## **Databáze obsahuje informace o studijních výsledcích studentů geografických oborů na PrF UK studujících tam první ročník v roce 2004.**

## **Proměnné:**

* **Obor**: obor studia
  + FYZG (fyzická geografie),
  + KARTG (kartografie),
  + REGG (regionální geografie),
  + SOCG (sociální geografie)
* **Pohlaví**: pohlaví studenta - m (muž), z (žena)
* **Celprij**: celkový počet bodů u přijímaček
* **Zemprij**: počet bodů u přijímaček ze zeměpisu
* **Matprij**: počet bodů u přijímaček z matematiky
* **TypM**: typ zadání přijímaček z matematiky
* **Matzem**: známka z prvního ročníku matematické geografie
* **Meteo**: známka z prvního ročníku meteorologie
* **Geol**: známka z prvního ročníku geologie
* **Mat**: známka z prvního ročníku matematika
* **Stat**: známka z prvního ročníku statistika
* **Matur**.drive: informace, jestli student maturoval dříve než v roce nastupu na VS
* **Ss2**: průměrná známka z 2. ročníku SS
* **Ss3**: průměrná známka z 3. ročníku SS

## **How to okecat ( Alexovo svaté patero )**

## kde jsi to našel

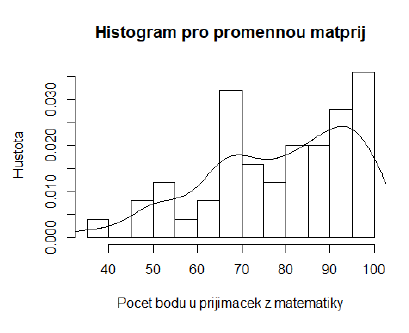
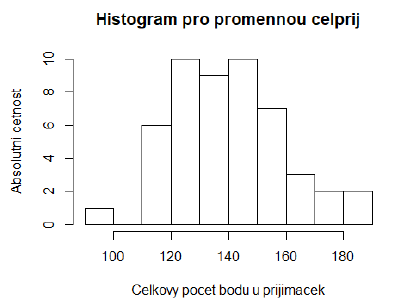
## jak se to jmenuje

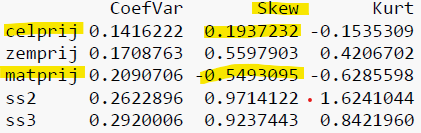
## jaká je hodnota

## co nám to říká

## co to znamená

## **1. Popište a porovnejte zešikmení proměnných celrpij a matprij**





* celprij
  + šikmost pravostranná, ovšem velmi nízká (viz. skew 0.19)
  + špičatost velmi nízká (viz. kurt -0.154)
* matprij
  + šikmost levostranná, výraznější (viz. skew -0.55)
  + špičatost je výrazně negativní (viz. kurt -0.629)

Z histogramu a hodnot zešikmení můžeme odvodit, že proměnná celprij reprezentující celkový počet bodů u přijímaček, má sice pravostranné zešikmení, ovšem velmi nízké. Můžeme o této proměnné říct, že se blíží normální rozdělení. Naproti tomu(tahle dvě kouzelná slovíčka se počítají jako porovnání) proměnná matprij reprezentující počet bodů jen z matematiky má levostrannou šikmost (koncentrace dat je vpravo).

## 

## **2. Jak velký je interval, kde se nachází prostřední polovina dat ss3**

## 

## **🙂**

## 

Interval, kde se nachází prostřední polovina dat ss3 je velký 0.62

Tuto hodnotu jsem zjistil z IQR - to je hodnota rozdílu 3. a 1. kvartilu, což je hodnota kde se nachází prostřední polovida dat ss3.

Hodnotu IQR najdeme ve výstupu “vystup.num”

## 

## 

## **3. Kde byste hledali modus/modusy proměnné zemprij? kolik jich má a kde leží**

## 

## 

## 

## 

## 

Z grafu můžeme vyčíst, že v proměnné zemprij můžeme nalézt modus. Jedná se o sloupec, který je nejvyšší, určuje nám interval o největší četnosti měření.

Tento modus se nachází v intervalu 60, 65. Pokud bychom uvažovali o proměnné, že je spojitá, vypočítali bychom přesný modus podle této rovnice:

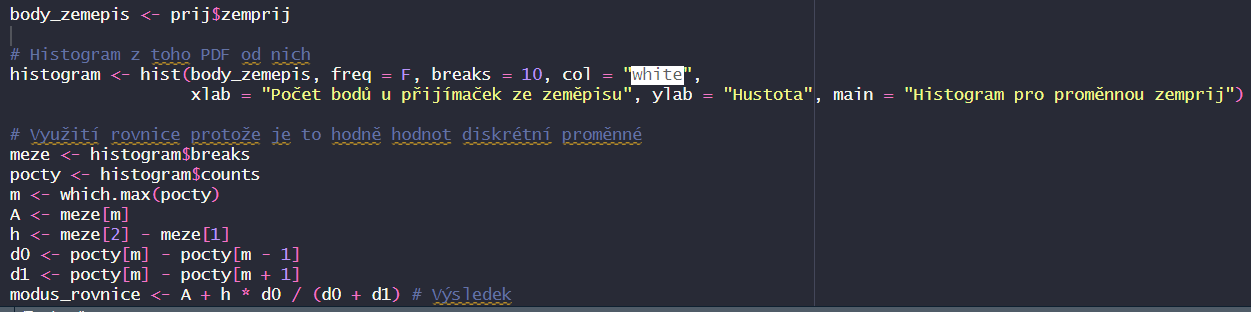
A + h\*d0/(d0+d1)

Po uplatnění této rovnice, nám modus vyšel 61,5.

**Přesná hodnota Modusu - how to (na to se neptají v týhle otázce✨)**

"Přeš DescTools() pokud je málo a podobných hodnot. " - x.val[which.max(ac)]

"Přeš Rovnici pokud je hodně různých hodnot " - rovnice

"Přeš Mode() - nejpočetnější hodnota." - Mode(prom)

## 

## **4. Existuje nějaký vztah mezi oborem studia a pohlavím? Volí studenti stejné obory v závislosti na pohlaví, nebo jsou mezi muži a ženami rozdíly? Pokud si myslíte, že jsou rozdíly, zkuste je popsat vlastními slovy.**

## 

## 

## 

## 

## 

## 

## 

FYZG (fyzická geografie),

KARTG (kartografie),

REGG (regionální geografie),

SOCG (sociální geografie)

Z grafu “Složení oboru podle Pohlaví”, popisující relativní četnost zastoupení pohlaví u jednotlivých oborů, můžeme vidět mírnou závislost volby oboru na pohlaví studenta. Největší rozdíly nalezneme u oboru sociální geografie (SOCG), kde z grafu vyplývá převažující počet žen. Naproti tomu, obor kartografie (KARTG) studují převážně muži. Domnívám se, že sociální obory jsou více oblíbené u ženského pohlaví, naproti tomu více technické obory jsou primárně výběrem mužů (ze zkušenosti studia IT).

Nechci tímto nikoho diskriminovat, pouze analyzuji data.

(Scatter-plot se tady nepoužije, protože je na číselné spojité proměnné.)

## **5. Existuje souvislost mezi známkami ze třetích ročníků SS a celkovým počtem bodů u přijímaček (proměnné ss3 a celprij)? Pripadnou závislost popište: je silná, nebo slabá a přímá, nebo nepřímá? O co se opíráte?**

nepřímá úměra (záporná) - čím víc, tím míň

přímá úměra (kladná) - čím víc, tím víc

lineární, exp, log…

-0.3

čím víc bodů - tím menší známka

## 

## 

## 

V korelační matici jsem našel hodnotu -0.30. Tato hodnota značí slabou zápornou závislost, to v tomto případě znamená, že pokud měl student více bodů u přijímacích zkoušek tak měl “nižší” známku (nižší znamená lepší). Z matice bodových grafů jde vidět náznak tvaru kvadratické nebo exponenciální závislosti

## **6. Porovnejte známky z matematické geografie a z geologie (proměnné matzem a geol) pomocí jejich tří kvartilů. Liší se od sebe kvartily těchto proměnných?**

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

Mezikvartilové rozpětí (IQR) je u obou proměnných stejný - tzn. u obou případů je rovno 1.

Ovšem, rozdíl mezi proměnnými je takový, že Q1 = 2 a Q3 = 3 u proměnné geol a u proměnné matzem je Q1 = 1, Q3 = 2. Po vizualizaci boxplot grafu si můžeme všimnout, že jsou si grafy vůči sobě opačné. Velikost 1. a 3. kvartilu je u proměnné matzem nulový. Velikost 2. a 4. kvartilu je u proměnné geol nulový.

Pomocí kvartilů tak můžeme vypozorovat, že známky z geol, kde 75% počtu známek jsou v intervalu 2, 3, jsou výrazně horší než známky z matzem, kde 75% počtu známek jsou v intervalu 1, 2.

## 

## **1. Je možné považovat proměnnou ss3 za normálně rozdělenou? K posouzení normality použijte vhodné grafy a případné odchylky od normality popište.**

## 

## 

## 

## 

## 

## 

*Velmi špičatá - velký kurt, v qq plot mnoho hodnot se skupí na jednom místě kolem nuly, normální by to mělo rovnoměrně*

*Velký ocas vpravo (šikmost) - vysoká hodnoty v qqplot na konci, v histogramu jde vidět taky, vysoký Skew blíží se hodnotě jedna*

*Kdyby to bylo normální rozdělení je Skew a Kurt blízko nule, připadně by byly viditelné silné outliery které zkreslují - tady nejsou, má extrémně asmyterické rozdělení*

Z histogramu pro proměnnou ss3 a qqplotu pro proměnnou ss3 můžeme odvodit, že rozdělení určitě není normální. Na histogramu si můžeme všimnout velmi silné šikmosti, která se propisuje i do qqplotu tím, že graf qqplotu není přímkou, ale je “prohnutý. Z grafu je viditelná i mírná špičatost, kterou potvrzuje qqplot tím, že je velká hustota dat uprostřed grafu.

## 

## **2. Spočtěte a interpretujte 95%-ní interval spolehlivosti pro celkový počet bodů u přijímaček (proměnná celprij).**

U celprij není pevně dán rozptyl - nutno vypočítat.

Příkaz do R:

# Neznámý rozptyl

mn <- mean(prij$celprij) # Výpočet průměru

sd\_unknown <- sd(prij$celprij) # Odhad standardní odchylky z dat

n <- length(prij$celprij) # Počet vzorků

alpha <- 0.05 # Hladina významnost (1 - interval, který hledáme)i

# Výpočet kvantilu Studentova t-rozdělení pro zadanou hladinu významnosti a stupně volnosti

q\_t <- qt(1 - alpha / 2, n - 1)

# Dolní mez intervalu spolehlivosti

(lower\_unknown <- mn - q\_t \* sd\_unknown / sqrt(n))  
(upper\_unknown <- mn + q\_t \* sd\_unknown / sqrt(n))

Nebo takhle:

MeanCI(prij$celpri, method="classic", conf.level = 0.95, sides="two.sided")

95% - ní interval spolehlivosti pro průměr proměnné celprij je: 135,027 - 146,3526.

95% -ní interval spolehlivosti pro medián proměnné celprij je: 134,477 - 145,523

Tento interval nám říká, že skutečný průměr a medián celkového počtu bodů leží se spolehlivotí 95% právě v tomto intervalu.

**Pokud chceme interval spolehlivosti pro celé rozložení dat (odhad horní a dolní meze hodnot), využijeme k tomu kvantily dat.**

**data <- prij$celprij**

**# Počet vzorků**

**n <- length(data)**

**# Kvantily pro dolní a horní mez (např. 2.5% a 97.5% pro 95% interval)**

**lower\_quantile <- quantile(data, 0.025)**

**upper\_quantile <- quantile(data, 0.975)**

**Tímto nám vznikne interval od 111,45 do 183,875. Díky tomuto intervalu můžeme se spolehlivostí 95% říct, že kdybychom měření provedli znovu, hodnoty budou ležet v tomto interval**

##

Pozn:

Interval spolehlivosti nebyl specifikován, na jaký parametr se váže (průměr, medián, smodch, …) to znamená, že u testu je lepší se zeptat, co tím autor vlastně zamýšlel za otázku. Kdyby ani Loukotová/Černíková nevěděla, udělej to podle sebe, asi tím, co ti je nejpříjemnější 🙂

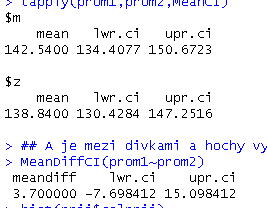
## **3. Liší se výsledky u přijímaček mezi pohlavími? A pokud ano, je rozdíl statisticky významný? Pracujte s proměnnými celprij a Pohlavi**

prom1 <- prij$celprij

prom2 <- prij$Pohlavi

tapply(prom1,prom2,MeanCI)

MeanDiffCI(prom1~prom2)



Můžeme si všimnout, že průměry celkového počtu bodů u mužů a u žen se liší, ovšem po vypočítání rozdílu intervalu spolehlivosti pro průměr 95%, nám vznikne interval -7,698 - 15.098. V tomto intervalu se nachází hodnota 0 - to znamená, že se výsledky výrazně statisticky neliší.

Pepovo úžasný how to:

1. Napsat ten kód co tu je pro dvě dané proměnné
2. Ukáže se tam interval (lwr.ci - upr.ci)
3. Pokud v tom intervalu je nula = není to významný rozdíl (je možné že budou mít stejný průměr)
4. Pokud v tom intervalu není nula = je to významný rozdíl (není šance že budou mít stejný průměr)

Poznámky:

1. MeanCI = interval odchylky od průměru
2. MeanDiffCI(prom1~prom2) to udělá rozdíl těch intervalů - vznikne ten interval co checkuješ o nulu

## 

## **Pravděpodobnost**

## U úloh z této části stačí uvést pouze výsledek. Výsledek vždy zaokrouhlete na 3 desetinná místa.

## **1. Klíčivost semen určitého druhu konopí je 93 %.**

## **(a) Jaká je pravděpodobnost, že z 15 semen vám vyklíčí právě 12 semen?**

-> pravděpodobnostní

-> binomické

dbinom(12,15,0.93)

## **(b) Jaká je pravděpodobnost, že z 10 semen vám vyklíčí 8 a více semen?**

-> distribuční

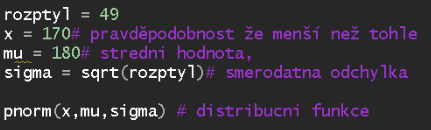
-> binomické

(1 - pbinom(7,10,0.93))

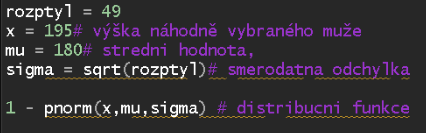
## 

## **2. Výška mužů má normální rozdělení se střední hodnotou 180 cm a rozptylem 49 cm2.**

## **(a) Jaká je pravděpodobnost, že náhodně vybraný muž je menší než 170 cm?**



## **(b) Jaká je pravděpodobnost, že náhodně vybraný muž bude vyšší než průměrný hráč NBA, jehož výška je 195 cm?**

****

## **(c) Na ulici potkáte v krátkém sledu dva muže. Jaká je pravděpodobnost, že první z mužů je vyšší než 195 cm a zároveň druhý z mužů je nižší než 170 cm? Předpokládejte, že jste nepotkali dvakrát stejného člověka.**

> (vyssi\_nez\_195) \* (nizsii\_nez\_170)  
(Je to pravidlo šance dvou nezávislých eventů - event A \* event B)