# CEFET/RJ
# Bacharelado em Ciência da Computação
# Inferência Estatística - Trabalho 01

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

setembro/2017

# Conteúdo

# 1 Estimativas de probabilidades

Considere o conjunto de dados `AdultDataset`[1]. Este arquivo contém informações pessoais como idade, renda, nível de escolaridade e outros de mais de 25000 pessoas. A variável remuneração, que está na última coluna do conjunto de dados, é classificada como $> 50000$ ou $\leq 50000$. A variável `Education-num` representa o número de anos estudados por uma pessoa. Utilize o `R` para realizar as seguintes tarefas:

(i) Nessa primeira atividade, você deve importar o `AdultDataset` para o `R`. Utilize os comandos fornecidos abaixo para fazer isso.

```
url <- "http://archive.ics.uci.edu/ml/machine-learning-
    databases/adult/adult.data"
adult <- read.csv(url, strip.white = TRUE, header =
    FALSE)
```

A seguir, altere os nomes das colunas do data frame, para facilitar a manipulação posterior dos dados. Faça isso com o comando abaixo.

```
#adicionando nomes às colunas do dataset
colnames(adult) <- c("age","workclass","final weight","
    education","education-num","martial-status","
    occupation","relationship","race","sex","capital-
    gain","capital-loss","hours-per-week","native-
    country","income")
```

(ii) Produza um gráfico box plot da variável Age; faça uma análise do resultado obtido determinando o primeiro quartil, a mediana ou segundo quartil, o terceiro quartil e a média. Existem valores extremos (*outliers*)? Se sim, o que eles significam neste conjunto de dados?

(iii) A freqüência relativa de ocorrência de um evento, observada em várias repetições da experimento aleatório correspondente, é uma medida da probabilidade desse evento. Essa é a concepção central da probabilidade na interpretação frequentista. Sendo assim, se $n_t$ for o número total de realizações de um experimento e $n_x$ é o número de testes em que o evento $x$ ocorreu, a probabilidade $\Pr(x)$ de que o evento ocorra pode ser aproximada pela freqüência relativa da seguinte maneira:

$$\Pr(x) = \frac{n_x}{n_t}$$

Claramente, à medida que o número de realizações aumenta, pode-se esperar que a frequência relativa se torne uma melhor aproximação da verdadeira frequência. Uma alegação da abordagem frequentista é que, no longo prazo, à medida que o número de realizações se aproxima do infinito, a freqüência relativa convergirá exatamente para a probabilidade real:

[1] Este conjunto de dados está disponível em `http://archive.ics.uci.edu/ml/datasets/Adult`

$$\Pr(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t}.$$

Voltando ao conjunto de dados `AdultDataset`, cada linha desse conjunto pode ser interpretada como resultante da realização de um experimento aleatório. Em cada realização desse experimento aleatório, é escolhido um valor para cada coluna (campo) dessa linha. Nessa interpretação, portanto, podemos calcular aproximações (estimativas) para alguns valores de probabilidades. Por exemplo, se considerarmos o evento $E =$ *pessoa é mulher e tem idade maior do que 80 anos*, se $n_M$ é a quantidade de linhas do conjunto de dados que correspondem a mulheres maiores do que 80 anos, e se $n_t$ é o total de linhas de conjunto de dados `AdultDataset`, então:

$$\Pr(E) \approx \frac{n_M}{n_t}.$$

Com a ajuda do `R`, você pode facilmente contar a ocorrência de eventos. Considere o exemplo a seguir.

```
adult_mulheres_maiores_80 <- adult[(adult["sex"] == "
    Female") & (adult["age"] > 80), ]
```

Com base no que foi descrito acima, e considerando as definições de probabilidades condicional e conjunta, determine estimativas para as probabilidades a seguir.

- Probabilidade de uma pessoa ter mais de 80 anos (ou seja, $Age > 80$) dado que tem $Salary > 50000$.
- Probabilidade de uma pessoa ter mais de 80 anos e salário maior que 50000. Ou seja, calcule $\Pr(Age > 80, Salary > 50000)$.
- $\Pr(Workclass = \text{State-gov}, Occupation = \text{Adm-clerical}, Sex = \text{Male})$.
- $\Pr(Workclass = \text{Self-emp-inc}, Occupation = \text{Exec-managerial}, Sex = \text{Male})$

## 2 Distribuição Normal

Nesta parte do trabalho, por meio de simulações realizadas com o `R`, você irá verificar a validade de algumas proporções relativas à distribuição normal. Para isso, utilize amostras da distribuição normal padrão para gerar estimativas para cada uma das probabilidades a seguir. Faça isso com duas quantidades diferentes de amostras, 10 e 10000. Compare seus resultados com os fornecidos abaixo.

- $\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.683$
- $\Pr(\mu - 1.282\sigma \leq x \leq \mu + 1.282\sigma) \approx 0.8$
- $\Pr(\mu - 1.645\sigma \leq x \leq \mu + 1.645\sigma) \approx 0.9$

- $\Pr(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) \approx 0.95$
- $\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.954$
- $\Pr(\mu - 2.57\sigma \leq x \leq \mu + 2.57\sigma) \approx 0.99$
- $\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.997$

# 3 Distribuições Amostrais

Considere uma população de sacos de batatas de 5Kg cada, para os quais a variável de interesse é o número de batatas em cada saco. Para configurar essa situação como um problema de Probabilidade em vez de um problema de Estatística, fornecemos a distribuição de probabilidades para essa variável: trata-se de uma distribuição uniforme discreta nos inteiros de 10 até 20.

Nesta parte do trabalho, considere que o plano amostral utilizado é uma amostragem aleatória simples feita com substituição (*simple random sample with replacement*).

(i) Esboce um gráfico (histograma) para indicar a forma desta distribuição do número de batatas.

(ii) Encontre a média, variância e desvio padrão desta distribuição. (Observe que esta é uma população discreta.)

(iii) Considere tomar amostras de tamanho 2 dessa população e calcular a média de cada amostra. Feito isso, para cada amostra, você vai ter calculado uma *estatística pontual* (*point statistic*). Se você fizer isso para todas as possíveis amostras aleatórias de tamanho 2, a distribuição destes é denominada *distribuição amostral da média* (*sampling distribution of the sample mean*) para $n = 2$. Para este caso particular, encontre essa distribuição gerando os valores e suas respectivas probabilidades. Em seguida, esboçe um gráfico para indicar a forma da distribuição da média da amostra para $n = 2$.

(iv) Encontre a média, variância e desvio padrão desta distribuição amostral da média para $n = 2$. (Mais uma vez, note que esta é uma população discreta. Computar esses parâmetros da mesma maneira que no item (ii). No entanto, você não receberá os mesmos valores que no item (ii), porque esta é uma distribuição diferente.)

(v) O Teorema do Limite Central apresenta uma teoria sobre o que o desvio padrão e a média da distribuição amostral da média da amostra deveriam ser. Use essa teoria e os resultados do item (ii) acima para encontrar o desvio padrão e a média da distribuição amostral da média para $n = 2$. Estes valores deveriam ser os mesmo encontrados no item (iv).

(vi) Suponha, por um momento, que você não conhecia o modelo de probabilidade para a população e que deseja estimar a média da população de

uma amostra de tamanho 2 tomada aleatoriamente. Qual estatística você calcularia dessa amostra para estimar a média da população? Você acha que essa estatística seria um bom estimador da população? O que poderia ser um melhor estimador? Por quê?

(vii) Em Estatística, a *amplitude* (*range*) de uma população é a diferença entre o maior e o menor valores encontrados nessa população. Note portanto que a amplitude é um parâmetro populacional, assim como também o são a média e a mediana pouplacionais. Considerando novamente a população original, qual é a sua amplitude?

(viii) Refaça todo o item (iii) para a distribuição amostral da amplitude amostral, para n = 2.

(ix) No seu gráfico da distribuição amostral da amplitude amostral para n = 2, indique o valor do parâmetro da população (i.e., a amplitude populacional). Você acha que a amplitude amostral é um bom estimador da amplitude da população? Em caso contrário, como você poderia construir uma estatística que seria um melhor estimador do alcance da população?

# 4 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[2].

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_T1.zip`. Esse arquivo compactado deve ser entregue pelo Moodle, até a data acordada.

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