CheckPoint3

Rodolfo Viana

06-05-2015

```
library(dplyr)
library(ggplot2)
library(Hmisc)

file <- read.table("students.data", header=TRUE)
```

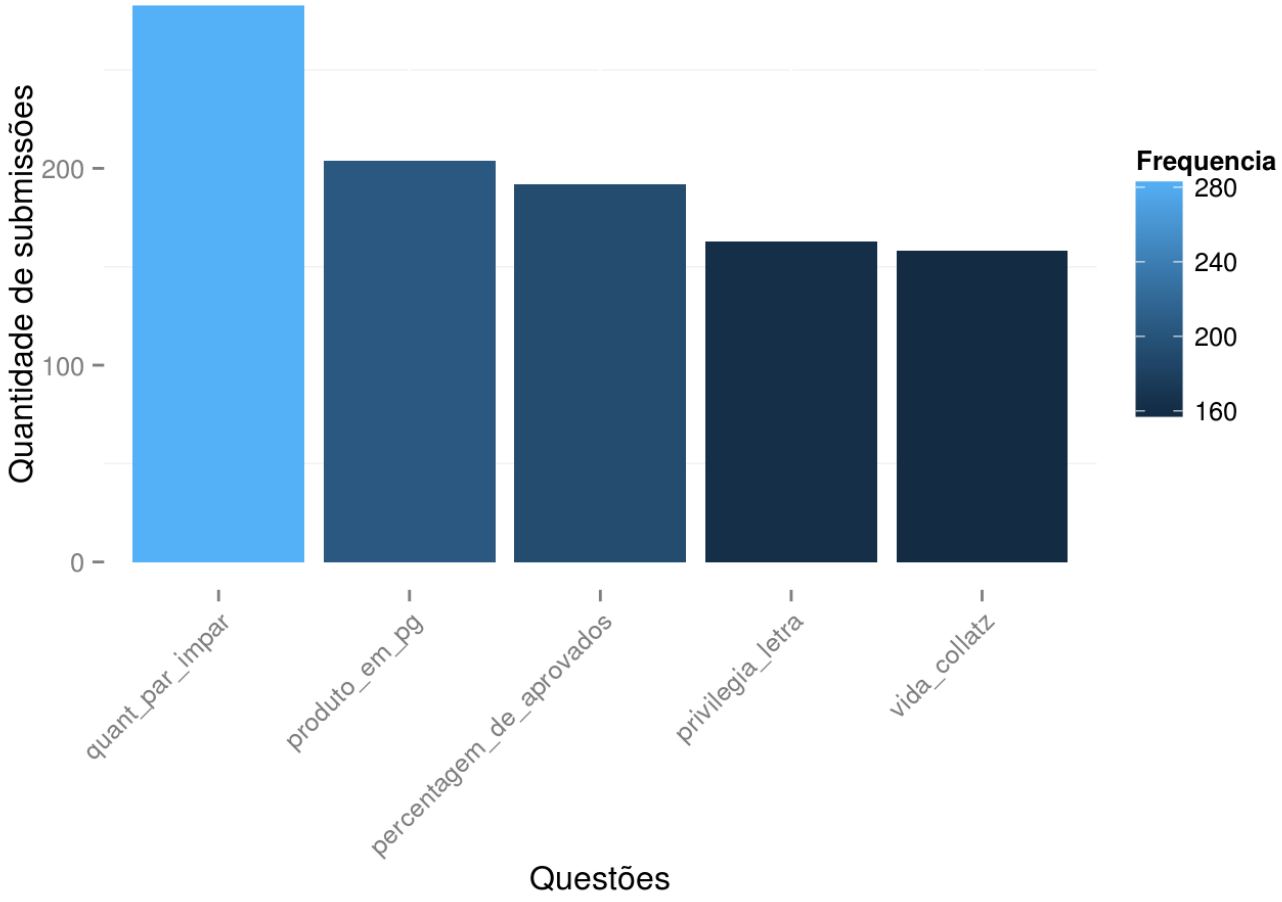
Em busca de saber o comportamento dos alunos de Programação 1 da Universidade Federal de Campina Grande, foram feitas analises nas submissões das atividades que os alunos submetem ao longo do período. Como a população dos alunos de programação 1 é muito grande, utilizei uma amostra para fazer inferências em relação aos alunos de programação 1 e as atividades. Para este estudo, coletamos as submissões dos alunos de Programação I da Universidade Federal de Campina Grande (UFCG) durante o período letivo 2014.2. No total, há 15148 submissões de 101 alunos para 427 questões.

Ao longo da disciplina, cada aluno dita o seu ritimo. O aluno resolve questões e a medida que vai acertando vai subindo de nível e recebendo mais questões. A correção das atividade é feita de forma automática. Após a submissão de uma atividade o aluno é informado se a resposta enviada é a correta ou não, caso não seja a resposta esperada, o aluno pode reenviar quantas vezes achar necessário.

Por essa razão, existem questões que possuem um grande número de submissões. Então achei importante descobrir a média de submissões das 5 questões com maior número de submissões.

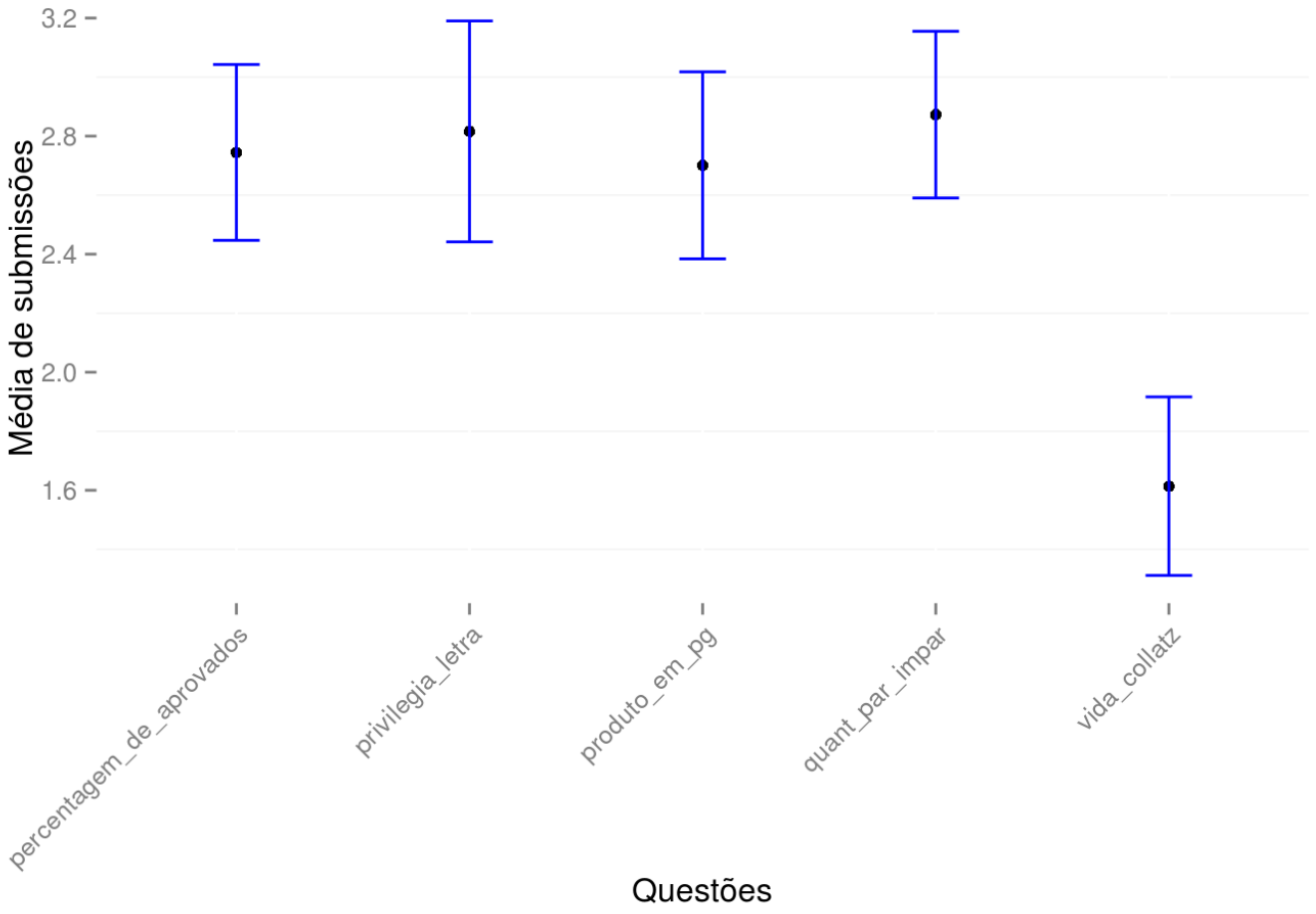
Às questões que possuim mais subsmissões são:

```
mediaQuestoes <- file %>%
  select(question) %>%
  table() %>%
  as.data.frame()

colnames(mediaQuestoes) <- c("Questao", "Frequencia")
newdata <- mediaQuestoes[order(-mediaQuestoes$Frequencia),]
newdata <- head(newdata, n = 5)

ggplot(newdata, aes(x=reorder(Questao, -Frequencia), y=Frequencia, fill = Frequ
encia)) +
  geom_bar(stat="identity") +
  labs(y='Quantidade de submissões', x='Questões') +
   theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.background=el
ement_blank())
```



Após descobrir às 5 questões que possuem o maior número de submissões, foi calculado a média do número de submissões da amostra para essas 5 questões. Com 95% de precissão temos o seguinte resultado.

```
condition <- newdata$Questao

subset <- file %>%
  filter(question %in% condition)

ggplot(subset, aes(x = question, y = attempt)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", colour = "blue", w
idth = 0.2) +
  labs(y='Média de submissões', x='Questões') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.background=ele
ment_blank())
```



Devido ao intervalo de confiança não podemos afimar qual questão que possui a maior média no número de submissões. Podemos apenas afimar que a questão **vida_collatz** é a que possui a menor média dentre as 5 primeiras.

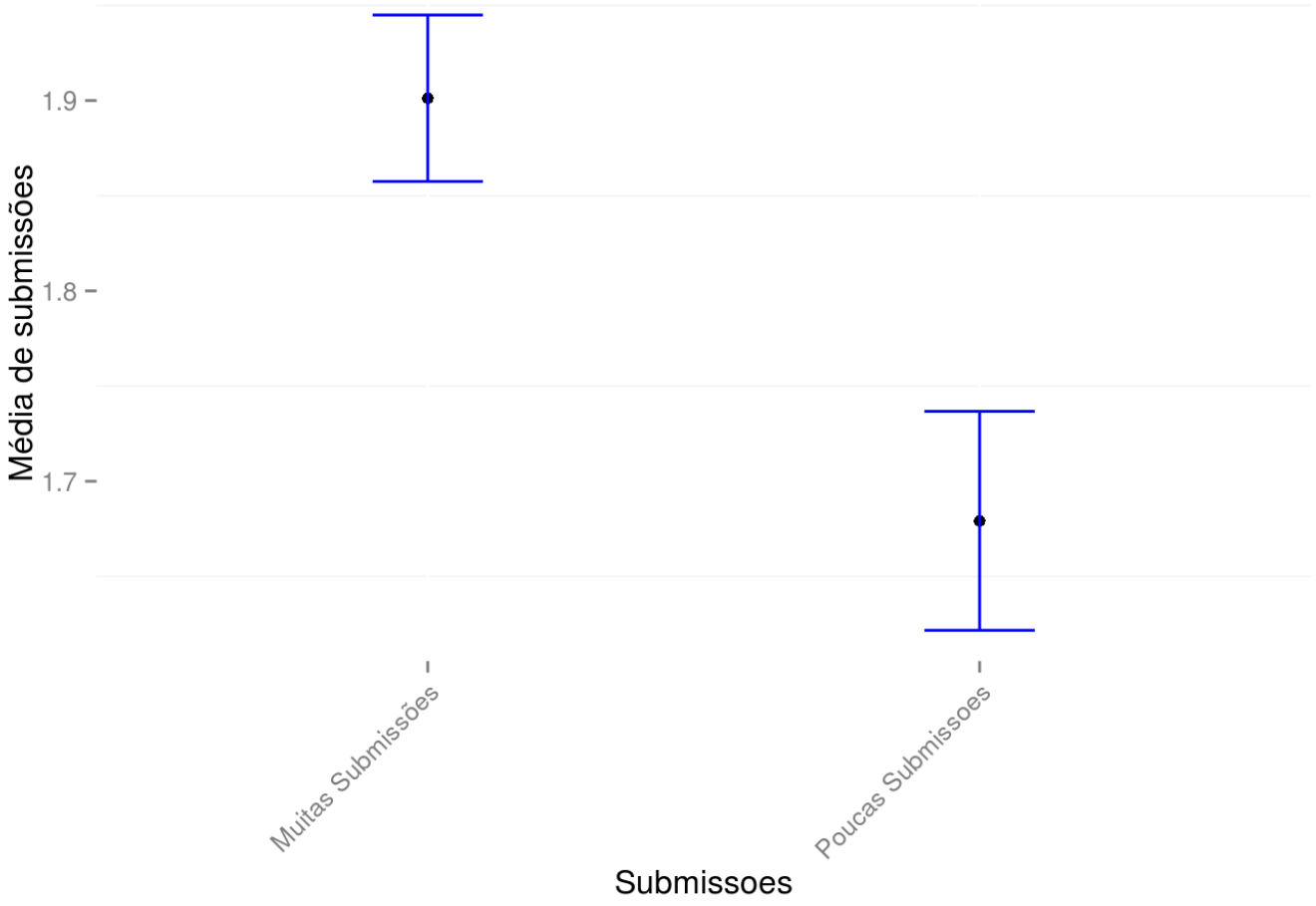
Foi observado também que existe uma grande variância no número de submissões por aluno. Para efeito de comparação dividimos os alunos em 2 grupos: Um grupo com poucas submissões (Menos de 150 submissões durante todo o período) um grupo com muitas submissões (Mais de 150 submissões durante todo o período)

O valor 150 não foi escolhido arbritariamente, ele representa o valor médio de submissões da nossa amostra.

Como os alunos vão passando de níveis ao longo da disciplina, achei importante saber (em média) quantas submissões são necessárias até que o aluno acerte a questão.

Foi então calculado a média do número de acertos da amostra para os dois grupos de alunos. Com 95% de precissão temos o seguinte resultado.

```
alunos <- file %>%
  select(student) %>%
  table() %>%
  as.data.frame()

colnames(alunos) <- c("Aluno", "Submissoes")
alunosOrder <- alunos[order(-alunos$Submissoes),]

condicao1 = head(alunosOrder, n=39)$Aluno
condicao2 = tail(alunosOrder, n=62)$Aluno

muitasSubmissoes <- file %>%
  filter(student %in% condicao1)

poucasSubmissoes <- file %>%
  filter(student %in% condicao2)

muitasSubmissoes["Submissoes"] <- "Muitas Submissões"
poucasSubmissoes["Submissoes"] <- "Poucas Submissoes"

fileSubmissoes <- union(muitasSubmissoes, poucasSubmissoes) %>%
  filter(result == "True")

ggplot(fileSubmissoes, aes(x = Submissoes, y = attempt)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", colour = "blue", w
idth = 0.2) +
  labs(y='Média de submissões') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.background=ele
ment_blank())
```



É possível notar que os alunos que possuem muitas submissões, demoram mais submissões para responder a questão corretamente. O que significa dizer que, os alunos que tem muitas submissões não necessariamente responderam muitas questões.