## 1. Introdução

### Dedicado ao Augusto Boal e ao DMQ

A escola é o lugar para a construção e vivência de experiencias e valores significativos (Leandro Karnal).

Esse texto é dedicado aos estudantes de estatística da UNIRIO e aos meus (espetaculares) colegas de departamento.

Minha convivência com eles me ajudou a perceber que um curso de estatística para uma turma de administração pública precisa ser muito mais do que fórmulas, questões de livros e aplicação de provas.

Essa minha experiência com eles me diz que há uma necessidade real de mostrar como a disciplina de estatística é uma parte essencial do currículo profissional dos novos gestores públicos. Isso me fez tentar traduzir as fórmulas matemática da área da estatística em objetivos pedagógicos aplicados. Assim, esse texto é o resultado da minha abordagem pessoal aos conceitos desenvolvidos no coração do Departamento de Métodos Quantitativos da UNIRIO.

Começo dizendo que tentei mudar o modo que a estatística fosse percebida pelos estudantes. Tentei fazê-la de forma totalmente aplicada. Tentei mostrá-la como uma importante ferramenta da administração pública.

O objetivo não era conseguir encontrar o valor correto de uma conta/calculo de estatística, mas sim utilizar o método quantitativo como um instrumento para orientar o processo de tomada de decisão na administração pública. Acredito que desse modo o estudante não só aprenderá a estatística, mas deverá preservar o pensamento estocástico e reconhecer como e quando aplicar esses conceitos na sua vida profissional décadas depois de passar por esse semestre.

#### O funcionário Johnnson e o professor João

Os casos apresentado aqui foram desenvolvidos a partir do livro de PETERS e GRAY (1994). Este livro apresenta dezoito casos de problemas de gestão resolvidos com a estatística e deve ser visto como a fonte primária (e complementar) desse texto. Só para dar um exemplo, o protagonista Johnson do caso de discriminação racial corporativa (*employee discrimination*) foi traduzido em um caso de racismo do professor João da Escola Oliveira Ltda.

#### O valor correto pouco útil da aplicação das formulas de estatística *que o computador faz para gente*.

.

Não é que a abordagem de aplicação de formulas e fazer contas estatística “na mão” seja inútil. O que estou querendo dizer é que o estudante não vai fazer mestrado em estatística e muitas vezes buscar o valor correto de uma conta estatística que o R ou o Python faz de forma automatica será de pouca importancia para um futuro gestor. Assim, ele estuda para passar na prova e depois esquece desse conteúdo.

Os livros didáticos tradicionais de estatística buscam enfatizar o processo de encontrar a resposta certa em um determinado intervalo de confiança (95%/99% de confiança). Muitas vezes, essa abordagem não tem nenhum contexto do motivo para a construção desse intervalo. Esse tipo de método tende a valoriza o calculo e a memorização das formulas. O objetivo geralmente é a aplicação correta de procedimentos estatísticos para encontrar a resposta numérica correta. Neste caso, o limite inferior e superior do intervalo de confiança. Você sabe o que estou dizendo. Todos nós passamos por isso como alunos e professores.

Nesta abordagem, no lugar do objetivo ser conseguir a resposta correta em uma aplicacao das contas e análise estatística, o objetivo, muitas vezes, é entender a situação problema da gestão e resolver um problema real no processo de tomada de decisão em situações tipicas da administração/gestão.

Neste projeto, busco dar aos estudantes a oportunidade de trabalhar com um caso de estudo representativo do mundo real que exige do estudante a identificação do problema a ser abordado, bem como indentificar quais são as informações relevantes e não relevantes. De forma ainda mais clara: esta abordagem precisa que o estudante “mergulhe direto nos dados” para definir o problema, desenvolver alternativas de gestão e recomendar uma solução.

#### A famosa pergunta: PARA QUE TENHO QUE APRENDER ISSO?

I have no data yet. It is a capital mistake to theorize before one has data.(Sherlock Holmes)

Acredito que esse destaque todo na busca pela resposta numérica correta falha com a importante contribuição que a estatistica pode dar ao processo de tomada de decisão pelo gestor público. PETERS e GRAY (1994) afirmam que se buscarmos a resposta correta do estudante, falhamos em entender o motivo pelo qual o conhecimento estatístico pode ser útil na area de administração (como conhecimento estatístico deve ser utilizado). Também dizem que falhamos em entender como esse conhecimento pode ajudar no processo de tomada de decisão em situações administrativas do mundo real. E completam: falhamos em mostrar o valor do conhecimento estatístico e isso sempre levará ao alunx a questionar a relevancia da estatística para a sua área (não poderia concorda mais com eles).

#### Para além da estatística

No lugar de encontrar a resposta numérica correta em uma prova de estatistica, o uso desta metodologia permite ao aluno não só estudar a Estatística em si, mas também aprender a programar na linguagem R, bem como adquirir experiência em uso de bancos de dados, planilhas, apresentações, e arquivos Markdown (arquivos para uso na web). Acredito que assim ele estará mais preparado para a vida cidadã, vida academica, e para a sua vida profissional.

Entender Estatística é essencial aos cidadãos das sociedades atuais: para ser crítico em relação à informação disponível na sociedade, para entender e comunicar com base nessa informação, mas, também, para tomar decisões, uma vez que uma grande parte da organização dessas mesmas sociedades tem por base esses conhecimentos. (Shaughnessy, 1992, 1996).

Em resumo, a vivência com os professores do DMQ e o livro do PETERS e GRAY (1994) me fizeram perceber que seria muito interessante abrir mão da aplicação do calculo numérico correto em uma prova em favor do conceito e da interpretação da estatística no mundo real.

O prof. Pimentel da escola de informatica da UNIRIO uma vez afirmou o método tradicional por prova valoriza uma leitura mecânica. Isto é, porque assim o desafio passa a ser a memorização do texto. Se o estudante consegue decorar, terá respondido ao desafio da disciplina.

#### Autoria em estatística

Busco promover o letramento em Estatística, isto é,desenvolver a capacidade de interpretar e de criar discursos em estatística para atuar no mundo.

Praticar o diálogo Estatística com a Gestão

oportunizar diálogos para desenvolver a linguagem e o pensamento em Estatística, empregar vocabulário da linguagem Estatística

Interpretar-avaliar discursos em estatística (as análises mentem, omitem, enviesam?)

Utilizar os conceitos de estatística e de probabilidade para criar o discurso em estatística.

Aqui a linguagem de programação R deixa de ser apenas um técnica e passa a ser parte do processo de construção do conceito.

A pergunta da prova geralmente é parecida com à dada em sala de aula. Apesar de apenas trabalhar com a memoria, o método tem um certo prestigio com os alunos. **a prova seria exatamente o esperado**. Fácil, seguro e quase inutil na minha visão. Ninguem achava aquelas informacoes uteis ou interessantes (Leandro Karnal p. 54).

### 1.1. Objetivo

Este capítulo tem como objetivo apresentar um projeto de análise de dados baseado em caso de ensino com a liguagem de programação **R**.

### 1.2. Habilidades a serem desenvolvidas

A atividade envolve a defesa de uma posição pelo aluno sobre a situação problema. Assim, nessa atividade, buscamos: 1. a identificação da situação-problema  
2. a formulação de perguntas de pesquisa/hipóteses,  
3. análise exploratória inicial dos dados,  
4. depuração de erros na base de dados/busca por outliers/limpeza e higenização dos dados, 5. refazer a análise exploratória após a limpeza dos dados,  
6. escolha de um teste de hipóteses adequado,  
7. avaliação das condições minimas para o teste de hipoteses  
8. aplicação de testes de hipóteses/ realização do procedimentos  
9. interpretação dos resultados.  
10. escrita de um parecer com base no teste de hipoteses

### 1.3. Público-alvo

Alunos das ciências humanas e sociais, com enfase nos cursos de Administração Pública e de Empresas, Ciencia Política, Economia e Sociologia.

cada turma é uma nação diferente!

## 2. Os objetivos pedagógicos.

## 2.1. Os objetivos pedagógicos.

## 2.2 Sequencia didática

## 3. Casos de ensino – Teoria,

Casos para ensino em Administração Pública são relatos de situações da vida organizacional, construídos com propósitos educacionais específicos (Roesch,2007). Assim, nesse caso, a estrutura e o processo de redação de casos para ensino fundamenta-se em 03 (três) grandes pilares:

* o desenvolvimento do caso-problema e
* a narração da contextualização da forma interessante
* o conflito entre os atores/personagens

O caso-problema não é apenas a descrição de eventos ou problema. Ele contém uma história. Por isso, apesar que criar um bom caso de estudo devemos recorrer à literatura de ficção. Pessoalmente, eu gosto muito da ideia de estudar teatro como Shakespeare, Agusto Boal e Grotowski para construir casos de estudos e resolver os problemas narrativos decorrentes da estrutura proposta pelo caso-problema.

Assim, em um caso, o relato da situação-problema pode ser orientado por perguntas, tais como: Qual o problema? Quais as partes envolvidas? Que reações estão sendo manifestadas? Que soluções estão sendo propostas? Que resultados foram atingidos? Que novos problemas se originaram de tais soluções? (Roesch,2007)

O método do caso para ensino em Administração é um material didático que não se popularizou na universidade brasileira e poucos professores utilizam casos-problema em sala de aula. (Roesch,2007)

### 3.1. O conflito entre os personagens e o problema de pesquisa

“as controvérsias que proporcionam diferentes interpretações”

Aqui devemos ter pelo menos duas posições antagônicas. Devemos ter um conflito entre os os dois personagens do caso de estudo. Buscamos com isso que o aluno se posicione por uma das posições sobre o assunto. Esse posicionamento deverá ser fundamentado pela análise estatística.

## 3.2. A autoria do relatório e o letramento estatístico

Aqui o aluno tem que escrever um parecer fundamentado na análise estatística para algum personagem do caso de ensino. geralmente, o personagem é o(a) seu chefe. Desse modo, busca-se proporcionar ao alunx uma viagem de descoberta pelo metodo cientifico para fundamentar a sua anáilse.

Nesse sentido, ao final do caso de esuno

. O parecer/posicionamento do alunxs

aqui o aluno deverá construir um plano de ação.

## 4. Casos de ensino – Exemplos Aplicados na estatística

### 4.1. Estudo da discriminação racial na empresa pública de processamento de dados,

O primeiro passo é a importação da base de dados.

#library(readxl)  
#banco\_de\_dados <- read\_excel('Escola\_Oliveira.xlsx')  
  
banco\_de\_dados <- read.csv2('https://raw.githubusercontent.com/DATAUNIRIO/Base\_de\_dados/master/Escola\_Oliveira\_p\_Livro\_DMQ.csv')

Depois disso uma etapa importante é inspecionar os dados. Isso pode ser facilmente observado no R com a função *str* e observando o dicionário de dados ( codebook).

# Estrutura de dados

str(banco\_de\_dados)

dicionario <- tibble::tibble(  
Nome\_da\_variavel=c('Grau de Pagamento','Gênero','Raça','Casado','Idade','Anos trabalho','Desempenho'),  
Descricao\_da\_variavel=c(  
 '1 = Professor Assistente (Grau 1) \n   
2 = Professor Adjunto (Grau 2) \n  
3 = Professor Titular (Grau 3)','0 = Feminino\n 1 = Masculino','1 = Não Branco/Minoria (preto/pardo/indígena), \n 0 = Branco','0 = Não,\n 1 = Sim','Anos de idade do funcionário','Anos de trabalho como representante de serviços na ServPro','Avaliação de desempenho do professor (escala de 0 a 10)'))

| Nome\_da\_variavel | Descricao\_da\_variavel |
| --- | --- |
| Grau de Pagamento | 1 = Professor Assistente (Grau 1)    2 = Professor Adjunto (Grau 2)   3 = Professor Titular (Grau 3) |
| Gênero | 0 = Feminino  1 = Masculino |
| Raça | 1 = Não Branco/Minoria (preto/pardo/indígena),   0 = Branco |
| Casado | 0 = Não,  1 = Sim |
| Idade | Anos de idade do funcionário |
| Anos trabalho | Anos de trabalho como representante de serviços na ServPro |
| Desempenho | Avaliação de desempenho do professor (escala de 0 a 10) |

Aqui, os alunxs devem perceber que há um erro no banco de dados. Temos variáveis qualitativas sendo tratadas como quantitativas. para corrigí-las devemos usar uma das estratégias abaixo. Todas essas estratégias são aceitaveis. O uso dessas estratégias (básica, intermediária, avançada) dependem do nível de letramento de programação R cada alunx.

# nível básico  
banco\_de\_dados$sexo<-as.factor(banco\_de\_dados$sexo)  
banco\_de\_dados$raca<-as.factor(banco\_de\_dados$raca)  
banco\_de\_dados$casado<-as.factor(banco\_de\_dados$casado)  
banco\_de\_dados$grau\_pagamento<-as.factor(banco\_de\_dados$grau\_pagamento)

# nível intermediário  
banco\_de\_dados$sexo<-ifelse(banco\_de\_dados$sexo==0,'Feminino',  
 ifelse(banco\_de\_dados$sexo==1,'Masculino',NA))  
  
banco\_de\_dados$raca<-ifelse(banco\_de\_dados$raca==0,'Branco',  
 ifelse(banco\_de\_dados$raca==1,'Minoria',NA))  
   
banco\_de\_dados$casado<-ifelse(banco\_de\_dados$casado==0,'Não',  
 ifelse(banco\_de\_dados$casado==1,'Sim',NA))

# nível avançado com uso da função case\_when   
library(dplyr)  
banco\_de\_dados<-banco\_de\_dados %>%  
 mutate(  
 grau\_pagamento = case\_when(  
 grau\_pagamento == 1 ~ "Professor Assistente (Grau 1)",  
 grau\_pagamento == 2 ~ "Professor Assistente (Grau 2)",  
 grau\_pagamento == 3 ~ "Professor Assistente (Grau 3)",  
 TRUE ~ "outra situação"))

Se você tiver mais tempo do que um semestre para trabalhar com os alunos, sugiro trocar o registro do salário da linha 90 (id=90). A trocar do salário de R$ 2.710,00 para R$ 27.100,00. Aqui você poderia discutir um pouco mais do tratamento das variáveis, *outliers*, erro de digitação, registros duplicados entre outros elementos da fase de limpeza/higienização dos dados. O código para isso seria algo do tipo

# discussao do outlier  
banco\_de\_dados$salario[90]<- 27100  
# ausencia de outlier  
banco\_de\_dados$salario[90]<- 2710

### 4.2. Estudo da discriminação racial na Escola Oliveira,

### 4.3. Estudo da desigualdade de gênero dos candidatos a vereador(a) no município de varre-e-sai

### 5. *exemplos de relatorios de alunxs*

com essa atividade, os alunos conseguiram fazer isso aquilo e isso outro falar que será um banco de dados

<https://dataunirio.github.io/Trabalhos_alunos_UNIRIO/avaliacao_da_disciplina/index.html>

## 5. Conclusão

no futuro, buscarei colocar algumas variaveis irrelevantes para que o aluno possa separar variáveis pouco importantes de variáveis mais fundamentais

##Referências 1.Roesch, Sylvia Maria AzevedoNotas sobre a construção de casos para ensino. Revista de Administração Contemporânea [online]. 2007, v. 11, n. 2 [Acessado 27 Outubro 2021] , pp. 213-234. Disponível em: <https://doi.org/10.1590/S1415-65552007000200012>. Epub 18 Jan 2008. ISSN 1982-7849. <https://doi.org/10.1590/S1415-65552007000200012>.

### 2.3. Pré-processamento dos dados no R

Após a leitura dos dados, é a etapa de pré-processamento. Nesta fase o aluno colocará em prática conhecimentos teóricos, como a classificação de variáveis, e avaliará a qualidade dos dados.

A partir da função [str](https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/str), é possível ver o tamanho do banco e como as variáveis foram codificadas e estão sendo interpretadas pelo R.

# Vendo a estrutura dos dados  
#str(avc)

A função mostra que o banco avc possui 5110 observações e 12 colunas. As variáveis foram interpretadas pelo R como:

1. **id**: int - número inteiro
2. **gender**: chr - caracter (cada elemento do vetor é uma sequência de um ou mais caracteres)
3. **age**: num - numérico (real ou decimal)
4. **hypertension**: int - número inteiro
5. **heart\_disease**: int - número inteiro
6. **ever\_married**: chr - caracter
7. **work\_type**: chr - caracter
8. **Residence\_type**: chr - caracter
9. **avg\_glucose\_level**: num - numérico (real ou decimal)
10. **bmi**: chr - caracter
11. **smoking\_status**: chr - caracter
12. **stroke**: int - número inteiro

Nesta etapa, o aluno deve-se questionar quais variáveis foram codificadas de forma errada e se alguma ou mais colunas podem ser eliminadas por não fazerem sentido ao projeto ou por fazerem apenas o papel de identificador das linhas. Neste caso, a coluna id desempenha este papel. Para isso, o aluno pode atribui-la como nome das linhas e, em seguida, exclui-la do banco de dados. A função [rownames](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/row+colnames) pode ser usada para este objetivo.

# Atribuindo a coluna id como nome das linhas  
#rownames(avc) <- avc$id  
  
# Excluindo a coluna id  
#avc$id <- NULL

Em seguida, o aluno deverá classificar as variáveis e identificar quais delas não foram interpretadas corretamente pelo R. A partir do conhecimento teórico adquirido, ele saberá que uma codificação errada desencadeará em erros consecutivos na análise de dados.

**Dica**: aproveite esta etapa do projeto para revisar as classificações de variáveis. Se preferir, este [vídeo](https://www.youtube.com/watch?v=IPRmXE5SREA&list=PLDoqBsM6cBJYzQHCWMCCSaKhXCk00ppfO&index=2) pode ser compartilhado com os alunos para ajudá-los a lembrar o conteúdo.

Assim, espera-se as seguintes classificações:

1. **gender**: variável qualitativa nominal binária
2. **age**: variável quantitativa razão contínua
3. **hypertension**: variável qualitativa nominal binária
4. **heart\_disease**: variável qualitativa nominal binária
5. **ever\_married**: variável qualitativa nominal binária
6. **work\_type**: variável qualitativa nominal policotômica
7. **Residence\_type**: variável qualitativa nominal binária
8. **avg\_glucose\_level**: variável quantitativa razão contínua
9. **bmi**: variável quantitativa razão contínua
10. **smoking\_status**: variável qualitativa nominal policotômica
11. **stroke**: variável qualitativa nominal binária

Ao comparar com o resultado fornecido pelo R, vemos que as variáveis **hypertension**, **heart\_disease**, **bmi** e **stroke** foram codificadas de forma errada.

Por terem sido representadas pelos números 0 e 1, o R interpretou as variáveis **hypertension**, **heart\_disease**, e **stroke** como quantitativas. Logo, devemos transformá-las para qualitativa. A função a ser usada poderá ser a [as.factor](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/factor). A codificação fator é a melhor opção para representar uma variável qualitativa e, por isso, também transformaremos as variáveis **gender**, **ever\_married**, **work\_type**, **Residence\_type** e **smoking\_status** de chr para factor.

Para facilitar a tarefa, a função [mutate\_all](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/mutate) do pacote dplyr pode ser utilizada, pois ela transforma várias variáveis ao mesmo tempo. A função str pode ser novamente usada para ver as modificações.

# Transformando para factor  
#library(dplyr)  
#avc[,c(1,3:7,10,11)] <- avc[,c(1,3:7,10,11)] %>% mutate\_all(as.factor)  
  
# Vendo a estrutura dos dados  
#str(avc)

**Dica**: para que as tabelas e gráficos a serem preparados em etapas posteriores sejam autoinformativos, sugere-se a modificação das categorias das variáveis qualitativas para nomes mais compreensíveis. Essa etapa pode ser realizada pela função [levels](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/levels). As mudanças devem seguir a ordem que o R apresenta as categorias. Por isso, recomenda-se rodar primeiramente a função levels para ver a ordem dos níveis antes de realizar a substituição. –>

```r  
# Mudando os nomes das categorias  
#<!-- levels(avc$gender) <- c("Feminino", "Masculino", "Outro") -->  
#<!-- levels(avc$hypertension) <- c("Não", "Sim") -->  
#<!-- levels(avc$heart\_disease) <- c("Não", "Sim") -->  
#<!-- levels(avc$ever\_married) <- c("Não", "Sim") -->  
#<!-- levels(avc$work\_type) <- c("Filhos", "Governo", "Nunca trabalhou", "Privado", "Autônomo") -->  
#<!-- levels(avc$Residence\_type) <- c("Rural", "Urbano") -->  
#<!-- levels(avc$smoking\_status) <- c("Ex-fumante", "Nunca fumou", "Fuma", NA) # a informacao nao disponivel foi codificada como NA -->  
#<!-- levels(avc$stroke) <- c("Não", "Sim") -->

Podemos observar que por conta do “N/A”, o R interpretou a variável como caracter. Nesta etapa, o aluno deve então ser capaz de transformar esse dado ausente em NA e, em seguida, transformar a variável em numérica com a função [as.numeric](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/numeric).

# Transformando os "N/A" da variavel bmi em NA  
#avc$bmi[avc$bmi == "N/A"] <- NA  
  
# Transformando a variavel bmi em numerica  
#avc$bmi <- as.numeric(avc$bmi)

Após o pré-processamento de dados, o aluno deverá ser capaz de chegar na seguinte estrutura:

#str(avc)

–> ```