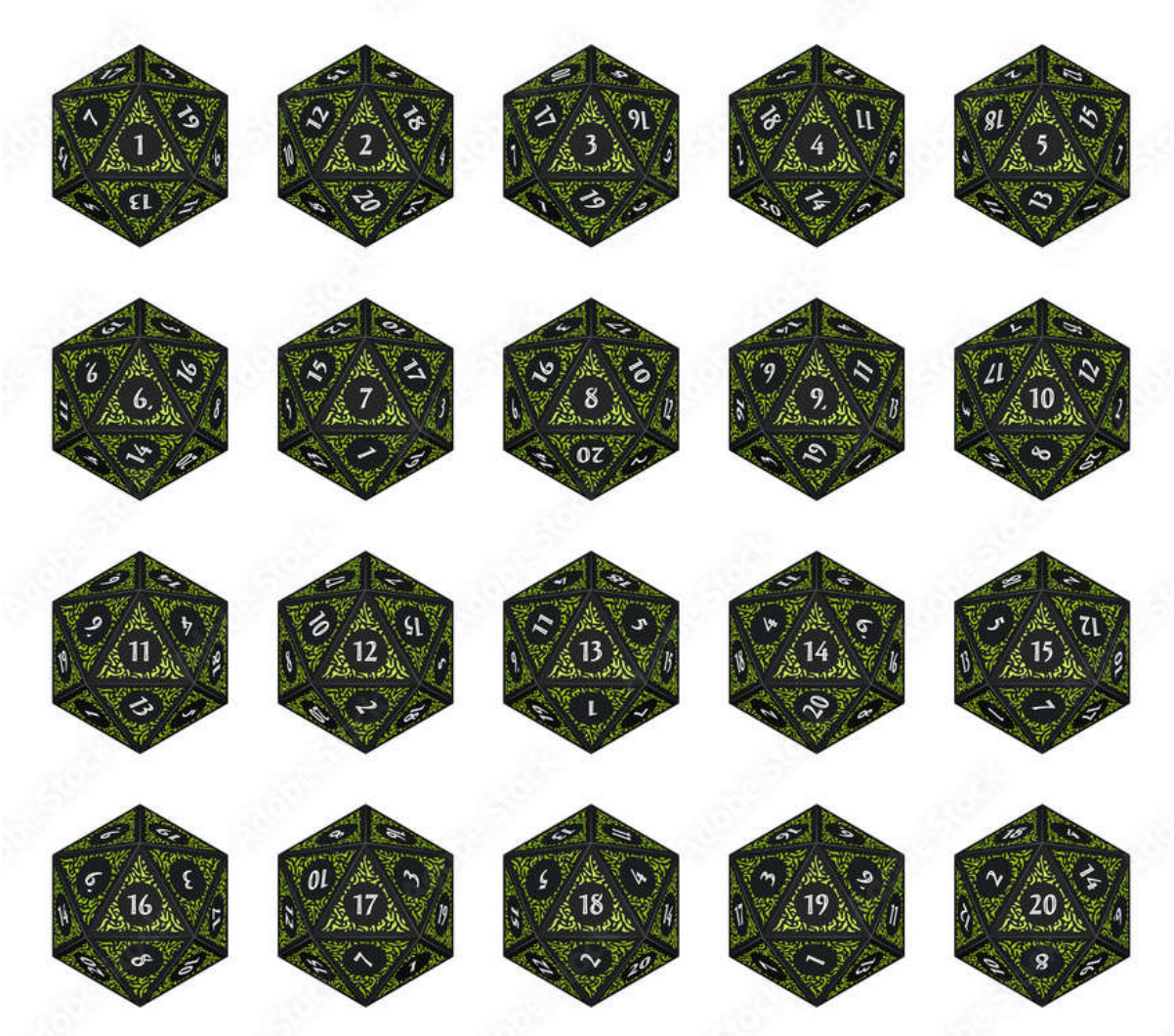
You rolled your d20 and got a nat 1.



Some more definitions

- **Sample space**

Collection of all possible outcomes from a chance experiment



The sample space when rolling your d20 is all integers from 1 to 20.

Some more definitions

- **Event**

Set of outcomes from a chance experiment

Event "nat. 20 or 1": all dice rolls that are either 20 or 1.

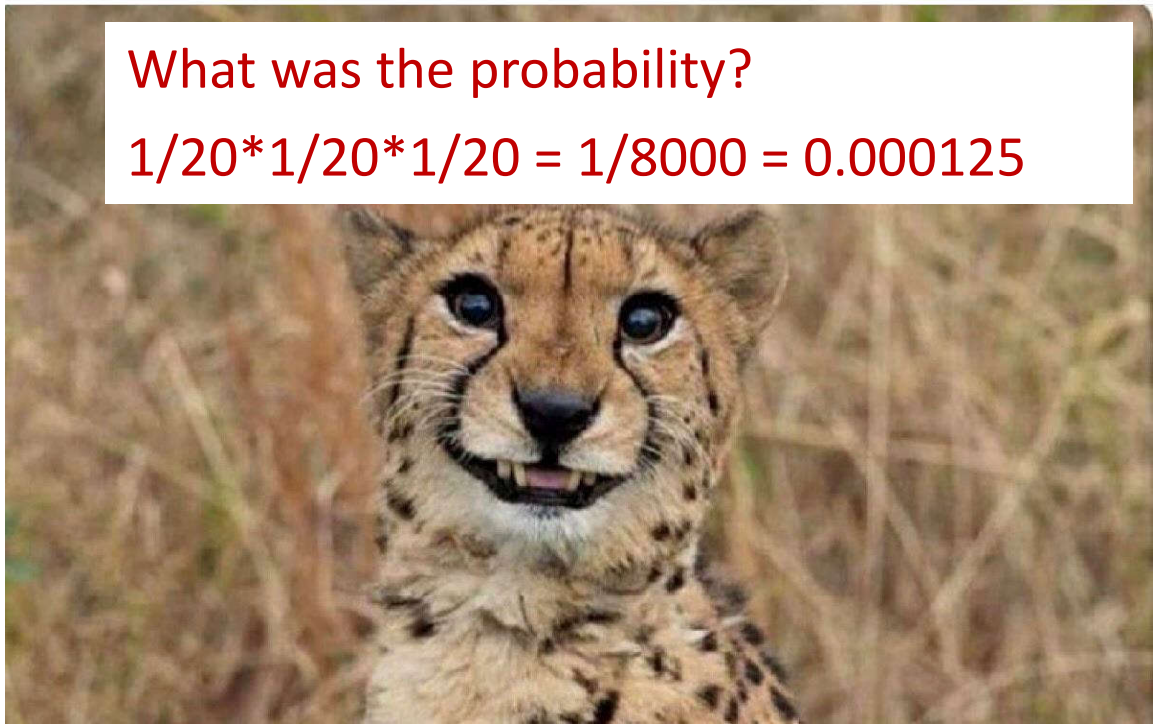
Some more definitions

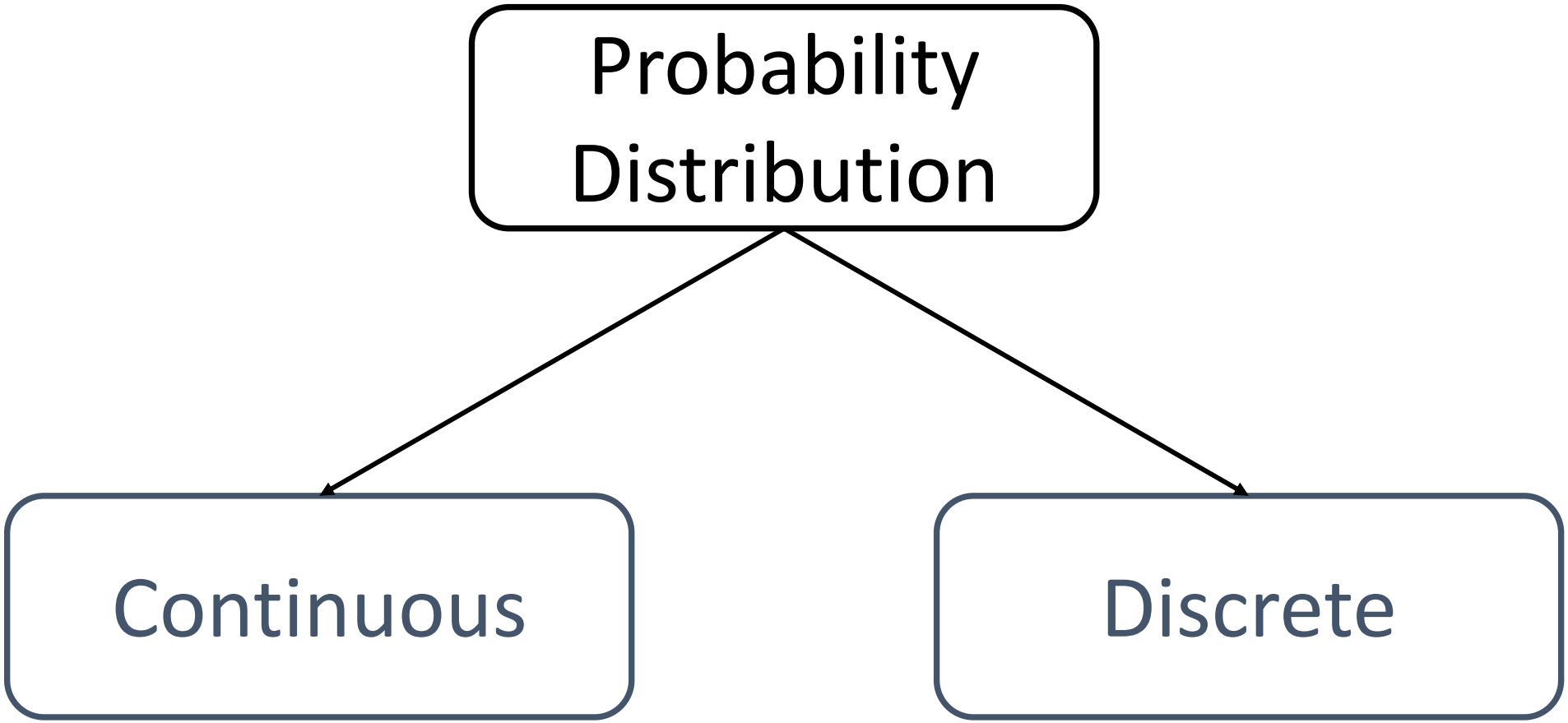
- **Probability**

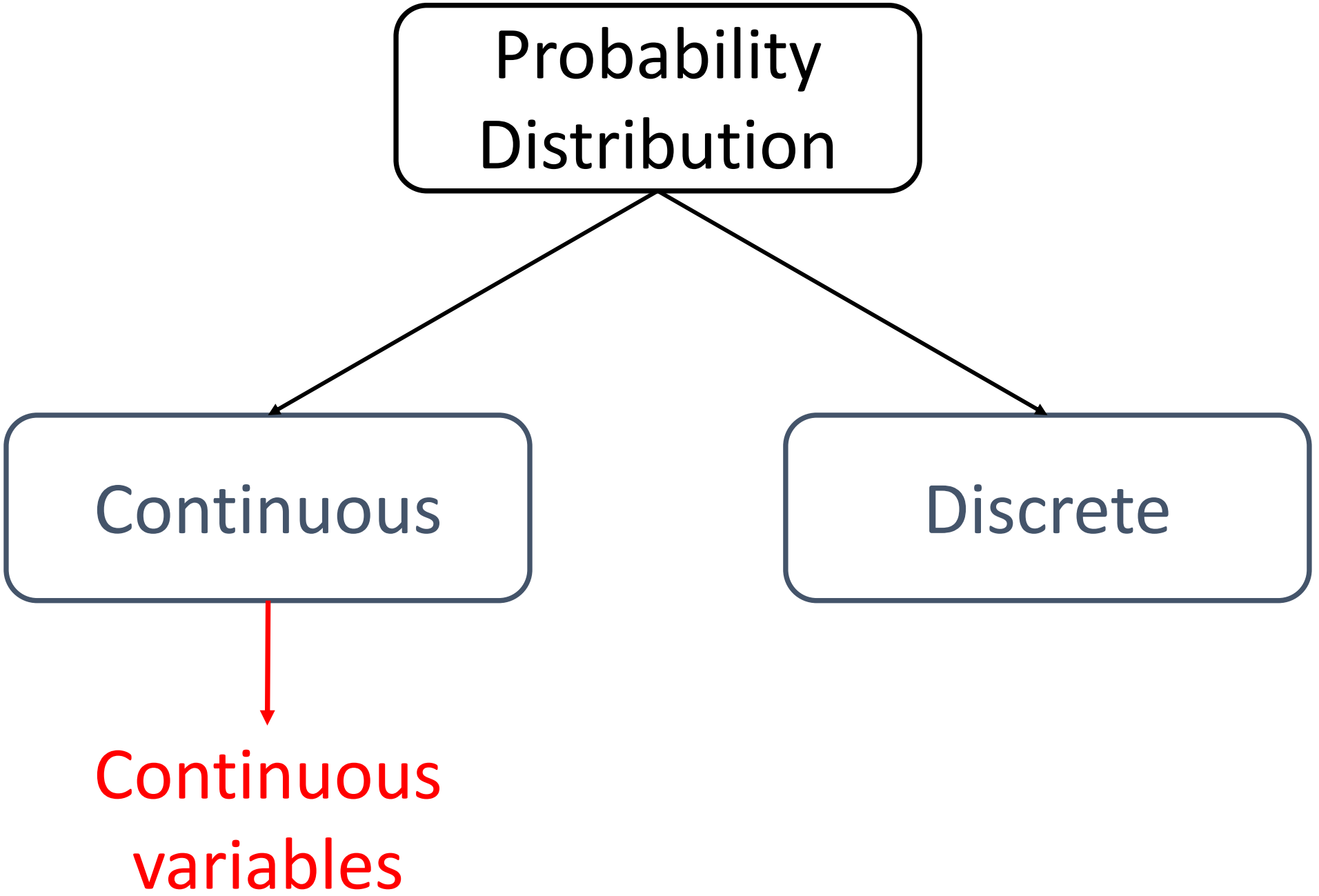
Mesure of likelihood of an event occurence

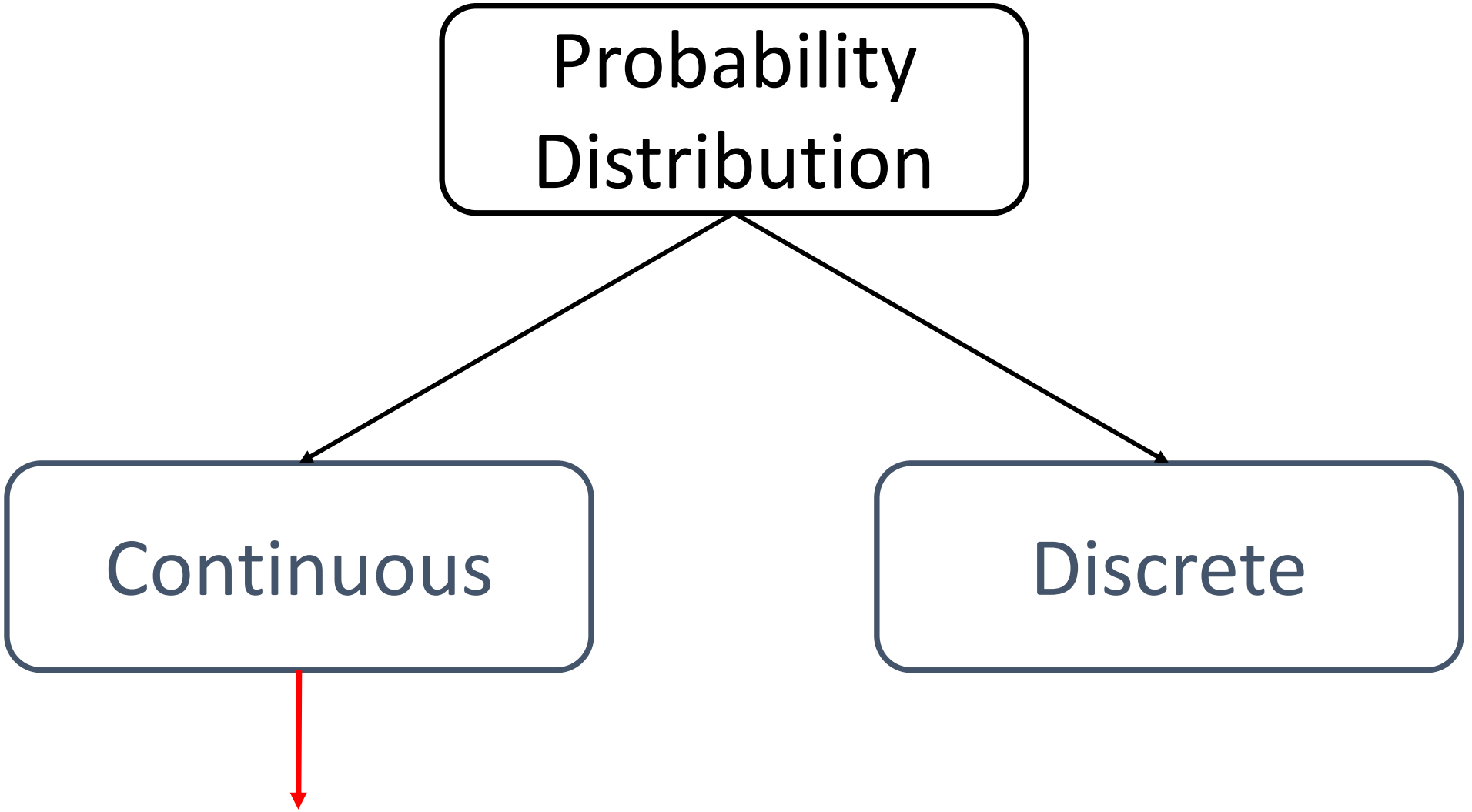
The Barbarian dragonborn rolls nat 1 three times in a row, almost killing off the party.

My tabaxi bloodhunter:



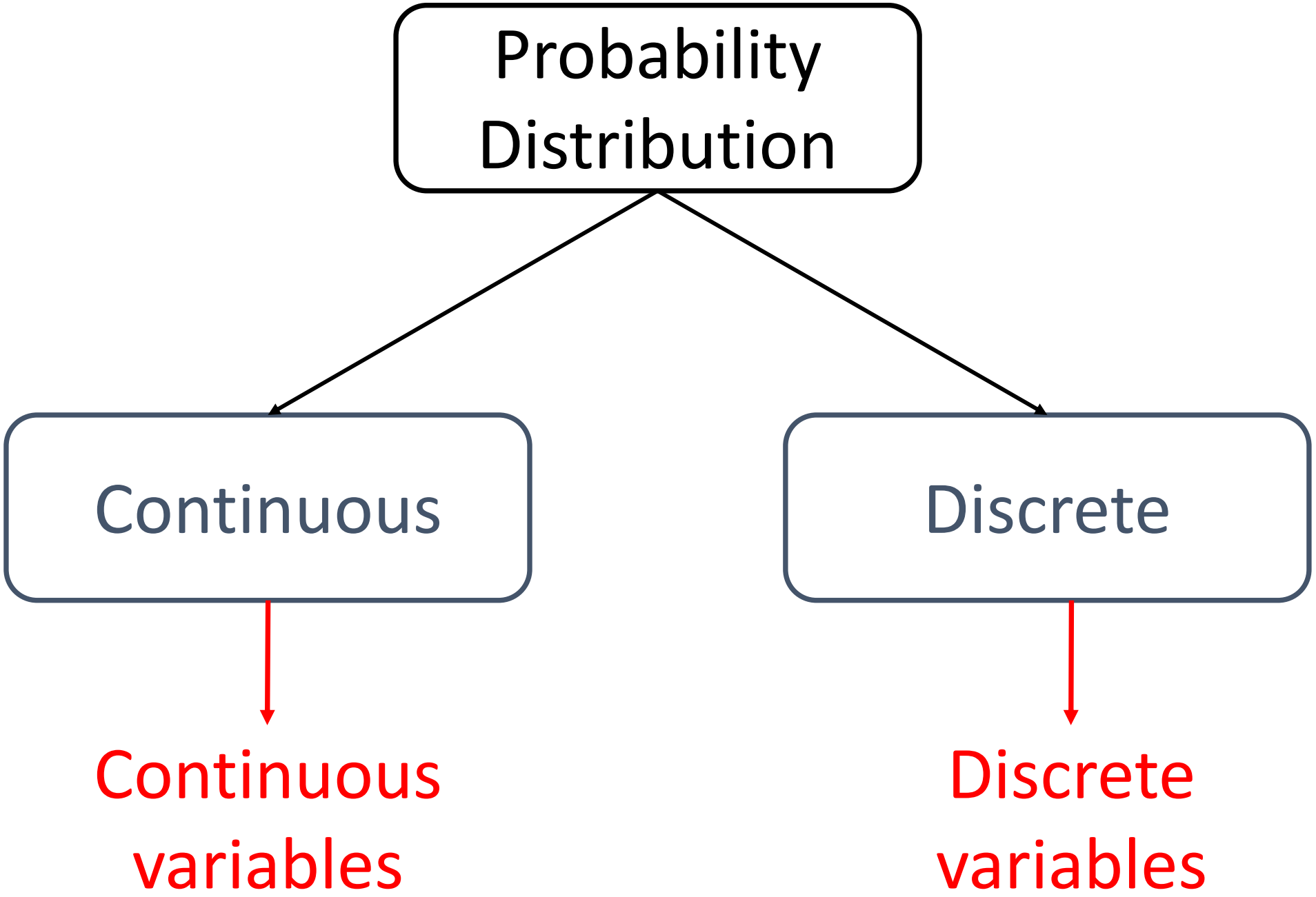

Probability Distribution
Continuous
Discrete


Probability Distribution
Continuous
Discrete
Continuous variables

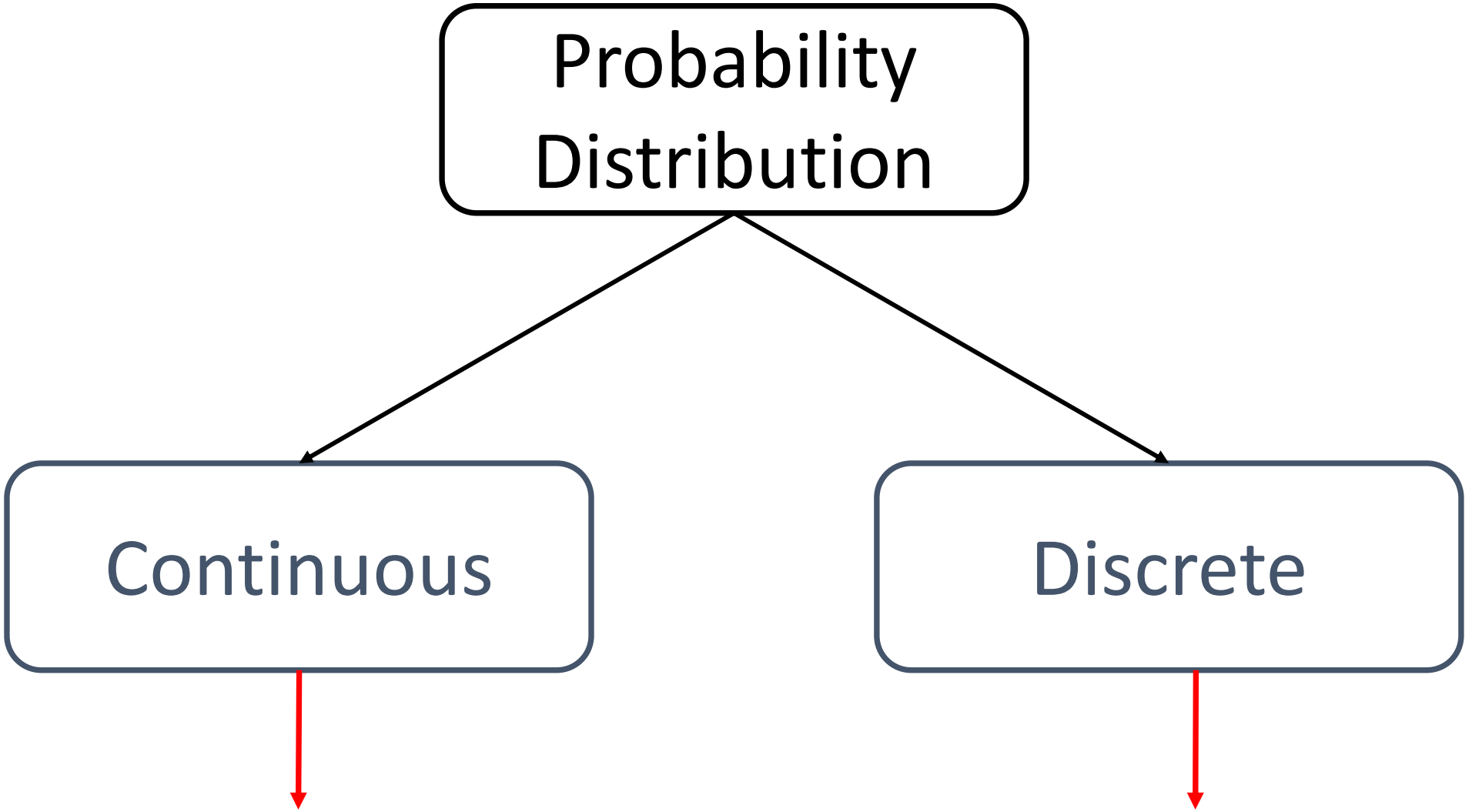


Continuous variables

Continuum of value within range

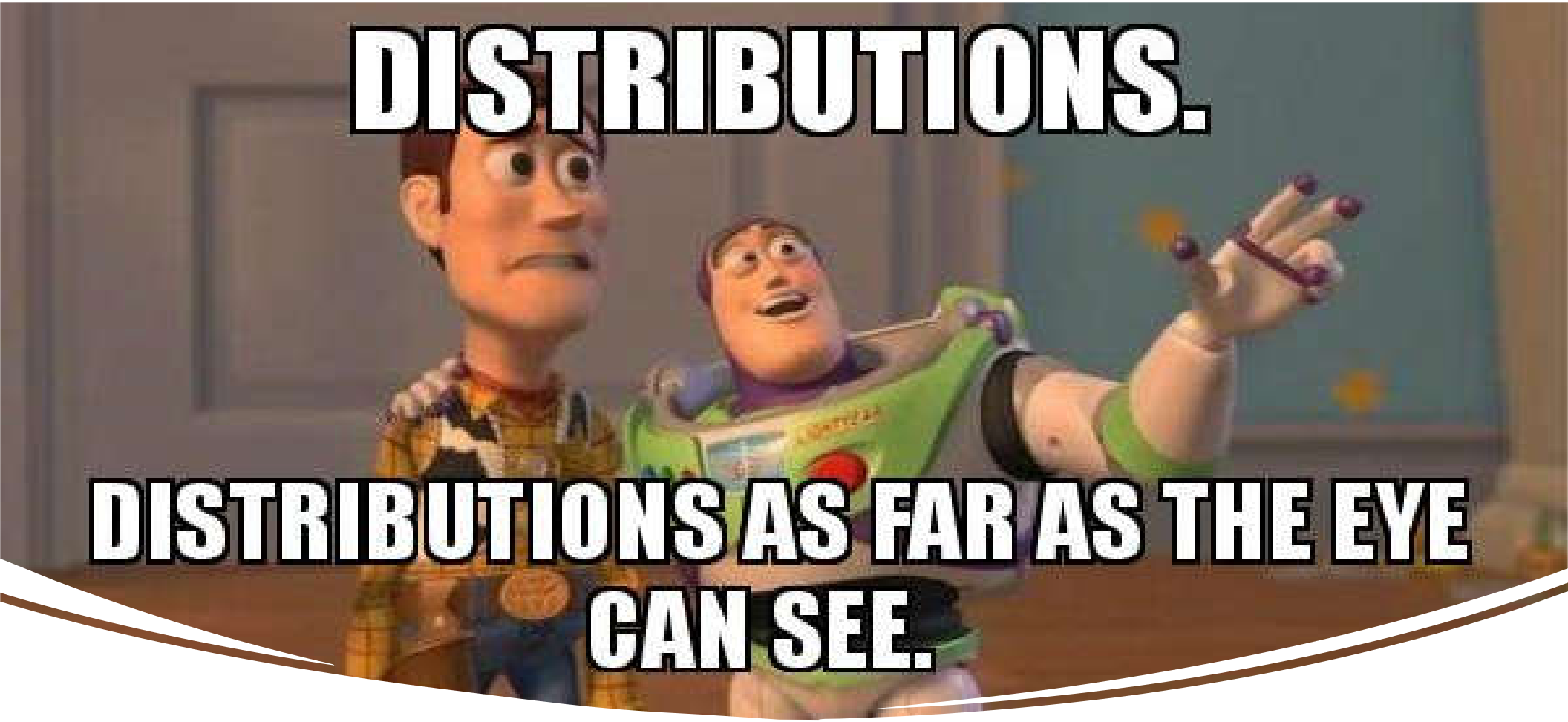


Continuum of value within range



Continuum of value within range

Only takes distinct values



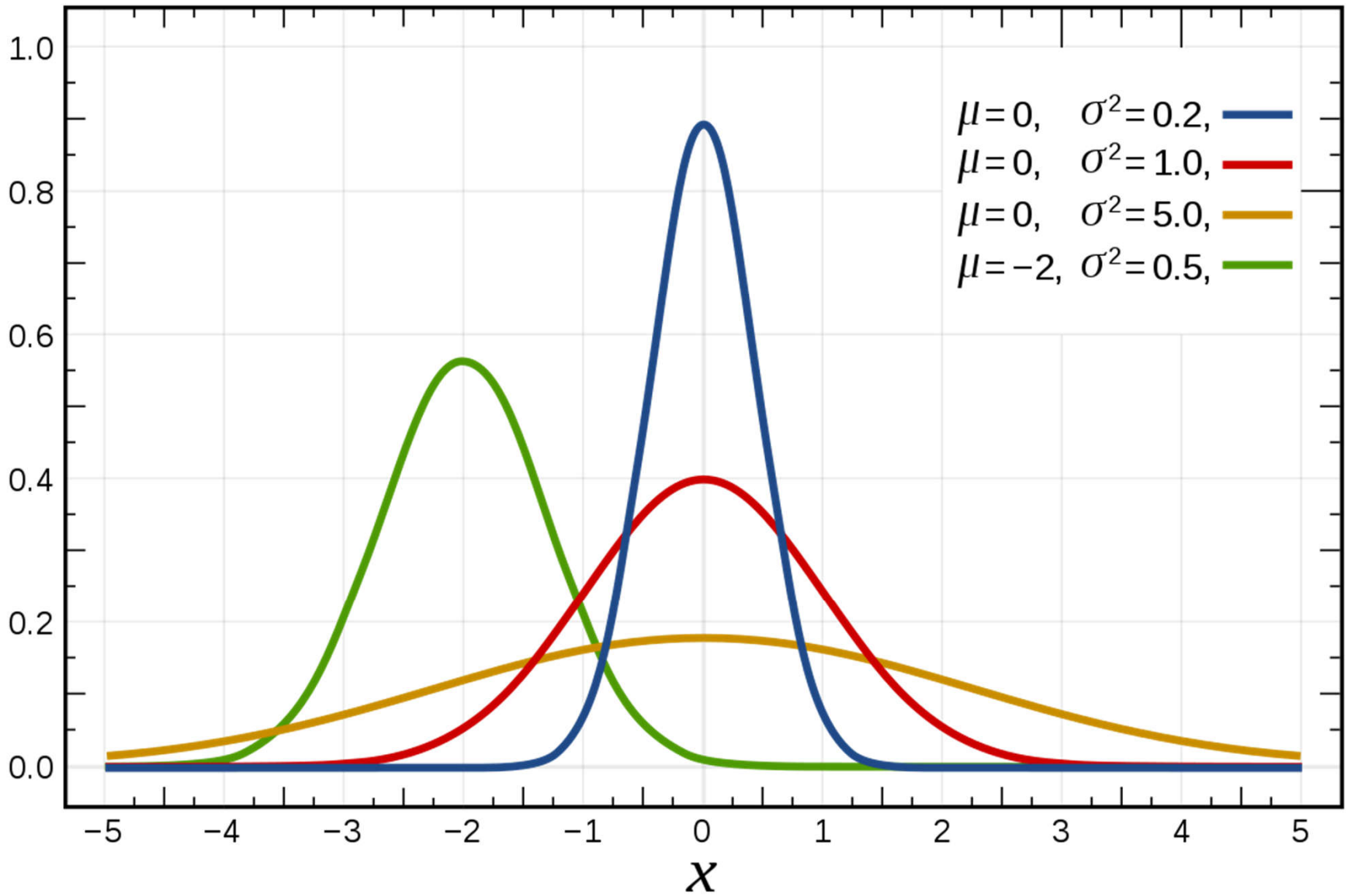
Most common distributions

Continuous variable:

- Normal (a.k.a. Gaussian)
- Uniform (a.k.a. rectangular)
- **Beta**
- gamma

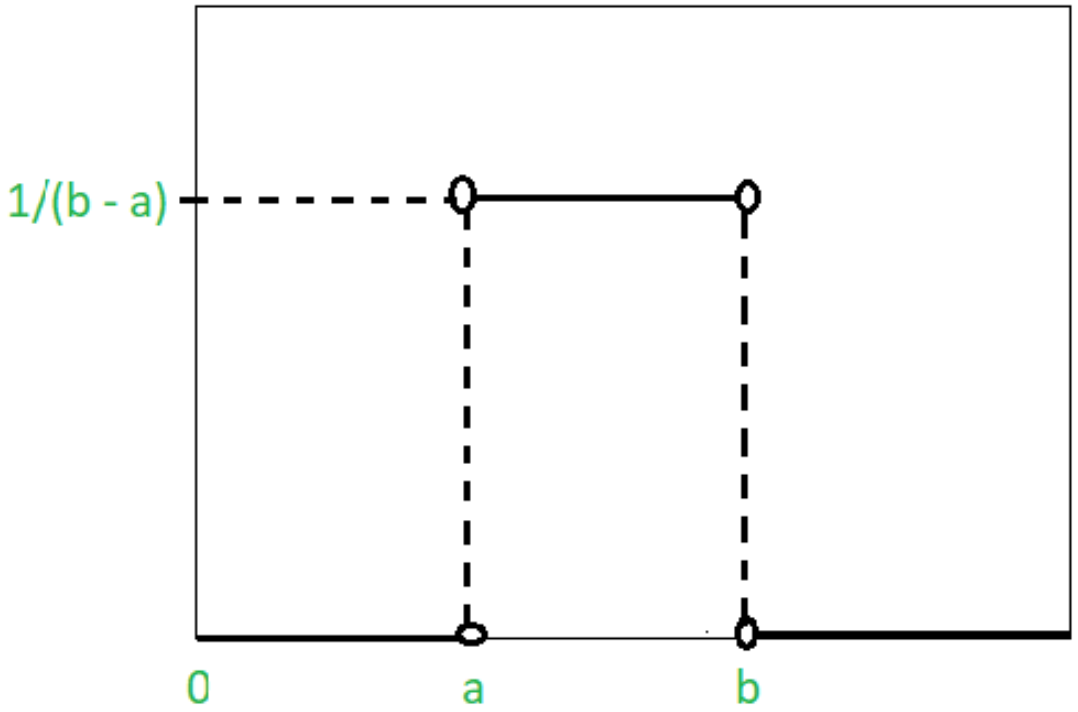
Gaussian or Normal distribution

- Symmetrical
- Defined by μ and σ



Uniform or rectangular distribution

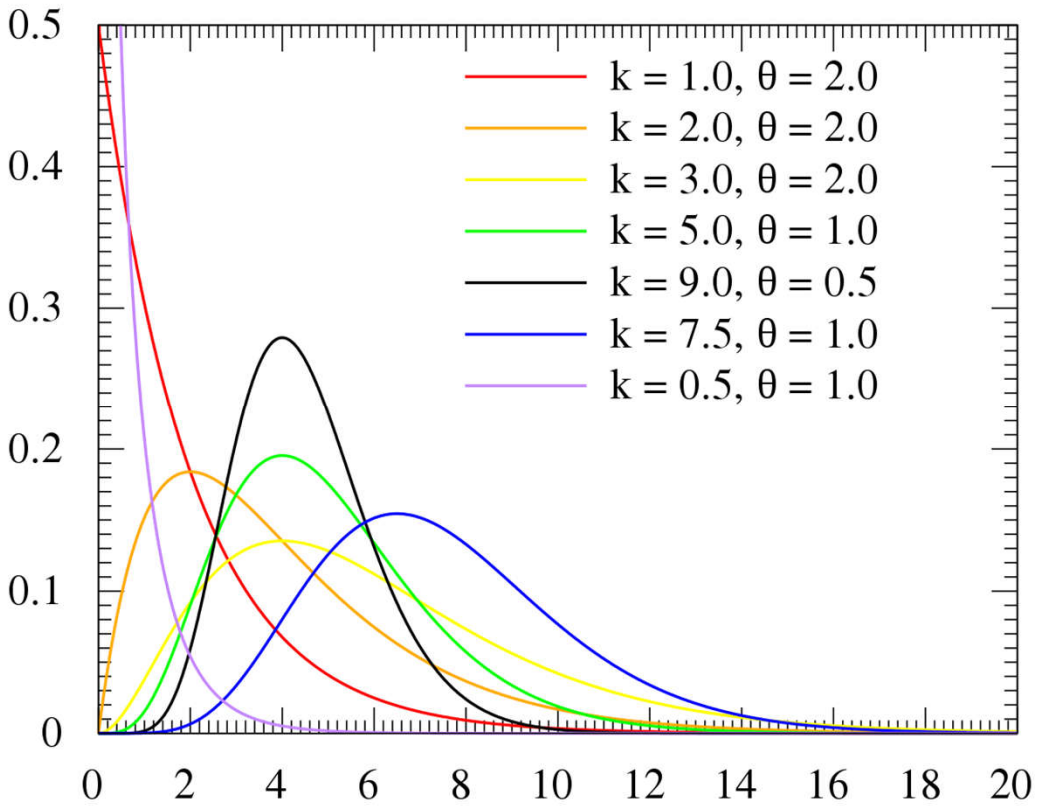
Defined by a (minimum value) and b (maximum value)



Constant probability over interval [a-b]

Gamma distribution

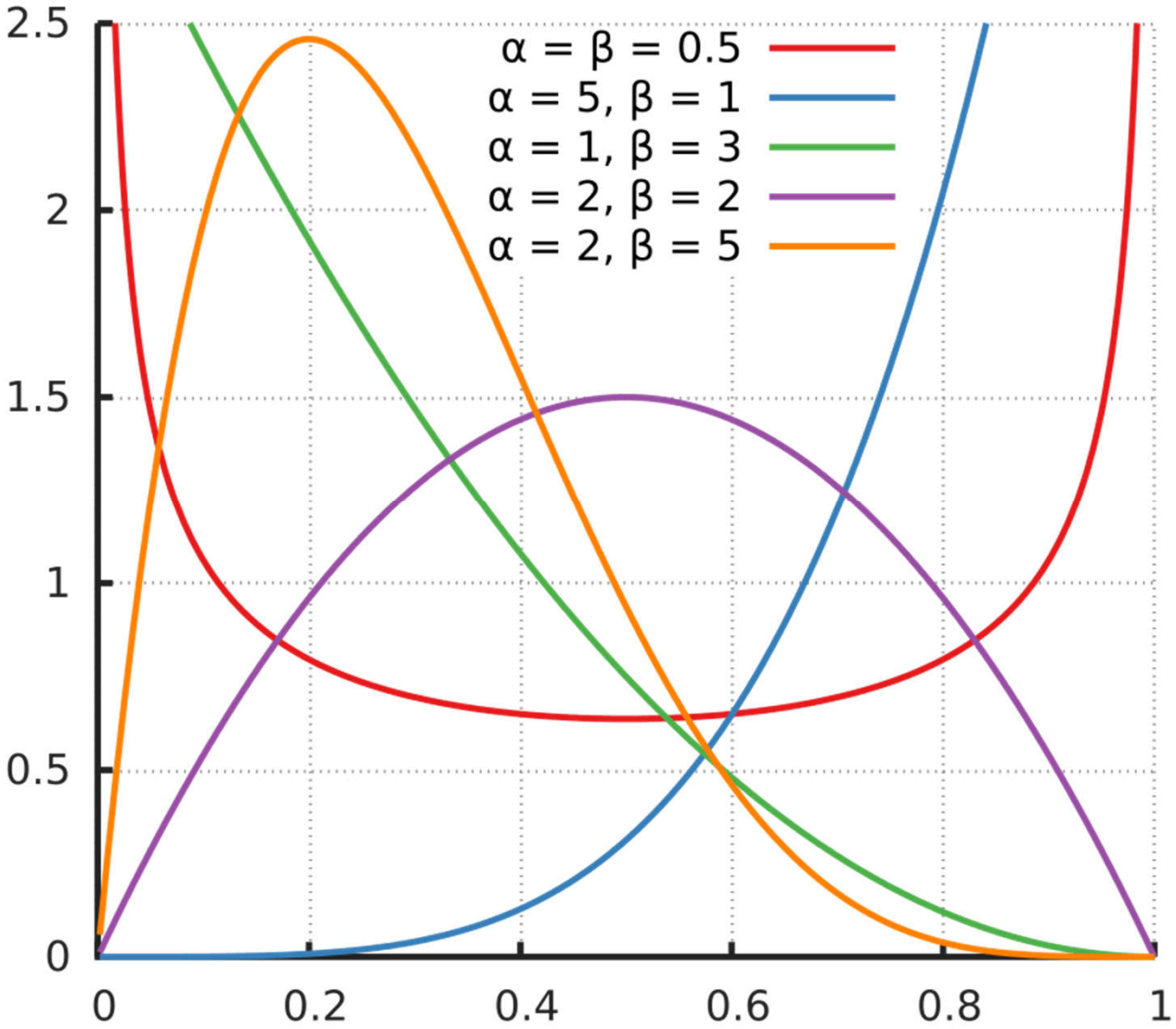
- Positive and right-skewed
- Defined by parameters k (shape) and θ (scale)



- Continuous, positive, right-skewed data with constant variance on the log-scale
- Survival data, rainfall,…
- χ^2 distribution: special case of gamma
- Used in χ^2 goodness-of-fit tests

Beta distribution

- Defined on [0,1]
- Defined by shape parameters α and β, both >0

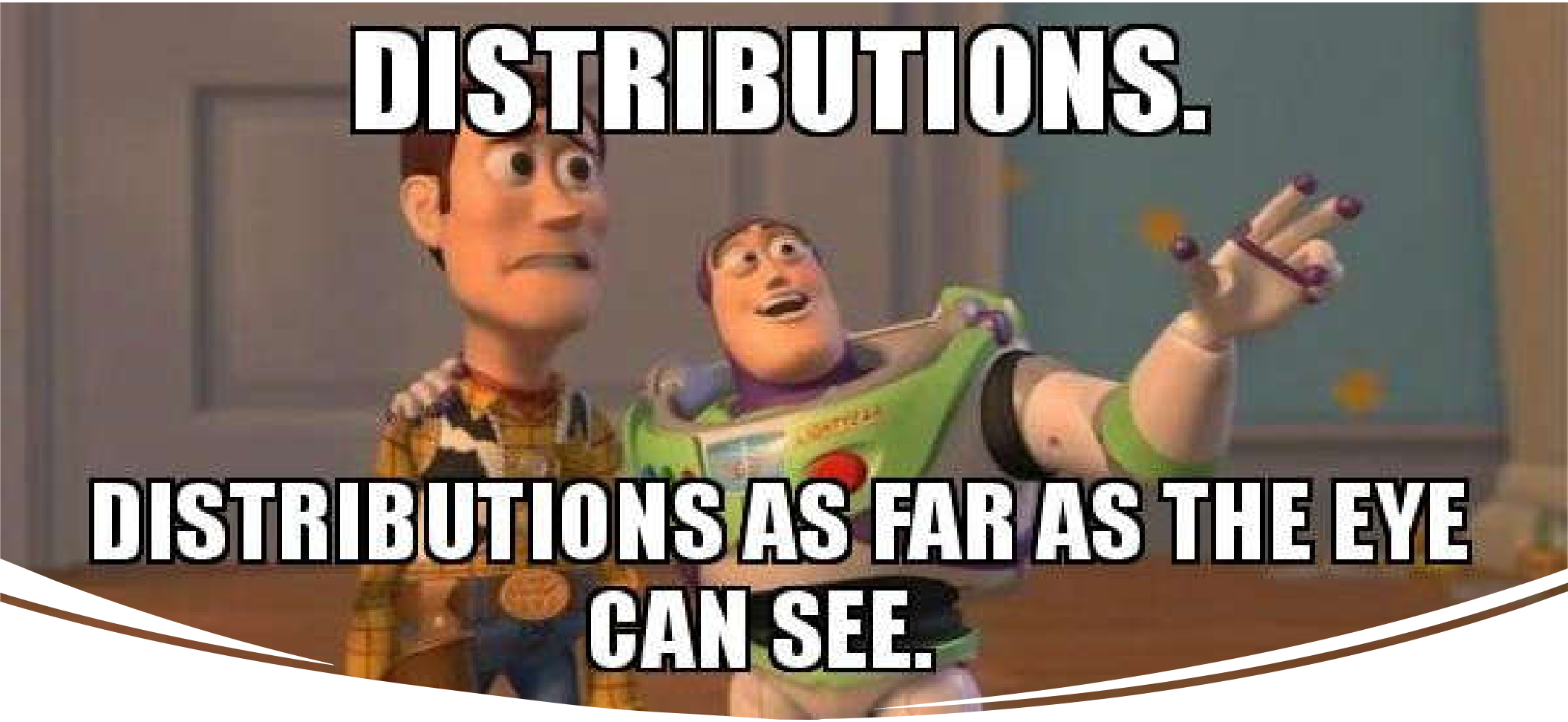


Beta distribution

- Derived variables, notably proportions
- Doesn't deal with exact 0s and 1s *(loglikelihood function contains ln(x) and ln(1-x), and thus unbounded since ln(1) = 0 and ln(0) = -∞)*

When 0 and 1 occur:

- Possible transformation: (y * (n−1) + 0.5) / n, *n = sample size*
- Zero-one inflated beta models



Most common distributions

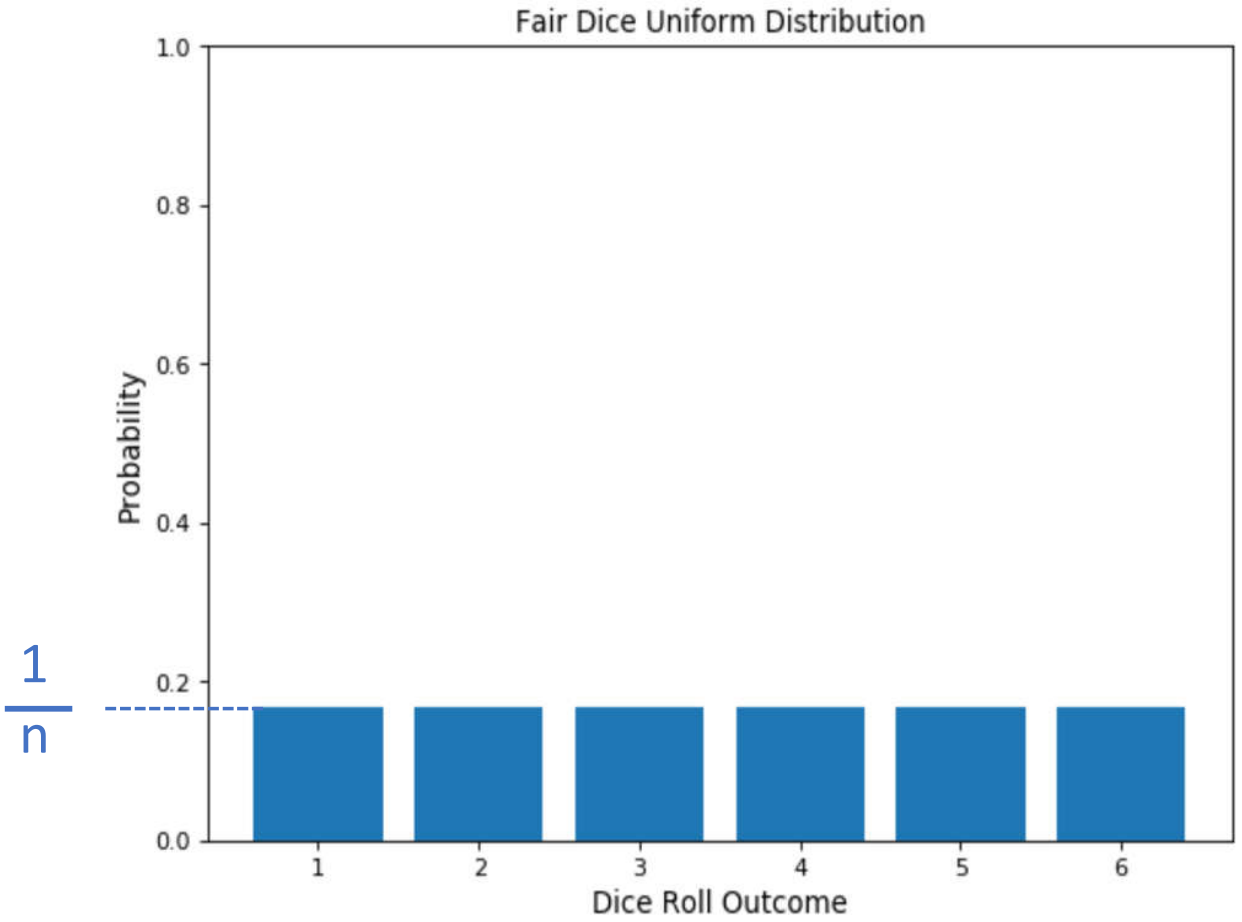
Discrete variables:

- Uniform (Yes, again!)
- Poisson
- Negative Binomial
- Binomial

Discrete uniform distribution

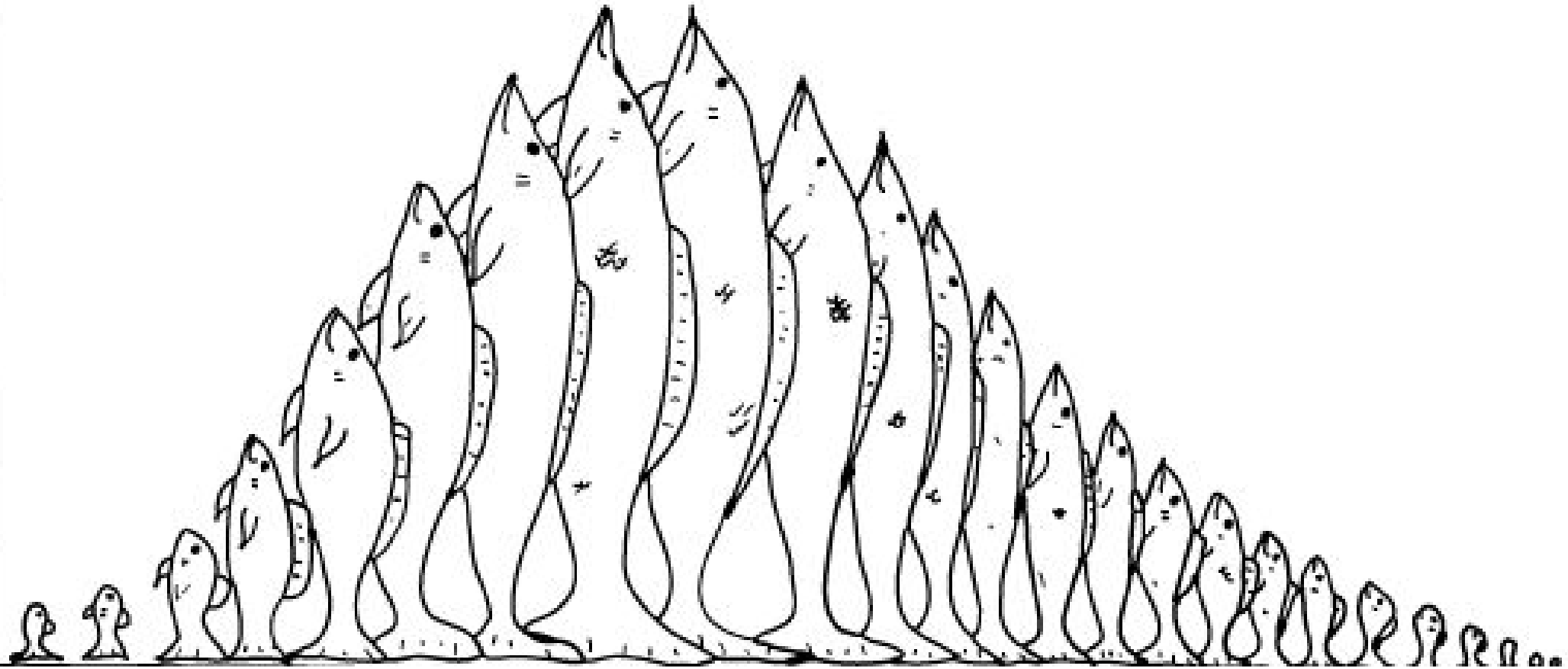
All outcome equally likely

Defined by a (min) and b (max) like the continuous



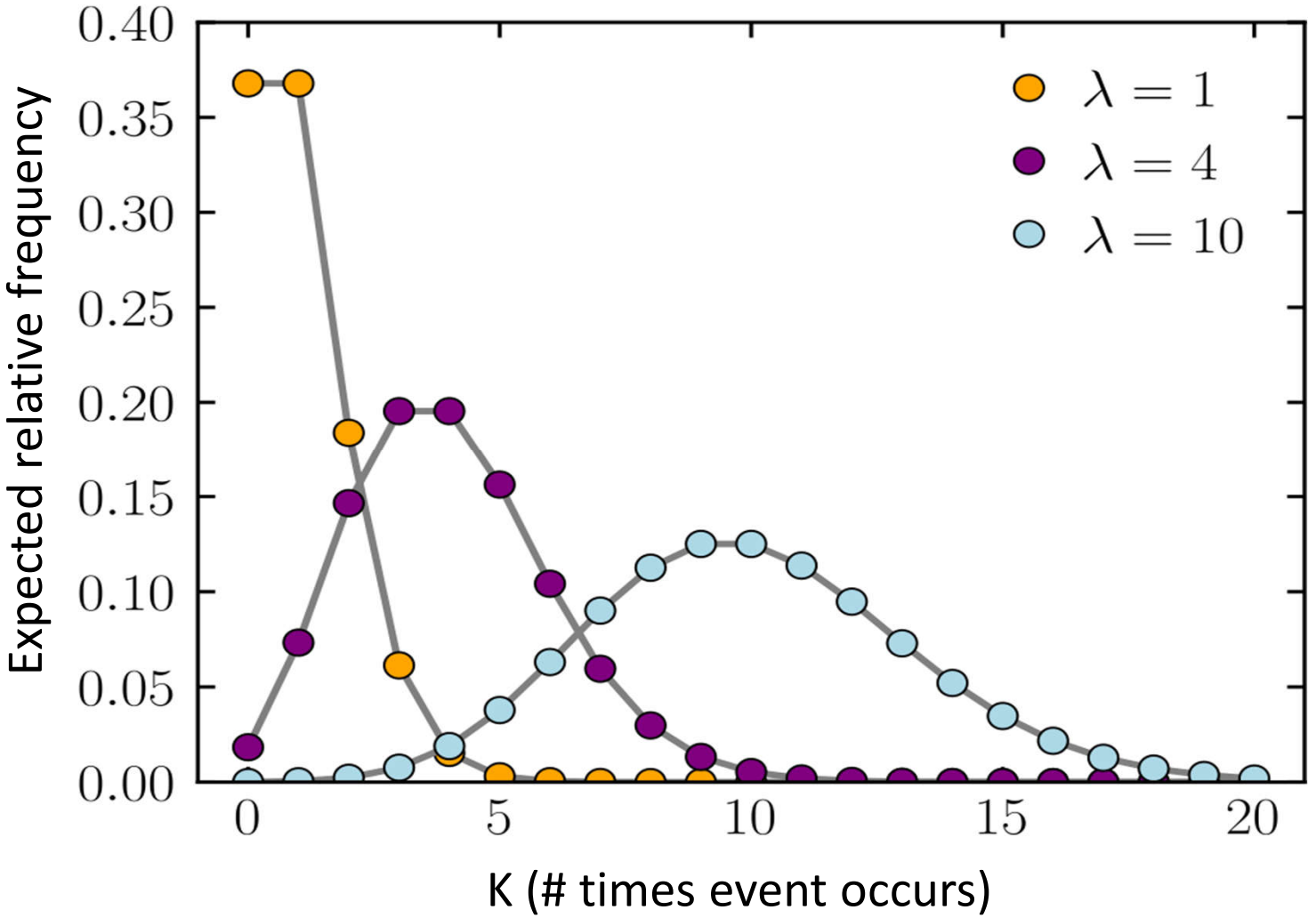
The Poisson distribution

- Count distribution
- Defined by 1 parameter λ mean number of events
- $\lambda = \mu = \sigma^2$ (lambda = mean = var)



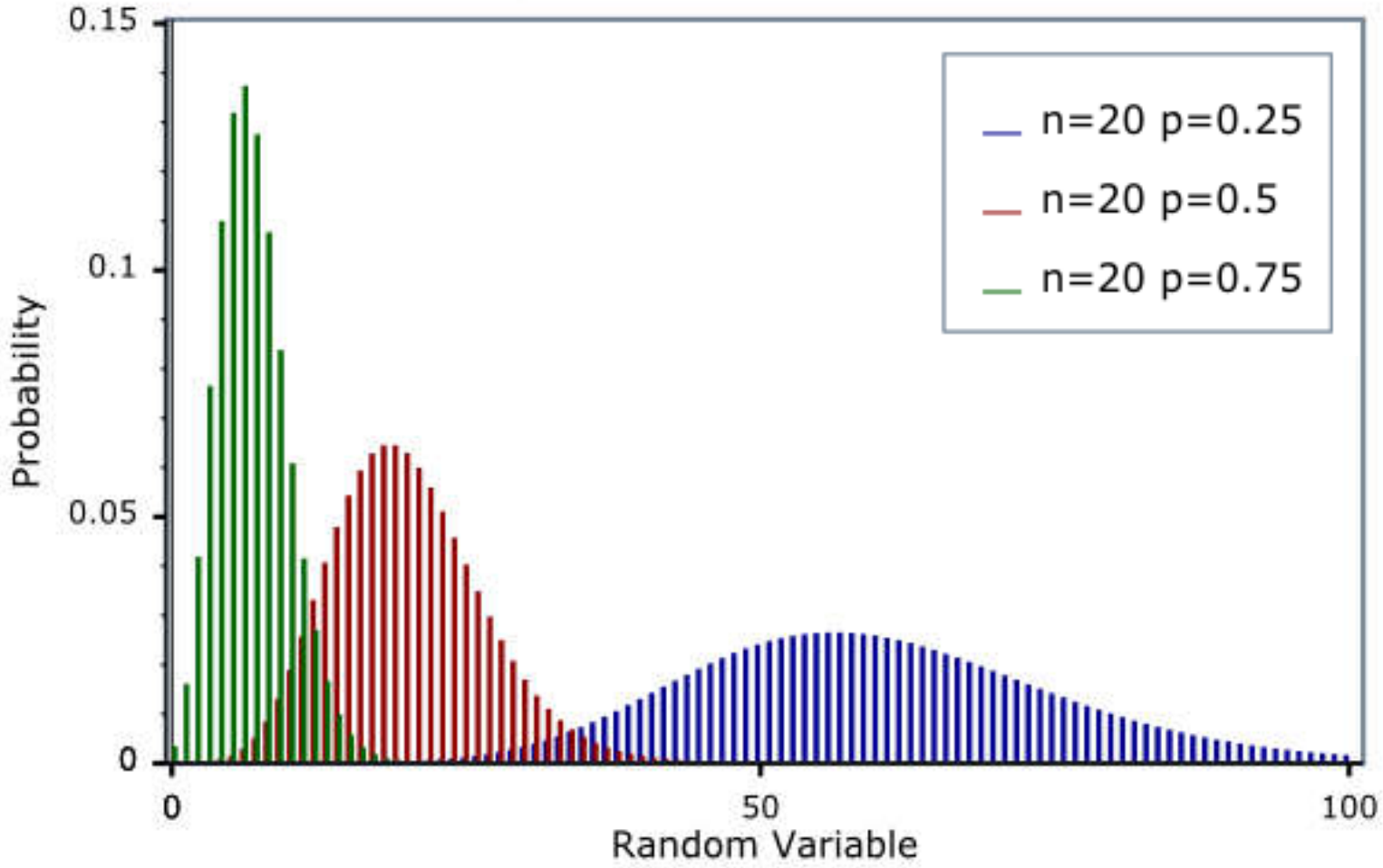
The Poisson distribution

- As mean increases, Poisson distributions approximate normal distribution
- Count data with large mean can often be modelled as continuous



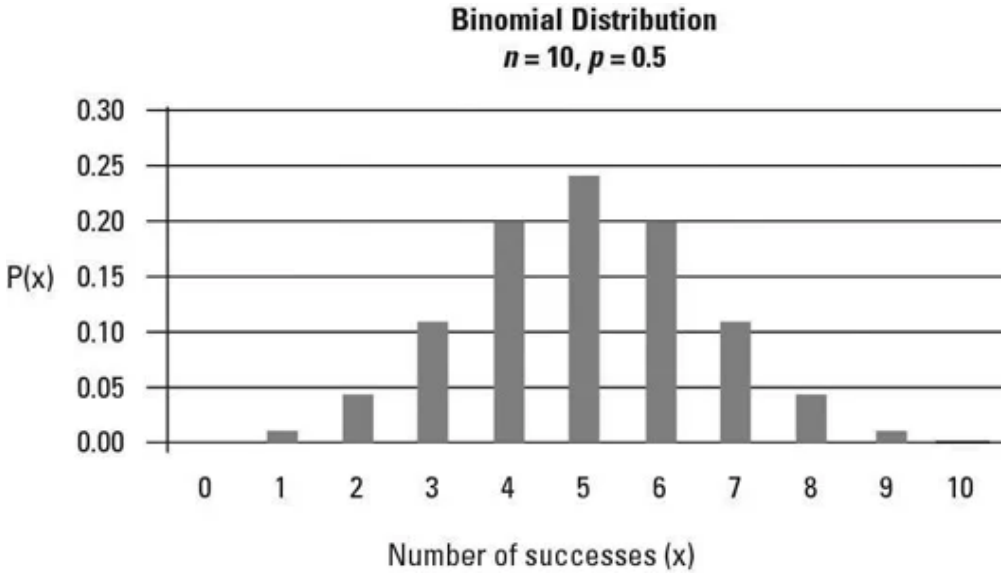
Negative binomial distribution

- In most cases $\mu = \sigma^2$ will not hold
- Overdispersion: $\mu < \sigma^2$
- 1 more parameter θ
- $\lambda = \mu$ and $\sigma^2 = \mu + \frac{\mu^2}{\theta}$



Binomial distribution

- Series of Bernoulli trials
- Bernouilli trials: only 2 outcomes
- Binary variable: 1/0
- # of 1 in n independent Bernoulli trials
- Defined by n (# trials) and p (probability of success in 1 trial)



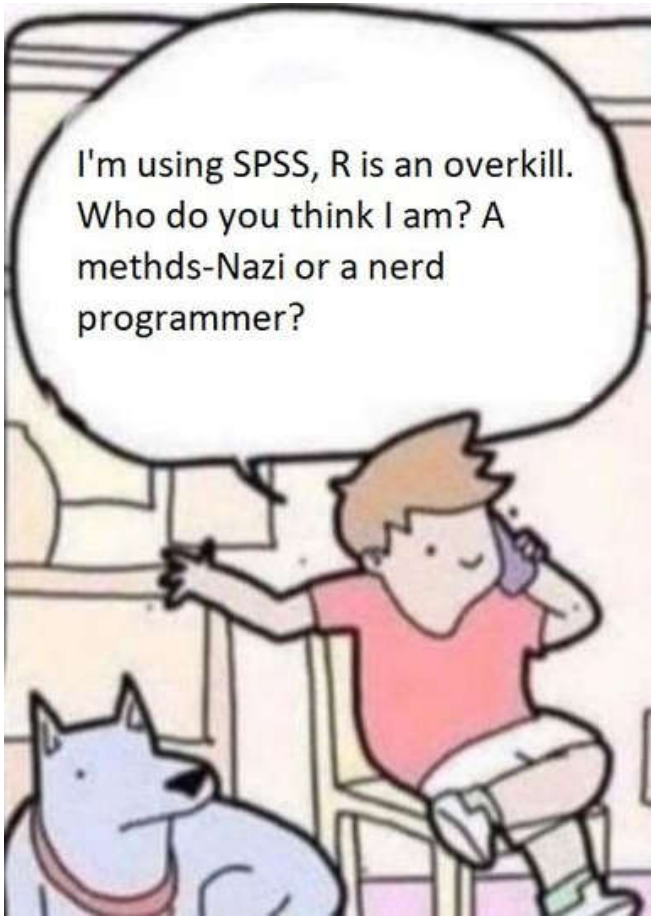
Categorical distributions

Which morphs are associated with which habitat?

- Coat color: response variable with k categories
- Multinomial distribution: possible results of a random variable that can take on one of K categories, given the specified probability of each category
- Generalizes the binomial distribution

Intro to R and RStudio

What is R?



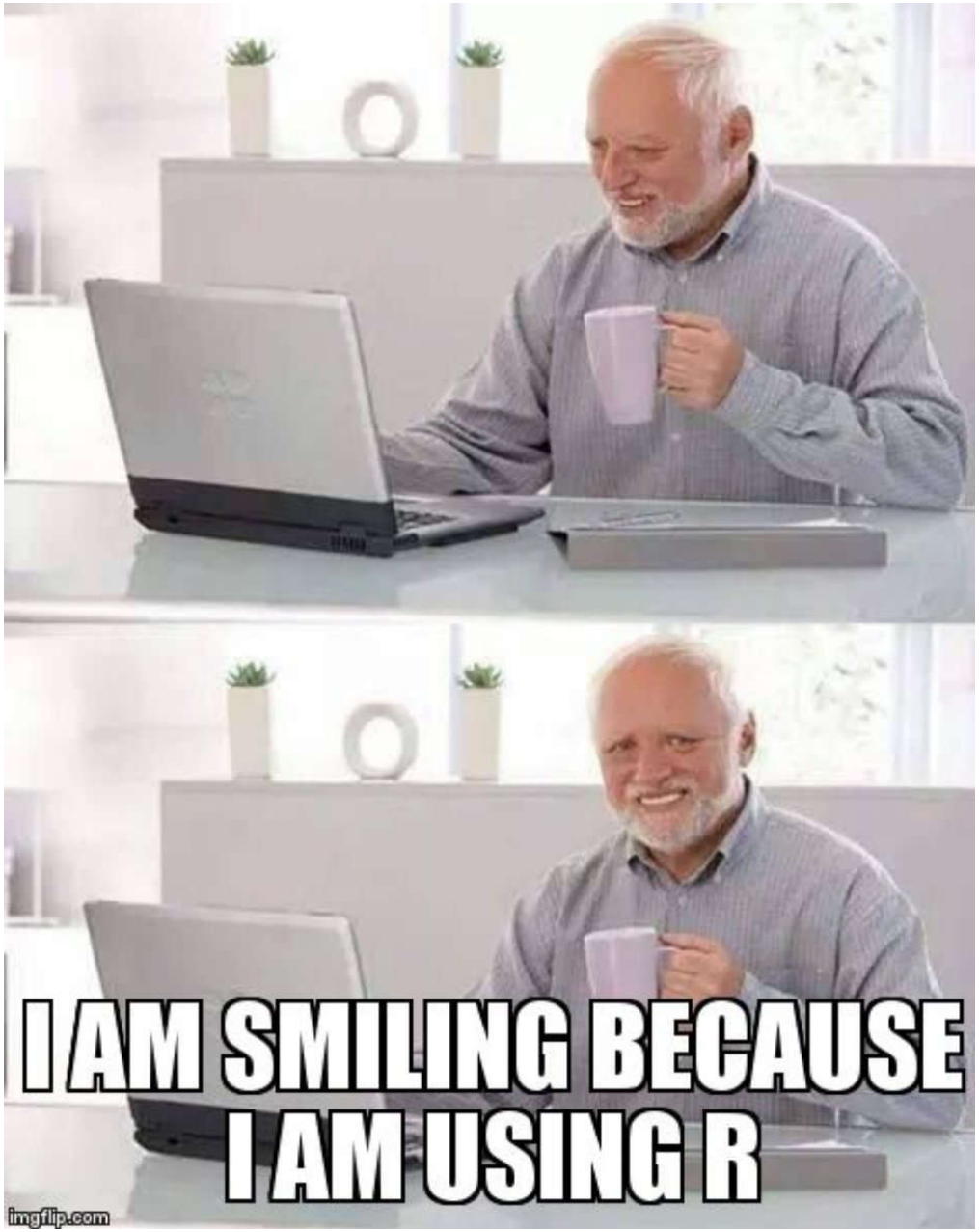
- A programming language
- A calculator
- An environment: integrated suite of software

- Open source
- Get under the hood, unlike most software: you know what you're doing and what's going on

Why is R?

- Data handling
- Computation
- Graphics

- Wide variety of:
  - Statistical techniques
  - Graphical techniques
  - Data-type it can handle



What is RStudio?

- integrated development environment for R
- A tool to make your life easier with R

Why is RStudio?

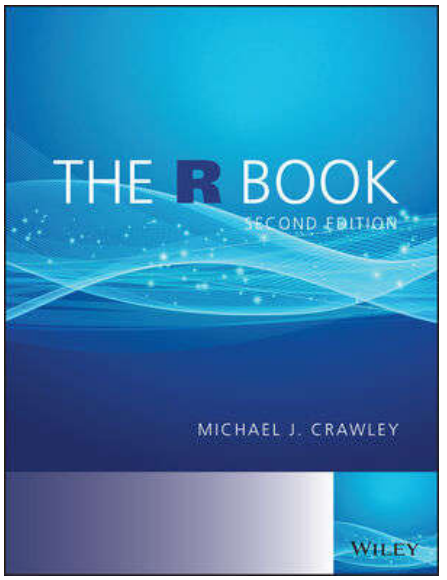
tools for:

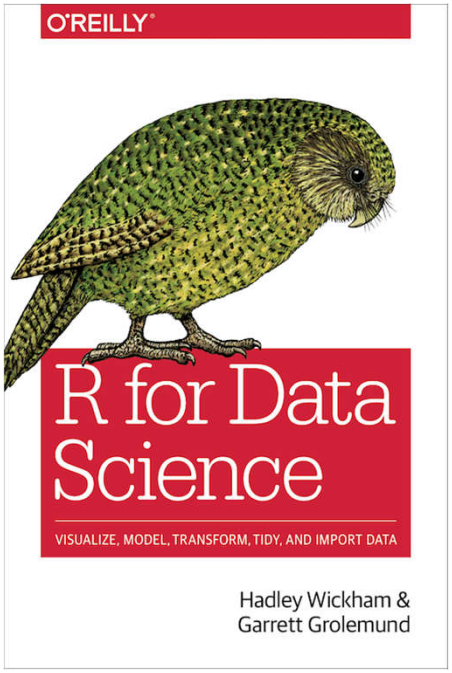
- plotting
- history
- debugging
- workspace management

sometimes you just need the comforting warm embrace of rstudio #rstats

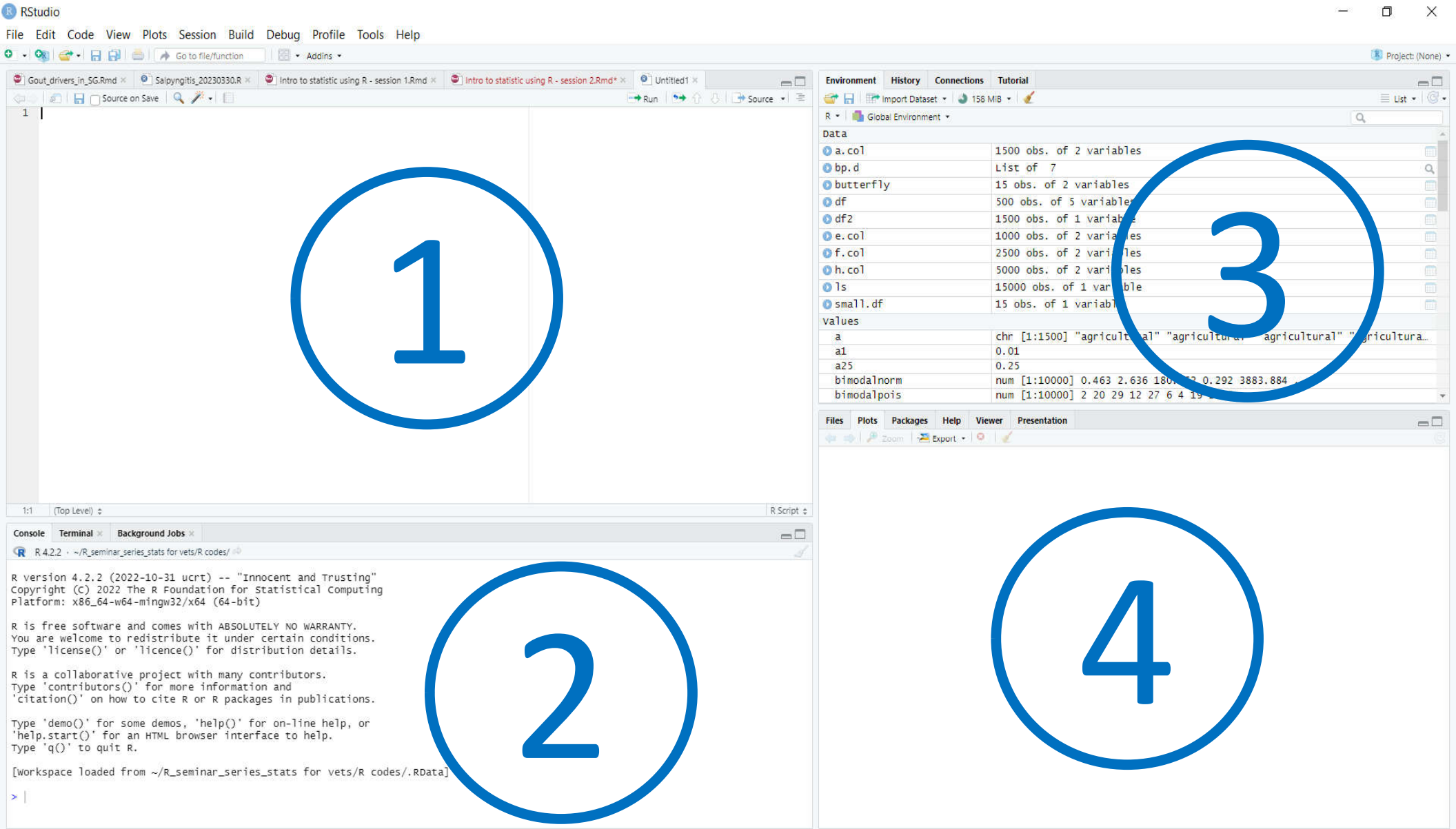
R resources

- https://r4ds.had.co.nz/ for data wrangling and graphs
- List of sites I provided during the last seminar
- I'm happy to help you trouble shoot your code, but from now on, **I'm expecting you to first try these resources**

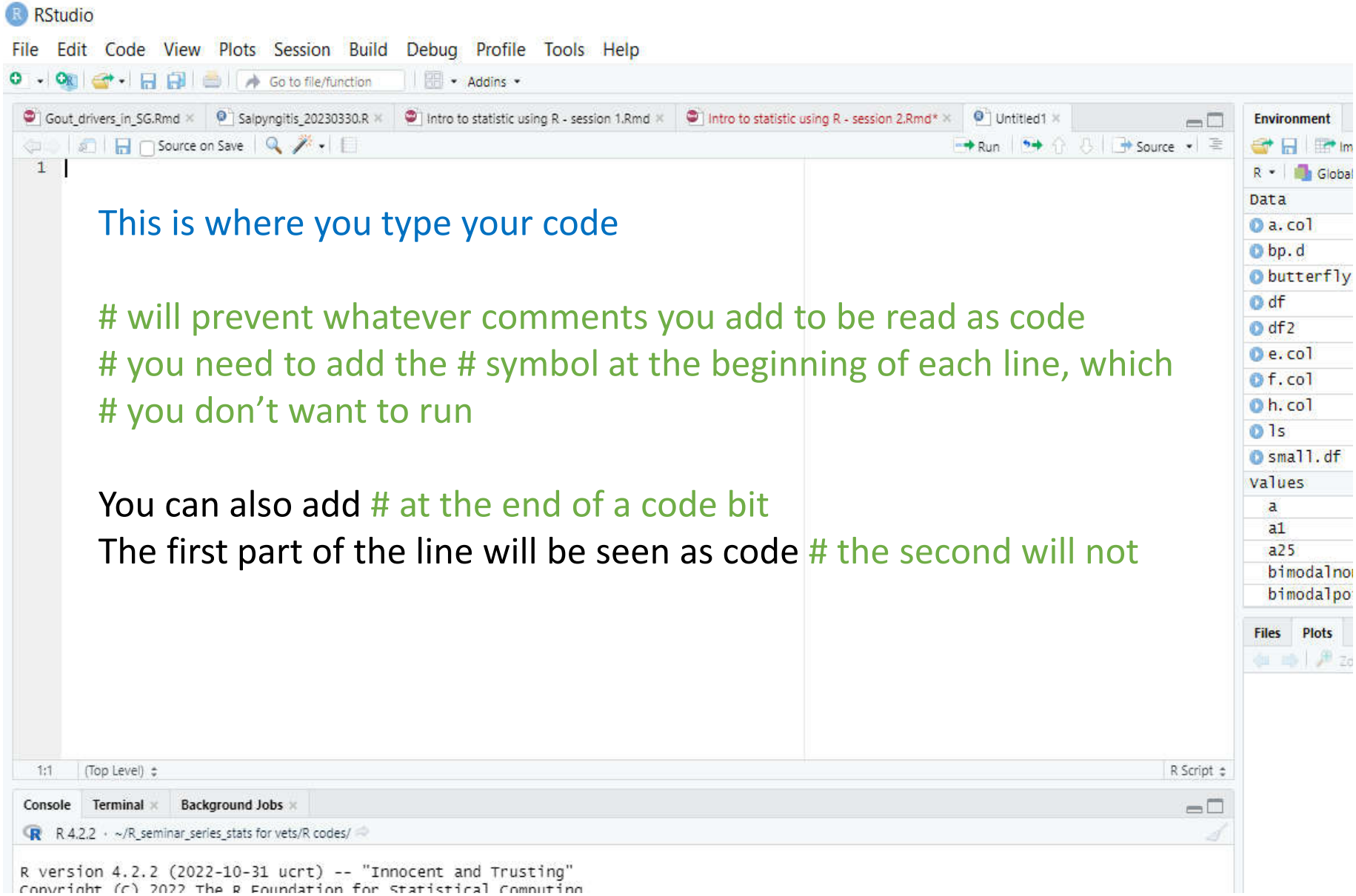




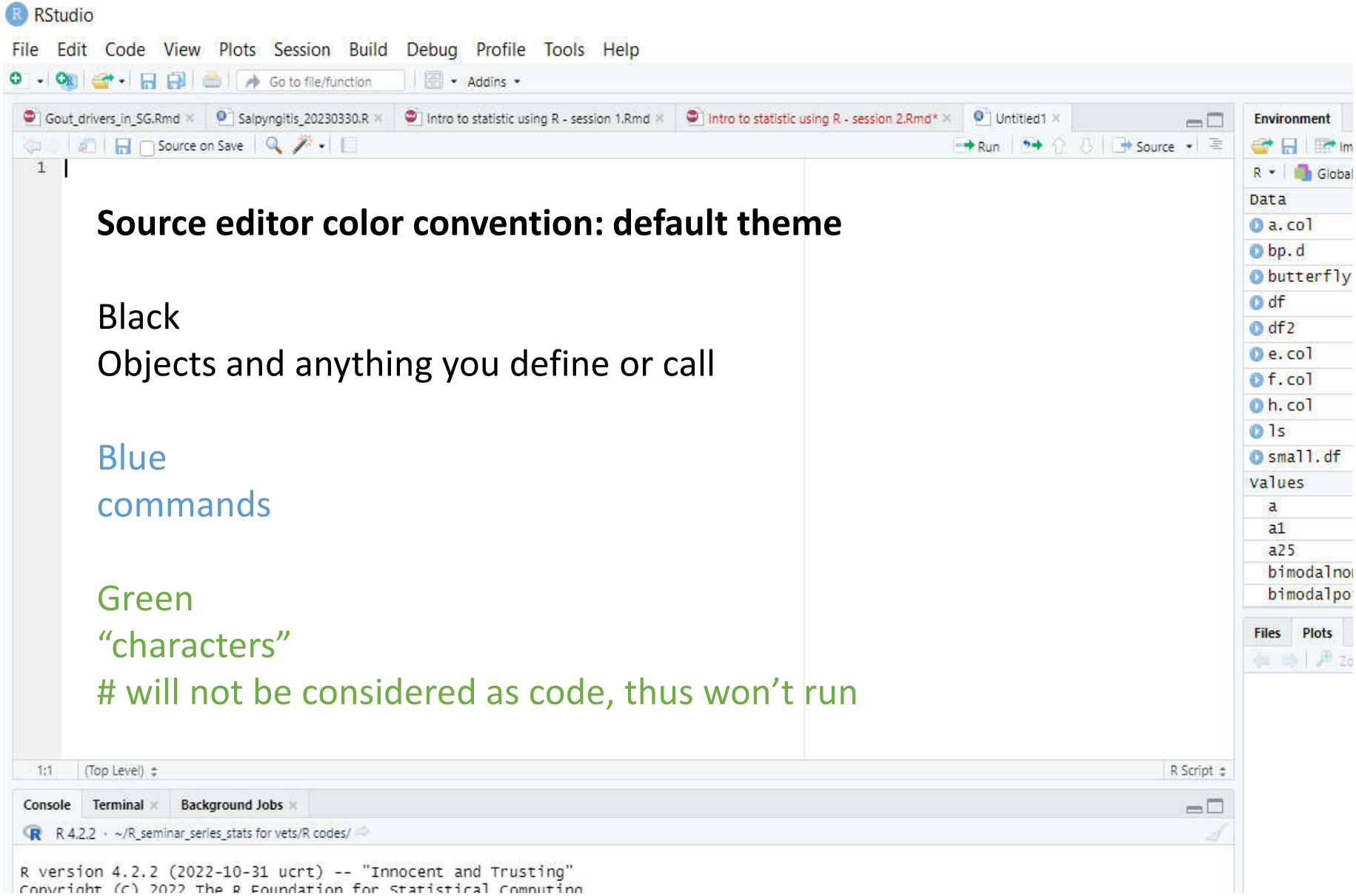
Panes in RStudio



1 Source editor

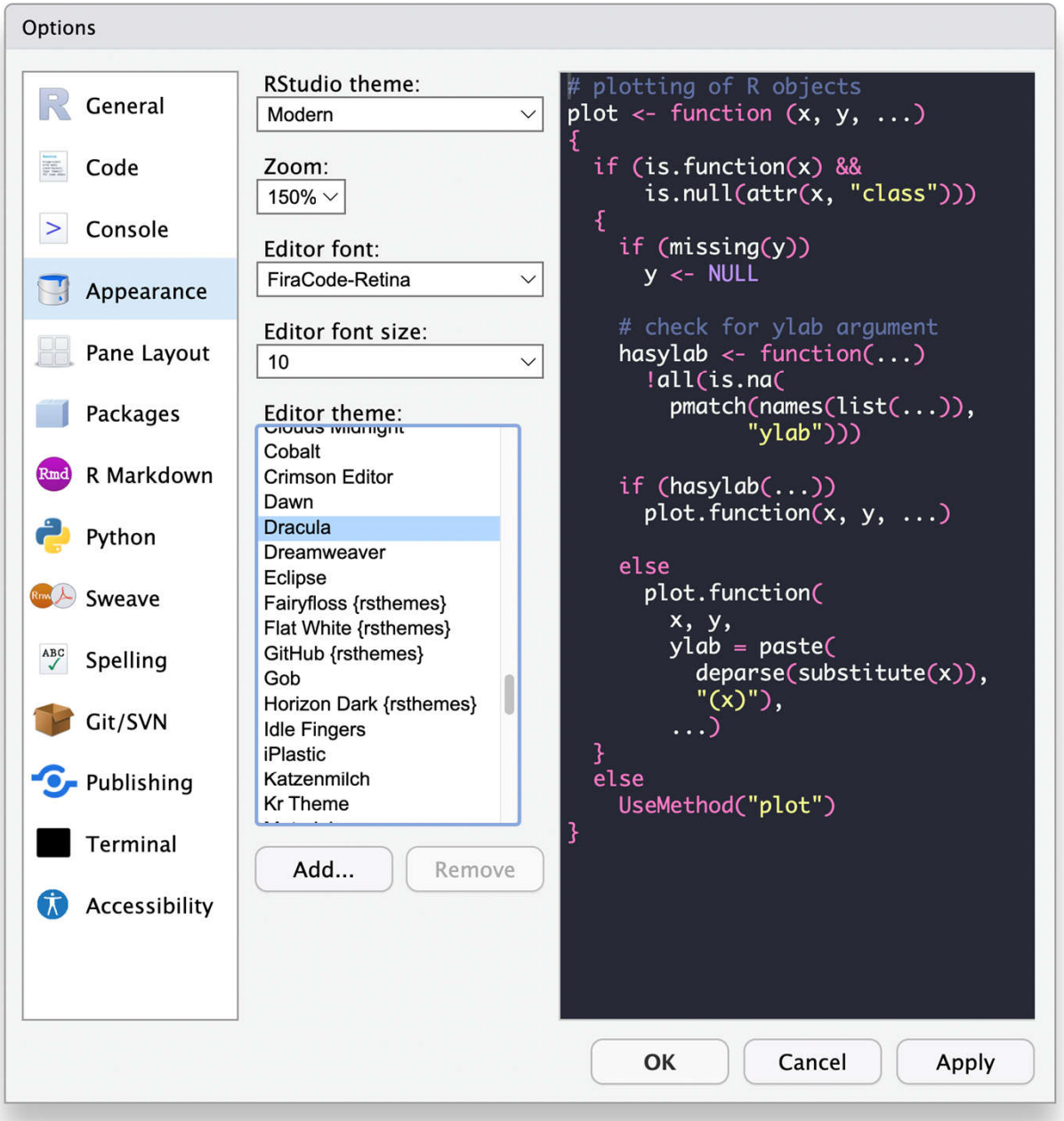


1 Source editor

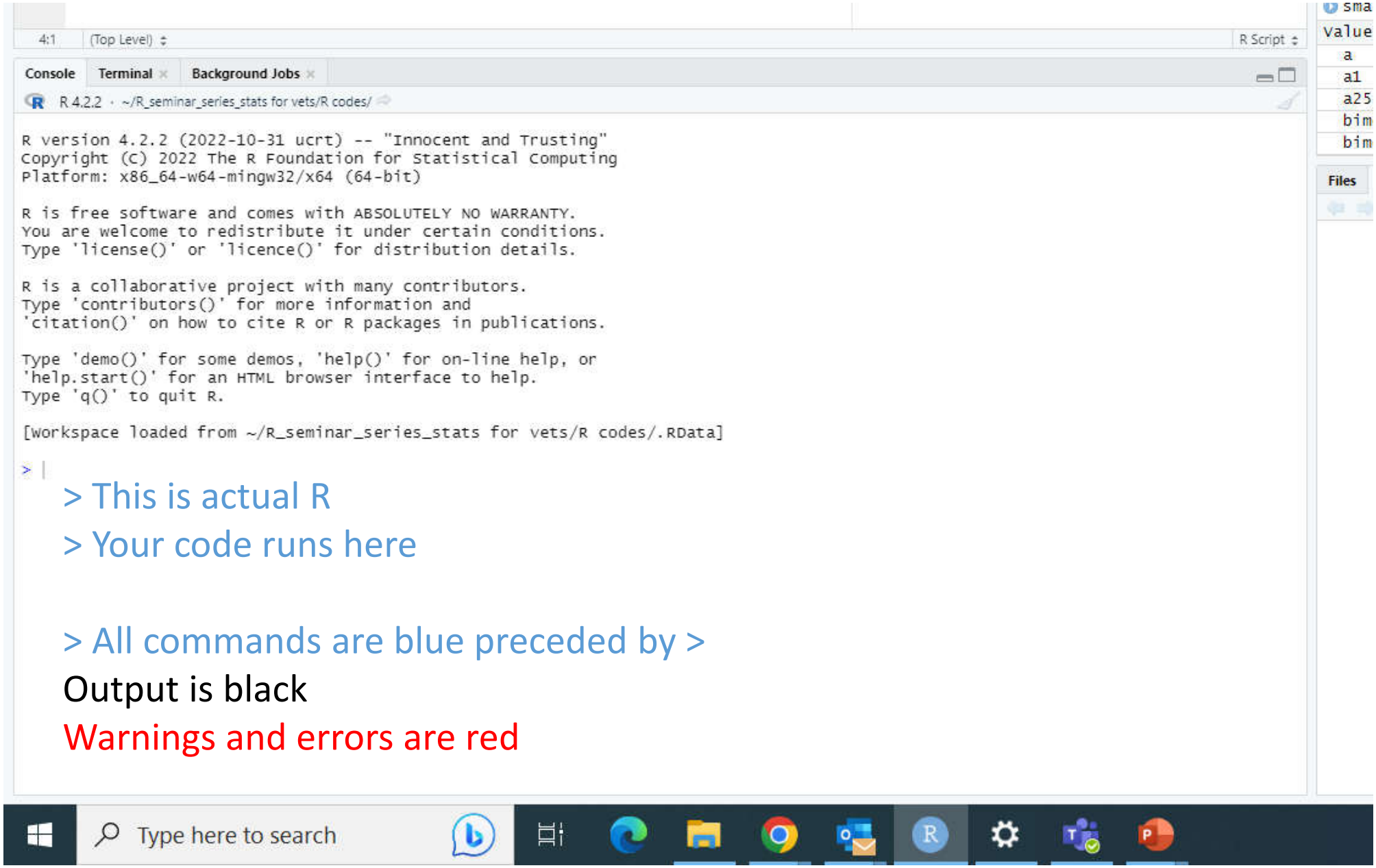


1 Source editor

But you can change the theme!



② R Console



> This is actual R

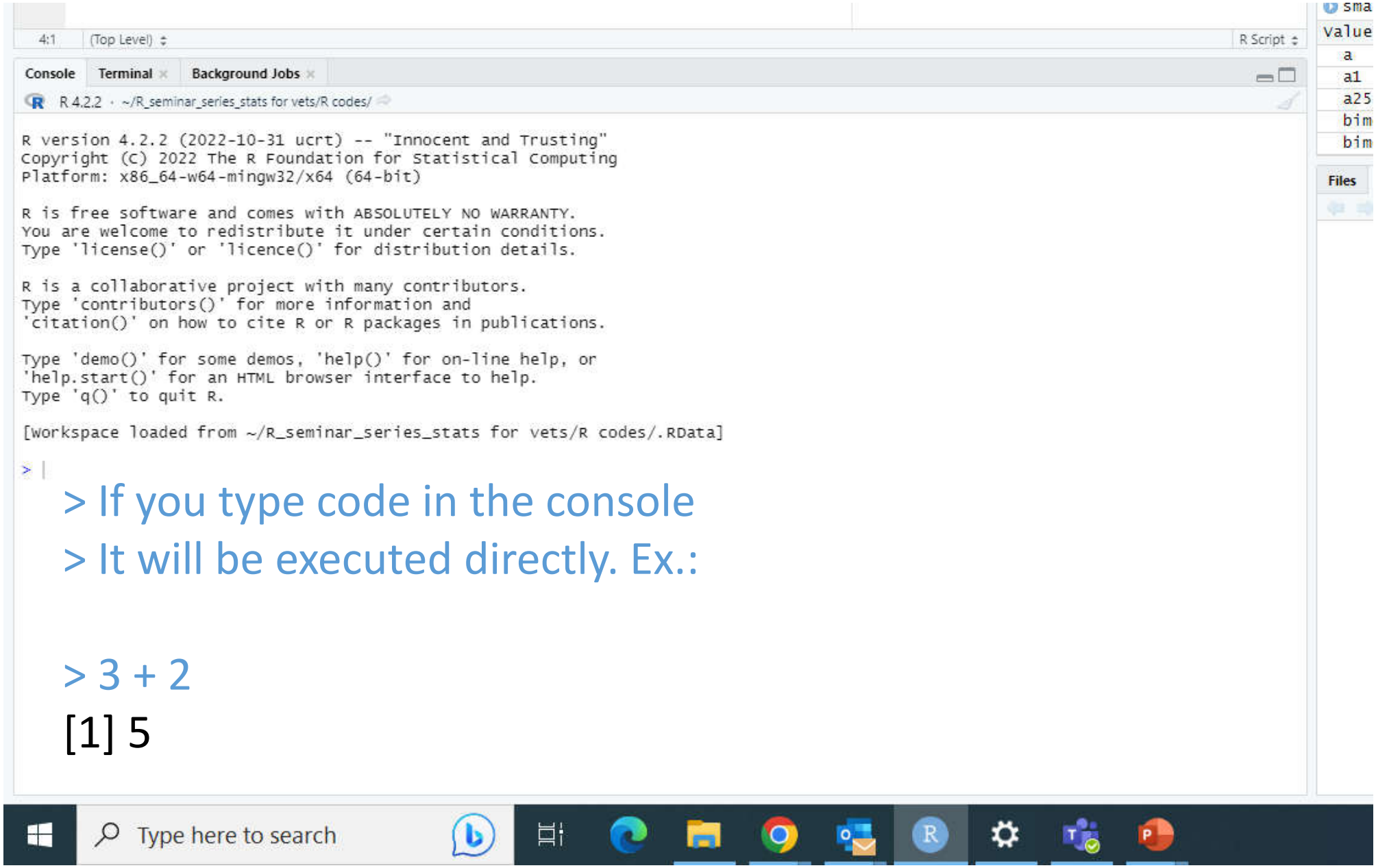
> Your code runs here

> All commands are blue preceded by >

Output is black

Warnings and errors are red

② R Console

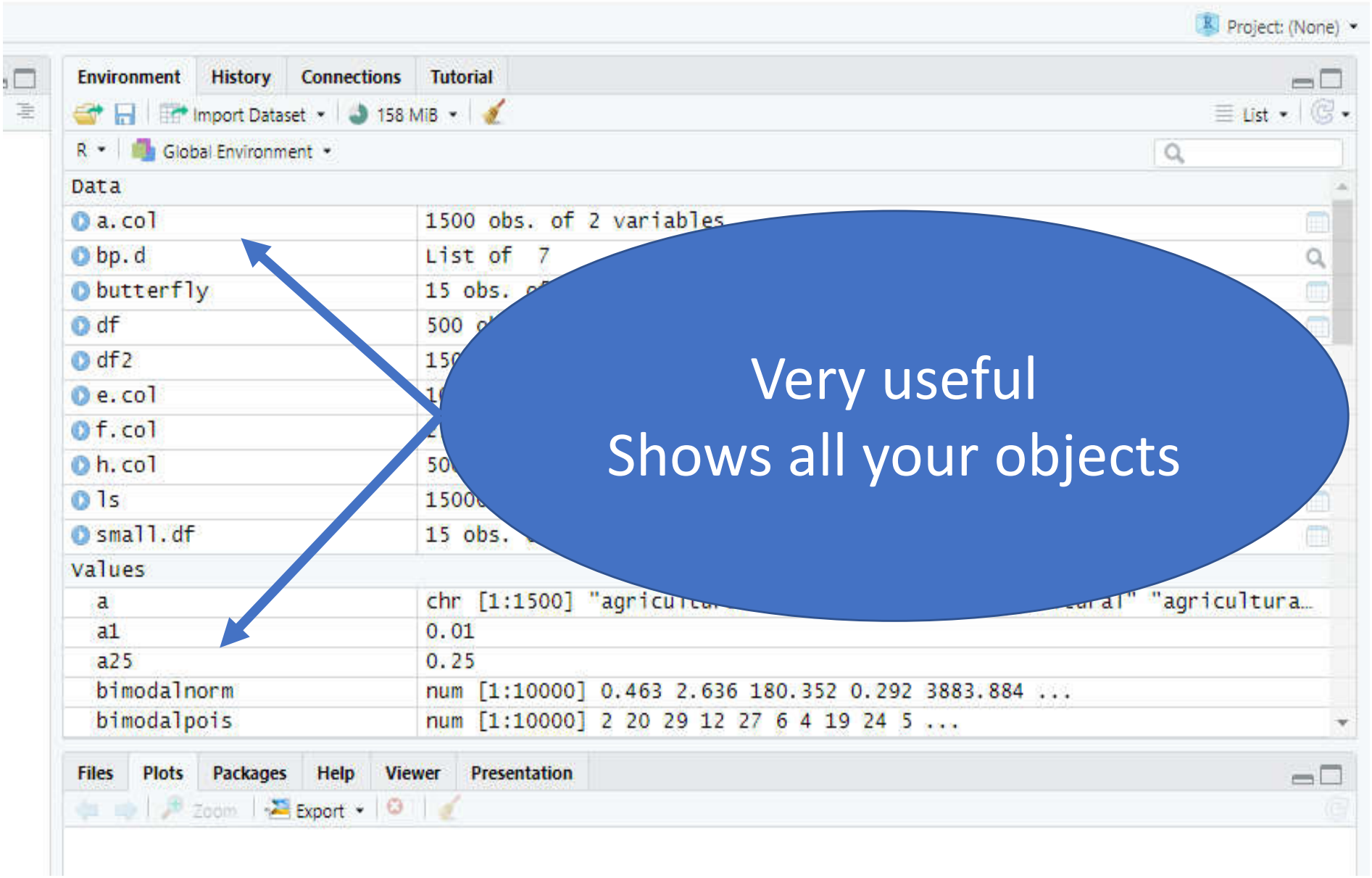


> If you type code in the console

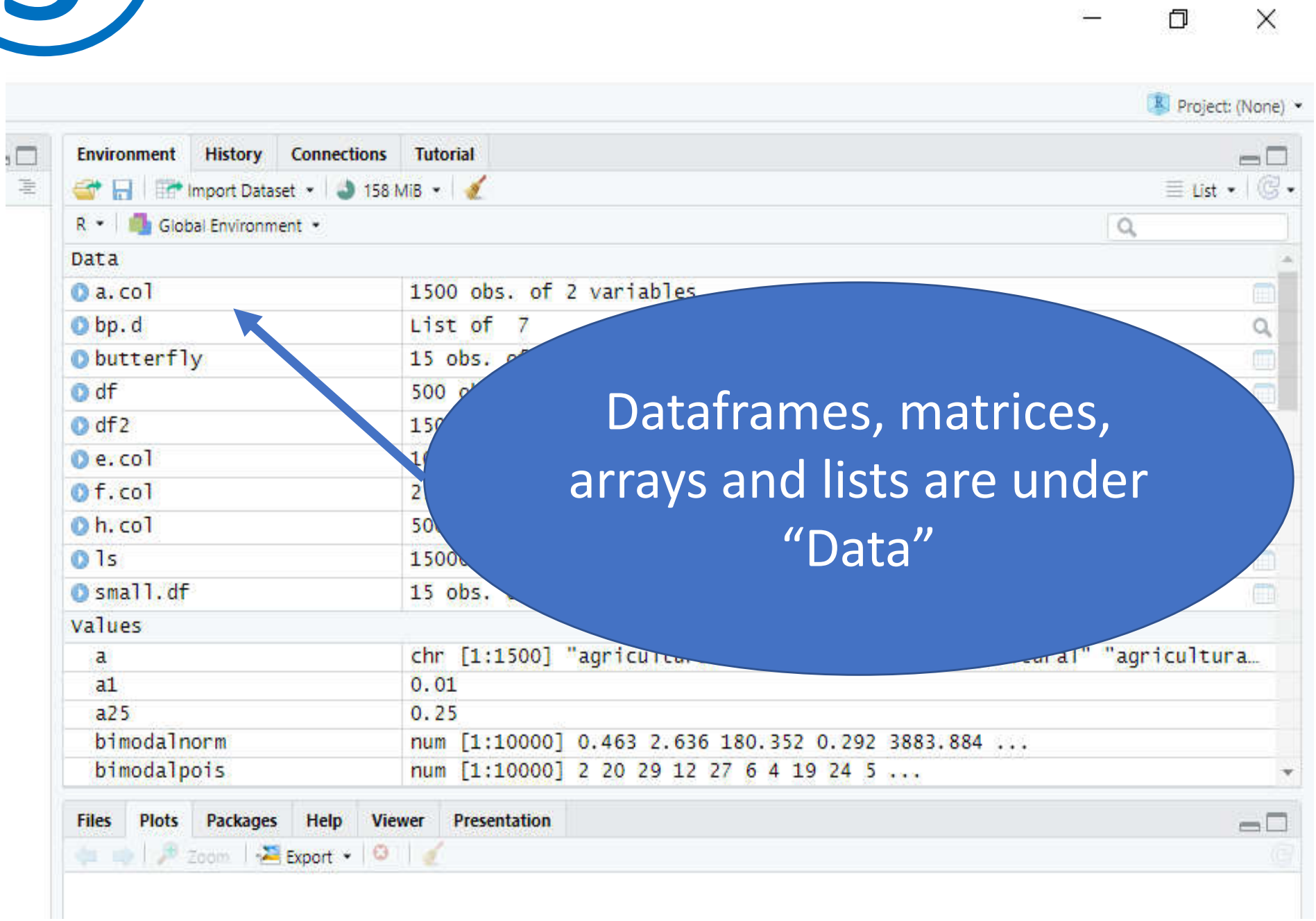
> It will be executed directly. Ex.:

```
> 3 + 2
[1] 5
```

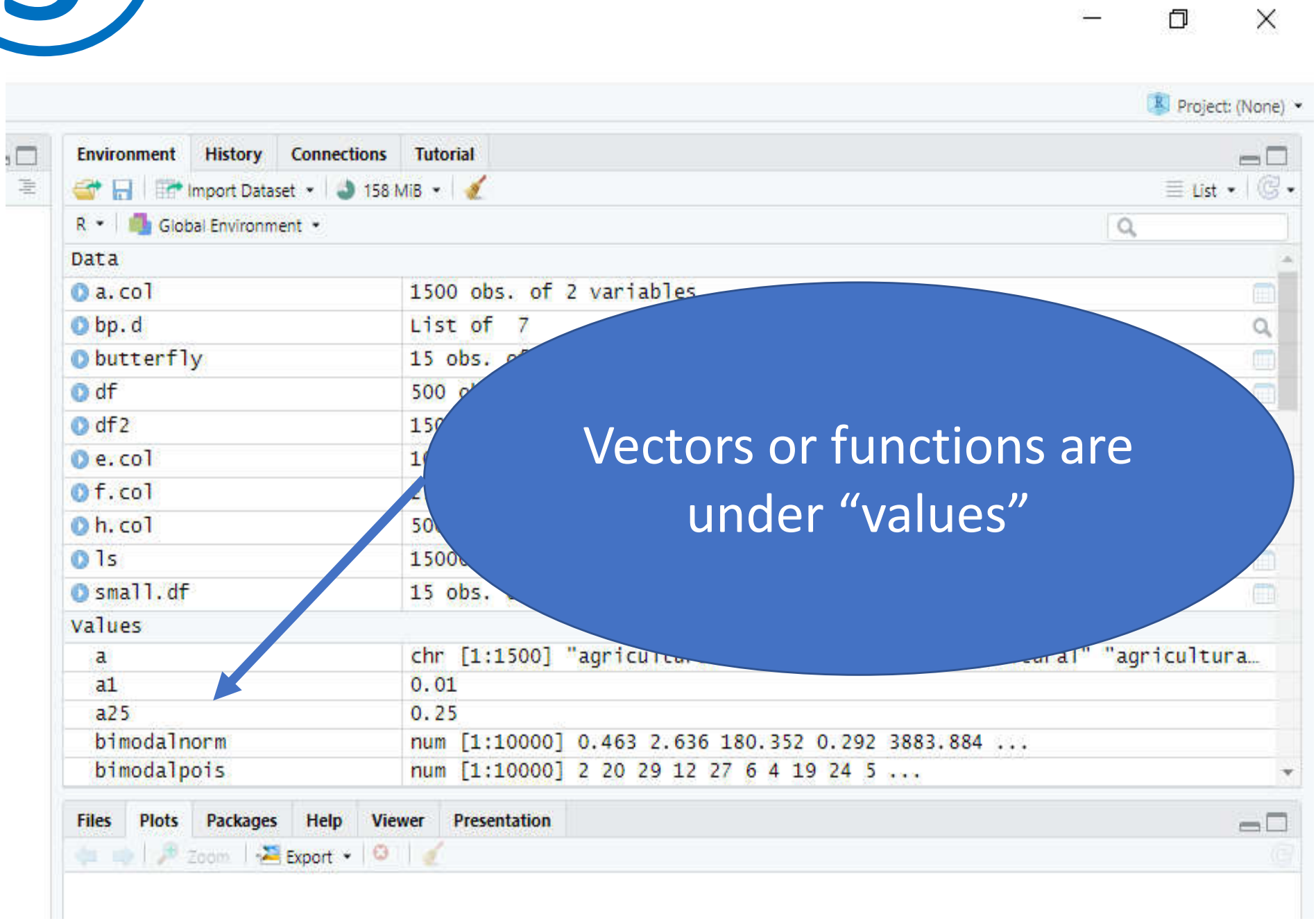
3 Environment pane



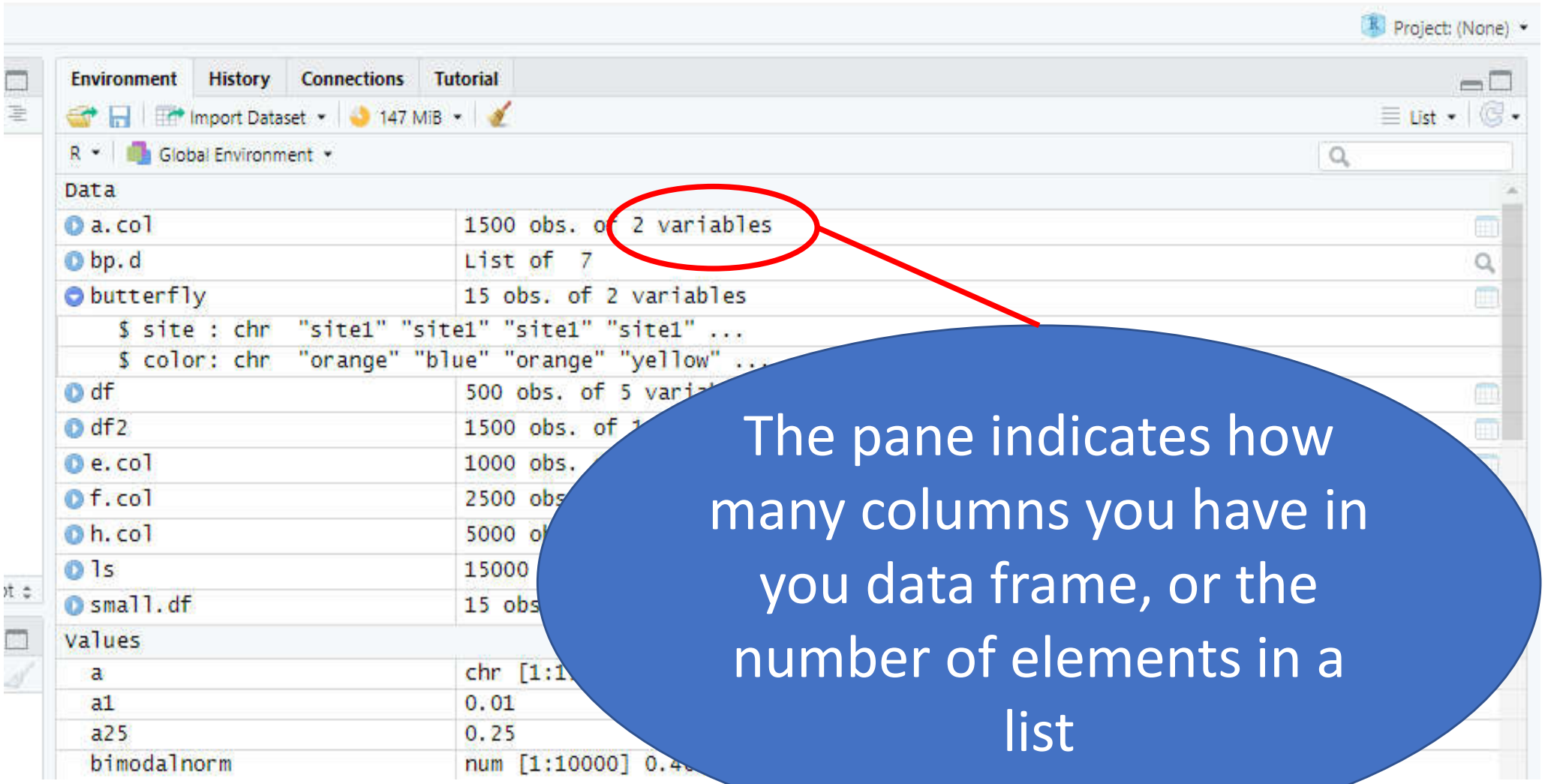
3 Environment pane



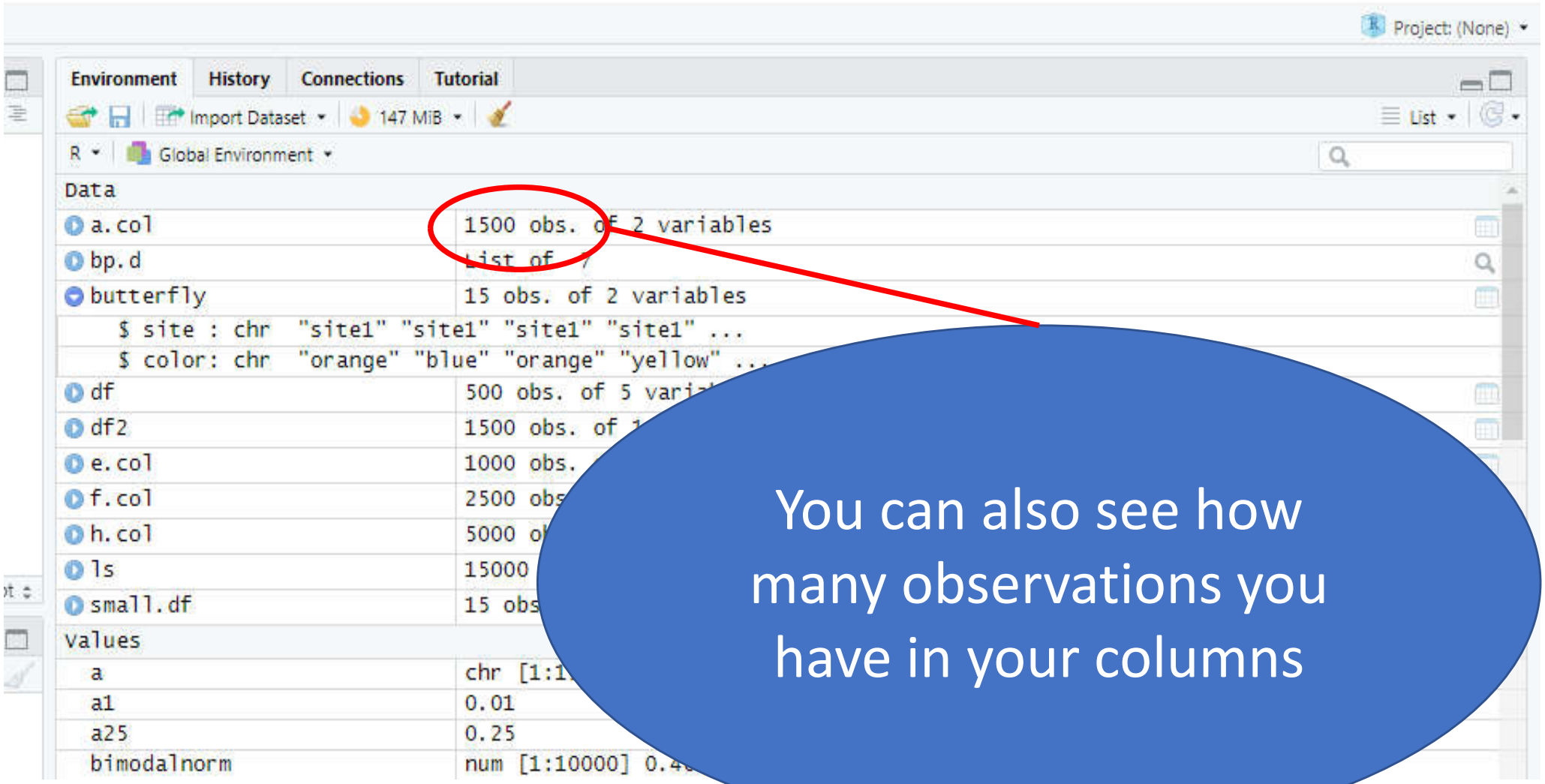
3 Environment pane



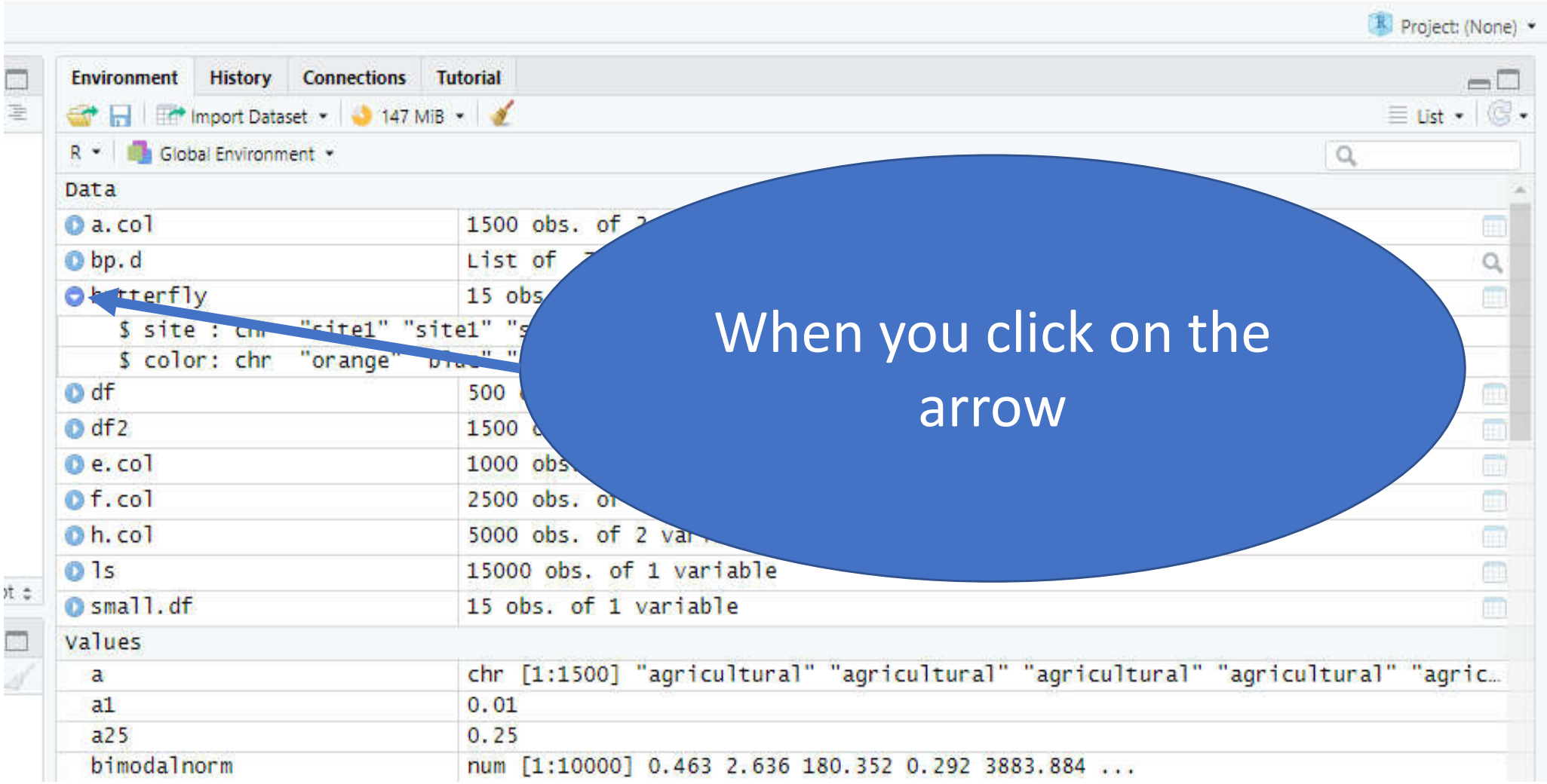
3 Environment pane



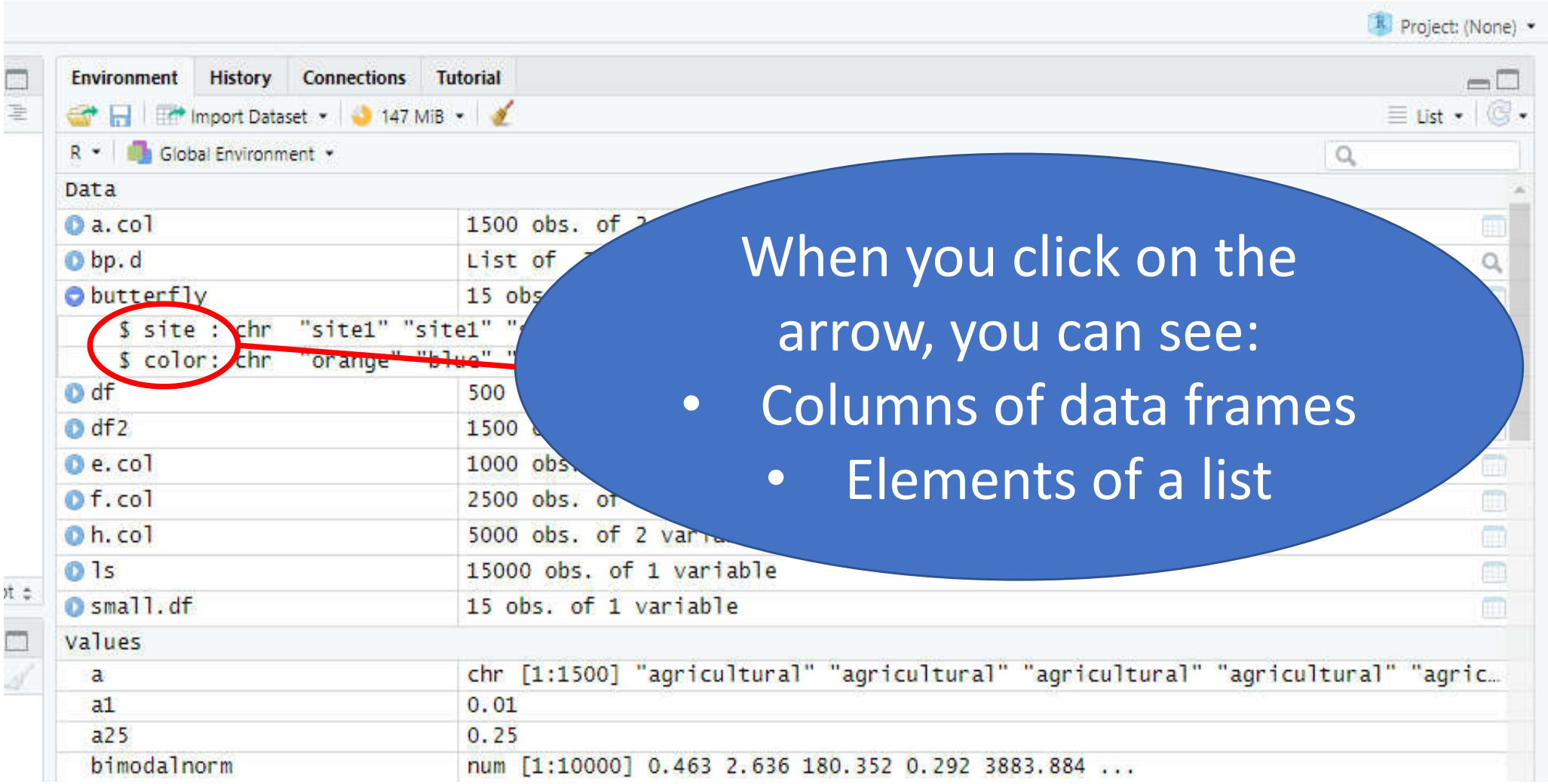
3 Environment pane



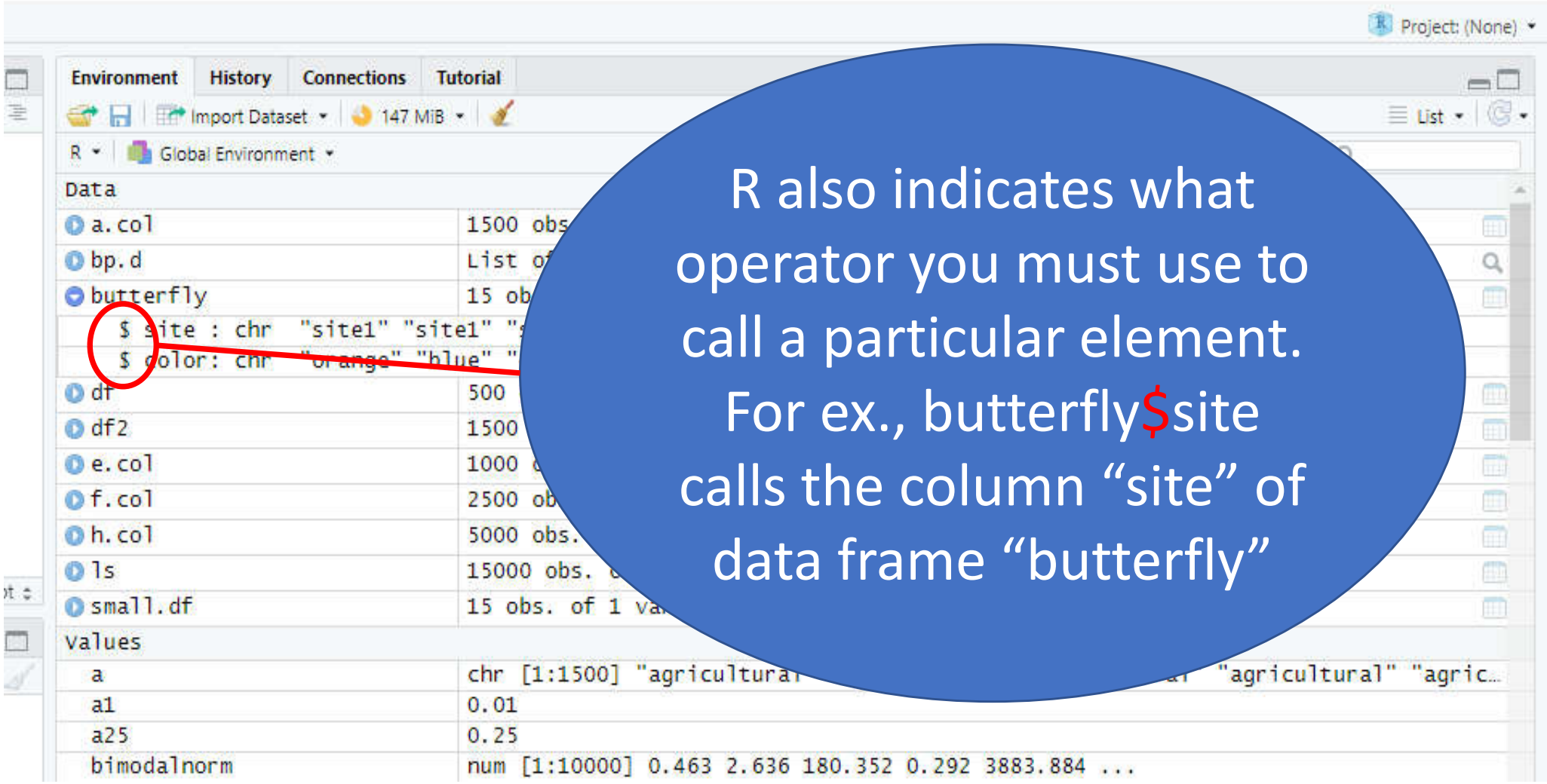
3 Environment pane



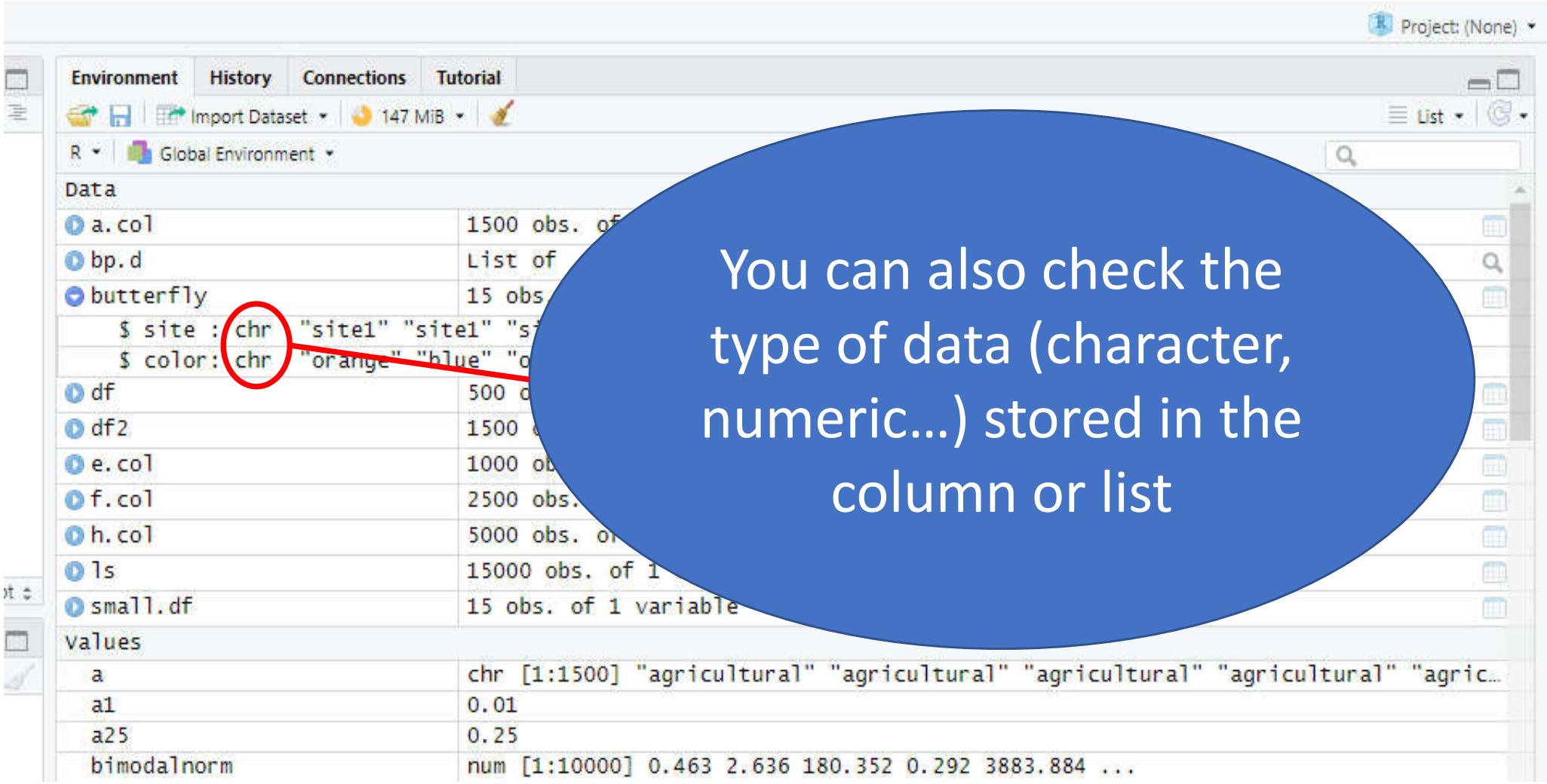
3 Environment pane



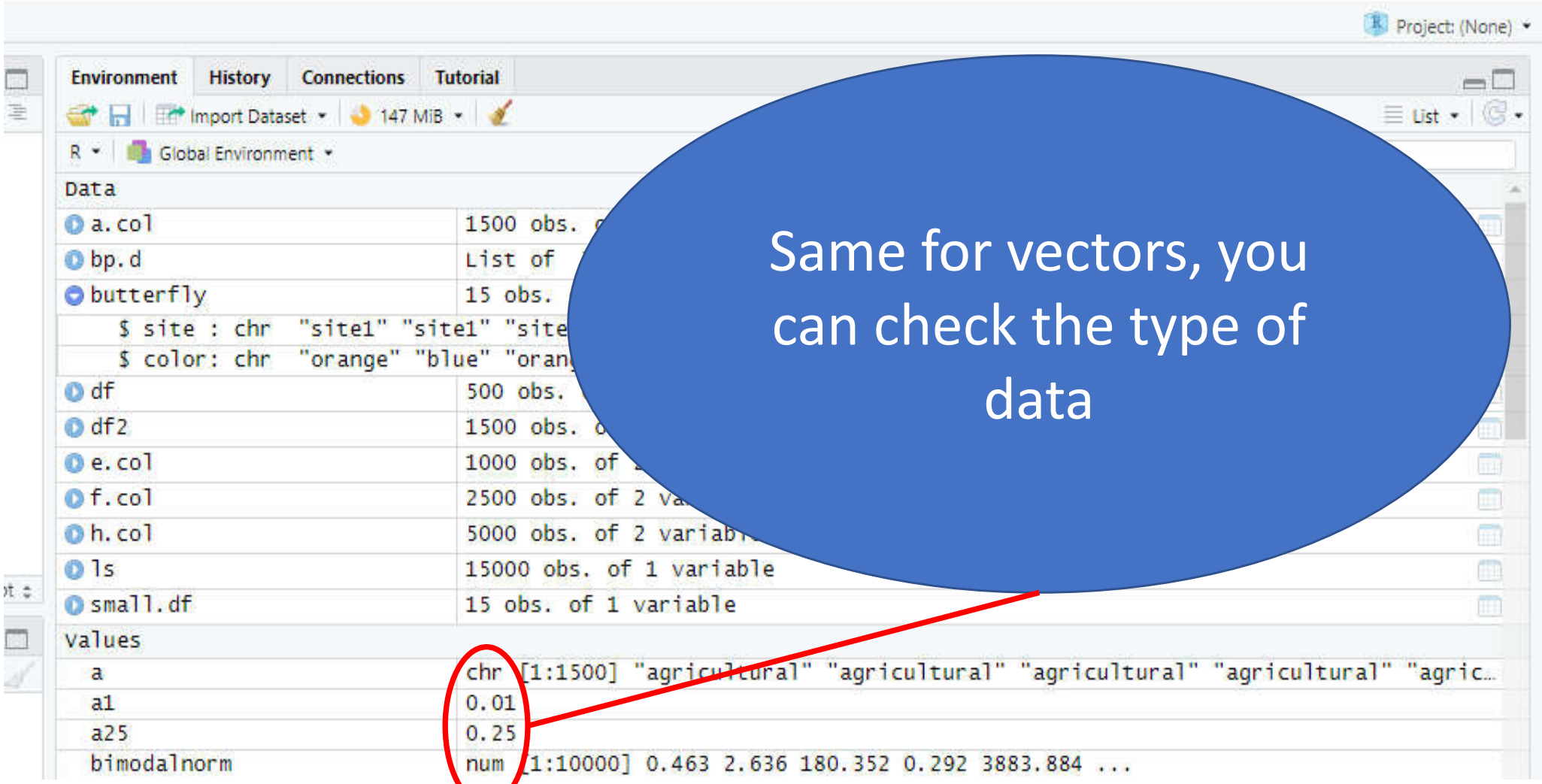
3 Environment pane



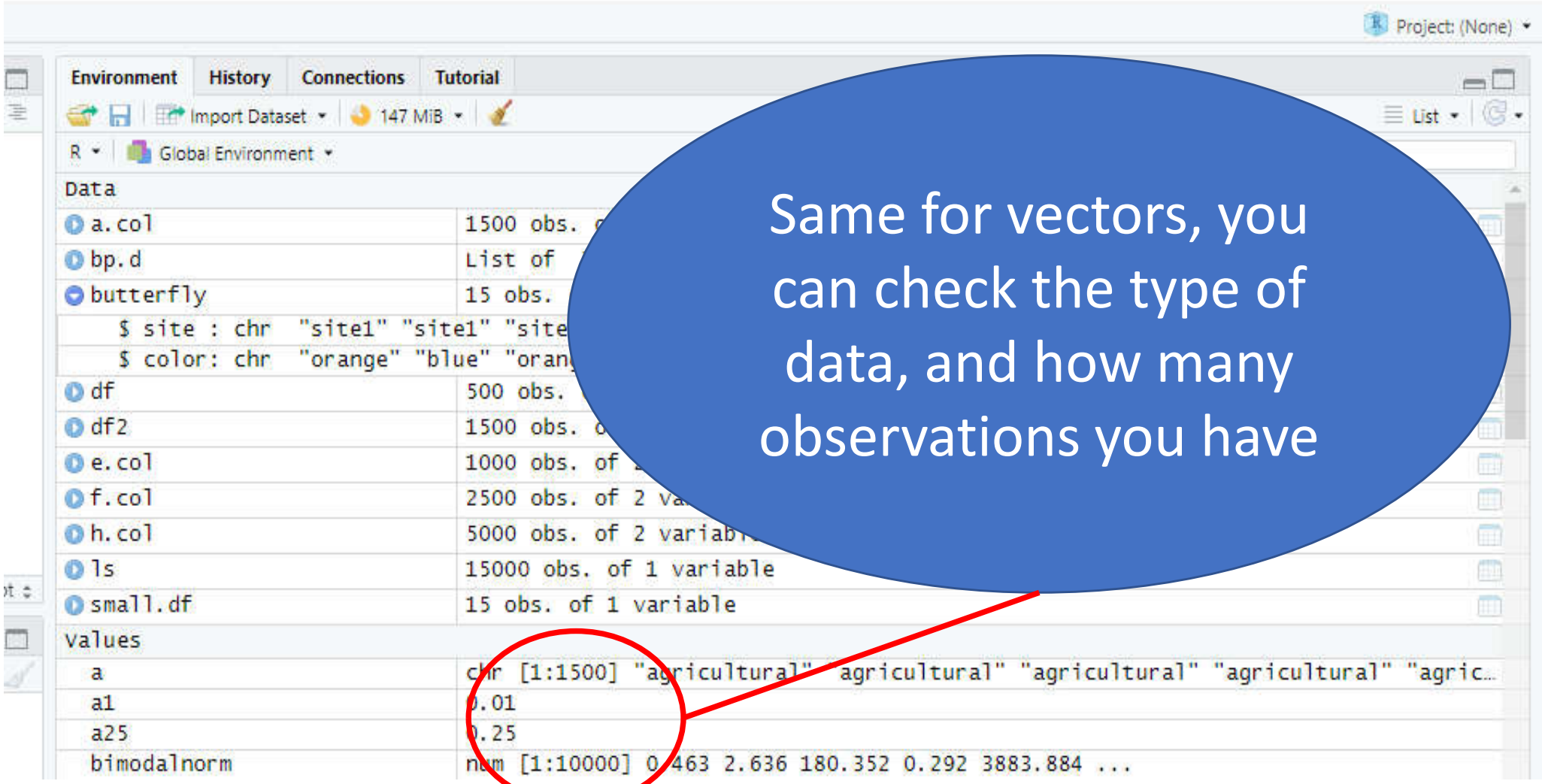
3 Environment pane



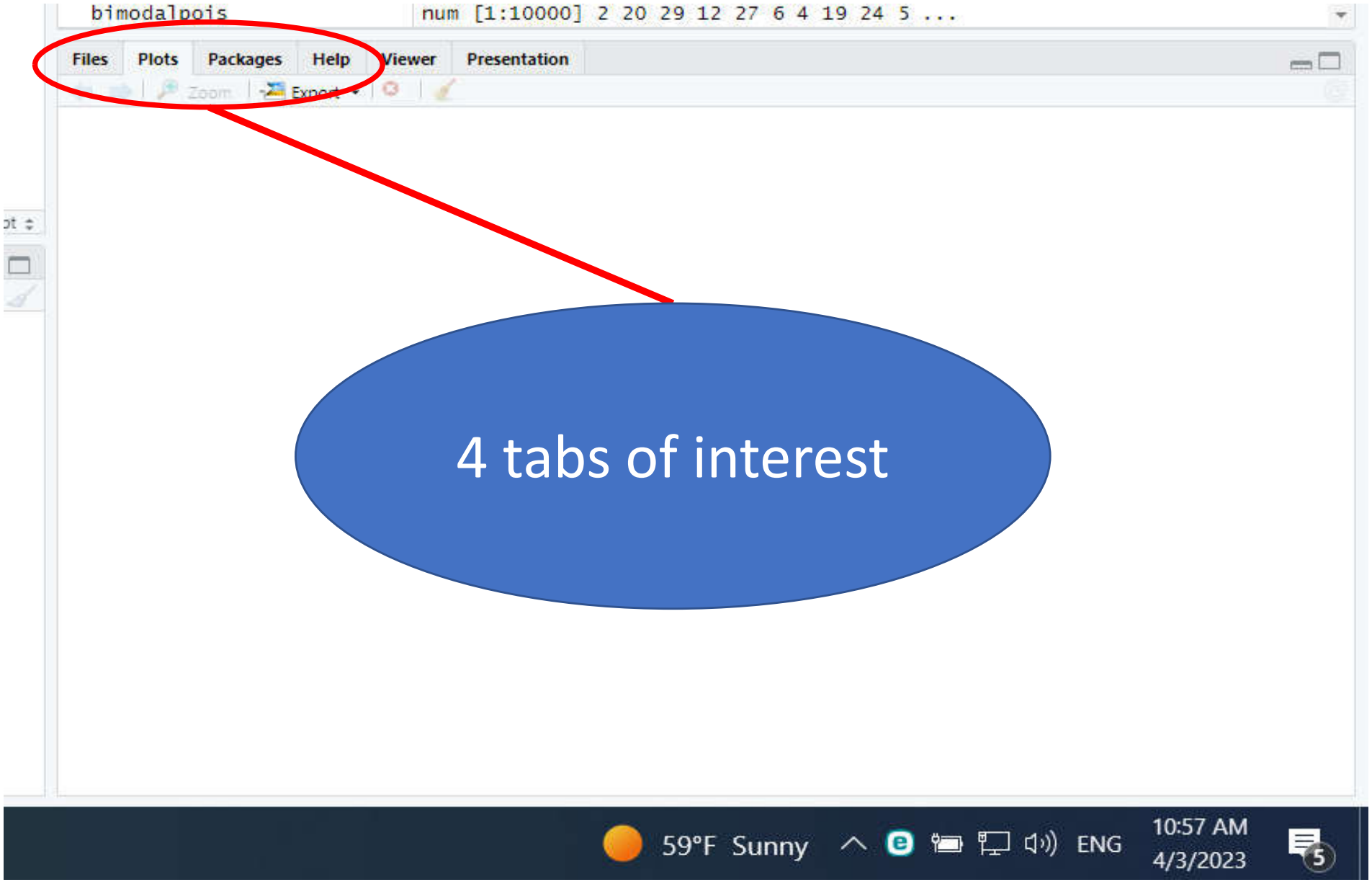
3 Environment pane



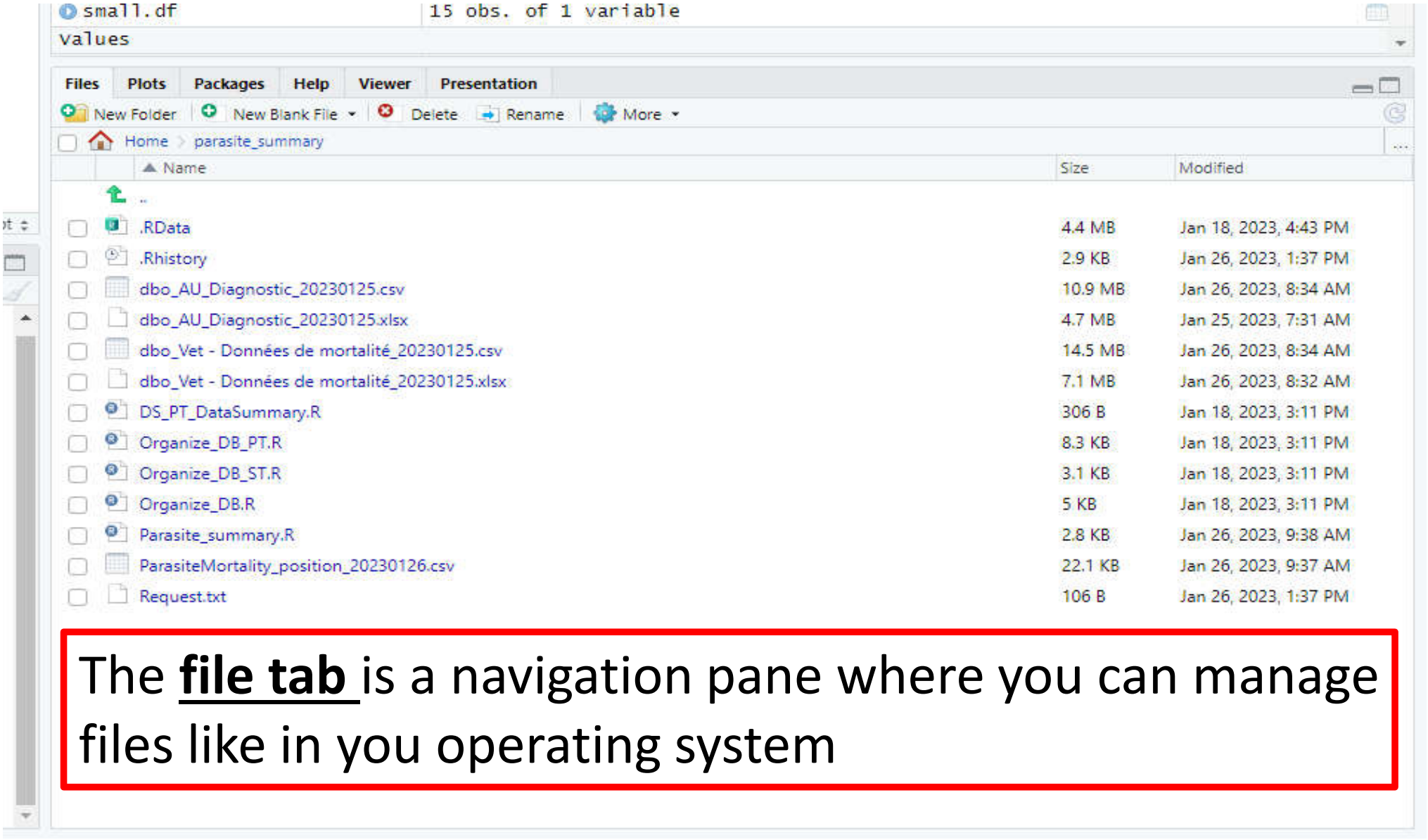
3 Environment pane



4 Output pane

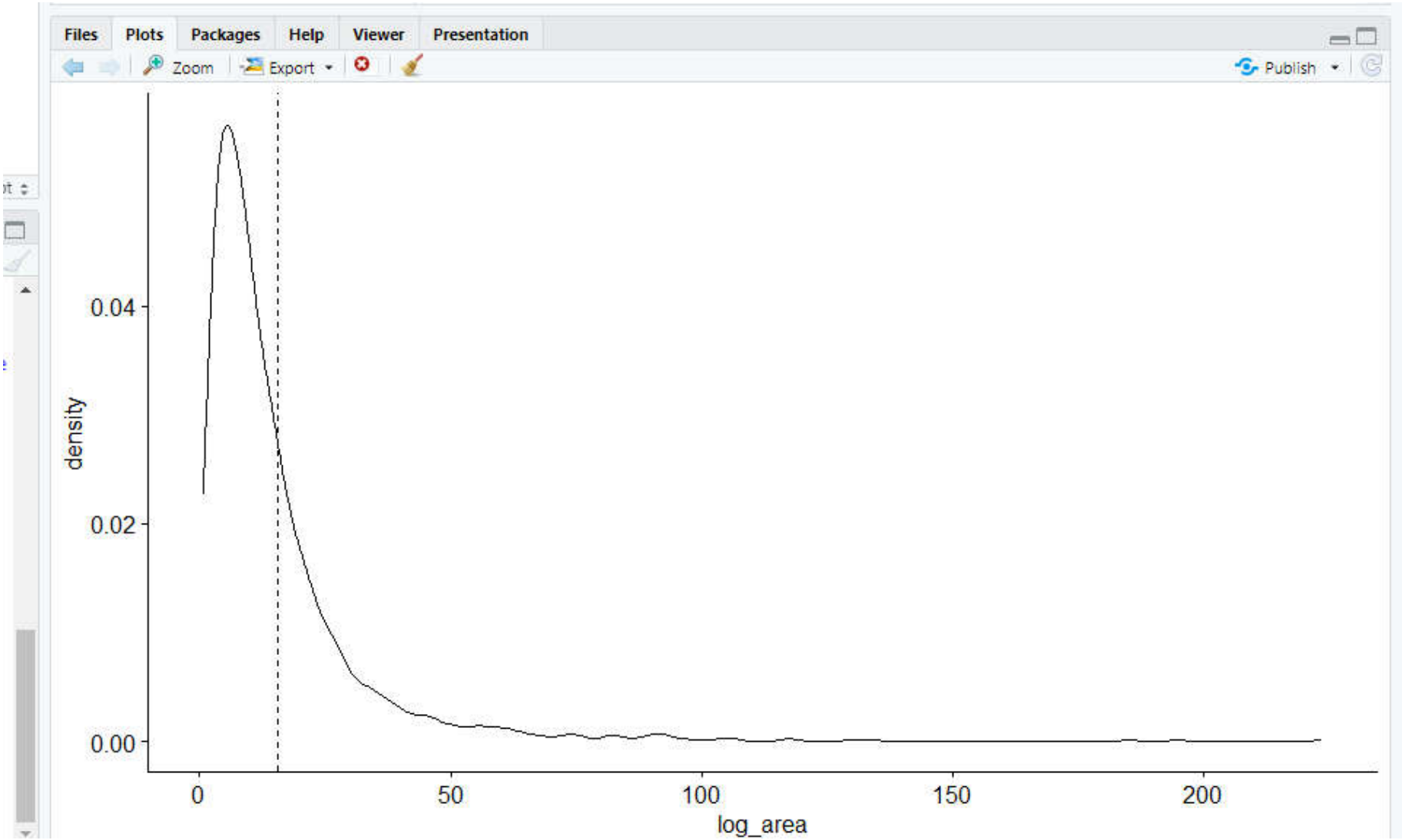


④ Output pane



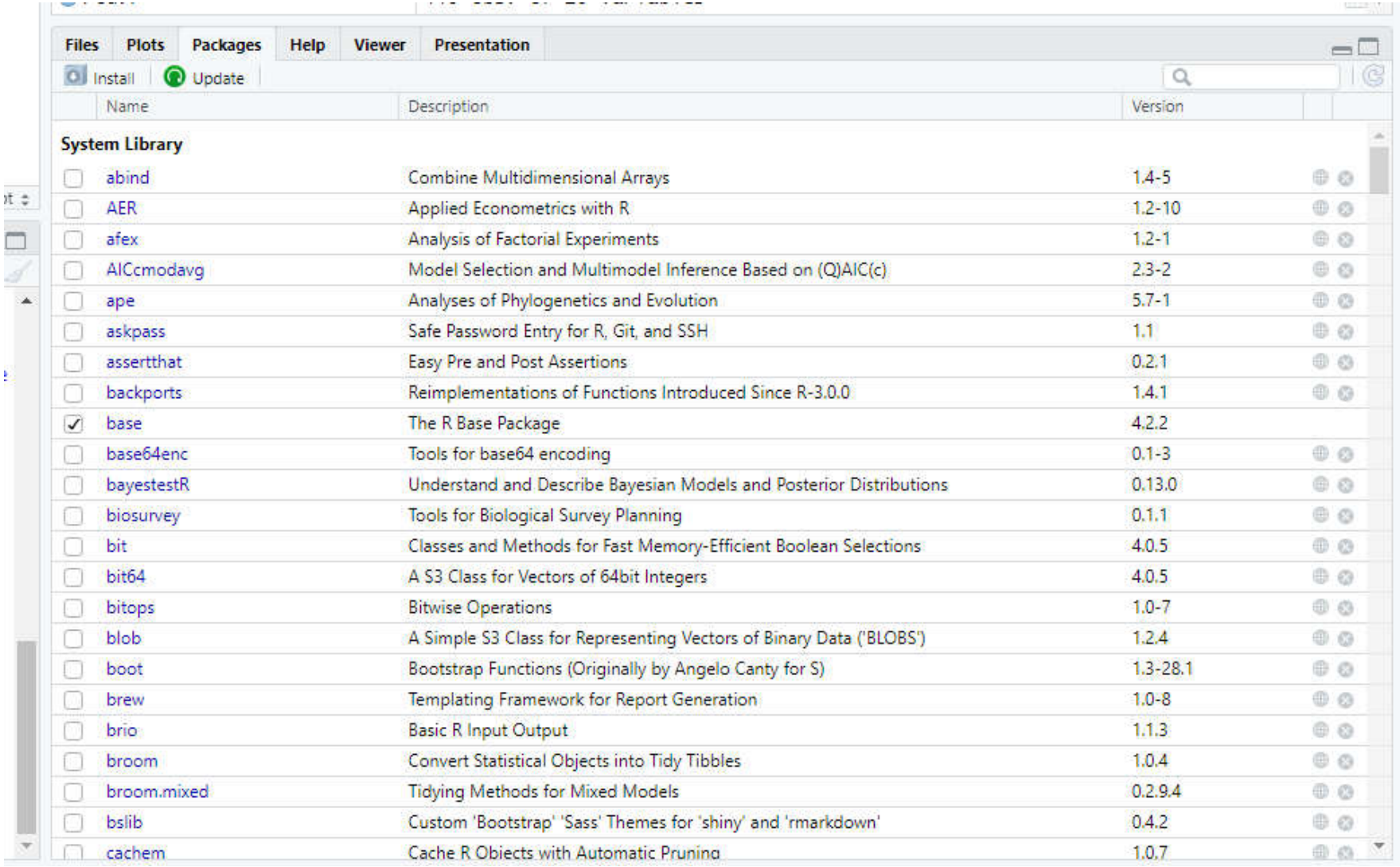
The **file tab** is a navigation pane where you can manage files like in you operating system

4 Output pane



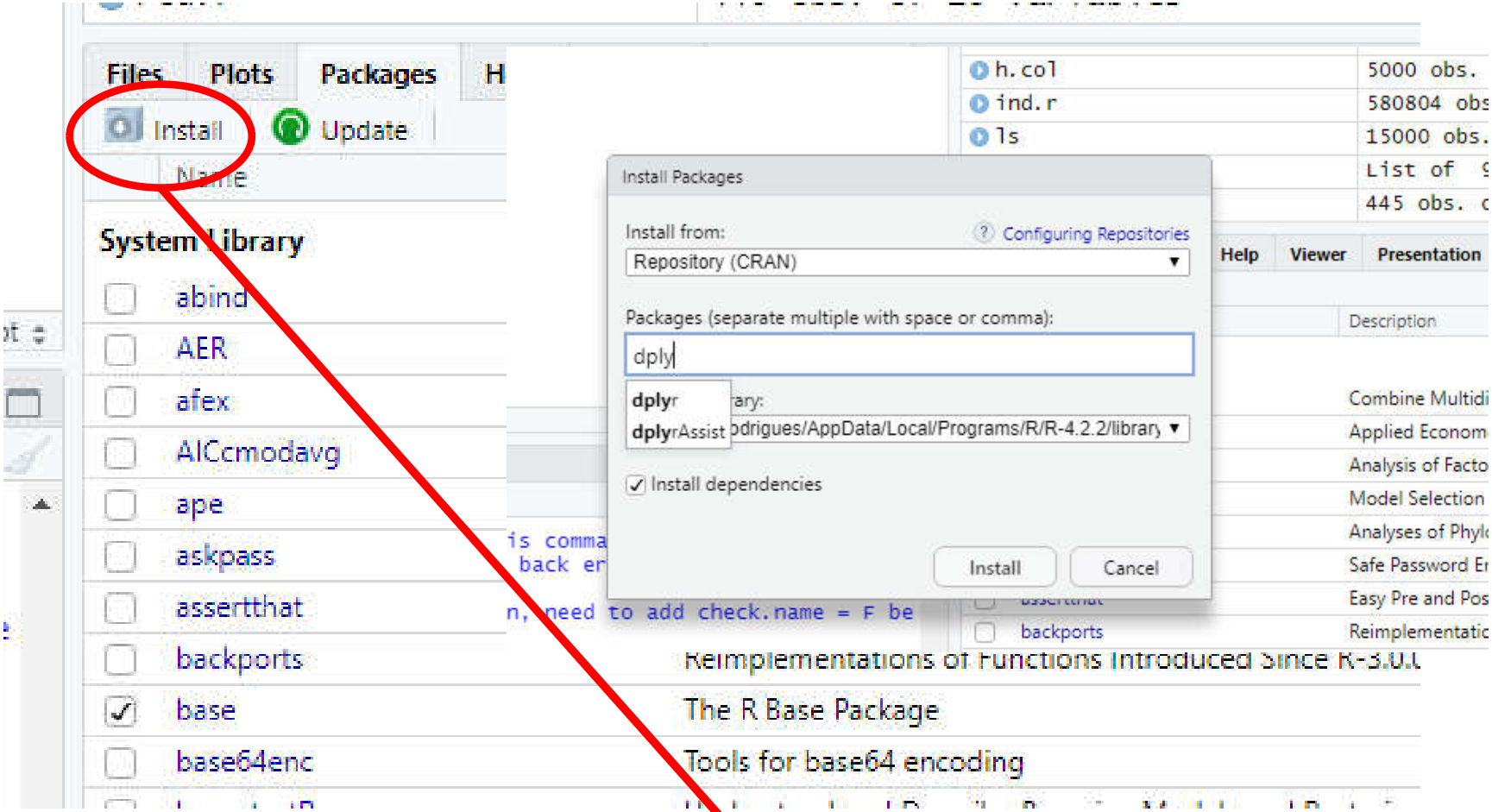
The **plot tab** is where you can see you plots

4 Output pane



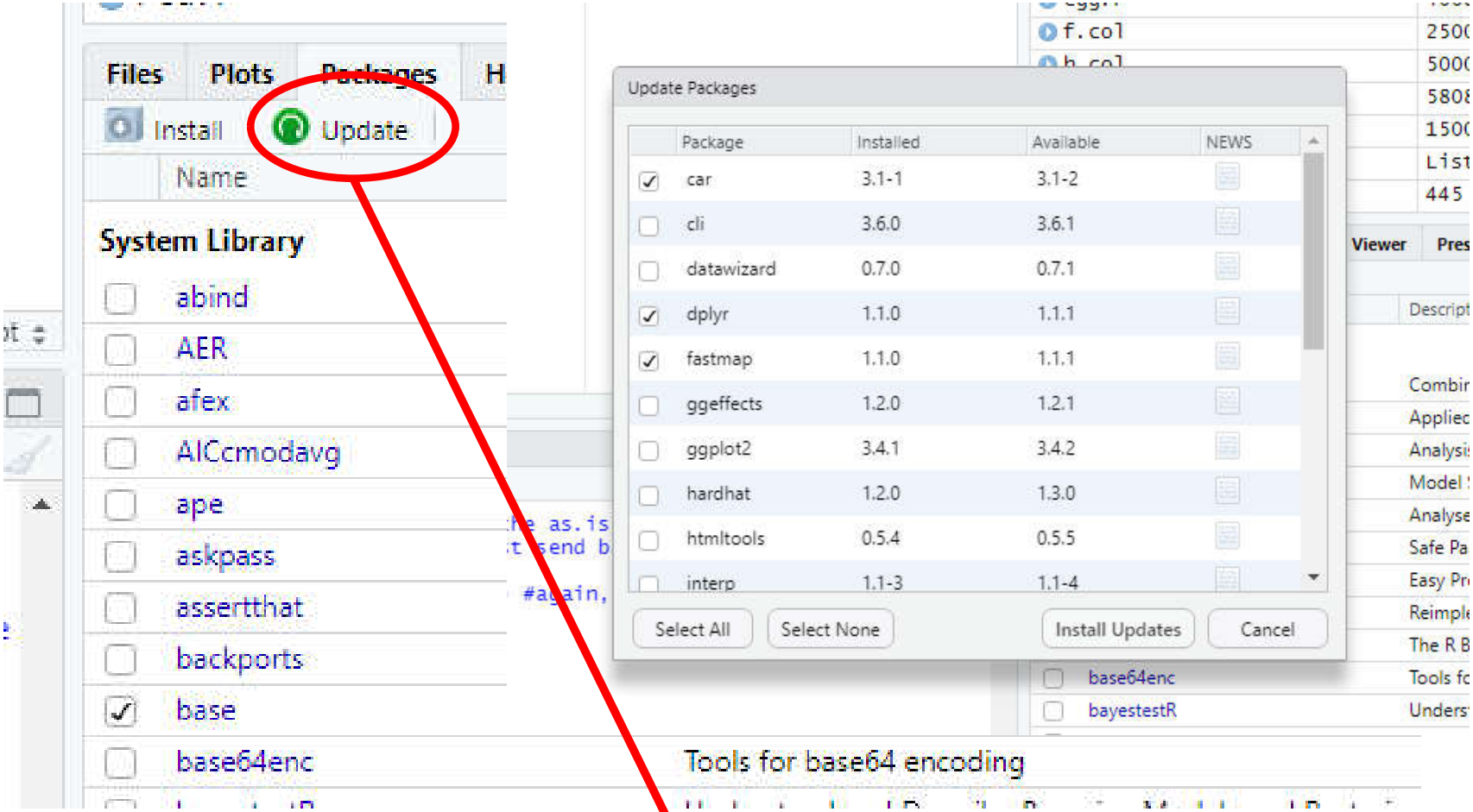
The **package tab** shows the packages stored in your computer

4 Output pane



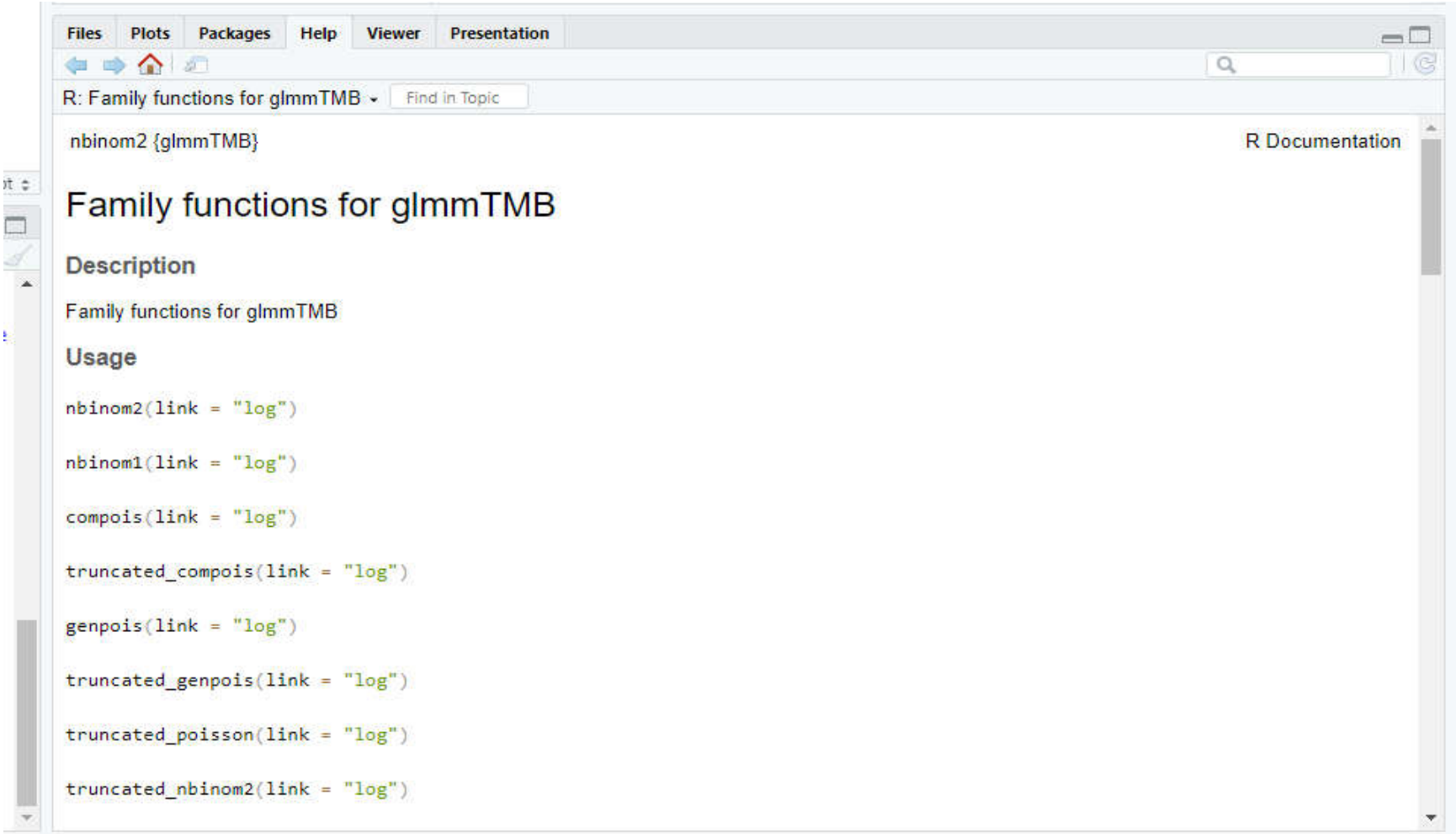
The **package tab** also has the install button from which you can directly install packages

④ Output pane



Same to update already installed packages

4 Output pane



The **Help tab** provides access to R documentation for all functions


Intermission