CheckPoint2

Arthur Sena

05/11/2015

```
library(ggplot2, quietly = T, warn.conflicts = F)
library(dplyr, quietly = T, warn.conflicts = F)
```

Descrição dos Dados

Os nossos dados são referentes aos gols do campeonato brasileiro de 2013. Onde, no total, nós temos um conjunto com 973 observações cada uma com 7 colunas/variáveis. Uma pequena amostra dos dados pode ser visualizada logo abaixo:

```
    gols<-read.csv("camp brasileiro 2013 - dados processados copy.csv")
    head(gols)
```

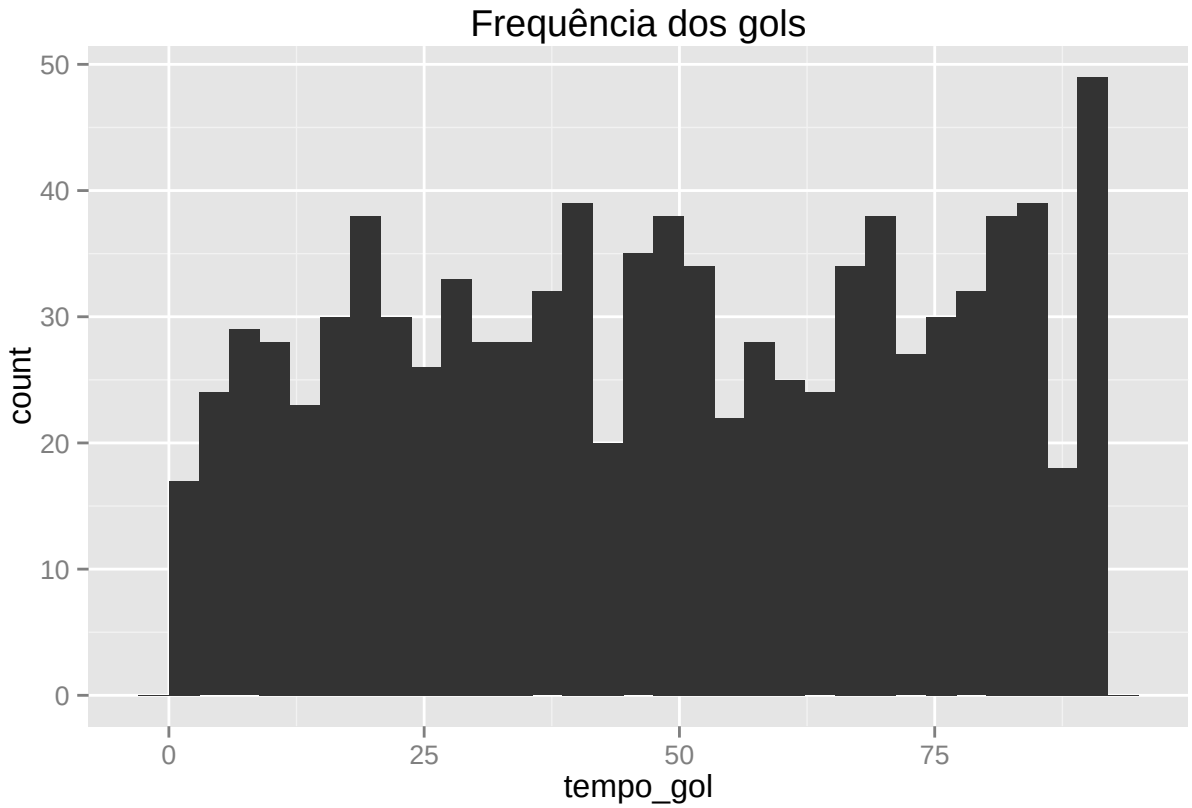
```
##   id_jogo        time_a        time_b tempo_gol      time_gol
## 1       1 Vasco da Gama    Portuguesa        47 Vasco da Gama
## 2       2       Vitória Internacional         2       Vitória
## 3       2       Vitória Internacional        11       Vitória
## 4       2       Vitória Internacional        29 Internacional
## 5       2       Vitória Internacional        63 Internacional
## 6       3   Corinthians      Botafogo        24      Botafogo
##   placar_time_a placar_time_b
## 1             1             0
## 2             1             0
## 3             2             0
## 4             2             1
## 5             2             2
## 6             0             1
```

Os dados acima apresentam uma coluna de "tempo_gol", na qual representa o minuto em que um determinado gol ocorreu naquele jogo. Para efeito de melhor visualização e entendimento dos dados, nós podemos criar um histograma com a frequência de gols em cada tempo do jogo.

```
       gols <- filter(gols,tempo_gol != "None")
       gols$tempo_gol <- as.numeric(gols$tempo_gol)

       tempo1 <- filter(gols,tempo_gol <= 45)
       tempo2 <- filter(gols,tempo_gol > 45)
       tempo2$tempo_gol <- tempo2$tempo_gol - 45

       qplot(data = gols, tempo_gol, geom = "histogram")+ggtitle("Frequência dos gols")
```



Podemos observar que o histograma se encontra em uma forma uniforme, onde não é possível reparar em algum intervalo de tempo que se sobressai em relação aos outros.

Análise dos Dados

Com os dados em mãos, nós podemos fazer algumas perguntas interessantes, como por exemplo: Existe diferença significativa entre o momento em que o gol acontece no 1o e 2o tempo das partidas, considerando o minuto do 1o ou 2o tempo em que ele acontece?

Para responder essa pergunta, nós precisamos dividir os gols entre os que ocorreram no primeiro tempo e os que ocorreram no segundo tempo do jogo. Feito isso, podemos aplicar um teste de hipótese no qual as hipóteses seriam:

Hipótese Nula: Não existe diferença significativa entre a média dos gols.
Hipótese Alternativa: Existe uma diferença significativa entre a média dos gols.

```
gols<-read.csv("camp brasileiro 2013 - dados processados copy.csv")
gols <- filter(gols,tempo_gol != "None")
gols$tempo_gol <- as.numeric(gols$tempo_gol)

tempo1 <- filter(gols,tempo_gol <= 45)
tempo2 <- filter(gols,tempo_gol > 45)
tempo2$tempo_gol <- tempo2$tempo_gol - 45

t.test(tempo1$tempo_gol,tempo2$tempo_gol)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  tempo1$tempo_gol and tempo2$tempo_gol
## t = -1.1699, df = 931.854, p-value = 0.2423
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.7279333  0.6902483
## sample estimates:
## mean of x mean of y
##  23.59361  24.61245
```

Observando o valor do "p-value", o qual se refere ao valor da probabilidade de refutar a hipótese nula, podemos refutar ou não a hipótese nula. Se o "p-value" apresentar um valor menor que o alfa, onde nesse caso é 5%, nós podemos refutar a hiótese nula. Caso contrário, não podemos refutá-la. No nosso caso, podemos notar que o valor do "p-value" é maior que o alfa, ou seja, não podemos refutar a hipótese nula. Outra informação valiosa é o intervalo de confiança da diferença das médias, onde obtemos um intervalo entre [-2.7279333, 0.6902483]. Observando esse intervalo podemos perceber que o valor zero se encontra nele, ou seja, com 95% de confiança dizemos que não existe uma diferença significativa entre as médias dos gols nos dois tempos.

Utilizando o "Wilcoxon test" nós podemos refazer o experimento anterior utilizando as mesmas hipótese, mas usando a mediana como parâmetro, onde esta não é influenciada por "outlier's" presentes em nossos dados.

```
      wilcox.test(tempo1$tempo_gol,tempo2$tempo_gol)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  tempo1$tempo_gol and tempo2$tempo_gol
## W = 103877, p-value = 0.2089
## alternative hypothesis: true location shift is not equal to 0
```

Mais uma vez notamos que o valor do "p-value" é maior que 5% (alfa), ou seja, continuamos não podendo refutar a hipótese nula.

Após calcularmos os resultados dos testes de hipóteses com os Intervalos de Confiança podemos concluir que as duas técnicas se auto complementam, no sentido que sem o IC eu não consigo afirmar o tamanho da diferença das médias, caso exista tal diferença. Com o mesmo raciocinio, podemos dizer que sem o teste de hipóteses não temos o valor do p-value, ou seja, não tenho a probabilidade de rejeitar a hipótese nula.

Assim sendo, podemos dizer que as duas técnias permitem chegar na mesma conclusão, por exemplo, o gráfico abaixo representa o intervalo de confiança das médias dos gols no primeiro e segundo tempo. Observando tal gráfico, é facilmente notado que tais intervalos, praticamente, sobrepõem um ao outro. Ou seja, não podemos dizer que exista uma diferença significativa entre eles com 95% de confiança.

```
        gols<-read.csv("camp brasileiro 2013 - dados processados copy.csv")
        gols <- filter(gols,tempo_gol != "None")
        gols$tempo_gol <- as.numeric(gols$tempo_gol)

        gols$tempo_jogo <- 0

         for (i in seq(1,nrow(gols))){
           if (gols[i,4] > 45){
```

```
      gols[i,8] <- 2
      gols[i,4] <- gols[i,4] - 45
    }
    else{
      gols[i,8] <- 1
    }
  }

ggplot(gols, aes(x = tempo_jogo, y = tempo_gol)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", colour = "blue",) + ggtitle("Inter
```