Basics of Data Wrangling

ADA

Miguel Salema

Católica-Lisbon SBE

February 21, 2024

Section 1

Review

Resources

- DataCamp
- The ADA website
- The R website

Review of Last Class

Exercise 1

What is the average of a sequence starting at 9, ending at 100, and going 5 by 5?

Extra Exercise

What is the average of the previous vector *after* taking the square root of all the numbers/elements?

Review of Last Class

Exercise 1

What is the average of a sequence starting at 9, ending at 100, and going 5 by 5?

```
mean(seq(9, 100, 5))
## [1] 54
```

Extra Exercise

What is the average of the previous vector *after* taking the square root of all the numbers/elements?

```
mean(sqrt(seq(9, 100, 5)))
## [1] 7.061317
```

Quarto

Have to rendered last class Quarto document? Let's review it again. Specially:

- yaml (yaml ain't markdown language)
- Lists
- R chunks options
- Inline R code
- Cross-references
  - To Figures
  - Sections
  - ...
- Citations

Assignment

In the website:

https://afidalgo.org/ada/

Deliver on moodle.

Section 2

Tidyverse

What is the Tidyverse?

{tidyverse} is a bundle of packages. The packages are essentially a redesign of the core functions of R with a coherent philosophy. The bundle comes close to become a complete and self-contained system on its own thanks to all the functions included in its packages.

```
#load
library(tidyverse)
```

Tibbles

The `read_csv()` function automatically read the data into a `tibble`. "A `tibble` is a more advanced and capable version of `R`'s default format for data (`data.frame`)."[1]. This is the data format we will use during this class.

Usage:

```
df <- as_tibble(df)
```

[1] From tibble.tidyverse.org

Piping

Figure 1: *The Treachery of Images* by René Magritte (1929)

Package

```
library(magrittr) #already inside tidyverse
```

```
# Ctrl + Shift + M (Windows)
# Cmd + Shift + M (Mac)
%>%
```

Why pipe?

What piping does:

$$f(x, y) = x \,\%>\%\, f(y)$$

Piping substitutes the 1^{st} element of a function.

This is very helpful for chained operations:

$$j(h(g(f(x)))) = x \,\%>\%\, f() \,\%>\%\, g() \,\%>\%\, h() \,\%>\%\, j()$$

Becomes less cumbersome and more intuitive because we read the operation by order.

Why pipe?

Basic Example

```
vec1 <- c(1:1000, NA) # vector with missing value

# these 2 lines are equivalent:
round(sqrt(sd(vec1, na.rm = TRUE)), digits = 2)
## [1] 16.99
vec1 %>% sd(na.rm = TRUE) %>% sqrt() %>% round(digits = 2)
## [1] 16.99
```

Exercise

Use a pipe to find out the names of every column in the dataset.

Why pipe?

Basic Example

```
vec1 <- c(1:1000, NA) # vector with missing value

# these 2 lines are equivalent:
round(sqrt(sd(vec1, na.rm = TRUE)), digits = 2)
## [1] 16.99
vec1 %>% sd(na.rm = TRUE) %>% sqrt() %>% round(digits = 2)
## [1] 16.99
```

Exercise

Use a pipe to find out the names of every column in the dataset.

```
df %>% names()
```

Main Functions of Dplyr

The `dplyr` package is part of the `tidyverse` family. Its main functions are:

- `select`: to choose what columns to keep;
- `arrange`: to order the dataset;
- `filter`: sub-setting what rows to keep;
- `mutate`: create a new variable;
- `group_by`: group the data;
- `summarize`: produce summary data responsive to groups.

```
# load the package
library(dplyr)

# alternativly, load the full tidyverse
library(tidyverse)
```

select()

Usage:

```
select(df, var1, ..., varn)

#Use always the pipe
df %>%
  select(var1, ..., varn)
```

arrange()

The order will change with the position of variables.

```
df %>%
  arrange(var1, ..., varn)
```

filter()

The `filter` function is from the `dplyr` package, which is part of the `tidyverse` family.

```
df %>%
  filter(condition)
```

Logical Operators

- ==
- > ; <; >= ; <=
- %in%
- !

```
TRUE & TRUE
## [1] TRUE
```

```
TRUE & FALSE
## [1] FALSE
```

```
TRUE | FALSE
## [1] TRUE
```

```
FALSE | FALSE
## [1] FALSE
```

filter()

Exercise 2

How many workers are older than 50 years old?

Exercise 3

What is the percentage of workers with more than 50 years old that have a college degree?

filter()

Exercise 2

How many workers are older than 50 years old?

```
df %>%
  filter(age > 50) %>%
  summarise(n())
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1 57590
```

Exercise 3

What is the percentage of workers with more than 50 years old that has a college degree?

```
df %>%
  filter(educ %in% c("Bachelor_3", "Bachelor_4", "Master", "Doctorate") & age > 50) %>%
  summarise(n())
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1  4808
```

mutate()

```
df %>%
  mutate(new_variable = expression)
```

group_by() + summarize()

```
df %>%
  group_by(group) %>% #groups the dataset
  summarize(new_variable = expression) %>%
  ungroup() #remove the grouping
```

Section 3

Why does the Portuguese Average Wage Seems Higher in Reality than in the Data?

Av. Wage in the Portuguese Private Sector

Year after year, students get surprised about how low the average portuguese wage is. Why is that? I guess that is due to:

- Informality: at least 20% of the Portuguese GDP is informal;
  A economia informal tem representado entre cerca de 22% e 25% do PIB, quando comparada, por exemplo, com a França (entre 13% e 14%) ou os EUA (9% - 10%) (Antunes e Cavalcanti 2006)
- "The Lisbon College" bubble: you think about your parents and the parents of your friends.

"The Lisbon College" bubble

- Given you are in Lisbon and in college, these adults generally fulfill this conditions:
  1. They have a college degree (people with a college degree send their kids to college more often);
  2. They are older than 45 years old ;
  3. They work in the Lisbon metropolitan area.

`qp_sample.csv`

This file is a small sample (10%) from *Quadros de Pessoal* (QP), a Portuguese LEED (linked employer-employee dataset). Each observation represents a real worker. It includes only some worker-level variables. QP contains information about all Portuguese private firms with at least 1 employee. The sample includes only observations from 2018.

Implementation

```
# The average wage in 2018
df %>%
  summarise(mean(total_wage))
## # A tibble: 1 x 1
##   `mean(total_wage)`
##                <dbl>
## 1              1363.

# The average wage in 2018 for workers with a college degree
df %>%
  filter(educ %in% c("Bachelor_3", "Bachelor_4", "Master", "Doctorate")) %>%
  summarise(mean(total_wage))
## # A tibble: 1 x 1
##   `mean(total_wage)`
##                <dbl>
## 1              2215.
```

Implementation

```
# The average wage in 2018 for workers with a college degree and older than 45
df %>%
  filter(educ %in% c("Bachelor_3", "Bachelor_4", "Master", "Doctorate") &
           age > 45) %>%
  summarise(mean(total_wage))
## # A tibble: 1 x 1
##   `mean(total_wage)`
##                <dbl>
## 1              3457.

# The average wage in 2018 for workers with a college degree, older than 45 and wor
df %>%
  filter(educ %in% c("Bachelor_3", "Bachelor_4", "Master", "Doctorate") &
           age > 45 &
           nut2_firm == "Lisboa") %>%
  summarise(mean(total_wage))
## # A tibble: 1 x 1
##   `mean(total_wage)`
##                <dbl>
## 1              3948.
```

Implementation

```
# The average wage in 2018 for workers with a college degree, older than 45, work i
df %>%
  filter(educ %in% c("Bachelor_3", "Bachelor_4", "Master", "Doctorate") &
           age > 45 &
           nut2_firm == "Lisboa" &
           male == 1) %>%
  summarise(mean(total_wage))
## # A tibble: 1 x 1
##   `mean(total_wage)`
##                <dbl>
## 1              4700.
```

We can see that being a men also had an effect on the average wage. In one of our next classes we will explore more the gender wage gap. What explains it?

Section 4

The Normal Distribution

The normal distribution is the most common distribution used in statistics.

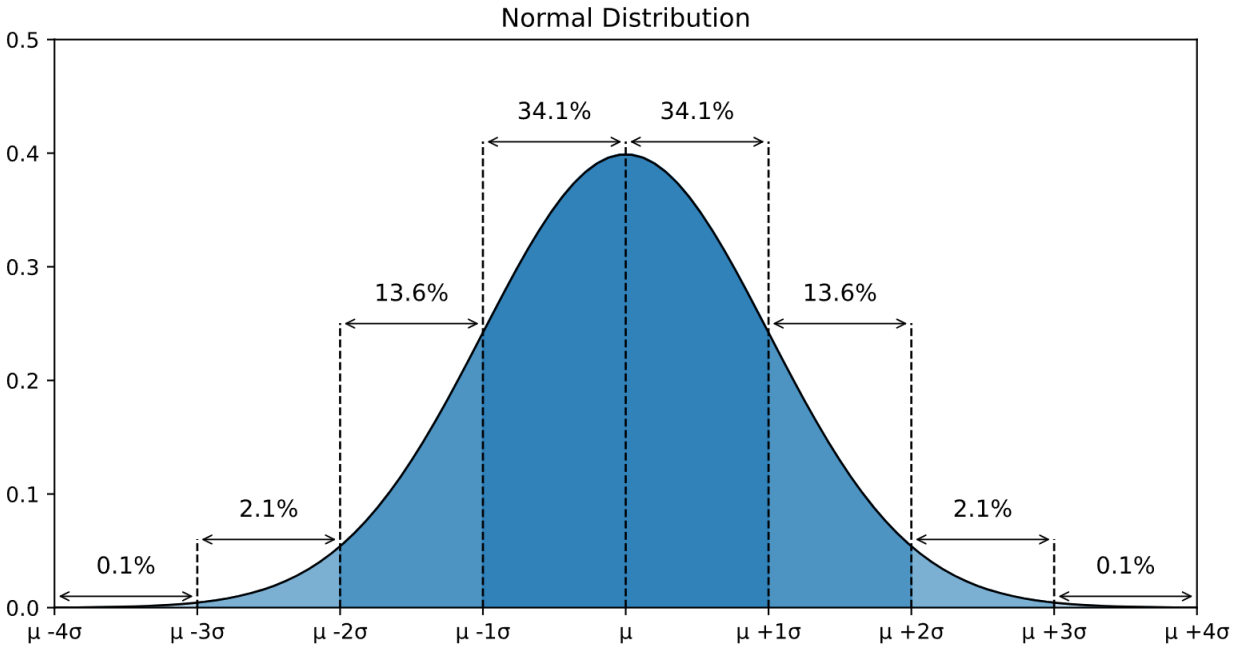


Figure 2: Benchmark values of the Normal Distribution

Probability Density Function

The PDF for the for the normal distribution is given by:

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

where μ is the mean and σ^2 is the standard deviation.

Properties:

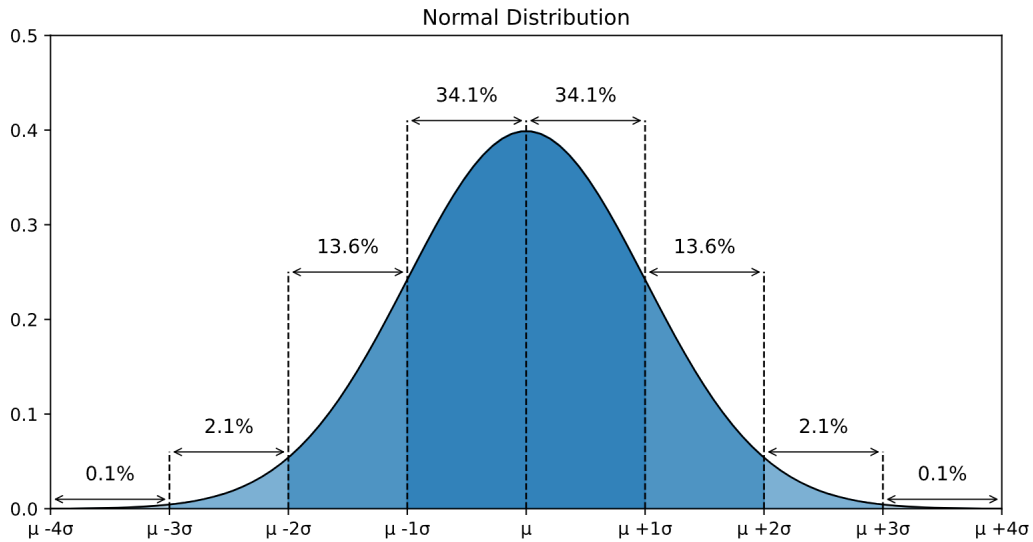
- Bell shaped;
- Symmetrical;
- Unimodal (it has one "peak");
- Mean = Median.

Example: IQ

In modern IQ tests the raw score is transformed into a normal distribution with (Gottfredson 2009):

- $\mu = 100$
- $\sigma = 15$

What is the probability of a random person having an IQ higher than 130?



Standard Normal Distribution

The standard normal is $\mu = 0$ and $\sigma = \sigma^2 = 1$.

$$x \sim N(0, 1)$$

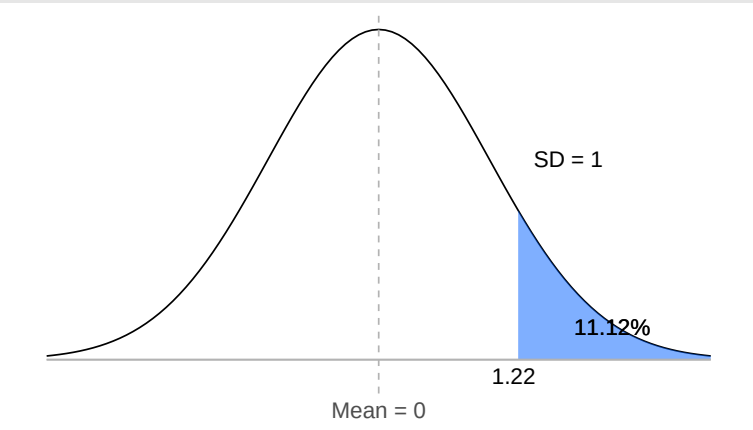
The standard normal's PDF is often denoted by $\phi(x)$ and its CDF by $\Phi(x)$.

We can convert **any** random variable following a normal distribution with any μ and σ into a random variable following a standard normal distribution. Doing so will help us with probability calculations.

What is an Unlikely Draw?

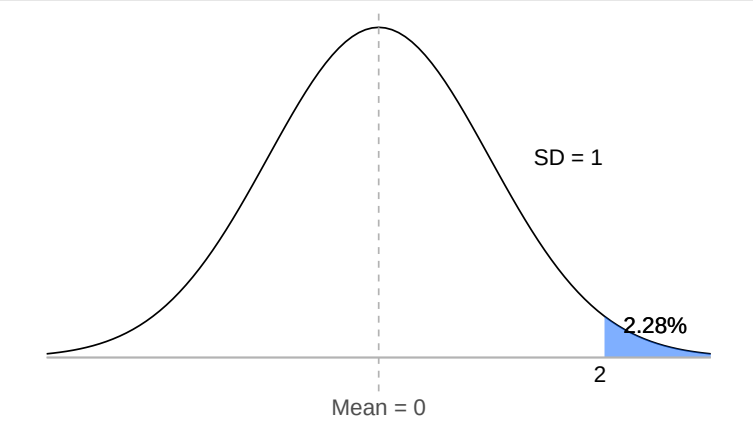
What is an extreme value? If we assume that x was drawn from a normal distribution we can know if it's likely/unlikely that a number higher than z was just randomly selected. Using the PDF, we can know the probability of having achieved a number higher than z by chance alone.

$z = 1.22$

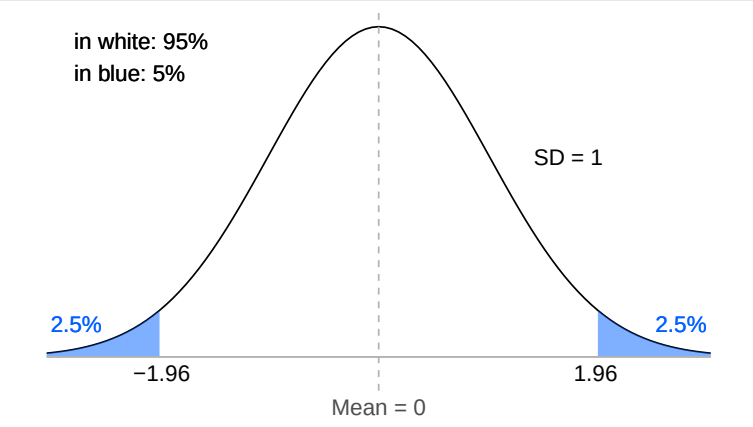


What is an Unlikely Draw?

$z = 2$



Not Between -1.96 and 1.96

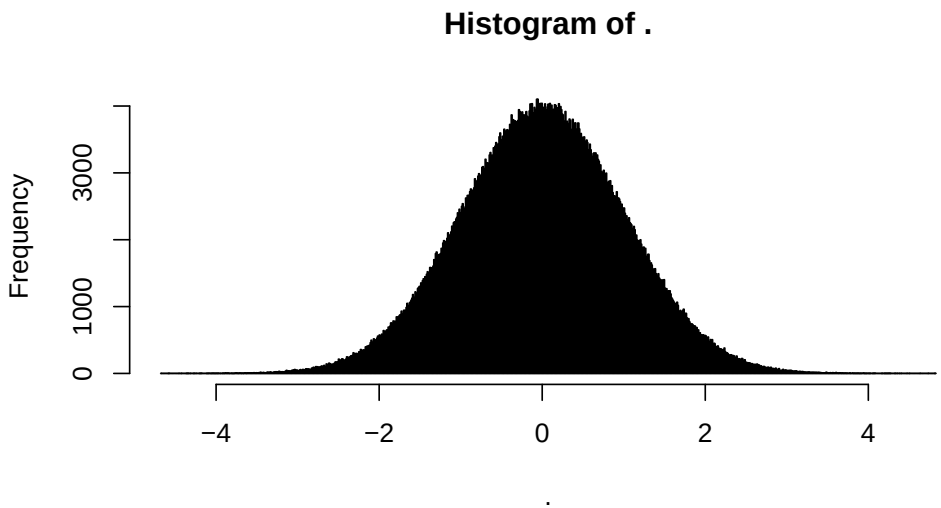


Demonstration in R

The function `rnorm` draws "randomly" from the normal distribution:

```
# notice the output is different
rnorm(1)
## [1] 1.370958
rnorm(1)
## [1] -0.5646982
rnorm(1)
## [1] 0.3631284
```

```
rnorm(1000000) %>%
  hist(breaks = 1000)
```



Demonstration in R

```r
set.seed(123) #always get the same result
i <- 1000000 # number of random draws
logical_vec <- rnorm(i) > 2 # TRUE if the random draw is bigger than 2
n <- logical_vec %>% sum() # sum all the TRUEs
n/i*100 # compute the shares
## [1] 2.2581
```

Section 5

Annex

Bibliography

Gottfredson, Linda S. 2009. "Logical Fallacies Used to Dismiss the Evidence on Intelligence Testing."