Relação entre SAT e GPA

dados = read\_csv(here::here("data/sat-gpa.csv"),   
 col\_types = "dddd")  
  
glimpse(dados)

## Rows: 105  
## Columns: 4  
## $ math\_SAT <dbl> 643, 558, 583, 685, 592, 562, 573, 559, 552, 617, 684, 568, 6~  
## $ verb\_SAT <dbl> 589, 512, 503, 602, 538, 486, 548, 536, 583, 591, 649, 592, 5~  
## $ comp\_GPA <dbl> 3.76, 2.87, 2.54, 3.83, 3.29, 2.64, 2.86, 2.03, 2.81, 3.41, 3~  
## $ univ\_GPA <dbl> 3.52, 2.91, 2.40, 3.47, 3.47, 2.37, 2.40, 2.24, 3.02, 3.32, 3~

L4P1: Exercícios com inferência

1: Relação entre SAT e GPA Nesse cenário, usaremos dados de resultados de duas avaliações feitas em alunos que se formaram em Computação em uma universidade pública americana. O primeiro é o SAT: “The SAT exam is a standardized test that most U.S. colleges use to determine whether students should be accepted into the institution.” e nossos dados tem a nota em matemática (math\_SAT) e expressão verbal (verb\_SAT) no SAT dos estudantes. A segunda nota disponível é a GPA dos alunos, que você pode interpretar como um agregado do desempenho deles no curso. Temos o GPA deles para as matérias de computação (comp\_GPA) e no geral (univ\_GPA). Temos os dados de uma amostra de alunos de computação a partir da qual queremos generalizar afirmações para alunos de Computação em geral (ao menos dessa universidade). Nossas perguntas são: Como é a correlação entre o desempenho dos alunos em cada uma das dimensões do SAT (verbal e matemática) e seu desempenho nas matérias de computação do curso? Como se comparam entre si as correlações dos dois critérios do SAT com o desempenho nas matérias de Computação? Um dos critérios é claramente mais correlacionado com o desempenho nas matérias de computação que o outro? Os resultados apontam que essa diferença é grande? Pequena? O dados estão aqui. Esse notebook tem código que pode lhe ajudar. Faça clone do repositório e experimente com ele. As visualizações de ICs dele são um bom modelo para lhe inspirar.

kkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkkk

1: Relação entre SAT e GPA Nesse cenário, usaremos dados de resultados de duas avaliações feitas em alunos que se formaram em Computação em uma universidade pública americana. O primeiro é o SAT: “O exame SAT é um teste padronizado que a maioria das faculdades dos EUA usa para determinar se os alunos devem ser aceitos na instituição.” e nossos dados tem a nota em matemática (math\_SAT) e expressão verbal (verb\_SAT) no SAT dos estudantes. A segunda nota disponível é a GPA dos alunos, que você pode interpretar como um agregado do desempenho deles no curso. Temos o GPA deles para as matérias de computação (comp\_GPA) e no geral (univ\_GPA). Temos os dados de uma amostra de alunos de computação a partir da qual queremos generalizar afirmações para alunos de Computação em geral (ao menos dessa universidade). Nossas perguntas são: Como é a correlação entre o desempenho dos alunos em cada uma das dimensões do SAT (verbal e matemática) e seu desempenho nas matérias de computação do curso? Como se comparam entre si as correlações dos dois critérios do SAT com o desempenho nas matérias de Computação? Um dos critérios é claramente mais correlacionado com o desempenho nas matérias de computação que o outro? Os resultados apontam que essa diferença é grande? Pequena? O dados estão aqui. Esse notebook tem código que pode lhe ajudar. Faça clone do repositório e experimente com ele. As visualizações de ICs dele são um bom modelo para lhe inspirar.

dados\_math = dados %>%  
 mutate (materia\_SAT = "math\_SAT", nota\_SAT = math\_SAT, computacao\_GPA = comp\_GPA)  
  
dados\_verb = dados %>%  
 mutate (materia\_SAT = "verb\_SAT", nota\_SAT = verb\_SAT, computacao\_GPA = comp\_GPA)  
  
  
dados\_finais = bind\_rows (dados\_math, dados\_verb) %>%   
 select(materia\_SAT, nota\_SAT, computacao\_GPA)

#dados %>%  
# select(math\_SAT, verb\_SAT) %>%  
# pivot\_longer(c(math\_SAT, verb\_SAT))

Visualizações das informações

total\_linhas = dados %>% count()  
  
dados\_medias = dados %>%   
 summarise(media\_math\_SAT = mean(math\_SAT),   
 media\_verb\_SAT = mean(verb\_SAT),   
 media\_comp\_GPA = mean(comp\_GPA),   
 .groups = "drop")  
  
dados\_variancia = dados %>%   
 summarise(variancia\_math\_SAT = var(math\_SAT),   
 variancia\_verb\_SAT = var(verb\_SAT),   
 variancia\_comp\_GPA = var(comp\_GPA),   
 .groups = "drop")  
  
dados\_desvio\_padrao = dados %>%   
 summarise(desvio\_padrao\_math\_SAT = sd(math\_SAT),   
 desvio\_padrao\_verb\_SAT = sd(verb\_SAT),   
 desvio\_padrao\_comp\_GPA =   
 sd(comp\_GPA), .groups = "drop")  
  
dados\_coef\_var = dados %>%   
 summarise(coef\_var\_math\_SAT = sd(math\_SAT) / mean(math\_SAT) \* 100,  
 coef\_var\_verb\_SAT = sd(verb\_SAT) / mean(verb\_SAT) \* 100,   
 coef\_var\_comp\_GPA = sd(comp\_GPA) / mean(comp\_GPA) \* 100,  
 .groups = "drop")  
  
  
dados\_max = dados %>%   
 summarise(max\_math\_SAT = max(math\_SAT),   
 max\_verb\_SAT = max(verb\_SAT),   
 max\_comp\_GPA = max(comp\_GPA),   
 .groups = "drop")  
  
dados\_min = dados %>%   
 summarise(min\_math\_SAT = min(math\_SAT),   
 min\_verb\_SAT = min(verb\_SAT),   
 min\_comp\_GPA = min(comp\_GPA),   
 .groups = "drop")  
  
dados\_assimetria = dados %>%   
 summarise(assimetria\_math\_SAT = skewness(dados$math\_SAT),   
 assimetria\_verb\_SAT = skewness(dados$verb\_SAT),   
 assimetria\_comp\_GPA = skewness(dados$comp\_GPA),   
 .groups = "drop")  
  
#dados\_medias  
#dados\_variancia  
#dados\_desvio\_padrao  
#dados\_coef\_var   
dados\_min

## # A tibble: 1 x 3  
## min\_math\_SAT min\_verb\_SAT min\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 516 480 2.03

dados\_max

## # A tibble: 1 x 3  
## max\_math\_SAT max\_verb\_SAT max\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 718 732 4

dados\_assimetria

## # A tibble: 1 x 3  
## assimetria\_math\_SAT assimetria\_verb\_SAT assimetria\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 0.214 0.184 -0.674

A média para math\_SAT foi igual a 623.0762. A méida para verb\_SAT foi igual 598.6. E a média para comp\_GPA foi de 3.128.

dados\_medias

## # A tibble: 1 x 3  
## media\_math\_SAT media\_verb\_SAT media\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 623. 599. 3.13

Também foram calculados a variância, o desvio padrão e o coeficiente de variação para cada variável como medidas de dispersão.

dados\_variancia

## # A tibble: 1 x 3  
## variancia\_math\_SAT variancia\_verb\_SAT variancia\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 2890. 3964. 0.259

dados\_desvio\_padrao

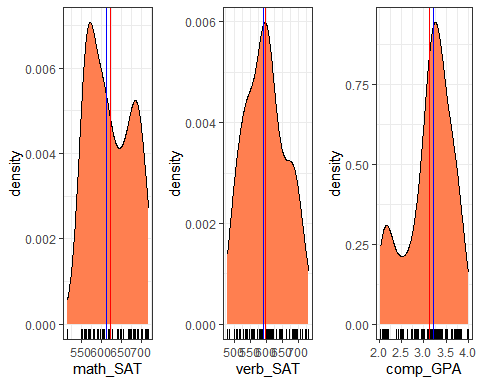
## # A tibble: 1 x 3  
## desvio\_padrao\_math\_SAT desvio\_padrao\_verb\_SAT desvio\_padrao\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 53.8 63.0 0.509

A variância do desvio padrão são as duas medidas de dispensão mais usudas. Neste caso, tanto o a variância quanto o desvio padrão indentificou um variabilidade na dispersão das variáveis em torno da média. Entretanto, como a variância quanto o desvio padrão dependem da escala assumida de cada variável, calculou-se o o coeficiente de variação para poder fazer uma análise comparada da varibilidade. o Coeficente de Variáção é um medida de varibilidade padronizada, ou seja, expressa percentualmente a avalição dos dados em relação a média. De acordo com os coeficinete de variação calculados para as três variáveis, pode-se dizer que a variável math\_SAT mostrou-se menor concentrada relativamente às demais variáveis. Já a variável comp\_GPA mostrou-se relativamente mais concentrada em torno da média do que a demais variáveis.

dados\_coef\_var

## # A tibble: 1 x 3  
## coef\_var\_math\_SAT coef\_var\_verb\_SAT coef\_var\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 8.63 10.5 16.3

p1 = dados %>%   
 ggplot(aes(x = math\_SAT)) +  
 geom\_density(fill = "coral", color = 'black') +  
 geom\_rug() +  
 geom\_vline (xintercept = mean (dados$math\_SAT), color = "red") +  
 geom\_vline (xintercept = median (dados$math\_SAT), color = "blue")   
   
 p2 = dados %>%   
 ggplot(aes(x = verb\_SAT)) +  
 geom\_density(fill = "coral", color = 'black') +  
 geom\_rug() +  
 geom\_vline (xintercept = mean (dados$verb\_SAT), color = "red") +  
 geom\_vline (xintercept = median (dados$verb\_SAT), color = "blue")   
  
 p3 = dados %>%   
 ggplot(aes(x = comp\_GPA)) +  
 geom\_density(fill = "coral", color = 'black') +  
 geom\_rug() +  
 geom\_vline (xintercept = mean (dados$comp\_GPA), color = "red") +  
 geom\_vline (xintercept = median (dados$comp\_GPA), color = "blue")   
   
 grid.arrange(p1, p2, p3, ncol = 3)



Há basicamente três tipos de assimetria: a) assimetria à direita - ocorre quando a moda > mediana > média (a cauda a direita) b) assimetria à esquerda - ocorre quando média < mediana < moda (a cauda a esquerda) c) simetrico - ocorre quando média = mediana = moda

A medidas de posição (média) podem ser influenciada pelo grau de simetria da distribuição dos dados amostrais. Isso pode ocorrer devido à existência de valores discrepantes (outlines) que podem influenciar o valor da média. Uma forma de ter uma idéia da assimetria é comparando as medidas de posição acima mencionadas. No gráfico de densidade para cada variável, é possível observar os valores da média (linha em vermelho) e mediana (linha em azul). A partir do gráfico de densidade, pode-se dizer que a densidade de probabilidade das váriaveis math\_SAT e verb\_SAT indicam ter distribuição assimétrica à direita ou positiva. Já a a variável comp\_GPA mostra-se ter densidade de probabilidade assimêstrica à esquerda ou negativa. Esses resultados para a simetria é confirmado pelo cálculo direto da medidad de assimetria para cada variável descrito abaixo.

dados\_assimetria

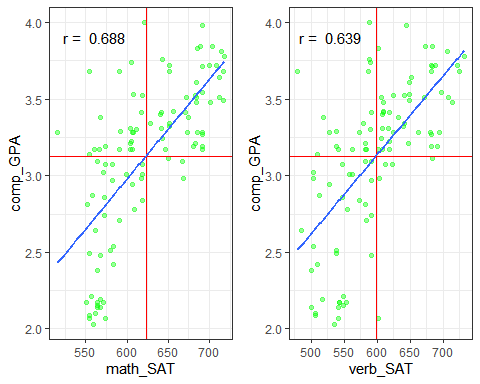
## # A tibble: 1 x 3  
## assimetria\_math\_SAT assimetria\_verb\_SAT assimetria\_comp\_GPA  
## <dbl> <dbl> <dbl>  
## 1 0.214 0.184 -0.674

**Curtose**

A curtose é uma medida de dispersão que caracteriza o “achatamento” da curva da função de distribuição. Quanto maior for a dispersão dos dados em torno da média, ou seja, quanto maior o desvio padrão a função de distribuição é mais “achatada”. O contrário também vale, quanto menor o desvio padrão mais ponteaguda é o formato da função de distribuição de probabilidade. Na ausência de uma medida da curtose, pode-se ter uma indicação do formato relativo das curvas de distribuição das probabilidades. Observavou-se que o desvio padrão da variável verb\_SAT é maior que o desvio padrão math\_SAT. E o desvio padrão comp\_GPA é menor que as outras. Portanto, isto pode que o formato da curva de distribuição de verb\_SAT parece ser relatavidamente mais achatada.

correlacao\_math\_SAT\_comp\_GPA = dados %>%  
 summarise(correlacao = cor(math\_SAT, comp\_GPA, method = "pearson"))%>%  
 pull(correlacao)  
  
correlacao\_verb\_SAT\_comp\_GPA = dados %>%  
 summarise(correlacao = cor(verb\_SAT, comp\_GPA, method = "pearson"))%>%  
 pull(correlacao)  
  
  
p1 = dados %>%   
 ggplot(aes(x = math\_SAT, y = comp\_GPA)) +   
 geom\_point(alpha = .4, color = "green") +  
 geom\_vline (xintercept = mean (dados$math\_SAT), color = "red") +  
 geom\_hline (yintercept = mean (dados$comp\_GPA), color = "red") +  
 geom\_smooth(method=lm, se=FALSE) +  
 annotate(geom = "text", x = 560, y = 3.9, label = paste("r = ", round(correlacao\_math\_SAT\_comp\_GPA, 3)))  
  
  
  
p2 = dados %>%   
 ggplot(aes(x = verb\_SAT, y = comp\_GPA)) +   
 geom\_point(alpha = .4, color = "green") +  
 geom\_vline (xintercept = mean (dados$verb\_SAT), color = "red") +  
 geom\_hline (yintercept = mean (dados$comp\_GPA), color = "red") +  
 geom\_smooth(method=lm, se=FALSE) +  
 annotate(geom = "text", x = 530, y = 3.9, label = paste("r = ", round(correlacao\_verb\_SAT\_comp\_GPA, 3)))  
   
  
  
grid.arrange(p1, p2, ncol = 2)

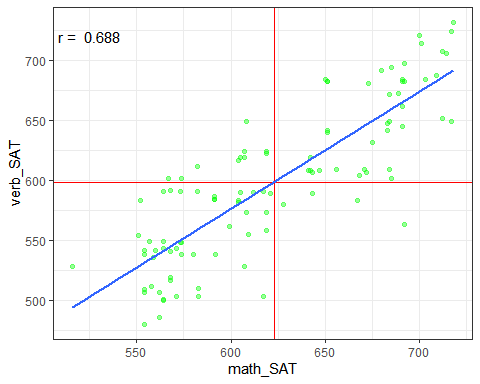
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



1. Para medir para medir a correlação linear entre duas variáveis, utilizou-se utilizou-se o coeficiente de pearson. O coeficiente de Pearson para as variáveis math\_SAT e comp\_GPA foi igual a 0,688. Isso indica que há uma relação linear direta e forte entre essas variáveis. O coeficiente de Pearson para as variáveis verb\_SAT e comp\_GPA foi igual a 0,639.
2. Isso também está indicando que há uma relação linear direta e forte entre essas duas últimas variáveis. O Coeficiente é um número que varia entre -1 e +1, quanto mais próximo de 1 ou -1, indica que há uma correlação forte seja negativa ou positiva. Quanto mais próximo de 0 indica uma correlação fraca entre duas variáveis. Portanto, pode-se infererir que math\_SAT é mais correlacionado com comp\_GPA do que verb\_SAT relativo a comp\_GPA. Entretanto os valores estimados dos coeficiente de relação não se diferenciam muito.

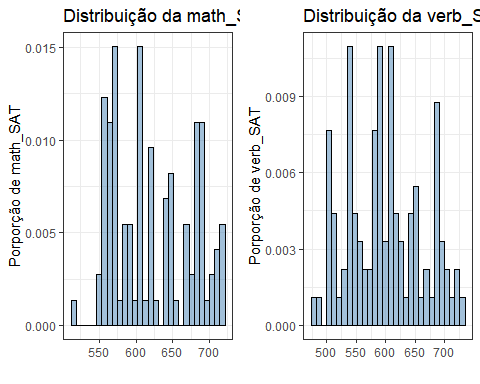
correlacao\_math\_SAT\_comp\_GPA = dados %>%  
 summarise(correlacao = cor(math\_SAT, comp\_GPA, method = "pearson"))%>%  
 pull(correlacao)  
  
 dados %>%   
 ggplot(aes(x = math\_SAT, y = verb\_SAT)) +   
 geom\_point(alpha = .4, color = "green") +  
 geom\_vline (xintercept = mean (dados$math\_SAT), color = "red") +  
 geom\_hline (yintercept = mean (dados$verb\_SAT), color = "red") +  
 geom\_smooth(method=lm, se=FALSE) +  
 annotate(geom = "text", x = 525, y = 720, label = paste("r = ", round(correlacao\_math\_SAT\_comp\_GPA, 3)))

## `geom\_smooth()` using formula 'y ~ x'



p1 = dados %>%   
 ggplot(aes(x = math\_SAT, y = after\_stat(density))) +   
 geom\_histogram(fill = "steelblue", color = "black", alpha = .5) +   
 labs(  
 y = "Porporção de math\_SAT",   
 x = "",  
 title = "Distribuição da math\_SAT"  
 )  
  
p2 = dados %>%   
 ggplot(aes(x = verb\_SAT, y = after\_stat(density))) +   
 geom\_histogram(fill = "steelblue", color = "black", alpha = .5) +   
 labs(  
 y = "Porporção de verb\_SAT",   
 x = "",  
 title = "Distribuição da verb\_SAT"  
 )  
  
grid.arrange(p1, p2, ncol = 2)

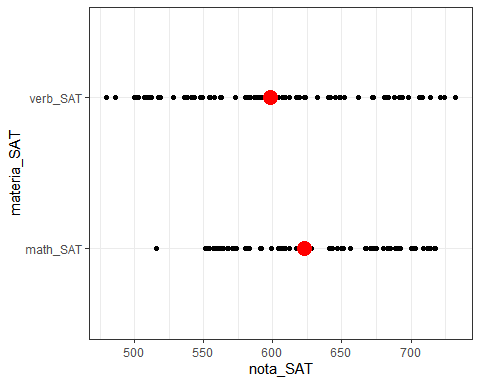
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



A amostra da distribuição mat\_SATH indica maior concetração

comparacao = dados\_finais   
  
comparacao %>%   
 #filter(materia\_SAT == "math\_SAT") %>%  
 ggplot(aes(x = nota\_SAT, y = materia\_SAT)) +   
 geom\_quasirandom(width = .1) +  
 stat\_summary(geom = "point", fun = "mean", color = "red", size = 5)

## Warning in f(...): The default behavior of beeswarm has changed in version  
## 0.6.0. In versions <0.6.0, this plot would have been dodged on the y-axis. In  
## versions >=0.6.0, grouponX=FALSE must be explicitly set to group on y-axis.  
## Please set grouponX=TRUE/FALSE to avoid this warning and ensure proper axis  
## choice.



##### Qual a diferença na amostra

comparacao1 = dados\_finais   
 agrupado = comparacao1 %>%   
 group\_by(materia\_SAT) %>%   
 summarise(media1 = mean(nota\_SAT))  
  
agrupado

## # A tibble: 2 x 2  
## materia\_SAT media1  
## <chr> <dbl>  
## 1 math\_SAT 623.  
## 2 verb\_SAT 599.

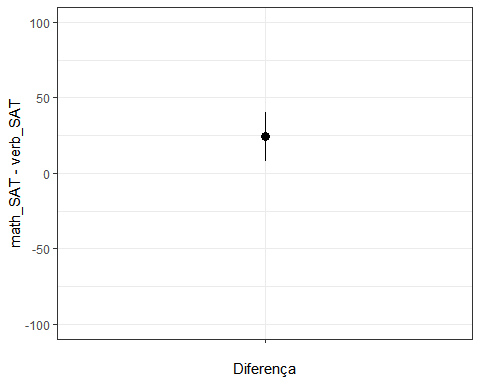
library(boot)  
comparacao1 = dados\_finais  
  
  
  
theta <- function(d, i) {  
 agrupado = d %>%   
 slice(i) %>%   
 group\_by(materia\_SAT) %>%   
 summarise(media = mean(nota\_SAT))  
  
   
 b = agrupado %>% filter(materia\_SAT == "math\_SAT") %>% pull(media)  
 l = agrupado %>% filter(materia\_SAT == "verb\_SAT") %>% pull(media)  
 b - l  
}  
  
theta(comparacao1, i = 1:NROW(comparacao1))

## [1] 24.47619

ci1 = boot(data = comparacao1,  
 statistic = theta,  
 R = 2000) %>%  
 tidy(conf.level = .95,  
 conf.method = "bca",  
 conf.int = TRUE)  
  
ci1

## # A tibble: 1 x 5  
## statistic bias std.error conf.low conf.high  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 24.5 -0.146 8.14 8.08 40.0

ci1 %>%  
 ggplot(aes(  
 x = "",  
 y = statistic,  
 ymin = conf.low,  
 ymax = conf.high  
 )) +  
 geom\_pointrange() +  
 geom\_point(size = 3) +   
 ylim(-100, 100) +   
 labs(x = "Diferença",   
 y = "math\_SAT - verb\_SAT")



Para amostra foi utilizado as médias de math\_SAT - verb\_SAT, com isto, para a realização do desempenho às matérias entre math\_SAT e verb\_SAT observou-se que o resultado a confiança de 95%, em que apontam para um intervalos com conf.low = 8.456478 e 39.20849, podendo ser com pouco efeito sendo pequeno positiva diferença para math\_SAT.