Neparametricke testy

S Vondracek

2022-11-09

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Cílem tohoto projektu je pomocí vhodného neparametrického testu ověřit, zda je bakalářský obor Energetika obtížnější, než obor Strojní inženýrství. Testování bude probíhat poměrně jednoduchým způsobem, který jsem zvolil pouze pro účely tohoto úkolu, a nelze jej považovat za příliš sofistikovaný

Naše hypotéza stojí na předpokladu, že oba obory navštěvují studenti se stejnými předpoklady pro skládání zkoušek. Porovnáním známek v obou skupinách tak můžeme zodpovědět výzkumnou otázku, zda se známky napříč obory liší.

# data  
data\_stud <- as\_tibble(read\_csv("./data.csv")) %>% map\_df(~round(., 2))

## Rows: 20 Columns: 2  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (2): m\_eng, energ  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data\_stud

## # A tibble: 20 x 2  
## m\_eng energ  
## <dbl> <dbl>  
## 1 2.2 2.09  
## 2 1.46 2.19  
## 3 2.01 2.13  
## 4 2.3 2.57  
## 5 2.19 2.56  
## 6 2.43 1.75  
## 7 2.16 2.49  
## 8 2.28 2.68  
## 9 1.57 2.08  
## 10 2.79 2.26  
## 11 1.64 3.57  
## 12 2.01 2.82  
## 13 1.75 3.13  
## 14 2.39 NA   
## 15 2.04 NA   
## 16 2.24 NA   
## 17 2.43 NA   
## 18 2.04 NA   
## 19 1.86 NA   
## 20 2.06 NA

summary(data\_stud)

## m\_eng energ   
## Min. :1.460 Min. :1.750   
## 1st Qu.:1.972 1st Qu.:2.130   
## Median :2.110 Median :2.490   
## Mean :2.092 Mean :2.486   
## 3rd Qu.:2.285 3rd Qu.:2.680   
## Max. :2.790 Max. :3.570   
## NA's :7

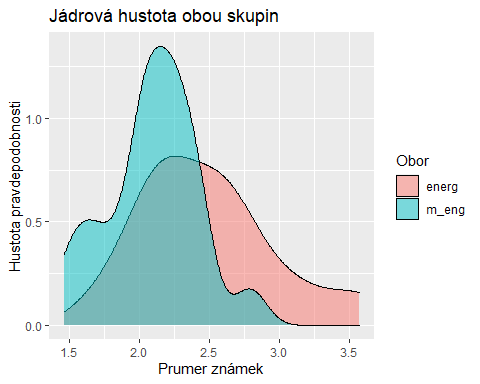
map\_df(data\_stud, ~var(., na.rm=TRUE))

## # A tibble: 1 x 2  
## m\_eng energ  
## <dbl> <dbl>  
## 1 0.105 0.240

Vidíme, že studenti Energetiky mají větší variabilitu co se týče známek

data\_stud %>%   
 mutate(id = 1:nrow(data\_stud)) %>%   
 pivot\_longer(cols = -id) %>%   
 ggplot(data = .,  
 aes(x = value, fill=name)) +   
 geom\_density(alpha = 0.5) +  
 ggtitle("Jádrová hustota obou skupin") +  
 scale\_x\_continuous("Průměr známek") +  
 scale\_y\_continuous("Hustota pravděpodobnosti") +  
 labs(fill = "Obor")

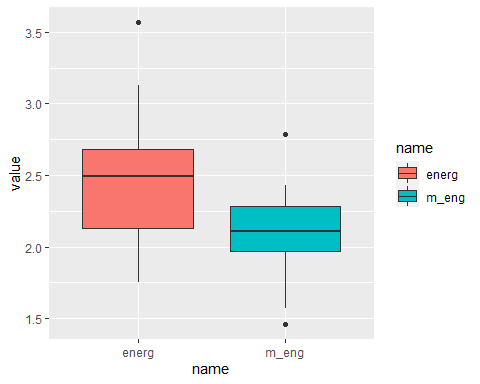
## Warning: Removed 7 rows containing non-finite values (stat\_density).



Pro porovnání obou skupin studentů bude vhodné vykrelit jejich jádrovou hustotu. Vidíme, že obor energetika se vyznačuje tím, že studenti mají variabilnější průměr svých známek.

data\_stud %>%   
 mutate(id = 1:nrow(data\_stud)) %>%   
 pivot\_longer(cols = -id) %>%   
 ggplot(data = .,  
 aes(y = value, x=name, fill=name)) +   
 geom\_boxplot()

## Warning: Removed 7 rows containing non-finite values (stat\_boxplot).



Rovněž boxplot, nikoliv překvapivě, podporuje naše zjištění.

## Neparametrické testy

wilcox.test(data\_stud$m\_eng, data\_stud$energ, alternative="two.sided")

## Warning in wilcox.test.default(data\_stud$m\_eng, data\_stud$energ, alternative =  
## "two.sided"): cannot compute exact p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: data\_stud$m\_eng and data\_stud$energ  
## W = 65, p-value = 0.01743  
## alternative hypothesis: true location shift is not equal to 0