**Disciplina Introdução à Pesquisa Científica – Análise estatística com RStudio**

**Prof. Dr. Guilherme Giannini Artioli**

Neste guia, vocês encontrarão as instruções para rodar os testes estatísticos que serão necessários para a elaboração do trabalho final da disciplina. Recapitulando, vocês deverão rodar 2 testes, a saber:

* Teste *t* de Student para amostras dependentes, também conhecido como teste *t* pareado.
* Teste de correlação linear de Pearson, também conhecido como teste *r* de Pearson.

O teste *t* será utilizado para comparar as médias das alturas dos 10 saltos verticais entre as condições COM *vs.* SEM aquecimento. O valor de *p* deverá guiar nossa interpretação. Se *p*<0,05, então as condições são diferentes (do ponto de vista estatístico). Nesse caso, pode-se afirmar que existe um efeito do aquecimento sobre o desempenho. Por outro lado, se *p*>0,05, então as condições não são diferentes. Nesse caso, pode-se afirmar que não existe um efeito do aquecimento sobre o desempenho.

Já o teste *r* de Pearson é utilizado para avaliar o grau de associação entre duas variáveis. No caso do nosso trabalho, vamos testar se a variável “peso corporal” está associada com a variável “desempenho” (altura dos saltos). Nossa interpretação deve ser feita com base nos valores de ***r*** e de ***p***. O valor de r indica o grau de associação, ou correlação. Variando de -1 a +1, o sinal indica se a correlação é direta (valores positivos; quanto maior uma variável, maior a outra) ou inversa (valores negativos; quanto maior uma variável, menor a outra). O valor de r também indica se a associação é forte, moderada ou fraca. Quanto mais próximo de Ι1Ι (repare no símbolo “módulo”), mais forte é a correlação; quanto mais próximo de 0, mais fraca é a correlação.

Para realizar as análises, vocês precisarão ter em mãos os programas R e RStudio instalados em seus computadores. Logo baixo vocês encontrarão um guia de instalação desses programas. Basta seguir o passo-a-passo. Vocês também irão precisar da planilha de Excel contendo os seguintes valores para cada um dos 37 participantes:

1. Soma dos 10 saltos na condição COM aquecimento
2. Soma dos 10 saltos na condição SEM aquecimento
3. Média das somas dos 10 saltos em ambas as condições
4. Peso corporal

Certifiquem-se de que os dados estejam todos alinhados para cada um dos participantes, como mostra a imagem abaixo:



Agora vamos baixar e instalar o R e o RStudio para podermos analisar esses dados.

**Guia para instalação e uso do software de análises estatísticas Rstudio**

1. Instale o programa R. Vá ao link: <https://vps.fmvz.usp.br/CRAN/>
2. Instale o programa RStudio (versão gratuita para Desktop). Vá ao link: <https://rstudio.com/products/rstudio/download/#download>
3. Após instalar ambos os programas acima, abra o RStudio, como mostra a imagem abaixo:



1. Se tudo deu certo, você verá uma tela como demonstrado abaixo, e poderá começar a digitar seus comandos:



Aqui é onde você irá visualizar gráficos

Aqui é onde você irá importar (e ver) dados do Excel

Aqui é onde você irá digitar os comandos

1. Agora estamos prontos para começar nossas análises!

**Guia para utilização de funções básicas do RStudio**

**(comandos estarão em vermelho)**

**(comentários explicando os comandos estarão em azul)**

Comece importando sua planilha do Excel ao RStudio. Lembre-se de que ela deve estar salva como .xls. Basta clicar em “import dataset”, como mostra a imagem abaixo:



Selecione seu arquivo do Excel e, se tudo estiver certo, o RStudio reconhecerá linhas, colunas e os títulos das variáveis, como na imagem abaixo. Basta então clicar em “import”.



Ao clicar em “import”, você verá que seu conjunto de dados aparecerá na tela principal do RStudio, como mostra a imagem abaixo. Os dados apresentados no R devem ser iguais aos que você organizou no Excel.



Daqui para frente, iremos utilizar comandos (na linguagem de programação R) para realizar análises estatísticas simples, e também construir alguns gráficos (gráficos são opcionais para o trabalho final). Tenha em mente que você precisará constantemente fazer referência ao arquivo que você importou do Excel – isso é feito por meio do nome que você deu ao arquivo.

Neste meu exemplo, o nome do arquivo é: Dados2020.xls. Portanto, sempre que vir Dados2020, substitua pelo nome do seu arquivo. Tenha também em mente que você precisará constantemente fazer referência às variáveis de sua planilha. No meu caso, eu as nomeei como “SEM” , “COM”, “MediaCOMSEM” e “Peso”. Portanto, quando vir esses nomes, basta substituí-los pelos nomes que estão na sua planilha.

Vamos começar calculando as médias dos saltos COM e SEM aquecimento. Para isso usaremos o comando mean() dentro dos parênteses você deve indicar qual conjunto de dados o software deve considerar

Digite no RStudio:

 mean(Dados2020$SEM)

Note: o símbolo $ indica qual coluna (pesopre) de qual arquivo (teste\_R\_xls)

Note: o resultado será impresso logo abaixo do seu comando

Agora calcule a média dos saltos COM aquecimento.

Digite no RStudio:

 mean(Dados2020$COM)

Agora calcule o desvio padrão dessas mesmas variáveis. Lembre-se de que desvio padrão é uma medida que indica a dispersão dos dados. Quanto maior o desvio padrão, mas os dados se afastam da média. Para isso, use o comando sd() dentro dos parênteses você deve indicar qual conjunto de dados o software deve considerar

Digite no RStudio:

 sd(Dados2020$SEM)

e depois digite:

sd(Dados2020$COM)

Agora faça um gráfico do tipo boxplot. Para isso, use o comando boxplot(x, y, n...) dentro dos parênteses, indique todos os conjuntos de dados que deseja incluir em seu boxplot, separando-os por vírgula.

Digite no RStudio:

 boxplot(Dados2020$SEM, Dados2020$COM)

O gráfico deverá aparecer na janela ao lado de sua tela de comandos. Você poderá exportar esse gráfico e salvá-lo em seu computador. Se quiser, pode em seu trabalho final usar este gráfico ao invés daquele de barras que fizemos no Excel.

Agora rode o teste t de student pareado. Esse teste serve para comparar se duas médias são significantemente diferentes ou não. No caso desse estudo, vamos testar se o salto é mais alto COM aquecimento em comparação ao salto SEM aquecimento. Para saber se esse aumento é sistemático (real, causado pelo aquecimento), ou meramente efeito de variação normal dos dados (não causada pelo aquecimento), usamos test t pareado. O termo “pareado” refere-se ao fato dados serem em “pares”, isto é, as mesmas pessoas avaliadas em duas ocasiões diferentes (como em estudos cruzados).

Para rodar o teste t pareado, você deverá usar o comando t.test(x, y, paired=TRUE) dentro dos parênteses, indique o par de conjunto de dados que devem ser testados, separando-os por vírgula. Indique também que o teste deve ser pareado usando a função paired=TRUE

Digite no RStudio:

 t.test(Dados2020$SEM, Dados2020$COM, paired=TRUE)

Você deverá ver os resultados logo abaixo do seu comando, como mostra a imagem abaixo:



Como a intenção deste curso é apenas familiarizá-los(as) com os conceitos mais básicos em estatística (e não dar um curso completo sobre o assunto – vocês terão semestre que vem...), vamos nos focar em apenas um parâmetro desses resultados: o valor de p (p-value). Assim, olhando a imagem acima, podemos ver que:

p = 0,1759

O valor de p indica a probabilidade de você estar errado ao afirmar que os grupos são diferentes. Normalmente, qualquer probabilidade (de estar errado) menor do 5% (0,05) é considerada aceitável. Ou seja, se o valor p é menor do que 0,05, eu posso afirmar que os grupos são diferentes, e que essa diferença é sistemática, resultante do aquecimento. Por outro lado, valores de p maiores do que 0,05 indicariam que diferenças entre as médias (PRE vs. PÓS treinamento) seriam meramente fruto de variações aleatórias, e não decorrentes do aquecimento.

No nosso exemplo, o valor de p foi 0,1759. Logo, p > 0,05. Logo, podemos afirmar que as diferenças entre as médias **NÃO** são significantes. Em outras palavras, a hipótese de que o aquecimento melhora o desempenho de salto é **falsa**.

Por fim, você irá rodar um último teste: a correlação linear. Nesse teste, você calcula o coeficiente de correlação entre duas variáveis. Ou seja, você testa se duas variáveis estão associadas ou não. Lembre-se de que associação ou correlação não permitem estabelecer relações de causa-e-efeito!

Para rodar a correlação, utilize o comando cor.teste(x, y, method=c(“pearson")) dentro dos parênteses, indique o par de conjunto de dados que você quer testar a correlação, separando-os por vírgula. Indique também qual tipo de teste deve ser usado (função method). Em nosso caso, usaremos o teste de Pearson.

Digite no RStudio:

 cor.test(Dados2020$MediaCOMSEM, Dados2020$Peso, method=c("pearson"))

Você deverá ver os resultados logo abaixo do seu comando, como mostra a imagem abaixo:



Novamente, não se preocupem com esses termos estatísticos. Vocês aprenderão a ler esses resultados no semestre que vem. Por enquanto, vamos nos focar apenas em dois desses parâmetros: o valor de p (p-value) e o coeficiente de correlação (valor de r, ou cor). Assim, olhando a imagem acima, temos que:

p = 0,1005

r = 0,2955427

O valor de p deve ser interpretado da mesma forma que fizemos no teste anterior: quando ele for menor do que 0,05, assumimos que a correlação é significativa (com chance de erro de 5%). Se p for maior do que 0,05, então assumimos que qualquer correlação seja fruto do acaso/aleatoriedade.

Já o valor de r, ou coeficiente de correlação, indica a força de correlação entre as duas variáveis. Esse valor pode variar entre -1......0......+1.

- Valores **negativos** de r indicam relação **inversa** entre as variáveis. Isso significa que quanto **maior** uma variável for, **menor** será a outra.

- Valores positivos de r indicam relação direta entre as variáveis. Isso significa que quanto **maior** uma variável for, **maior** será a outra. Ou que quanto **menor** uma variável for, **menor** será a outra.

- Quanto mais próximo de 1 (ou -1), mais forte é a correlação entre as variáveis. Quanto mais próximo de 0, mais fraca é a correlaçãp

Em nosso exemplo, a correlação entre peso e altura do salto **NÃO** é significante (p>0,05), além de ser fraca-moderada (r=0,29). Isso indica que o peso de uma pessoa parece não ter influência sobre a altura que essa mesma pessoa alcança quando salta.