CEFET/RJ
Bacharelado em Ciência da Computação
Inferência Estatística - Trabalho 02

Prof. Eduardo Bezerra (ebezerra@cefet-rj.br)

outubro/2017

Conteúdo

1 Teorema do Limite Central

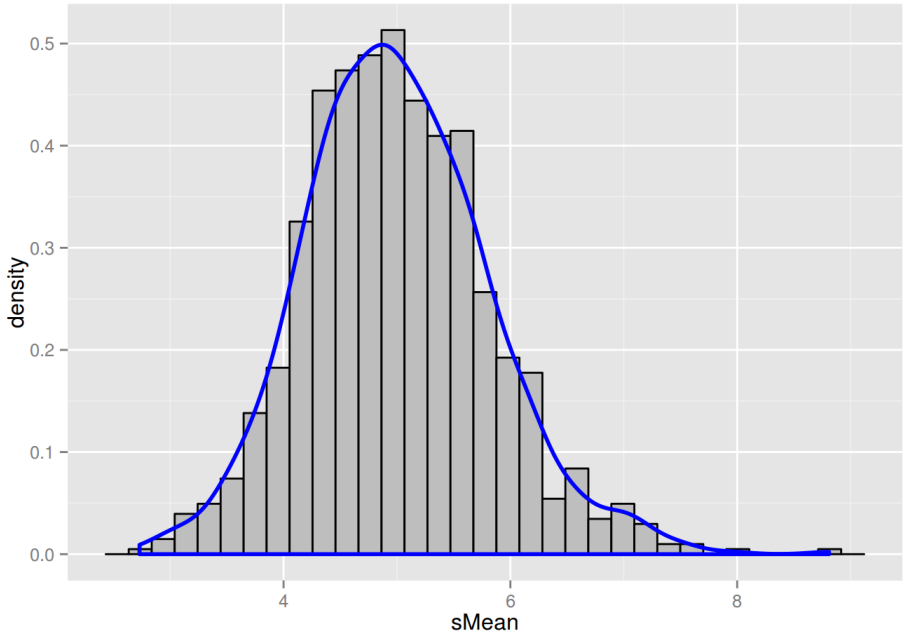
Nesta parte, você irá investigar a distribuição exponencial usando o `R` e o Teorema do Limite Central como ferramentas. A distribuição exponencial pode ser simulada em `R` por meio do comando abaixo (onde lambda é o parâmetro da distribuição):

```
rexp(n, lambda)
```

A média da distribuição exponencial é $1/lambda$ e o desvio padrão é também $1/lambda$.

Em sua investigação, defina o valor de lambda como igual a 0,2 para todas as simulações que você realizar. Sua investigação deve abranger a distribuição de médias de 40 exponenciais i.e., o tamanho de suas amostra deve ser $n = 40$). Além disso, sua investigação deve usar 1000 simulações (i.e., a quantidade de amostras deve ser igual a 1000).

Apresente um histograma da média da amostra (*sampling distribution*) e compare-a com a média teórica da distribuição. O gráfico que você deve produzir deve ser semelhante ao apresentado abaixo. Como sugestão, use a biblioteca `ggplot` para produzir essa gráfico. Repare que, assim como a figura abaixo, seu gráfico deve mostrar que a distribuição amostral é aproximadamente normal.



2 Média e Variância Amostrais

Nesta parte do trabalho, a variável de interesse é X com $X \sim N(\mu = 2, \sigma^2 = 4)$. Você irá investigar algumas propriedades da média e da variância amostrais ao tomar uma amostra aleatória de tamanho n, para diferentes valores de n. Em

outras palavras, você deverá usar o R para gerar n amostras aleatórias simples de X.

(i) Execute 10 simulações e tome uma amostra de tamanho $n = 10$ em cada etapa de simulação. Salve a média e a variância da amostra em um vetor. Dessa forma, você é capaz de investigar os resultados de todas as etapas da simulação. Os comandos para você executar essa parte do trabalho são fornecidos abaixo.

```
mu<-2
sigma<-2
n<-10
asim<-10
xbar<-rep(NA,asim)
xvar<-rep(NA,asim)
for(i in 1:asim)
{
  print(i)
  set.seed(i)
  # x contém uma amostra aleatória de tamanho n
      da variável X
  x<-rnorm(n,mu,sigma)
  xbar[i]<-mean(x)
  xvar[i]<-var(x)
}
```

(ii) Em aula, vimos que a média da amostra ($\bar{X}$) é um estimador não enviesado (*unbiased estimator*) para a média da população (μ). Explique essa propriedade por meio dos resultados obtidos no item (i).

(iii) A variância da amostra também é um estimador não enviesado para a variância populacional? Explique por meio dos resultados obtidos no item (i).

(iv) Aumente o número de simulações para 100 e, finalmente, para 1000. O que você percebe?

(v) Compare a variância da amostra com a variância da média da amostra.

(vi) Em cada etapa de simulação, produza uma amostra aleatória com $n = 10$, $n = 100$ e $n = 1000$. Os comandos em R para produzir essas amostras são fornecidos a seguir. Compare as médias dessas amostras e suas variâncias. Apresente uma discussão sobre os resultados obtidos?

```
mu<-2
sigma<-2
n<-10
asim<-1000
xbar<-rep(NA,asim)
xbar2<-rep(NA,asim)
xbar3<-rep(NA,asim)
xvar<-rep(NA,asim)
```

```
xvar2<-rep(NA,asim)
xvar3<-rep(NA,asim)
for(i in 1:asim)
{
        print(i)
        set.seed(i)
        x<-rnorm(n,mu,sigma)
        x2<-rnorm(n*10,mu,sigma)
        x3<-rnorm(n*100,mu,sigma)
        xbar[i]<-mean(x)
        xbar2[i]<-mean(x2)
        xbar3[i]<-mean(x3)
        xvar[i]<-var(x)
        xvar2[i]<-var(x2)
        xvar3[i]<-var(x3)
}

mean(xbar)
mean(xbar2)
mean(xbar3)

var(xbar)
var(xbar2)
var(xbar3)

mean(xvar)
mean(xvar2)
mean(xvar3)
```

3 Intervalo de Confiança para Médias

É comum estimar os parâmetros de uma população com base em dados de uma amostra aleatória simples. Você deve realizar as etapas dessa parte do trabalho com um *dataframe* (`http://www.r-tutor.com/r-introduction/data-frame`) chamado *survey*. Esse conjunto de dados contém o resultado de uma pesquisa feita com estudantes em uma universidade australiana.

O conjunto de dados *survey* pertence ao pacote `MASS`, que deve ser pré-carregado no espaço de trabalho `R` antes da sua utilização. Para isso, utilize os comandos a seguir.

```
library(MASS)       # faz a carga do pacote MASS
```

Consulte também mais detalhes acerca desse conjunto de dados por meio dos comandos abaixo.

```
head(survey)
help(survey)
```

Essa parte do trabalho envolve computar intervalos de confiança sobre as alturas dos estudantes (coluna `survey$Height` do conjunto de dados). Inicialmente, você deve eliminar valores faltantes (*missing values*) na coluna `survey$Height`. Para isso, pesquise sobre a função `na.omit` do `R`.

A seguir, usando a distribuição t de Student, calcule um intervalo de confiança bilateral no nível de 95% para a altura média dos estudantes da universidade.

Agora, repita o cálculo usando o z-score (em vez do t-score que você usou anteriormente). Apresente uma análise comparativa dos dois intervalos de confiança obtidos.

4 Intervalo de Confiança para Proporções

Nesta parte do trabalho, você deve produzir um intervalo de confiança, no nível de 95%, para a proporção de alunos da EIC (Escola de Informática e Computação) que estão realizando estágio atualmente. Para isso você deve realizar uma pesquisa junto a seus colegas para coletar dados necessários à produção desse intervalo de confiança. Para cada aluno que você entrevistar, colete seu nome completo e o período letivo no qual ele ingressou no curso.

Esteja certo de consultar a pelo menos 30 alunos, para que você possa realizar a construção do intervalo de confiança tomando como base o Teorema do Limite Central.

Esteja certo também de que você montou sua amostra da forma mais adequada possível. Por exemplo, sua amostra estaria enviesada se você coletasse sua amostra consultando apenas alunos do primeiro período, ou apenas alunos dos últimos períodos do curso.

Outra condição que você deve verificar é se a amostra coletada segue a distribuição normal. Para isso, produza um histograma de sua amostra e analise visualmente esse gráfico.

Após coletar os dados, transcreva-os para um arquivo de tal forma que eles possam ser manipulados com o `R`. A seguir, usando o `R`, produza o intervalo de confiança solicitado. Junto com o resultado, fornece também uma análise do resultado encontrado.

5 O que deve ser entregue

Você deve preparar um único relatório para a apresentar sua análise e conclusões sobre as diversas partes desse trabalho.

Alternativamente à entrega do relatório em PDF, você pode entregar um notebook Jupyter[1].

Independente de escolher entregar um relatório em PDF ou na forma de um notebook Jupyter, entregue também todos os arquivos em R que você criou para cada parte deste trabalho. Todos os arquivos em R deve estar na mesma pasta.

Crie um arquivo compactado que contém o relatório (ou notebook Jupyter) e os arquivos (*scripts*) em R. Esse arquivo compactado deve se chamar `SEU_NOME_COMPLETO_T2.zip`. Esse arquivo compactado deve ser entregue pelo Moodle, até a data acordada.

[1] `http://jupyter.org/`