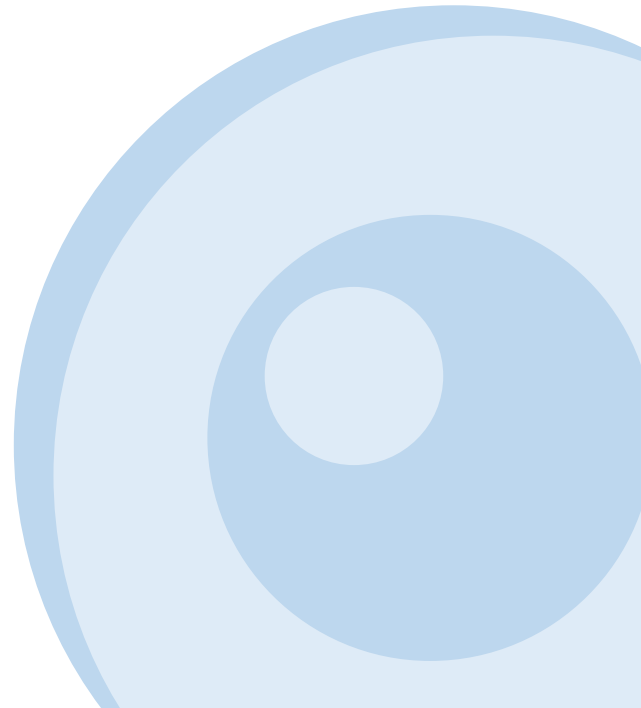


PSYC 193R:

Data Analysis & Visualization with R

Lecture 2: Basic Probability with R

Jonas S.-H. Lau [silau -et- ucsd.edu]
psyc193r.ucsd.edu



Agenda

- Basic R Operations
- Built-in functions and customized functions
- The Normal Distribution:
 - Converting percentile to quantile
 - Converting quantile to percentile



Two Components

Programming Language



“The Engine”

Programming Environment



“The Steering Wheel”

Why R?

- Open source: Free with strong community support
- Standard in Psychology, and in Industry
- Automation: minimize errors
- Flexibility: e.g., functions, machine learning
- Graphics & data visualization capabilities



Untitled1 x

Source

1

1:1

(Top Level)

R Script

Environment

History

Import Dataset

List

Global Environment

Environment is empty

Console ~/

Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in
publications.

Type 'demo()' for some demos, 'help()' for on-
line help, or
'help.start()' for an HTML browser interface
to help.
Type 'q()' to quit R.

> |

Files

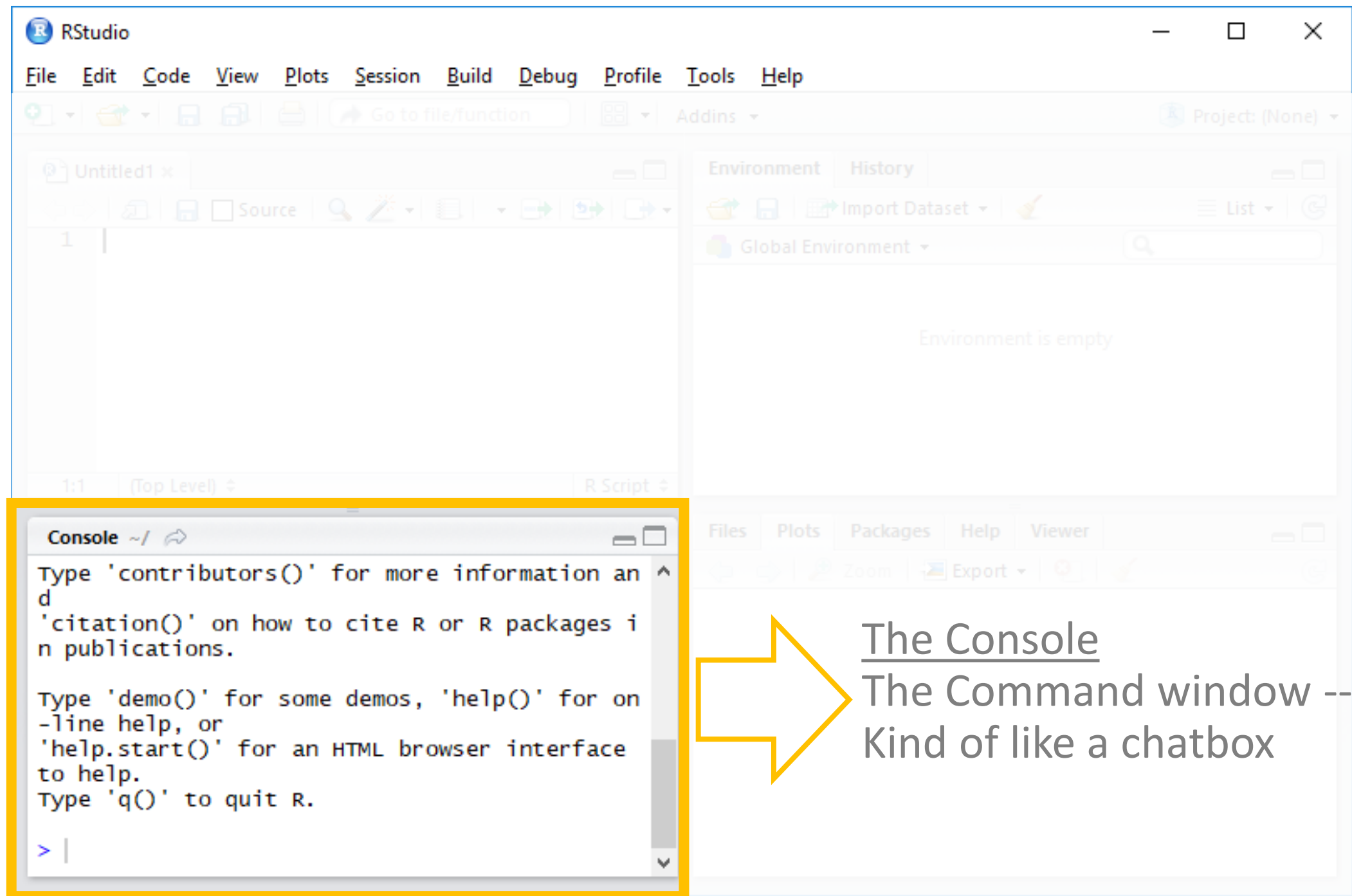
Plots

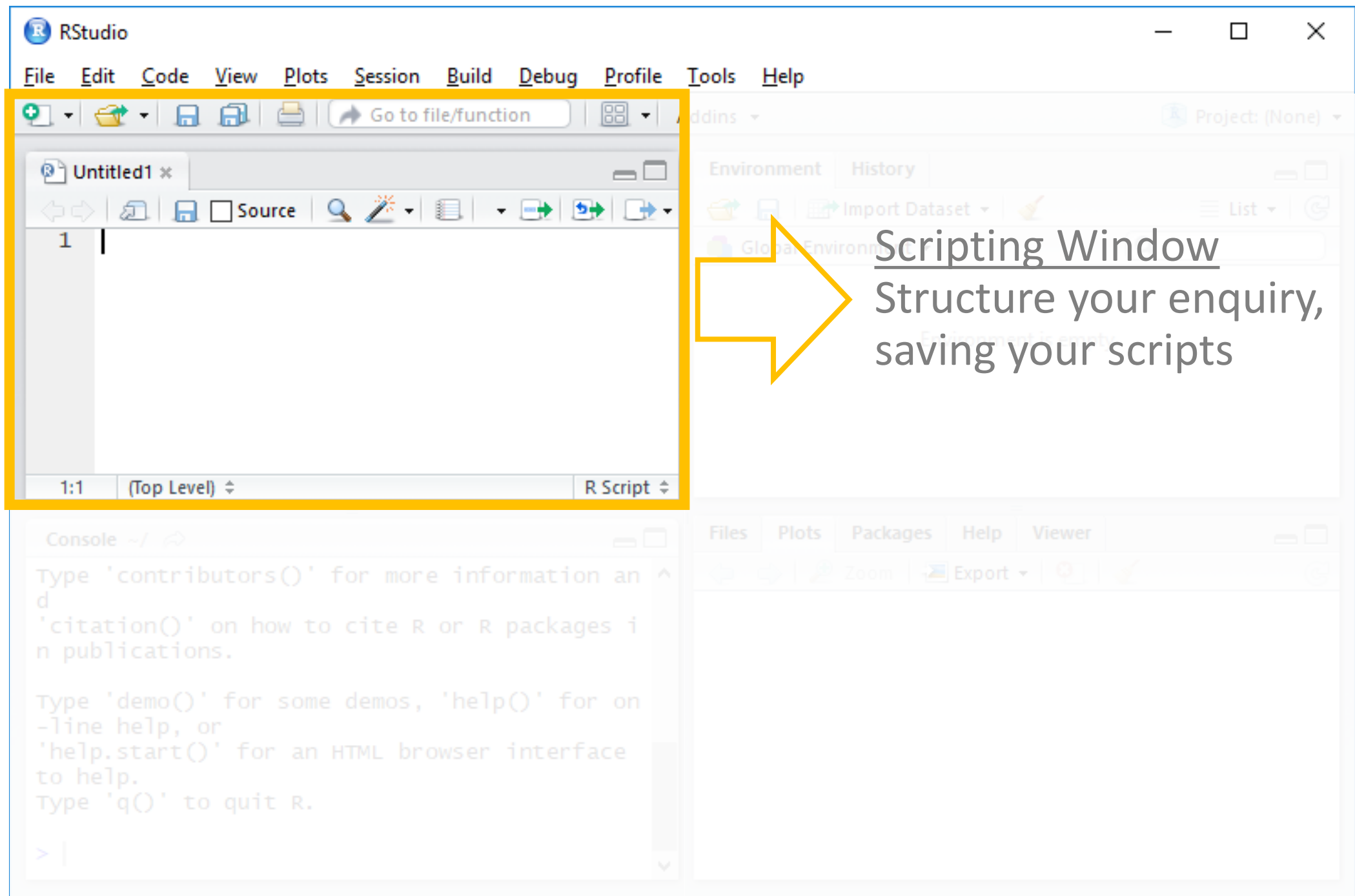
Packages

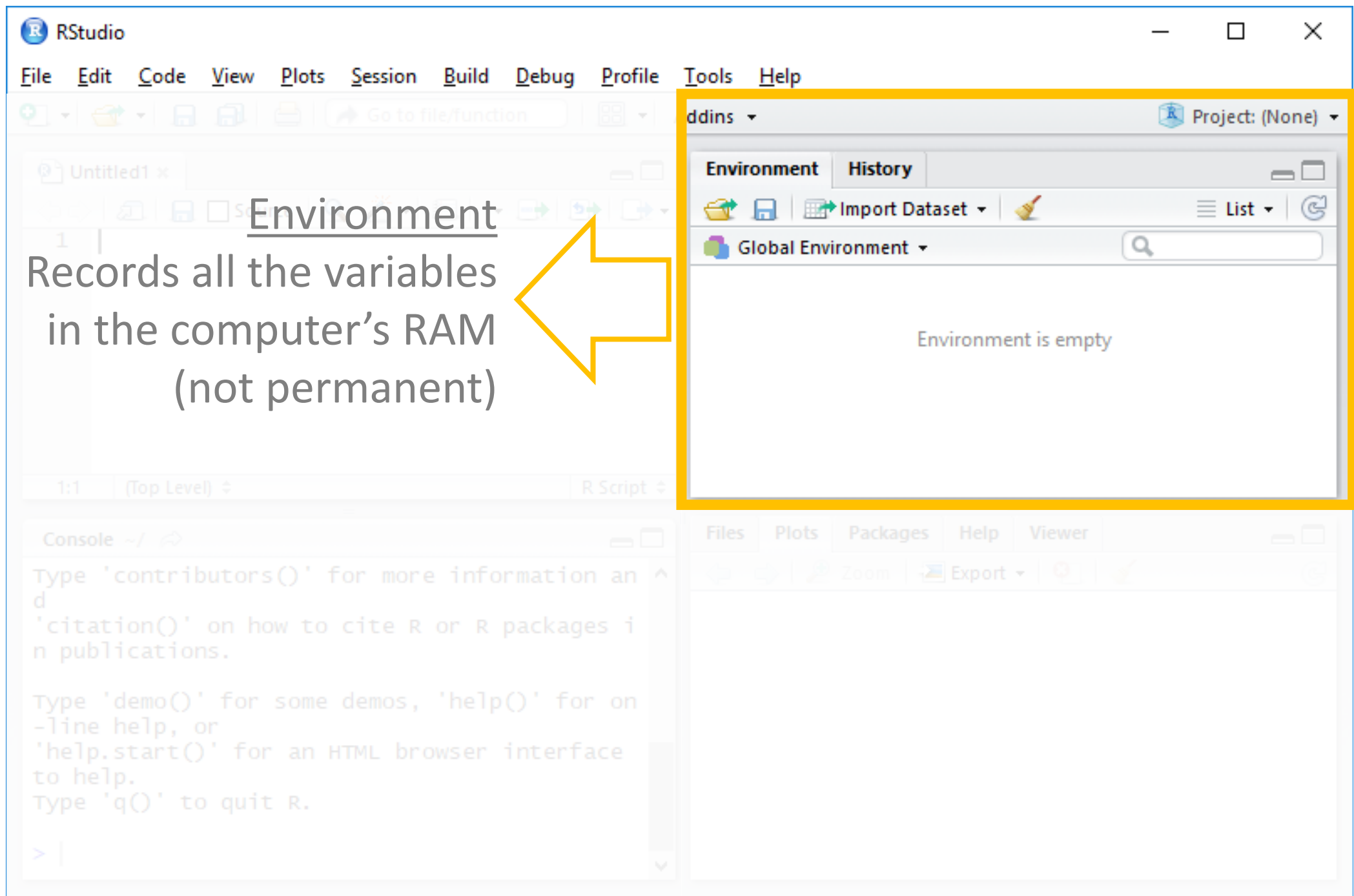
Help

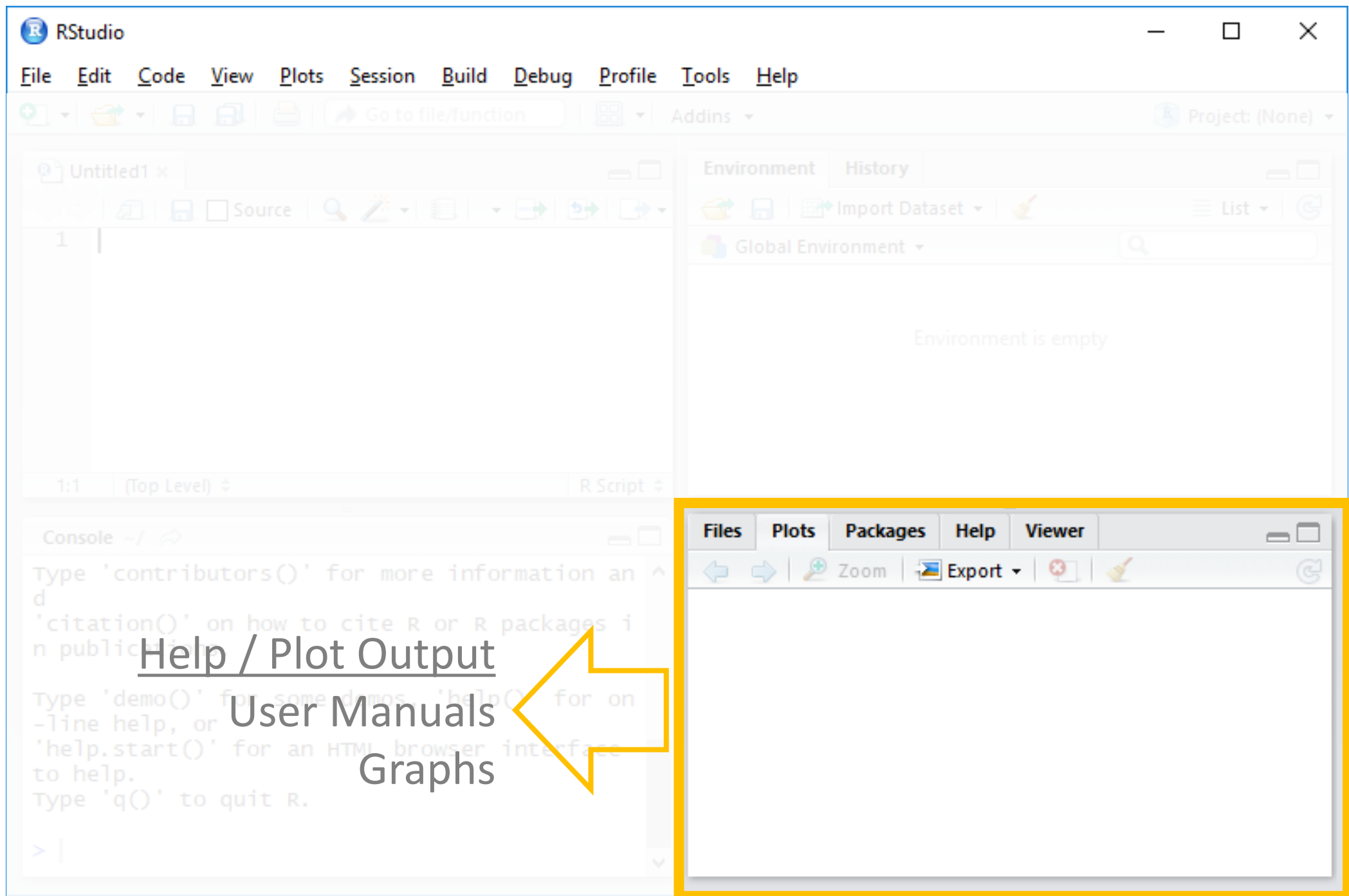
Viewer

Zoom Export









Options



General



Code



Appearance



Pane Layout



Packages



R Markdown



Sweave



Spelling



Git/SVN



Publishing

Choose the layout of the panes in RStudio by selecting from the controls in each quadrant.

Source

Environment, History, Connect

☒ Environment

☒ History

☐ Files

☐ Plots

☒ Connections

☐ Packages

☐ Help

☒ Build

☒ VCS

☐ Viewer

Console

Files, Plots, Packages, Help, View

☐ Environment

☐ History

☒ Files

☒ Plots

☐ Connections

☒ Packages

☒ Help

☐ Build

☐ VCS

☒ Viewer

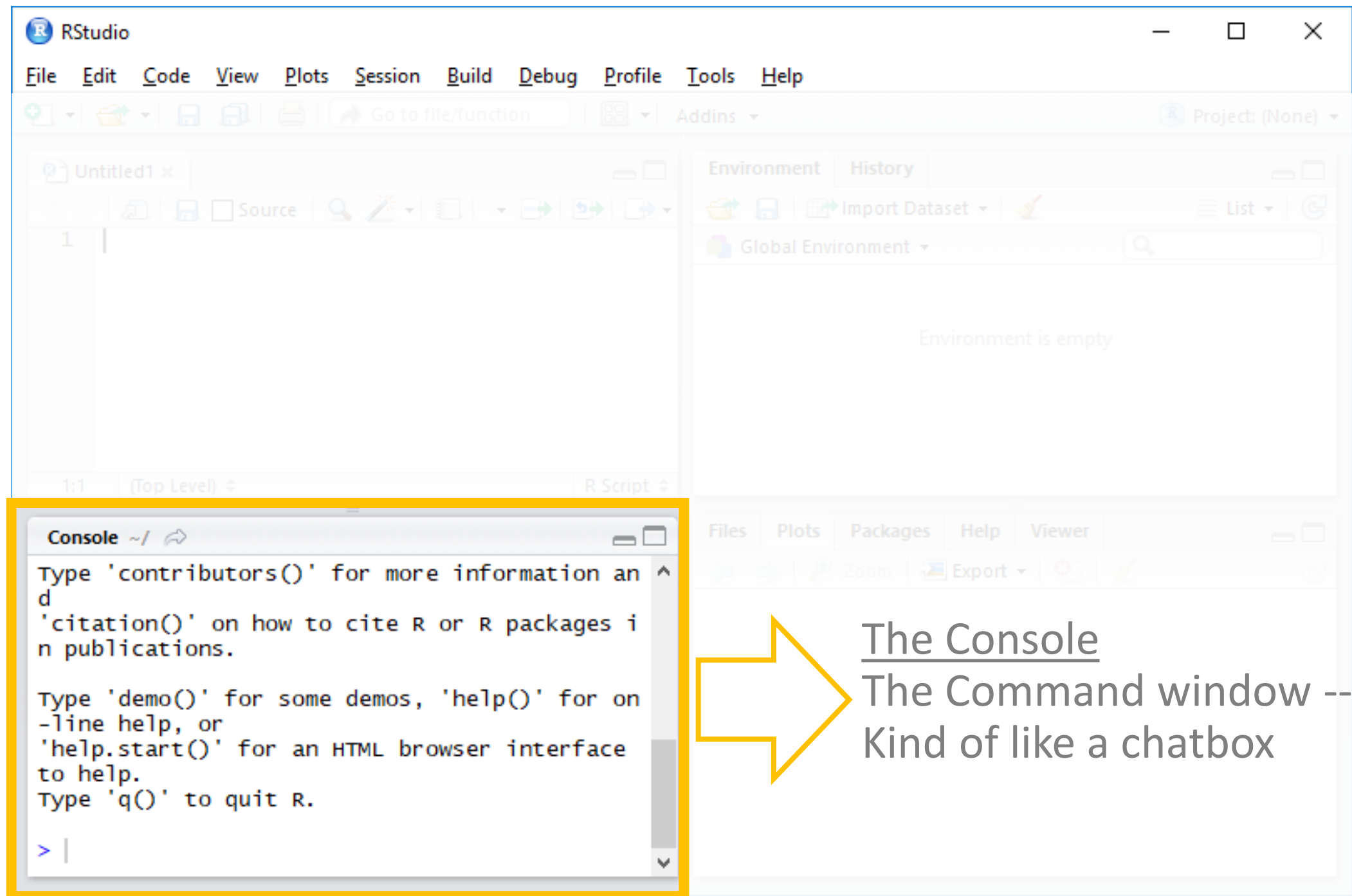
OK

Cancel

Apply

To Change the Layout:
Tools → Global Options →
Pane Layout

R as a calculator



The Console

The Command window --
Kind of like a chatbox

Basic Calculation

- Addition

```
>3 + 4  
[1] 12
```

- Subtraction

```
>8 - 6  
[1] 2
```

- Multiplication

```
>8 * 6  
[1] 48
```

- Division

```
>20 / 3  
[1] 6.666667
```

- Precedence

```
>(20 + 4) / 3  
[1] 8
```

- Remainder

```
>20 %% 3  
[1] 2
```

- Square

```
>3 ^ 2  
[1] 9
```

Variable Assignment (= or <-)

- A variable is a collection of value(s)
- Variable names should be meaningful
- Variable names are also case sensitive
- Variables are lost after you restart R

Do something:
assign x to 1, save
to environment

```
>x = 1
```

(same thing)

```
>x <- 3
```



Overwrites the variable
in the memory

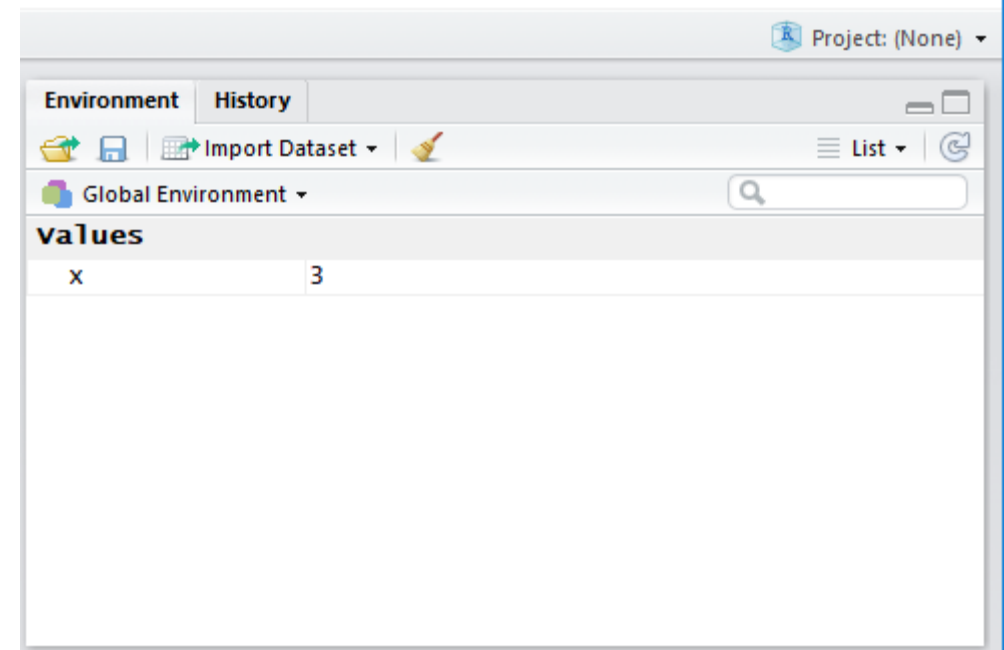
```
>x = 5 + 4
```

```
>x
```

```
[1] 9
```

1. Calculate the
answer

2. Assign the value
to the variable



Data Type (ref. Scale of Measurement)

- Numeric

```
>x = 3
```

- Class query

```
>x = 1  
>class(x)  
[1] "numeric"
```

- Factor

```
>x = factor(x)
```

- Character

```
>x = "Jonas"
```

Use ' or " to
indicate a string /
character

- Boolean

```
>x = TRUE
```

Or FALSE, always in
uppercase

- List of __

```
>x = c(1, 2, 3, 4, 5)
```

Arrays c()

- Numeric

```
>age = c(18, 28, 24)
```
 - Character

```
>name = c("Jonas",  
"Mary", "Tom")
```
 - Boolean

```
>male = c(TRUE,  
FALSE, TRUE)
```
 - Accessing elements in an array

```
>age[1]  
[1] 18
```
 - Broadcasting

```
>age = age + 1
```

Same as 

```
>age = age + c(1,1,1)
```
- ```
>age
[1] 19 29 25
```

# Data Frame `data.frame()`

- Constructing a Data Frame
- Checking the structure (first 6 entries)

```
> students = data.frame(name, age, male)
```

```
> head(students)
```

```
> age = c(18,28,24)
> name = c("Jonas", "Mary", "Tom")
> male = c(TRUE, FALSE, TRUE)
> students = data.frame(name, age, male)
> head(students)
 name age male
1 Jonas 18 TRUE
2 Mary 28 FALSE
3 Tom 24 TRUE
> |
```

Each row is  
an entry



Each column  
is a variable



# Data Frame `data.frame()`

- Constructing a Data Frame
- Checking the structure
- Accessing individual column
- Accessing individual record
- Adding an entry
  - Create a new entry
  - "row bind" it to the data frame
- Check number of entries
- Remove individual variables

```
>students = data.frame(name, age, male)
```

```
>head(students)
```

```
>students$age
```

```
>students[students$name == "Jonas",]
>subset(students, name == "Jonas")
```

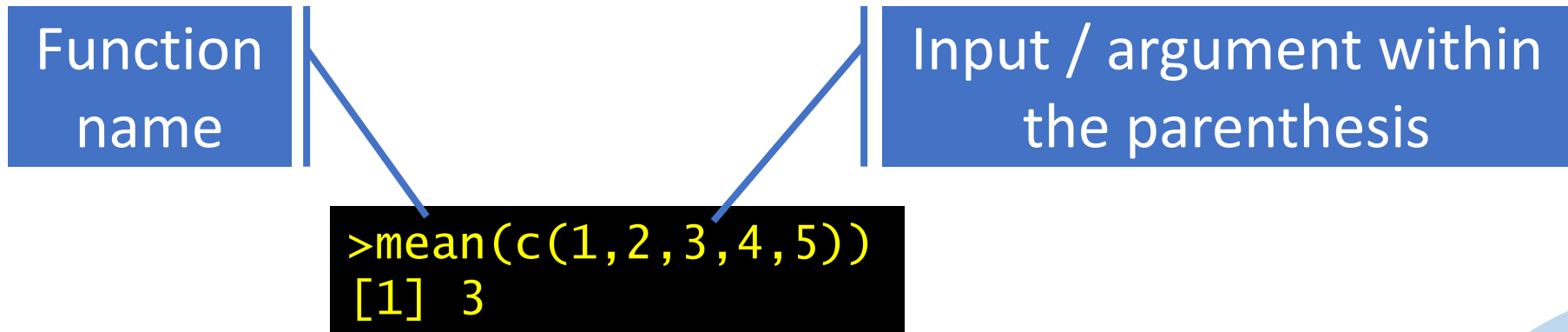
```
>new_student = data.frame(name = "Sue",
 age = 21, male = FALSE)
>students = rbind(students, new_student)
```

```
>nrow(students)
```

```
>rm(name, new_student, age, male)
```

# Using Built-in Functions

- Functions take in some input and return some output
- For example, the function `mean()` takes in an array of number and return the arithmetic average of the array



- For usage of a function, type `?`  followed by the function name, e.g.,

```
>?mean
```

# Some Common Built-in functions

We'll learn more functions in the quarter:

- Calculate the mean of an array, the input x is an array:

```
>mean(x)
```

- Or the sum of an array, the input x is an array

```
>sum(x)
```

- Return the number of items in an array:

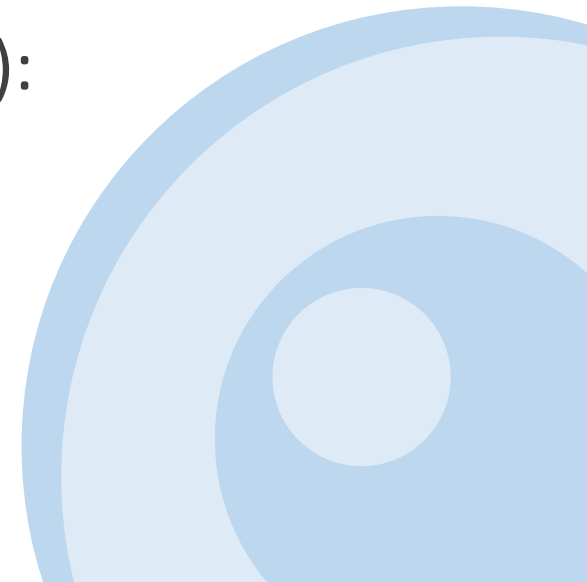
```
>length(x)
```

- Taking the square-root of a number / array (broadcasting):

```
>sqrt(x)
```

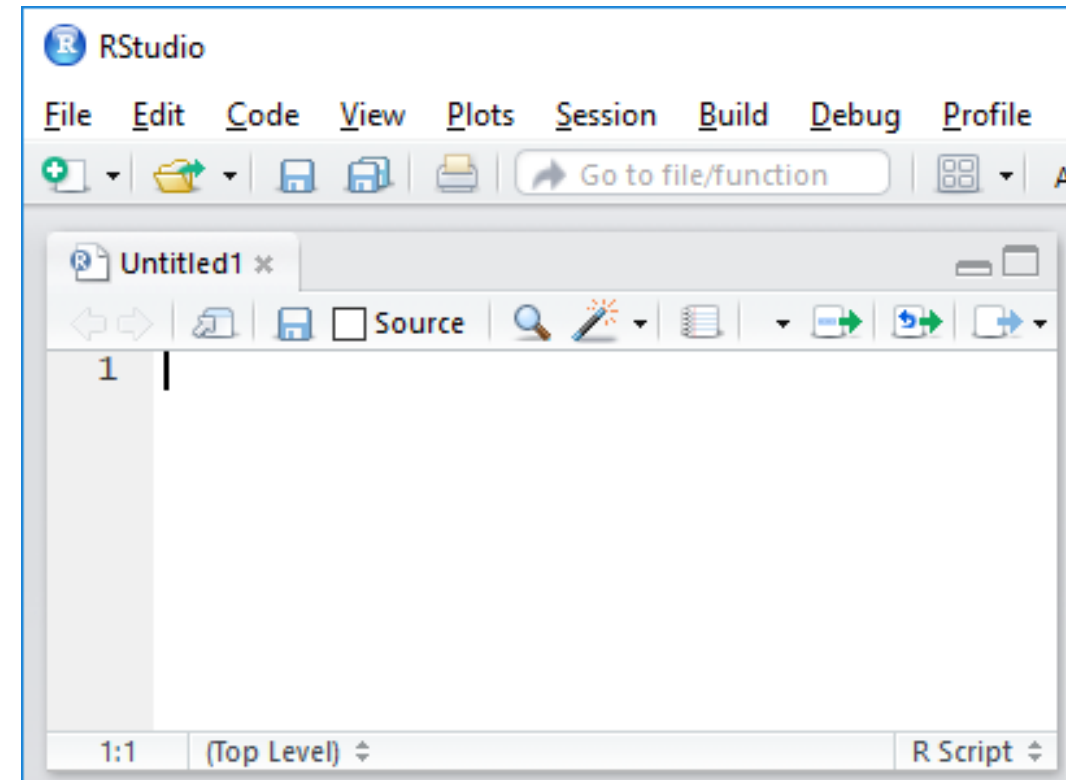
- Create an numeric array from 1 to 10, with a step of 0.5

```
>seq(from = 1, to = 10, by = 0.5)
```



# The Scripting Window (R Script / R Notebook)

- The console does not save your work after you close R
- You usually want to write your commands, save them, and be able to run them later
- R Notebook allows you to write codes, make notes (comments), and run the codes
- R Script is a list of commands  
File → New File → R Script
- R Notebook allows you to run your codes, show some graphs within a formatted document  
File → New File → R Notebook

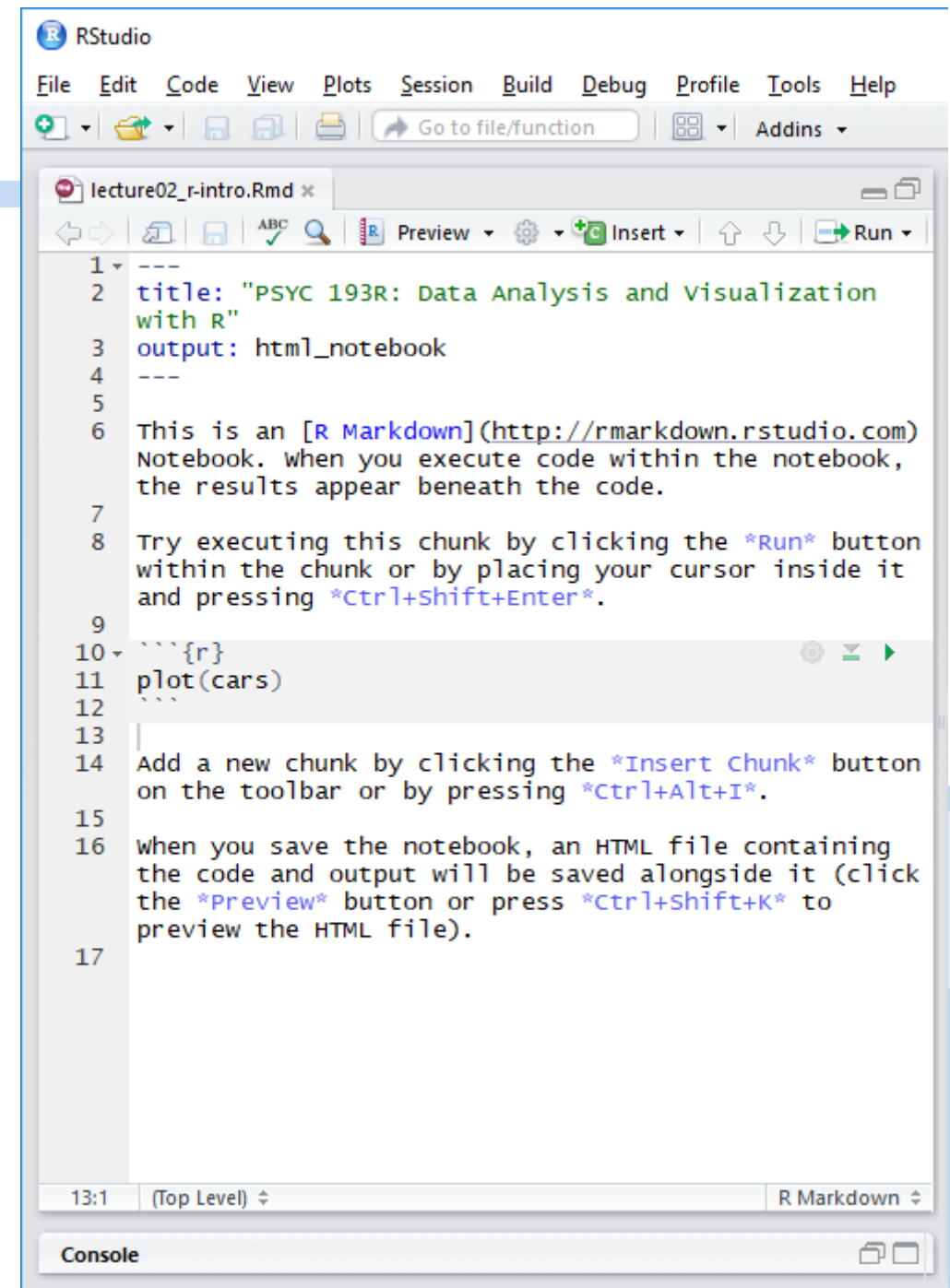


# R Notebook

- Anything outside the gray box will not be run

```
```{r}  
# Some code here  
```
```

- Anything in the gray box but start with `#` will not be run
- When you press `Ctrl+Shift+Enter`, the code in the gray box be transfer to the console and run



\* You will hand in the R Notebook and the html file for this class

# Write your own functions

- This function, named `addition()`, takes in two inputs / arguments and returns the sum of the two numbers.

Tell R this is a function

Some inputs, separate by commas if more than 1 input

Name of the function

```
addition = function(x,y){
 #sum of the two numbers
 z = x + y
 return (z)
}
```

Do some computations

Output of the function, followed by return keyword

```
>addition(4,5)
[1] 9
```

# General Programming Principles

Make sure: if you're reading your code again in 3 – 6 months, you would still understand yourself

Increasingly, codes are being shared, so make sure other people can understand you as well

- Name variables meaningfully  
(`x = 100` vs `population_mean = 100`)
- Show your steps
- Comments your code  
(outside the gray boxes, or with `#` inside the gray boxes)
- Write / use functions to perform repeated actions  
(Object-Based Programming)



# Your turn

- Let's write a function to convert temperature in degrees Fahrenheit into temperature in degrees Celsius:

$$T_{(^{\circ}\text{C})} = (T_{(^{\circ}\text{F})} - 32) \times 5/9$$

A default  
value  
(optional)

```
f_to_c = function(degrees_f = 0){
 #your code starts here (1 line)
 degrees_c = NA
 # your code ends here
 return (degrees_c)
}
```

```
> f_to_c(85)
[1] 29.4444
```

Expected  
Answer

In between:  
```\{r\}`` & ```\n```\nWrite the code  
Then press
Ctrl+Shift+Enter

Test your code,
run it in Console

Class Challenge

- Let's write a function to convert a raw score to a Z-score. It takes 3 arguments, (1) the raw score X , (2) the population mean μ , and (3) the population standard deviation σ

$$Z = \frac{X - \mu}{\sigma}$$

In between:
```\{r\} & ```\nWrite the code  
Then press  
Ctrl+Shift+Enter



```
raw_to_z = function(x, mu, sd){
 #your code starts here (1 line)
 z = NA
 # your code ends here
 return (z)
}
```

Test your code,  
run it in Console



```
> raw_to_z(110, mean = 100, sd = 15)
[1] 0.6666667
```



Expected  
Answer

# Class Challenge

- Write a function that takes in an array of numbers, and returns the population standard deviation of the array:

$$pop\_sd = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

- Test your code with:

```
> x = c(1,2,3,4,5)
pop_sd(x)
[1] 1.414214
```

```
pop_sd = function(x){
 # Mean of the scores (a number)
 M = NA

 # Deviation scores (a vector)
 deviation = NA

 # Squared deviations (a vector)
 sq_deviation = NA

 # Sum of squares (a number)
 SS = NA

 # Number of scores in the array (a number)
 N = NA

 # Calculate the population SD
 std_dev = NA

 # Return the output: the population SD
 return (std_dev)
}
```

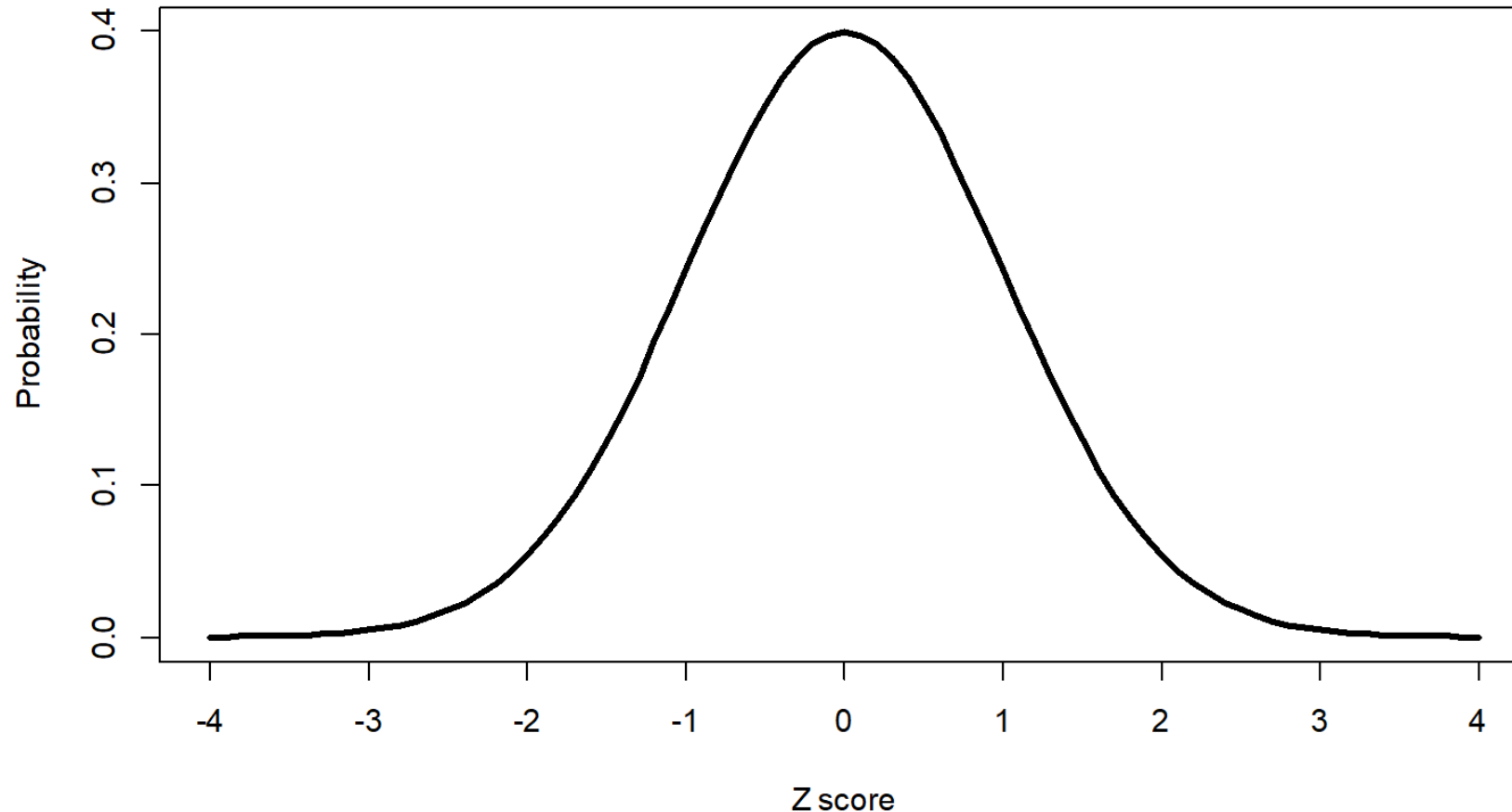
Probability

# Z-score and the Population

- Many distributions in the real world follow a "normal" distribution
- If we have every score in the population, we can summarize the distribution using:
  - a population mean  $\mu$
  - a standard deviation  $\sigma$
- Having every score of a population is unlikely, unless the population is small (e.g., the midterm scores of a class)
  - The more likely scenario is a Standardized Scale (e.g., IQ, SAT, GRE, or other well-established scales) with a well-defined mean and standard deviation

# The Unit Normal Distribution

- Population mean  $\mu = 0$ ; Population standard deviation  $\sigma = 1$
- Total area under the curve = 100%, or 1.0

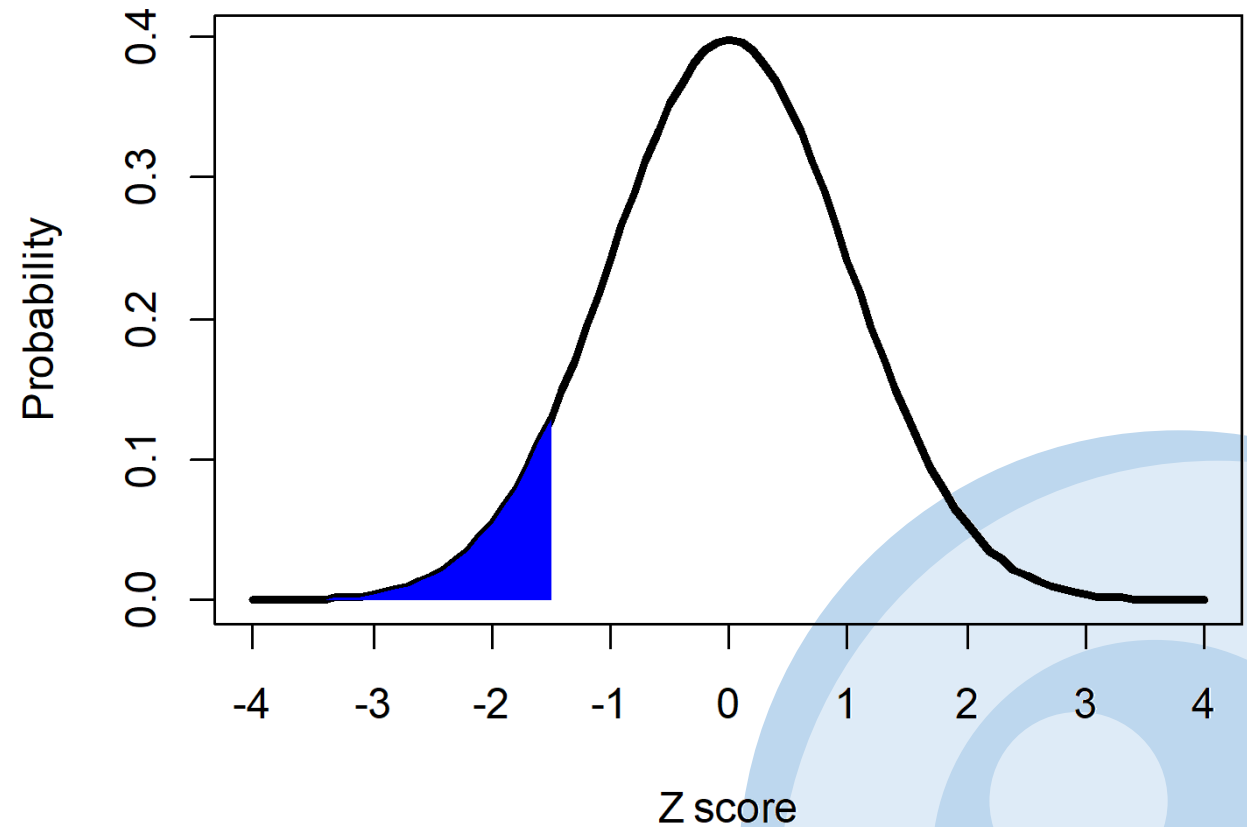


# The Unit Normal Distribution: $Z \rightarrow$ Probability

- Given a Z-score (e.g., -1.5), what is the probability of getting such a score or smaller?  
 $p(Z < -1)$

```
> pnorm(-1.5, mean = 0, sd = 1)
[1] 0.0668072
```

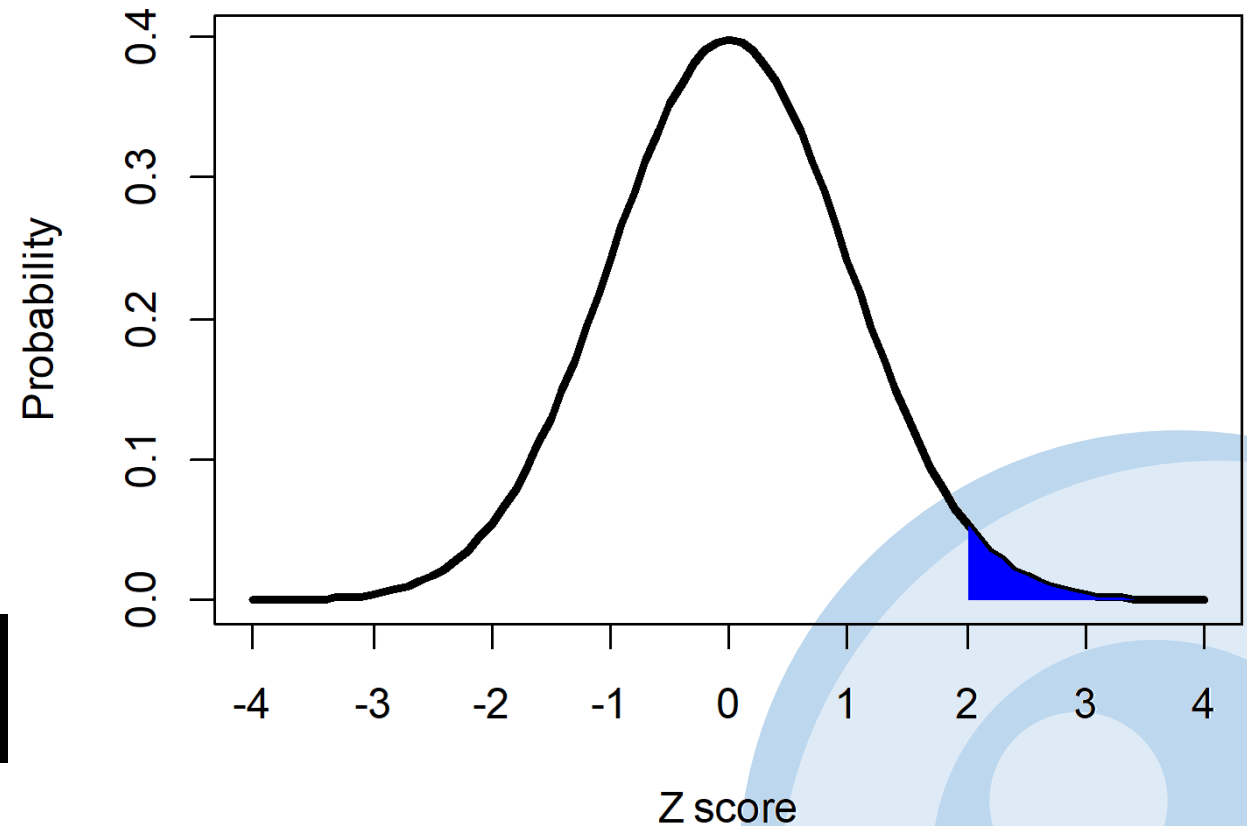
- This probability is also a *percentile*



# The Unit Normal Distribution: $Z \rightarrow$ Probability

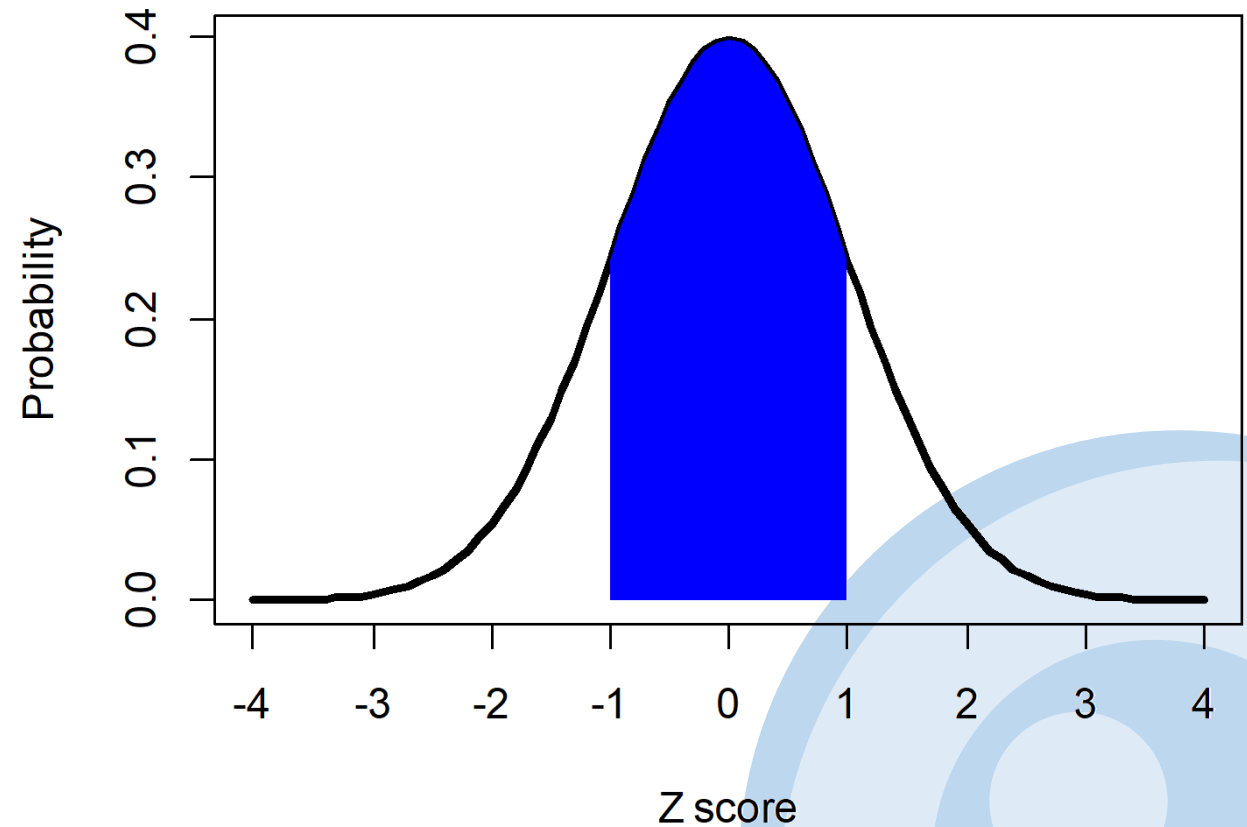
- `pnorm()` always gives you the probability from  $-\infty$  to the required score
- If you want to know the probability of getting a particular score or larger, use `1-pnorm()`
- The probability of getting a Z-score of 2 or larger is...  
 $p(Z > 2)$

```
> 1 - pnorm(2, mean = 0, sd = 1)
[1] 0.02275013
```



# Class Challenge: $Z \rightarrow$ Probability

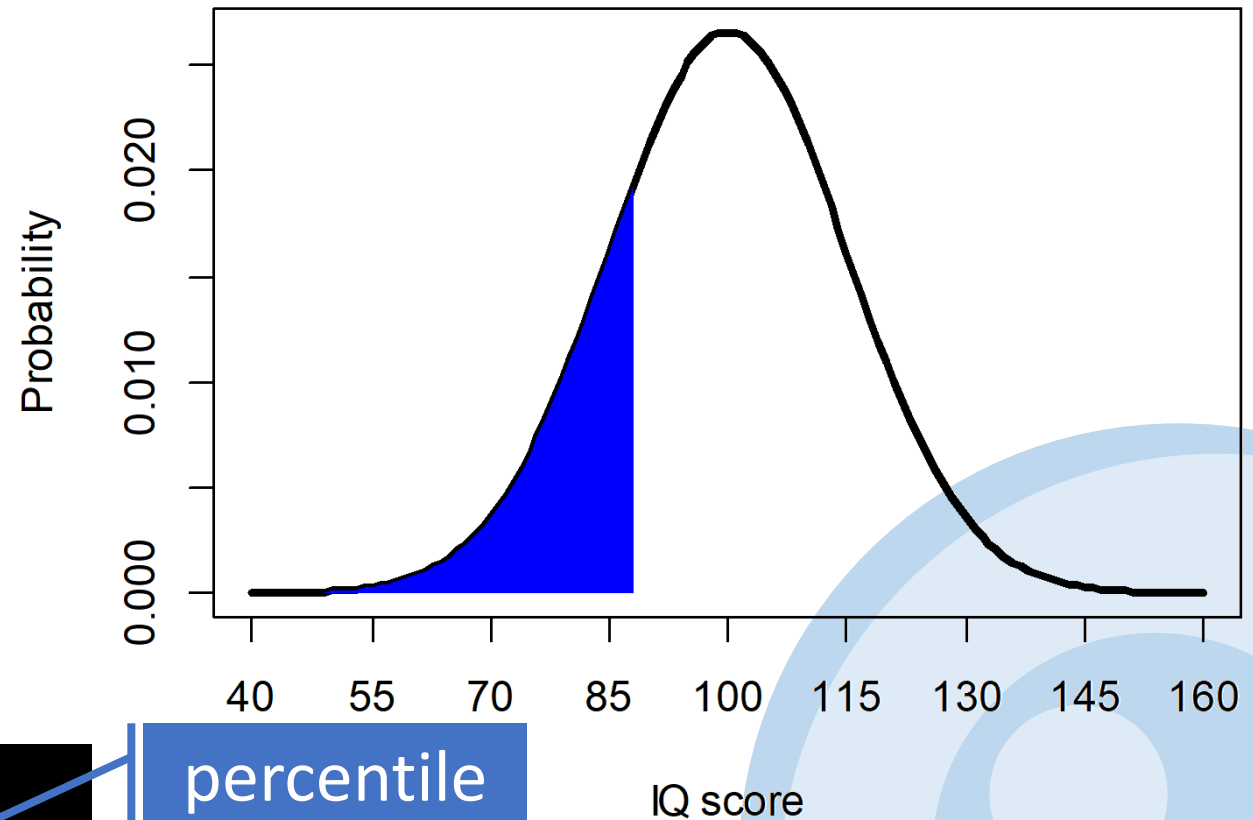
- Remember: probability is just area under the curve
- Find the probability of:
  - $p(Z < -1)$
  - $p(Z > -0.5)$
  - Z-score between -1 & +1, i.e.,  $p(Z > -1 \text{ \& } Z < 1)$



# IQ score distribution: Quantile → Probability

- We know that the general population has a mean IQ score of 100 and a standard deviation of 15.
- Jonas has an IQ score of 88, what is the probability of getting such a score or smaller?  
It is the same as asking: *How many percent of people in the population score the same, below him?*  $p(\text{IQ} < 88)$

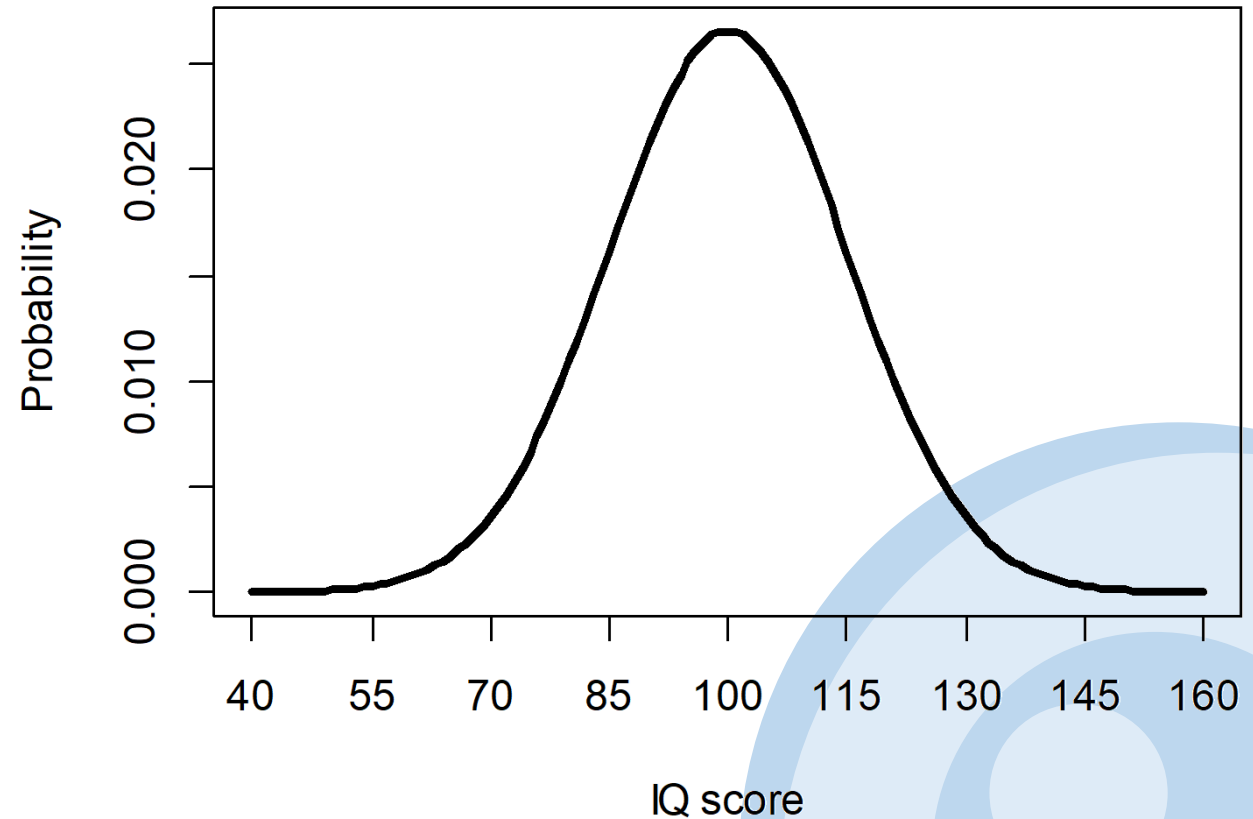
```
> pnorm(88, mean = 100, sd = 15)
[1] 0.2118554
```



# IQ score distribution: Quantile $\rightarrow$ Probability

- We know that the general population has a mean IQ score of 100 and a standard deviation of 15.
- Find the probability of the following:
- $p(\text{IQ} > 105)$
- $p(\text{IQ} > 85 \text{ \& } \text{IQ} < 115)$

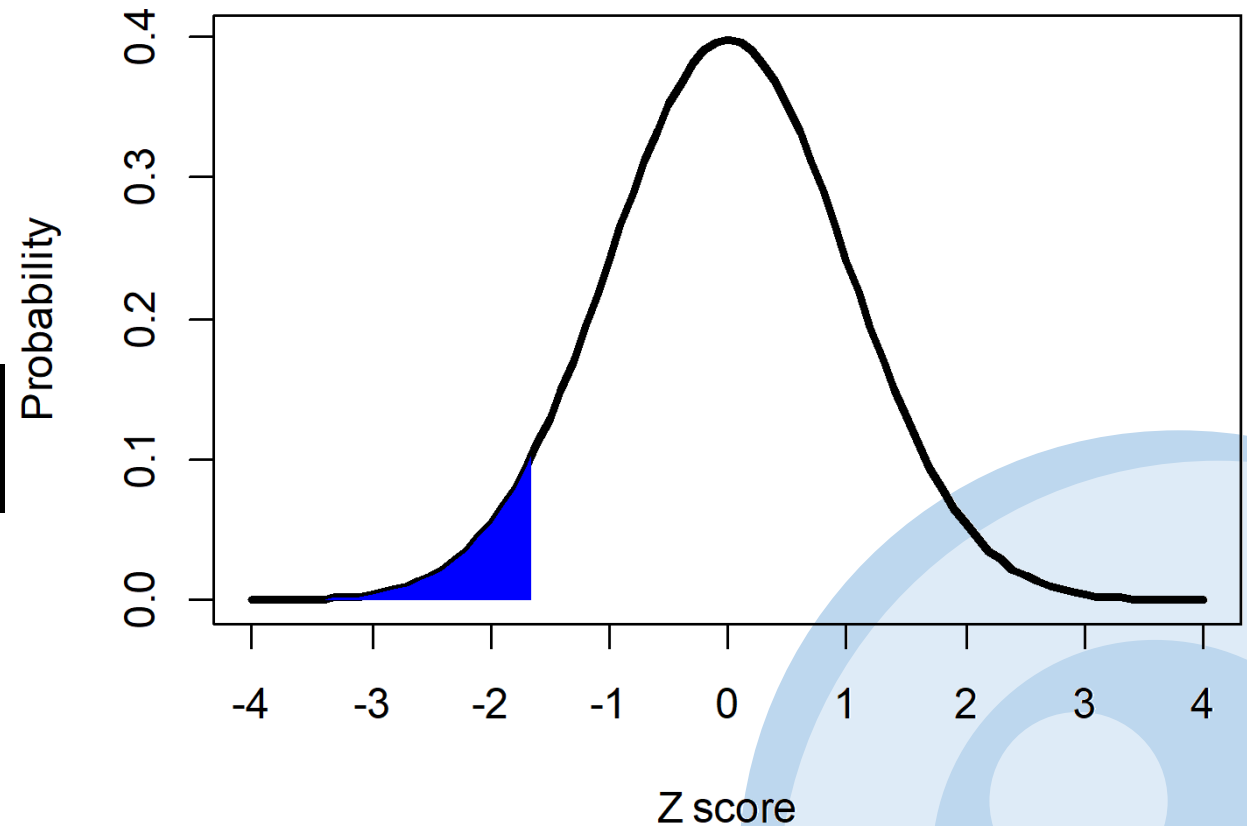
💡 Tips : Always draw the distribution



# Probability → Quantile: `qnorm()`

- Sometimes we're giving the probability and we want to know a cut-off point
- For example, we want to know the Z-score that corresponds to the lowest 5% of the distribution

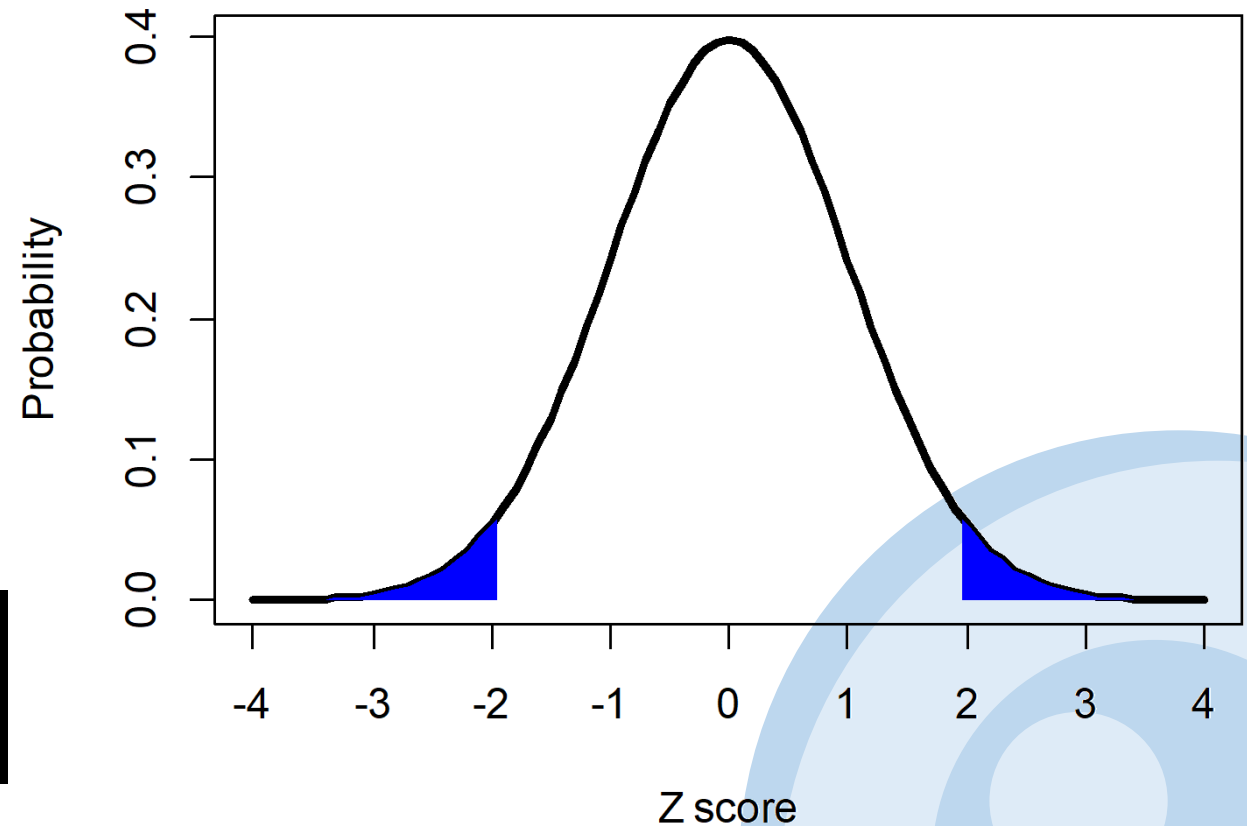
```
> qnorm(0.05, mean = 0, sd = 1)
[1] -1.644854
```



# Probability → Quantile: `qnorm()`

- Given a percentile, `qnorm()` returns the quantile
- If  $\text{mean} = 0$ , and  $\text{sd} = 1$ , the quantile is a Z-score
  - Note: Percentile is the probability of getting a score or a score smaller than it
- Z-scores associated with lower & upper 2.5%

```
> qnorm(c(0.025, 1-0.025), mean = 0, sd = 1)
[1] -1.959964 1.959964
```

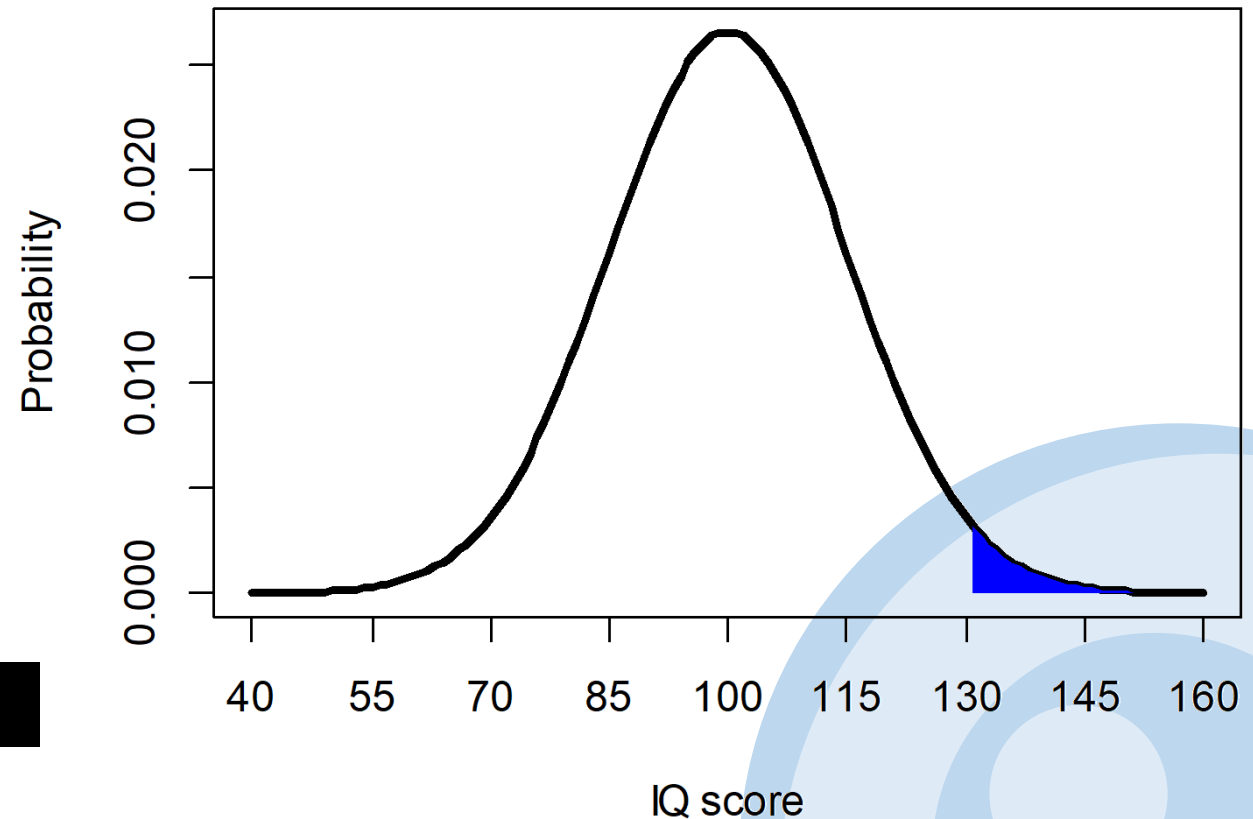


# Probability → Quantile: `qnorm()`

- You can change the mean and the sd arguments when you are working on some other normal distributions, e.g., IQ
- Mensa International accepts people who score at the 98th percentile or higher on a standardized IQ test. What is the lowest IQ score one can get to join Mensa?

```
> qnorm(NA, mean = NA, sd = NA)
```

Replace the NA's



# Class Challenge

💡 Tips: Decide: percentile → quantile /  
quantile → percentile first

1. People with an IQ score of 110 are generally regarded as intelligent enough for college education. What is the percentage of people in the general population who are eligible for college education?
2. People score the lowest 8% in an IQ test are considered mentally challenged. A clinical psychologist wants to know the cut-off score for the test. Please help her.
3. Jason considers himself a good writer. In the verbal section of SAT, he scored 720. The mean and standard deviation of the test is 500 and 100 respectively. What is the percentile of Jason's score.

# What we have achieved...

---

- Basic R Operations
- Built-in functions and customized functions ☆
- The normal distribution:
  - `pnorm()`: quantile → percentile ☆
  - `qnorm()`: percentile → quantile ☆
- Homework 1 will be post tomorrow

