

# ANÁLISES ESTADÍSTICAS



JASP



# UM GUÍA INTRODUTÓRIO

EMMANUEL GONÇALVES E SILVA  
SAULO BARROS DE MELO

CLENO COUTO  
ELAÍNE RABELO NEIVA

# **AUTORES | PRODUÇÃO | ORGANIZAÇÃO | REVISÃO TÉCNICA**

## **EMMANUEL GONÇALVES E SILVA**

Graduação em Administração pela Universidade de Brasília, mestrado em Psicologia Social, do Trabalho e das Organizações (PSTO) e é doutorando em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília. E-mail: emmanuelgoncalves@gmail.com.

## **CLENO COUTO**

Psicólogo pelo Centro Universitário de Brasília, mestre em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília (PSTO/UnB) e é Doutorando em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília (PSTO/UnB). E-mail: ccmn@proton.me.

## **SAULO BARROS DE MELO**

Graduação em Administração pela Universidade de Brasília (UnB). Mestrando em Controladoria e Contabilidade pela Universidade de São Paulo (USP). E-mail: saulobarros@gmail.com.

## **ELAINE RABELO NEIVA**

Graduação, mestrado e doutorado em Psicologia pela Universidade de Brasília e Estágio Doutoral no Exterior – Doutorado Sanduiche pela Universidade Complutense de Madri. Atualmente é professora da Universidade de Brasília e possui pós-doutorado no Programa de Pós-Graduação em Psicologia Social e Organizacional da Columbia University em Nova York, Teacher's College. E-mail: elaine@unb.br.

## DIREITOS AUTORAIS

Esta obra está licenciada com uma Licença Creative Commons Atribuição-NãoComercial 4.0 Internacional. Seu conteúdo, ou parte dele, pode ser copiado, adaptado ou redistribuído, desde que seja realizada a devida atribuição aos autores e não seja comercializado ou utilizado para quaisquer fins que envolvam interesse monetário. Este livro eletrônico pode ser duplicado em sua íntegra e sem alterações, distribuído e compartilhado para usos não comerciais, entre pessoas ou instituições sem fins lucrativos. Nenhuma parte isolada deste livro, que não seja a presente edição em sua íntegra, pode ser isoladamente copiada, reproduzida, ou armazenada em qualquer meio, ou utilizada para qualquer fim.

Brasília, abril de 2023



# SUMÁRIO

Apresentação .....	5
Interface do JASP .....	6
Análises Descritivas .....	13
Correlação.....	21
Regressão Linear Simples .....	27
Regressão Linear Múltipla .....	34
Regressão Logística Binária.....	38
Análise de Variância One-way.....	46
ANCOVA: Análise de COVariância.....	53
MANOVA: Análise Multivariada de Variância .....	59
Análise Fatorial Exploratória.....	64
Análise Fatorial Confirmatória .....	78
Referências.....	87

## APRESENTAÇÃO

O JASP é uma ferramenta de análise estatística que vem se tornando cada vez mais popular entre pesquisadores e profissionais em diversas áreas, como psicologia, biologia e economia. O software é conhecido por sua interface amigável e intuitiva, bem como por sua ampla gama de recursos e ferramentas avançadas.

Este manual oferece uma introdução ao JASP, incluindo informações sobre como baixar, instalar e usar o software. Ele também fornece uma visão geral das principais funcionalidades do JASP, como a importação de dados, a realização de testes estatísticos, a criação de gráficos e a exportação de resultados. Além disso, o manual inclui exemplos concretos e passo a passo de como aplicar o JASP em estudos reais, ajudando os leitores a compreender como usar a ferramenta para analisar seus próprios dados.

O eBook é uma ótima opção para pesquisadores, estudantes e profissionais que desejam aprender sobre o JASP e como usá-lo para realizar análises estatísticas de alta qualidade. Ele é escrito de forma clara e acessível, tornando-se uma leitura valiosa para qualquer pessoa interessada em aprender sobre o software e como ele pode ser aplicado em seus estudos ou trabalhos.

Este manual é resultado de um projeto conduzido por pesquisadores da Universidade de Brasília, com o objetivo de explorar e compreender a utilização do JASP na análise estatística.

## INTERFACE DO JASP

Desenvolvido por uma equipe de pesquisadores da Universidade de Amsterdã, JASP significa *Jeffrey's Amazing Statistics Program* (Incrível Programa de Estatística de Jeffrey, em português) como forma de reconhecimento ao Sir Harold Jeffreys, pioneiro da inferência Bayesiana. Este pacote estatístico foi desenvolvido em código aberto, está disponível em multiplataforma e é gratuito. O objetivo da equipe era desenvolver um programa de código aberto gratuito que incluía técnicas estatísticas padrão e mais avançadas com ênfase no fornecimento de uma interface de usuário simples e intuitiva.

A plataforma fornece uma interface simples que se denomina gratuita (*Free*), amigável (*Friendly*) e flexível (*Flexible*). A proposta, em contraste com os demais pacotes de softwares de análise de dados disponíveis no mercado, é apresentar uma área de interação mais simples, podendo arrastar e soltar os dados, menus intuitivos e de fácil visualização e resultados exibidos em tempo real aos cliques realizados na interface. Além disso, todos os gráficos e tabelas são apresentados no formato da American Psychological Association (APA).

O software, por ser gratuito, pode ser baixado direto no site (<https://jasp-stats.org/>), com versões disponíveis para Windows, MacOS e Linux, além da versão online que não exige a instalação em seu computador. A versão online, entretanto, apresenta algumas limitações, como estar desatualizada com relação a outras versões. As atualizações são realizadas regularmente apresentando melhorias e correções, além da inclusão de novas

técnicas e funções. Em geral, as atualizações são automáticas, mas também podem ser realizadas manualmente pelo usuário de forma gratuita.

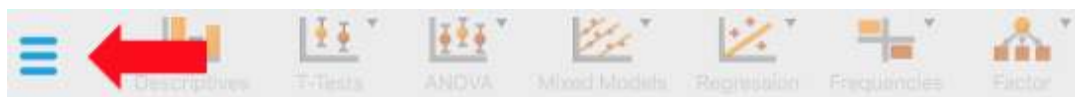
Para mais informações e atualizações sobre o software, análise de dados, vídeos demonstrativos e demais informações, consulte o site dos desenvolvedores.

## INTERFACE PRINCIPAL

Ao iniciar o JASP em seu computador você será apresentado a tela abaixo. Essa é a única tela que você utilizará e por onde terá acesso aos dados, configuração das análises e resultados.



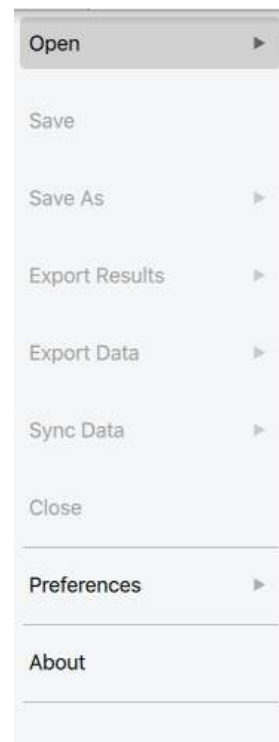
Para iniciar clique no menu “sanduíche” no canto superior esquerdo, conforme imagem abaixo.



Um menu lateral será aberto contendo as seguintes opções:

- *Open* (ou Abrir), onde é possível carregar os dados no JASP;

- *Save* (ou Salvar) e *Save As* (ou Salvar Como), que permite salvar a seção do JASP que você está utilizando com os dados, as análises e os resultados já gerados, de forma que o usuário possa continuar o processo de análise de onde parou em outro momento;
- *Export Results* (ou Exportar Resultados) que permite o usuário exportar os resultados gerados pelo programa em formato HTML, o qual é acessível por meio do navegador;
- *Export Data* (ou Exportar Dados) que permite que o usuário salve quaisquer alterações que tenham sido realizadas nos dados no mesmo ou em outro arquivo;
- *Sync Data* (ou Sincronizar Dados) para usuários que estejam utilizando base de dados online;
- *Close* (ou Fechar) para encerrar o programa e sair;
- *Preferences* (ou Preferências) que permite alterar algumas configurações do JASP, como idioma (o software já está disponível em Português), casas decimais e mais;
- *About* (ou Sobre) que apresenta uma tela com as informações sobre o software e a equipe que o desenvolveu.



### CARREGANDO OS DADOS

O JASP é compatível com diversos formatos de arquivo para leitura dos seus dados. Ao clicar em *Open* (ou Abrir), um novo menu lateral será aberto com as opções: *Recent Files* (ou Arquivos Recentes), onde você verá todos os dados ou arquivos carregados recentemente pelo JASP em seu computador; *Computer* (ou Computador), onde poderá navegar pelas pastas do seu computador para selecionar e carregar o arquivo de interesse; *OSF*, referente a *Open Science Framework* um repertório online para criação e armazenamento de arquivos sobre pesquisas científicas (mais informações em <https://osf.io>), para utilizar esse menu será necessário possuir acesso a OSF e fornecer







acesso ao JASP; *Database*; e, *Data Library* (ou Biblioteca de Dados), uma variedade de bancos de dados públicos disponíveis por meio do JASP para exemplo e testes.

Além do formato de arquivo proprietário “.jasp”, outros arquivos podem ser facilmente carregados pelo software por meio desse menu, como os que apresentam as seguintes extensões: .csv (valores separados por vírgulas); .txt (texto simples); .tsv (valores separados por tabulação); .sav (arquivo de dados do IBM SPSS); e .ods (planilha de documento aberto). Ao longo do tempo outros formatos podem ser adicionados, consulte a página de *Features* (ou Recursos) no site dos desenvolvedores (<https://jasp-stats.org/features/>).

Apesar do JASP não ser compatível com os formatos de planilhas do Microsoft Excel, estes dados podem ser facilmente convertidos para valores separados por vírgulas (mais recomendado), texto simples ou valores separados por tabulação no Microsoft Excel. Consulte a página de suporte do Microsoft Excel para mais informações (<https://support.microsoft.com/pt-br/office/importar-ou-exportar-arquivos-de-texto-txt-ou-csv-5250ac4c-663c-47ce-937b-339e391393ba>).

## GERENCIANDO O BANCO DE DADOS

	 extra	 group	 ID	
1	0.7	1	1	
2	-1.6	1	2	
3	-0.2	1	3	
4	-1.2	1	4	
5	-0.1	1	5	
6	3.4	1	6	
7	3.7	1	7	
8	0.8	1	8	
9	0	1	9	
10	2	1	10	

Ao abrir um banco de dados, ele será apresentado de forma similar a figura ao lado. Cada coluna representa uma variável (item da escala, fator, grupos, etc.), nomeie-as de forma que facilite sua memorização. A coluna *ID* pode ser importante para que você consiga identificar cada participante independente da ordenação do banco. Em um banco de dados *wide* cada linha vai corresponder a um participante, contudo em bancos de dados *long* um participante pode ocupar mais de uma linha. A diferenciação entre

os tipos de banco de dados extrapola o escopo deste livro, mas pode ser facilmente localizado em outros materiais.

O ícone de funil (▼) no canto superior esquerdo abre um menu para filtrar casos ou participantes com base nos valores de outras variáveis no banco. O ícone de soma (+) no canto superior direito abre uma janela e menu para criar uma nova variável (ou coluna)

com base na transformação das colunas já existentes; essa é uma opção importante para quando for realizar transformações nas variáveis ou calcular escores compostos de itens de escalas.

Ao lado do nome de cada variável é apresentado um ícone que representa o tipo de variável atribuído àquela variável pelo JASP. Os tipos de variáveis disponíveis no JASP são nominais (🗨️), ordinais (📊) e escalar (📏). Apesar do JASP detectar automaticamente qual tipo de variável é a mais apropriada para os dados, nem sempre essa definição será adequada. Nesses casos é importante que o usuário altere o tipo de variável para que isso não atrapalhe em análises futuras. Ao clicar no ícone do tipo de variável que aparece ao lado do nome da variável um menu será aberto abaixo apresentando as três opções.

No caso de variáveis nominais que estejam com dados apresentados em número, pode ser útil alterar também o rótulo (*label*) que representa cada valor numérico. Isso pode ser feito clicando no nome da variável. No exemplo do banco de dados apresentado na página anterior, ao clicar na variável *group* podemos atribuir rótulos a cada valor dela escrevendo o novo rótulo na linha referente a cada valor, conforme imagem abaixo. Esses rótulos serão utilizados na apresentação dos resultados das análises.



Esse menu também apresenta outras opções. A coluna *filter* (ou filtro) permite filtrar casos (ou participantes) dos dados que serão analisados clicando no *check* (✓) ao lado do valor que deseja filtrar. As setas permitem alterar a ordem dos rótulos e o X permite fechar o menu.

## PRINCIPAIS ANÁLISES



O JASP possui diversas análises estatísticas. As principais análises estão apresentadas na barra superior do software após o menu “sanduíche” logo após a instalação. Outras análises podem ser adicionadas (módulos ou *modules*) clicando no ícone de mais (+) no canto superior direito. Consulte o site para obter mais informações sobre módulos avançados disponíveis no software. As análises principais frequentistas

(estatísticas padrões paramétricas e não paramétricas) e seus testes Bayesianos alternativos são:

- Descritivas (*Descriptives*);
- Testes  $t^*$  (*T-Tests*): independente, pareado e de amostra única;
- ANOVA: independente, medidas repetidas, ANCOVA e MANOVA;
- Modelos Mistos\* (*Mixed Models*): Modelo Linear Geral e Modelos Mistos Lineares Generalizados;
- Regressão (*Regression*): correlação, regressão linear e regressão logística;
- Frequências\* (*Frequencies*): teste binomial, teste multinomial, Tabelas de Contingência e regressão log-linear;
- Análise Fatorial (*Factor*): Análise dos Componentes Principais, Análise Fatorial Exploratória e Análise Fatorial Confirmatória.

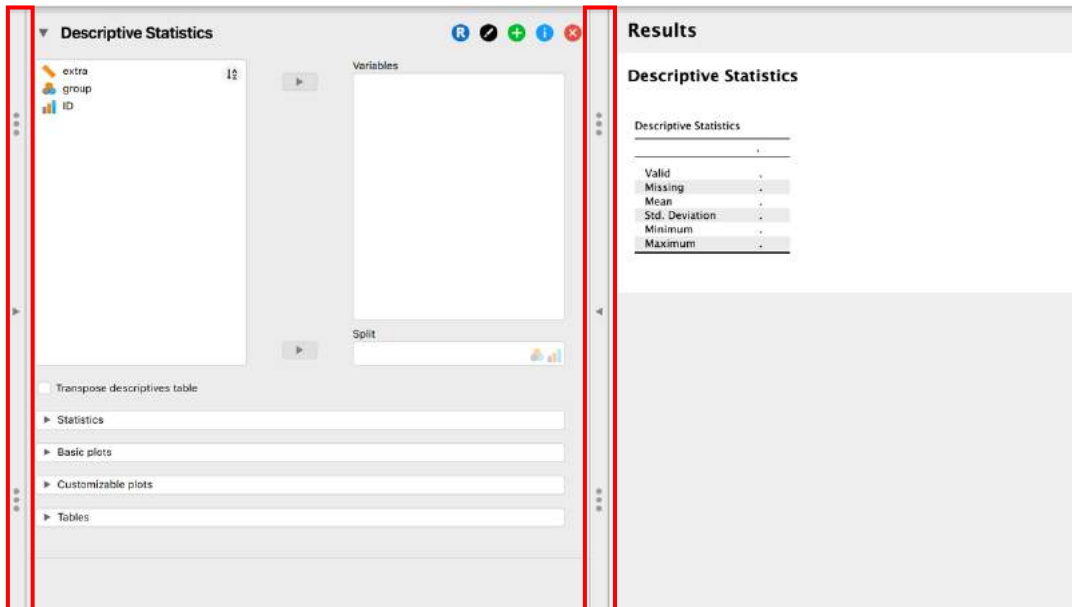
Análises marcadas com asteriscos (\*) não foram contempladas neste livro. Contudo, cabe destacar que os testes  $t$  são similares às respectivas ANOVAs, com exceção do limite de comparação entre dois grupos. Outras análises poderão ser melhor abordadas em manuais específicos.

## INTERFACE DE ANÁLISES E RESULTADOS

Ao selecionar uma das análises disponíveis no JASP, a tela será dividida em duas partes e os dados serão ocultados, conforme imagem abaixo. À esquerda é apresentada a tela para configuração da análise selecionada e à direita é apresentada a tela com os

## 12 ANÁLISES ESTATÍSTICAS COM JASP: UM GUIA INTRODUTÓRIO

resultados das configurações selecionadas, esses resultados serão atualizados a cada alteração realizada em suas configurações.



Alguns elementos da interface são comuns a todas as análises. As barras contornadas em vermelho permitem expandir ou ocultar as telas do JASP. A barra da esquerda expande ou oculta o banco de dados e a barra da direita expande ou oculta os resultados. Você pode interagir com essas barras pressionando e arrastando os três pontos ou clicando nas setas.

Além destas barras, cinco ícones são apresentados no canto direito próximo ao nome da análise selecionada. Cada item possui uma função, a saber:

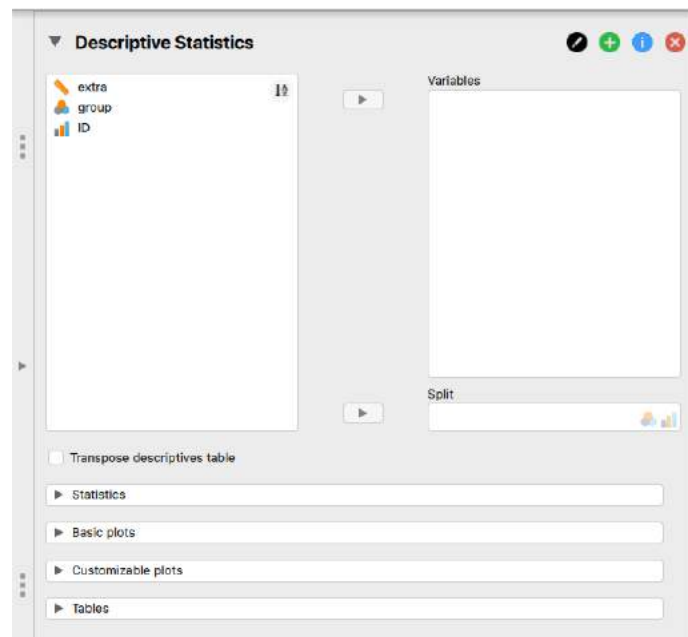


- R: apresenta a sintaxe em R da análise solicitada;
- Editar (lápiz): permite renomear a análise, o que pode facilitar a organização dos resultados;
- Duplicar (mais ou soma): cria uma cópia da análise;
- Informações (i): apresenta informações teóricas e referências relativas às técnicas utilizadas na análise selecionada;
- Remover (x): apaga a análise dos resultados.

A partir do próximo capítulo serão apresentadas algumas das análises estatísticas que podem ser realizadas no JASP.

## ANÁLISES DESCRITIVAS

As estatísticas descritivas envolvem medidas de tendência central, medidas de dispersão, análises da distribuição dos dados e tabelas de frequência. Para explorar essas medidas, abra o banco de dados no JASP e clique em *Descriptives* (Descritivos). Você será apresentado à tela ao lado.

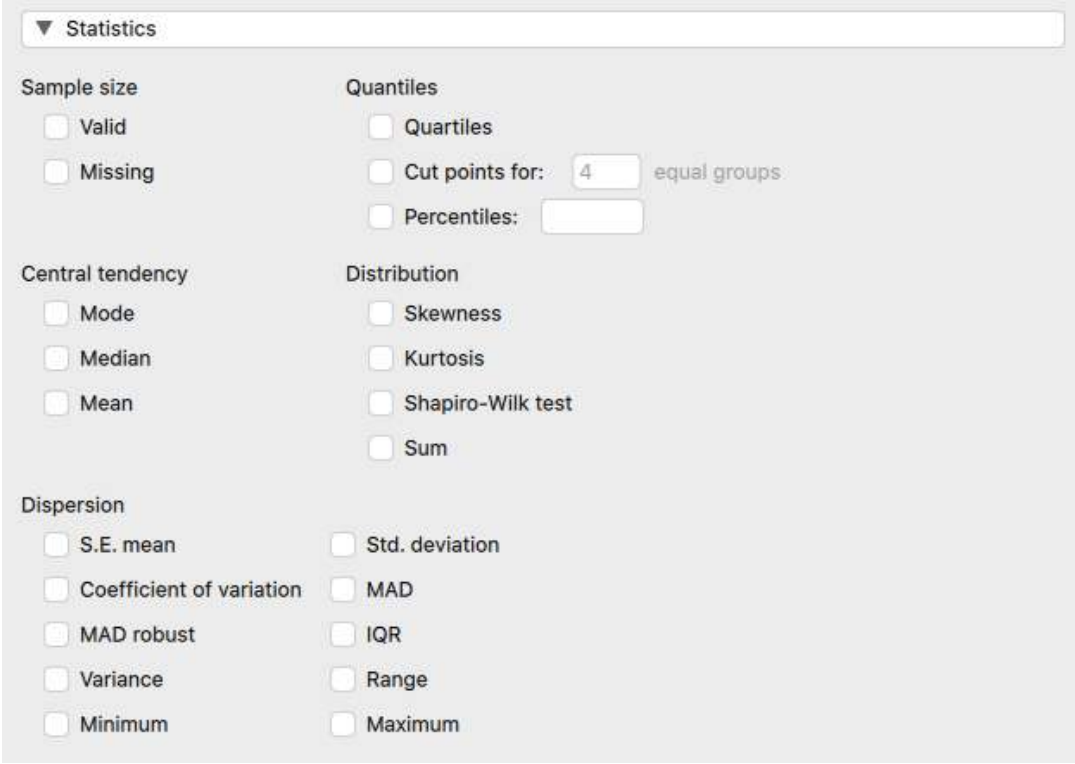


Na coluna à esquerda você terá acesso a todas as variáveis do seu banco de dados. Você pode selecionar as variáveis de interesse clicando nas variáveis e na seta que direciona a uma das colunas à direita ou pressionando e arrastando a variável para uma das colunas à direita. As estatísticas e análises serão geradas para todas as variáveis que você inserir na coluna *Variables* (Variáveis). A coluna *Split* (Dividir) pode ser utilizada para obter resultados divididos pelos valores de outra variável, contudo apenas variáveis nominais ou ordinais podem ser inseridas nesta coluna. Por exemplo, obter médias por grupo ou gênero.

Os quatro menus abaixo das colunas (*Statistics* ou Estatísticas, *Basic plots* ou Gráficos básicos, *Customizable plots* ou Gráficos customizáveis e *Tables* ou Tabelas) podem ser abertos ao clicar para acessar as diversas opções disponíveis para esse tipo de análise. Iniciando pelo menu de Estatísticas, são apresentados alguns grupos: *sample*

## 14 ANÁLISES ESTATÍSTICAS COM JASP: UM GUIA INTRODUTÓRIO

*size* (tamanho da amostra), *central tendency* (tendência central), *dispersion* (dispersão), *quantiles* (quantis) e *distribution* (distribuição).



The image shows a screenshot of the 'Statistics' panel in JASP. The panel is titled 'Statistics' and contains several sections of options, each with a dropdown arrow on the left. The sections are: 'Sample size' with 'Valid' and 'Missing' options; 'Central tendency' with 'Mode', 'Median', and 'Mean' options; 'Dispersion' with 'S.E. mean', 'Coefficient of variation', 'MAD robust', 'Variance', and 'Minimum' options; 'Quantiles' with 'Quantiles', 'Cut points for: 4 equal groups', and 'Percentiles' options; 'Distribution' with 'Skewness', 'Kurtosis', 'Shapiro-Wilk test', and 'Sum' options; and 'Std. deviation' with 'MAD', 'IQR', 'Range', and 'Maximum' options. All options are currently unchecked.

As opções disponíveis em cada grupo, assim como suas funções são:

- *Sample Size* (Tamanho da amostra): o usuário pode acessar dados descritivos dos dados válidos e valores omissos, assim verifica-se a integridade da amostra.
  - *Valid* (válidos): são considerados pelo software somente dados válidos, ou seja, variáveis com respostas;
  - *Missing* (ausentes): são considerados pelo software os dados faltosos, ou seja, casos em branco para cada variável.
- *Central tendency* (Tendência central): medida de valores numa amostra que identifica os diferentes pontos centrais nos dados, a identificação do valor central permite que outros valores sejam comparados a ele, mostrando a

dispersão ou o agrupamento da amostra, o que é conhecido como dispersão ou distribuição.

- *Mode* (moda): o valor mais frequente da variável na amostra;
- *Median* (mediana): depois de organizar os dados do menor para o maior, a mediana é o valor do meio, separando a metade inferior da metade superior do conjunto de dados;
- *Mean* (média): é igual à soma de todos os valores dividida pelo número de valores no conjunto de dados, ou seja, a média dos valores. É usada para descrever dados contínuos e fornece um modelo estatístico simples do centro de distribuição dos valores. É uma estimativa teórica do “valor típico”, no entanto, pode ser fortemente influenciado por pontuações “extremas”.
- *Dispersion* (dispersão): o tamanho da distribuição de valores num conjunto de dados, várias medidas de dispersão incluem um intervalo, como os quantis (por exemplo, quartis ou percentis) e os desvios padrão.
  - *Standard deviation* (desvio padrão): usado para quantificar a quantidade de dispersão dos valores dos dados em torno da média, um desvio padrão baixo indica que os valores estão próximos da média, ao passo em que um desvio padrão alto indica que os valores estão dispersos em uma faixa mais ampla;
  - *Variance* (variância): outra estimativa de quão longe os dados estão em relação a média, é calculado como o quadrado do desvio padrão;
  - *Standard error of the mean* (erro padrão da média): é uma medida de quão longe se espera que a média da amostra dos dados esteja da verdadeira média da população. A medida que o tamanho dos dados da amostra aumenta, o erro padrão diminui em comparação com o desvio padrão e a verdadeira média da população é conhecida com maior especificidade;
  - *MAD* (*median absolute deviation* ou desvio absoluto da mediana): medida robusta da disseminação de dados, similar ao desvio padrão

- que faz referência a mediana. É relativamente afetado por dados que não são normalmente distribuídos;
- *MAD Robust* (MAD Robusto): desvio absoluto mediano dos pontos de dados, ajustado por um fator para consistência assintoticamente normal;
  - *IQR* (intervalo interquartilico): é semelhante ao *MAD*, porém é menos robusto, apresenta o intervalo entre os quartis 1 e 3 (que dividem os dados em 25% e 75% respectivamente);
  - *Range* (alcance): representa a distância entre o maior e o menor valor existente nos dados;
  - *Minimum* (mínimo) e *Maximum* (máximo): representam, respectivamente, o menor e maior valor existente nos dados.
- *Quantiles* (quantis): são as separatrizes que dividem o conjunto em partes iguais;
    - *Quantiles* (quartis): opção mais comum, divide os dados em quatro partes. O primeiro quartil ou quartil inferior ( $Q_1$ ) é o valor do conjunto que delimita os 25% menores valores, em que 25% dos valores são menores do que  $Q_1$  e 75% são maiores do que  $Q_1$ . O segundo quartil ou quartil do meio é a própria mediana, que separa os 50% menores dos 50% maiores valores. O terceiro quartil ou quartil superior ( $Q_3$ ) é o valor que delimita os 25% maiores valores, em que 75% dos valores são menores do que  $Q_3$  e 25% são maiores do que  $Q_3$ .
  - *Distribution* (Distribuição): define uma curva (gráfico) e a área sob esta curva determina a probabilidade de ocorrer o evento associado à mesma;
    - *Skewness* (assimetria): descreve a mudança da distribuição de uma distribuição normal, a assimetria negativa mostra que a figura se move para a direita, resultando em uma cauda esquerda dominante e a assimetria positiva mostra que a figura se move para a esquerda, resultando em uma cauda direita dominante;
    - *Kurtosis* (curtose ou achatamento): descreve o quão pesadas ou leves são as caudas, a curtose positiva resulta em um aumento na “ponta” da distribuição com caudas pesadas (mais longas), ao passo



em que a curtose negativa exibe uma distribuição muito mais uniforme ou mais plana com caudas leves (mais curtas);

- *Shapiro-Wilk test* (teste de Shapiro-Wilk): para uma distribuição normal de dados ambos os valores devem ser próximos de zero, o teste de Shapiro-Wilk é usado para avaliar se os dados são ou não significativamente diferentes de uma distribuição normal, assim dados com distribuição normal apresentarão valor- $p$  não significativo ( $p > 0,05$ ).

Atualmente, o JASP também permite que sejam calculadas estatísticas inferenciais básicas durante as análises descritivas iniciais. O grupo *Inference* (ou inferência) apresenta as seguintes opções:

- *Standard error of the mean* (erro padrão da média): anteriormente apresentada no grupo dispersão;
- *Confidence interval* (intervalo de confiança): disponível para média, desvio padrão e variância, permitem

avaliar com base em um nível de confiança, que a medida obtida pela amostra se encontra dentro do intervalo provável em que se encontra a medida na população, tipicamente é utilizado o intervalo de 95% (equivalente ao valor- $p$  de 0,05), porém esse valor pode ser alterado.

- Para o intervalo de confiança da média, ainda é possível selecionar o método de cálculo deste intervalo, duas opções são fornecidas o modelo normal que utilizará os dados existentes para calcular o intervalo e a técnica de *Bootstrap* que irá gerar diversas amostras aleatórias a partir dos dados para cálculo desse intervalo. A quantidade de amostras a serem geradas pode ser definida em *Bootstrap samples*, sendo que mais amostragens irão representar maior confiança, porém também exigirão maior tempo para cálculo.

The image shows a screenshot of the 'Inference' settings panel in JASP. It contains several options with checkboxes and input fields:

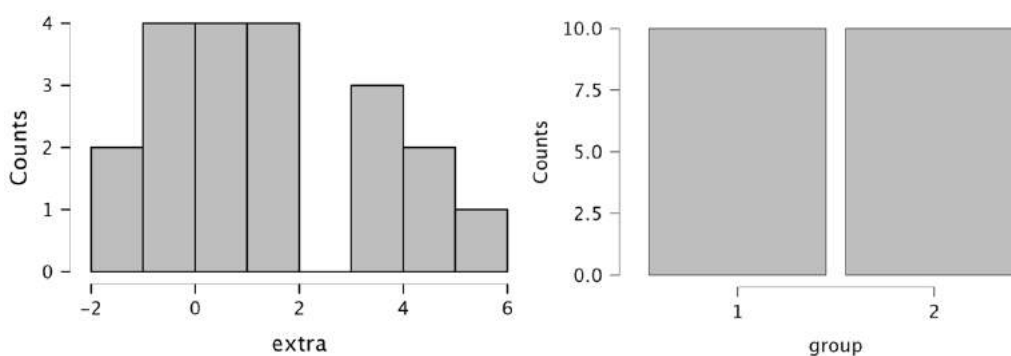
- S.E. mean
- Confidence interval for mean
  - Width: 95.0 %
  - Method: Normal model (dropdown menu)
- Confidence interval for std. deviation
  - Width: 95.0 %
- Confidence interval for variance
  - Width: 95.0 %
- Bootstrap confidence interval options
  - Bootstrap samples: 1000

Cabe ressaltar que o *Bootstrap* é considerado uma técnica mais robusta com relação ao modelo normal.

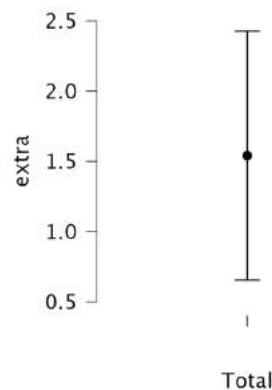
## GRÁFICOS

Após realizar os cálculos das estatísticas descritivas, é possível solicitar gráficos ilustrativos dos dados por meio dos gráficos básicos ou customizáveis. No menu de gráficos básicos será possível produzir gráficos de distribuição (*distribution*), intervalo (*interval*), pizza (*pie*), pontos (*dot*) e Q-Q. Também é possível criar gráficos categóricos, como o Pareto e *Likert*. Abaixo são apresentados alguns exemplos destes gráficos.

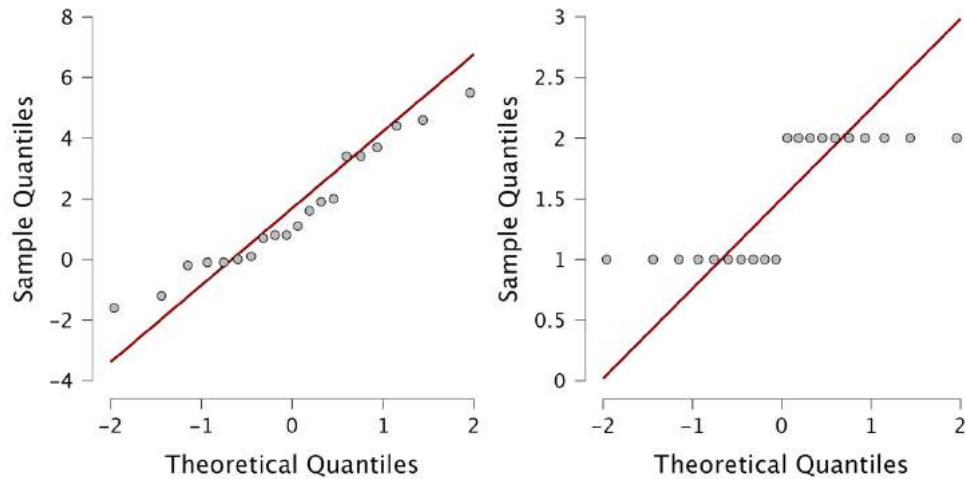
- Gráfico de distribuição: apresenta barras com frequência de cada categoria (dados nominais ou ordinais) ou intervalo (dados escalares), se solicitada a densidade (*display density*) será apresentada também a curva gerada por esses dados que permite comparação com a curva normal;



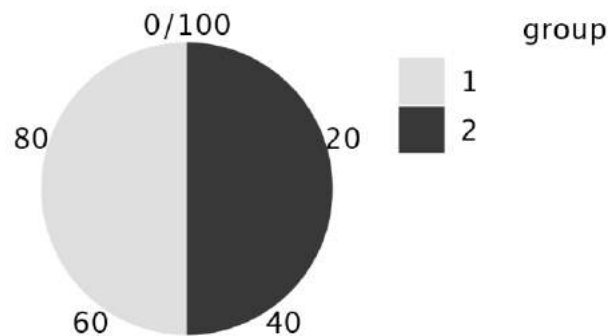
- Gráfico de intervalos: apresenta o intervalo considerando a média (ponto central), média menos desvio padrão (limite inferior) e média mais desvio padrão (limite superior);
- Gráfico Q-Q: apresenta a posição dos dados com relação aos quantis da amostra e teóricos, pode ser utilizado para verificar



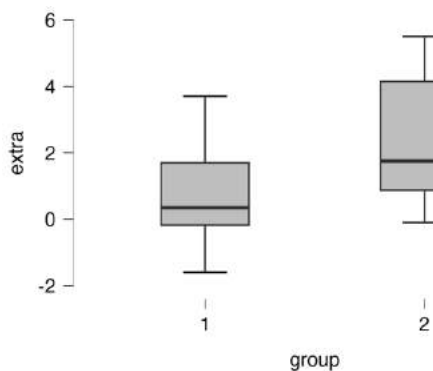
desvios da normalidade de dados escalares. Se os dados seguem uma distribuição normal, os pontos no gráfico Q-Q devem formar uma linha reta.



- Gráfico de pizza: apresenta a frequência relativa dos dados nominais;



No menu de gráficos customizáveis diversas opções avançadas estão disponíveis, contudo, destaca-se a possibilidade de gerar gráficos de caixa (*boxplots*) e gráficos de dispersão (*scatterplots*).

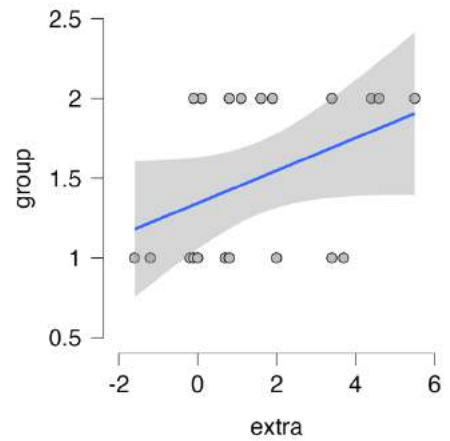


Os gráficos de caixa são úteis para comparação entre grupos, permitindo localizar os limites inferior e superior dos dados (linhas externas) e os três quartis: quartil inferior (linha inferior da caixa), mediana (linha que corta a divide a caixa) e quartil superior (linha superior da caixa), conforme imagem. As opções customizáveis são: *violin elemento* (elemento de violino), o qual adiciona elemento que representa de forma mais fidedigna a distribuição dos dados; *jitter elemento* (elemento de *jitter*), que adiciona os pontos dos dados ao gráfico;

de violino), o qual adiciona elemento que representa de forma mais fidedigna a distribuição dos dados; *jitter elemento* (elemento de *jitter*), que adiciona os pontos dos dados ao gráfico;

e *label outliers* (rotular outliers), que permite identificar quais são os casos extremos que podem influenciar as medidas de tendência central e dispersão dos dados.

Por fim, os gráficos de dispersão são interessantes para cruzamento de duas variáveis, este gráfico também pode ser customizável com a inclusão da reta de regressão (que pode ser linear ou suavizada [*smooth*]), intervalo de confiança indicado pela zona sombreada e gráficos de densidade ou histogramas para cada variável, apresentados acima e ao lado do gráfico principal.



## CORRELAÇÃO

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Correlação e seus pressupostos;
- Descrever as rotinas de comandos para a realização da análise de Correlação no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A correlação é uma técnica estatística que pode ser usada para determinar se, e com que intensidade, pares de variáveis estão associados. A correlação só é apropriada para dados quantificáveis nos quais os números são significativos, como dados contínuos ou ordinais. Não pode ser usada para dados puramente categóricos, para os quais temos que usar a análise de tabela de contingência (consulte a análise do qui-quadrado no JASP).

A correlação busca compreender como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas. Embora não implique em causalidade, o coeficiente de correlação exprime em números essa relação, ou seja, quantifica a relação entre as variáveis.

Para medir as relações entre as variáveis e o que elas estão representando são utilizados os métodos estatísticos chamados de coeficientes de correlação. O JASP possui três tipos de coeficientes de correlação, são eles:  $r$  de Pearson;  $\rho$  de Spearman; e  $\tau$  de Kendall.

O **coeficiente de correlação de Pearson** ( $r$ ), também chamado de correlação linear, exprime o grau de relação entre duas variáveis escalares por meio de um valor situado entre -1 e 1.

Quão mais próximo de -1 ou 1, mais forte será a relação entre as duas variáveis, sendo que uma correlação perfeita seria representada pelos valores -1 e 1. Correlações mais próximas de 0 indicam uma relação fraca entre as variáveis, sendo que 0 representa

a ausência de correlação. Correlações positivas indicam que o aumento do valor de uma variável leva ao aumento do valor da outra, assim como a redução no valor de uma variável leva a redução no valor da outra. Já correlações negativas indicam que o aumento do valor de uma variável leva a redução do valor da outra, assim como a redução no valor de uma variável leva ao aumento da outra.

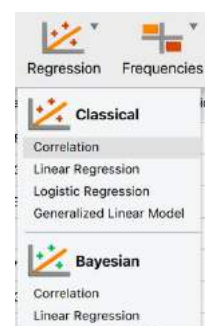
O **coeficiente de correlação de Spearman** ( $\rho$ ) é uma medida de correlação não paramétrica também avaliado entre -1 e 1. Esse coeficiente não exige que a relação entre as variáveis seja linear e permite a estimativa entre variáveis ordinais.

O **coeficiente de correlação de Kendall** ( $\tau$ ) é, assim como a correlação de Spearman, uma medida não paramétrica que não exige a linearidade dos dados e permite a análise de variáveis ordinais. Contudo, sua vantagem com relação ao coeficiente de Spearman está associado à sua possibilidade de utilização como um coeficiente de correlações parciais.

### ROTEIRO DE ANÁLISE: CORRELAÇÃO

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Big Five Personality Traits* que mensurou cinco traços de personalidade de 500 participantes. Nesse banco cinco variáveis, referentes aos traços de personalidade, estão disponíveis: *Neuroticismo* (neuroticismo), *Extraversão* (extroversão), *Openness* (abertura à experiência), *Agreeableness* (agradabilidade) e *Conscientiousness* (conscienciosidade).

Para executar a correlação, clique em *Regression* (regressão) no menu superior do JASP e no grupo *Classical* (clássico) selecione *Correlation* (correlação). Em seguida, na tela de configuração da análise, insira as variáveis de interesse para o quadro direito (*Variables* ou



variáveis), neste caso serão todas as variáveis. Abaixo são apresentadas as diversas opções para essa análise:

- **Sample Correlation Coefficient** (Coeficiente de Correlação da Amostra): permite determinar o coeficiente utilizado para o cálculo das correlações;

- **Alt. Hypothesis** (hipótese alternativa): permite determinar qual hipótese será testada: se há correlação (padrão e teste de significância bicaudal), se há correlação positiva (teste de significância unicaudal, aceita apenas correlações positivas) ou se há correlação negativa

(teste de significância unicaudal, aceita apenas correlações negativas);

- **Additional Options** (opções adicionais):
  - *Display pairwise* (mostrar par a par): altera a matriz de correlações nos resultados para uma tabela que apresenta cada correlação em uma linha;
  - *Report significance* (relatar significância): apresenta o teste de significância estatística (valor-p) da análise;
  - *Flag significant correlations* (marcar correlações significantes): será incluído nos resultados uma marcação, por meio de asteriscos, da significância estatística dos resultados;
  - *Confidence Intervals* (intervalos de confiância): solicita a apresentação do intervalo de confiança dos coeficientes de correlação e permite determinar o nível de confiança e se deve ser utilizada a técnica de Bootstrapping (reamostragem);
  - *Effect Size (Fisher's z)* (Tamanho de Efeito [z de Fisher]): apresenta o tamanho de efeito z de Fisher junto aos resultados, contudo cabe ressaltar

que o  $r$  de Pearson também pode ser considerada uma medida de tamanho de efeito;

- *Sample size* (Tamanho da amostra): apresenta o tamanho amostral para cada par de variáveis de acordo com os casos válidos e ausentes;
- *Plots* (gráficos): opções para apresentação gráfica dos resultados;
  - *Scatter plots* (gráficos de dispersão): apresenta gráficos de dispersão para cada par de variáveis no triângulo superior da matriz;
    - *Densities for variables* (densidade das variáveis): apresenta histograma e densidade para cada variável individual;
    - *Statistics* (estatísticas): apresenta o valor da correlação entre as variáveis no triângulo inferior da matriz;
    - *Confidence intervals* (intervalos de confiança): apresenta os intervalos de confiança da correlação;
    - *Prediction intervals* (intervalos de predição): apresenta os intervalos de predição da correlação;
  - *Heat map* (mapa de calor): exibe as correlações em formato de mapa de calor;

Dois menus de opções extras também estão disponíveis. O menu *Assumption Checks* (checagem de pressupostos) permite solicitar o teste multivariado (considerando todas as variáveis selecionadas) e o teste pareado (considerando apenas pares de variáveis) de normalidade de Shapiro-Wilk. Já o menu *Options* (opções) permite selecionar como tratar os dados ausentes (*Missing values*) com as opções *Exclude cases pairwise* (excluir casos par a par) e *Exclude cases listwise* (excluir casos em lista).

Para esse banco de dados iremos solicitar correlações de Pearson com a hipótese alternativa de que as variáveis são correlacionadas, a marcação da significância estatística, a apresentação do mapa de calor e os testes de normalidade multivariada e par a par de Shapiro-Wilk.

Dos resultados apresentados na tela, primeiro devemos verificar os testes de normalidade de Shapiro-Wilk para averiguar a normalidade dos dados, pressuposto da análise de correlação. O teste de normalidade multivariada de Shapiro-Wilk indicou desvio da normalidade com significância abaixo de 0,05 (0,99;  $p = 0,006$ ), contudo como as correlações são calculadas par a par, esse não é o teste mais adequado. A tabela abaixo apresenta o teste de Shapiro-Wilk para cada par de variáveis, na qual podemos verificar que apenas os pares de variáveis neuroticismo-conscienciosidade, extroversão-abertura à



mudança, extroversão-conscienciosidade e abertura à mudança-conscienciosidade apresentaram normalidade bivariada. Dessa forma, os outros pares precisariam ser transformados para garantir a normalidade necessária para a análise.

**Shapiro-Wilk Test for Bivariate Normality**

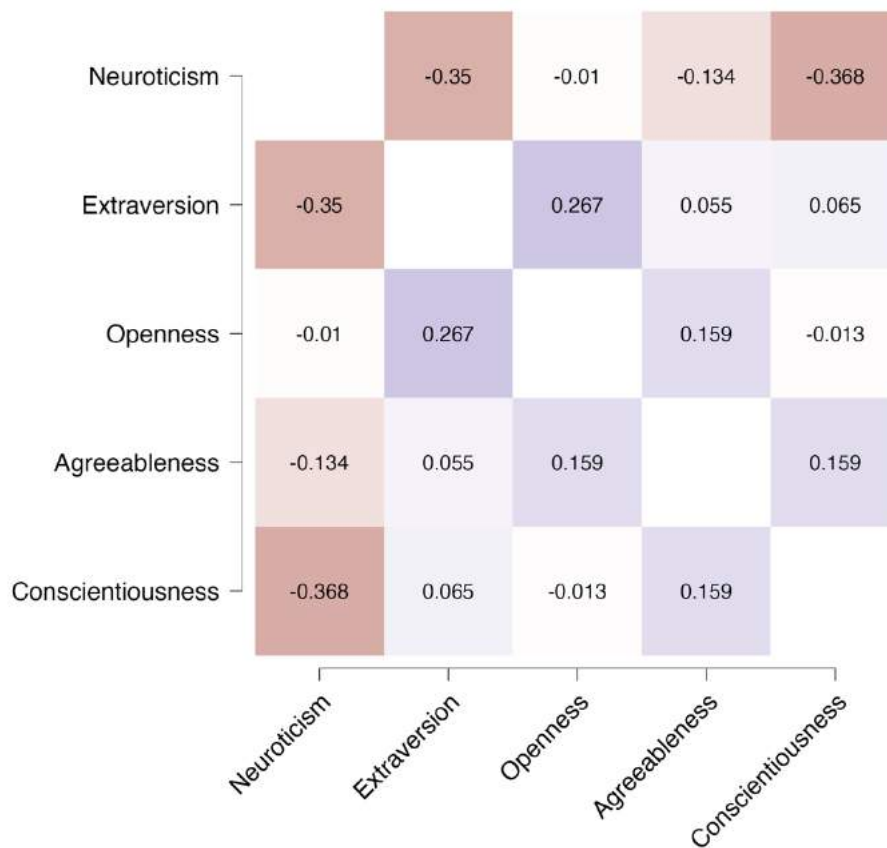
		Shapiro-Wilk	p
Neuroticism	- Extraversion	0.992	0.007
Neuroticism	- Openness	0.987	< .001
Neuroticism	- Agreeableness	0.985	< .001
Neuroticism	- Conscientiousness	0.998	0.924
Extraversion	- Openness	0.997	0.496
Extraversion	- Agreeableness	0.988	< .001
Extraversion	- Conscientiousness	0.994	0.068
Openness	- Agreeableness	0.991	0.003
Openness	- Conscientiousness	0.998	0.774
Agreeableness	- Conscientiousness	0.988	< .001

Os outros resultados referem-se a matriz de correlações e sua versão no mapa de calor. Os quais estão apresentados abaixo.

**Pearson's Correlations**

Variable	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
1. Neuroticism	—				
2. Extraversion	-0.350 ***	—			
3. Openness	-0.010	0.267 ***	—		
4. Agreeableness	-0.134 **	0.055	0.159 ***	—	
5. Conscientiousness	-0.368 ***	0.065	-0.013	0.159 ***	—

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$



## RELATO DOS RESULTADOS

A matriz de correlações apontou para associações positivas e significativas entre os traços de personalidade de extroversão e abertura a mudanças ( $r = 0,27$ ;  $p < 0,001$ ), abertura a mudanças e agradabilidade ( $r = 0,16$ ;  $p < 0,001$ ) e agradabilidade e conscienciosidade ( $r = 0,16$ ;  $p < 0,001$ ). Também foram identificadas associações negativas e significativas entre o traço de personalidade de neuroticismo com os traços de personalidade: extroversão ( $r = -0,35$ ;  $p < 0,001$ ), conscienciosidade ( $r = -0,37$ ;  $p < 0,001$ ) e agradabilidade ( $r = -0,13$ ;  $p < 0,01$ ).

## REGRESSÃO LINEAR SIMPLES

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a análise de Regressão Linear Simples e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Regressão Linear Simples no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A regressão linear é uma equação para se estimar a condicional (valor esperado) de uma variável  $y$ , dados os valores de outras variáveis  $x_n$ . A regressão, em geral, tem como objetivo tratar de um valor que não se consegue estimar inicialmente.

Enquanto os testes de correlação verificam a associação entre variáveis, a regressão é a próxima etapa comumente utilizada para análises preditivas. Ou seja, para prever o resultado de uma variável dependente a partir de uma (regressão simples) ou mais (regressão múltipla) variáveis preditoras independentes. O cálculo da regressão é a base para diversas outras análises comumente utilizadas e apresentadas neste livro.

A regressão linear simples apresenta como pressupostos: que a variável dependente (ou critério) seja escalar; linearidade, a variável dependente é associada de forma linear as variáveis independentes (ou preditoras); aditividade, em que os efeitos das variáveis independentes (ou preditoras) é aditivo; independência dos resíduos, na qual os resíduos não são correlacionados entre si, isto é, não há variação comum entre os dados que não seja explicada por outras variáveis do modelo; homogeneidade das variâncias (ou homoscedasticidade), em que a variância do erro de cada variável independente (ou

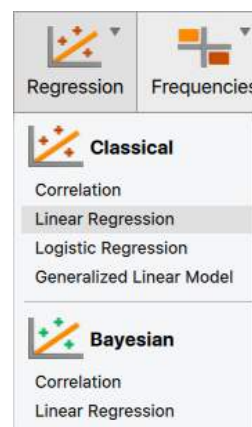
preditora) é constante para todos os valores dessa variável; e a normalidade dos resíduos, na qual os resíduos são normalmente distribuídos e apresentam média 0.

## ROTEIRO DE ANÁLISE: REGRESSÃO LINEAR SIMPLES

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Album Sales* que mensurou fatores que poderiam afetar a venda de 200 álbuns diferentes. Nesse banco quatro variáveis estão disponíveis: *adverts* (anúncios), o valor (em milhares de libras) gasto na promoção do álbum antes do lançamento; *sales* (vendas), as vendas (em milhares de cópias) de cada álbum na semana após o lançamento; *airplay*, quantas vezes as músicas do álbum foram tocadas em uma importante estação de rádio nacional na semana anterior ao lançamento; e *attract* (atração), quão atraente as pessoas acharam a imagem da banda (variando de 1 a 10).

Para realizar a análise, deve-se selecionar a opção *Regression* (regressão) no menu superior do JASP e em seguida selecionar *Linear Regression* (regressão linear) no grupo *Classical* (clássico). Na tela de configuração da análise, múltiplas opções estão disponíveis. Primeiro iremos inserir nossa variável dependente (*sales* ou vendas) no quadro denominado *Dependent Variable* (variável dependente) e para a regressão linear simples utilizaremos apenas uma variável independente (*adverts* ou anúncios) que será inserida no quadro *Covariates* (covariáveis). As outras opções não serão utilizadas na regressão linear simples, apenas na regressão linear múltipla.

O menu *Model* (modelo) nos permite definir os componentes e termos do modelo de regressão a ser estimado. Contudo, para o modelo simples não precisaremos realizar



nenhuma alteração nesse menu. O menu abaixo, *Statistics* (estatísticas), permite solicitar informações e resultados sobre os coeficientes e resíduos da análise.

- *Estimates* (estimativas): apresenta o coeficiente da regressão linear ( $b$ ) para cada variável independente, podendo ser solicitado seu cálculo a partir da técnica de reamostragem (*bootstraps*);
- *Confidence intervals* (intervalos de confiança): apresenta os intervalos de confiança para os coeficientes da regressão linear no nível determinado (padrão, 95%);
- *Covariance matrix* (matriz de covariância): apresenta a matriz de covariância para as variáveis independentes do modelo;
- *Model fit* (ajuste do modelo): apresenta uma tabela similar a ANOVA para cada modelo estimado, no caso da regressão simples será estimado apenas o modelo nulo e o modelo proposto;
- *R squared change* (mudança no  $R^2$ ): apresenta a mudança no valor do  $R^2$  para cada modelo estimado;
- *Descriptives* (descritivos): apresenta os dados descritivos para cada variável incluída no modelo (tamanho amostral, média, desvio padrão e erro padrão da média);
- *Part and partial correlations* (correlações semiparciais e parciais): essa opção não apresenta resultados para a regressão simples;
- *Collinearity diagnostics* (diagnósticos de colinearidade): opção utilizada apenas na regressão múltipla para avaliar (multi)colinearidade;
- *Residuals* (resíduos): grupo de estatísticas relacionadas aos resíduos (erros) do modelo.
  - *Statistics* (estatísticas): apresenta as estatísticas residuais (mínimo, máximo, média, desvio padrão e tamanho da amostra);
  - *Durbin-Watson*: apresenta os resultados do teste de Durbin-Watson que testa a autocorrelação dos resíduos, a independência dos

resíduos ocorrerá apenas caso o resultado do teste não seja significativo;

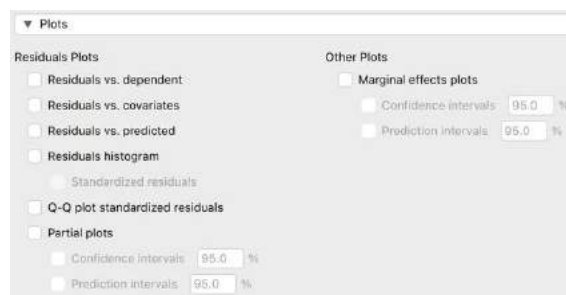
- *Casewise diagnostics*: permite realizar a análise de *outliers* (casos extremos) a partir da apresentação dos casos que tiverem resíduos padronizados acima de 3, distância de Cook acima de 1 ou todos os casos.

O menu *Method Specification* (especificação do método) também não será relevante para a análise de regressão simples, portanto não é necessário. Por fim, o menu *Plots* (gráficos) permite gerar diversos gráficos importantes para complementar a análise de regressão ou possibilitar a análise de seus pressupostos, suas opções estão apresentadas abaixo:

- *Residual plots* (gráficos de resíduos):

os resíduos devem se distribuir aleatoriamente sobre uma linha horizontal para que os pressupostos da regressão linear sejam aceitos, qualquer padrão ou agrupamento sistemático indica quebra dos pressupostos;

- *Residuals vs. dependent*: gráfico de dispersão dos resíduos em relação à variável dependente;
- *Residuals vs. covariates*: gráfico de dispersão dos resíduos em relação às variáveis independentes (ou preditoras);
- *Residuals vs. predicted*: gráfico de dispersão dos resíduos em relação aos valores previstos pelo modelo;
- *Residuals histogram*: histograma constando da distribuição dos resíduos, pode ser apresentado utilizando os resíduos padronizados;
- *Q-Q plot standardized residuals* (gráfico Q-Q dos resíduos padronizados): permite avaliar se os resíduos são normalmente distribuídos;
- *Partial plots* (Gráficos parciais) e *Marginal effects plots* (gráficos dos efeitos marginais): relevantes apenas na regressão múltipla.



Portanto, a presente análise será realizada utilizando o modelo inicialmente proposto e sendo solicitado: estimativas (*Estimates*), intervalos de confiança (*confidence intervals*), estatísticas descritivas (*descriptives*), estatísticas residuais (*statistics*), teste de Durbin-Watson, diagnósticos caso a caso (*casewise diagnostics*) por meio da distância de

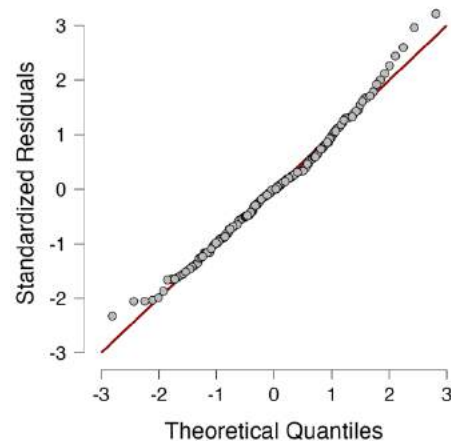
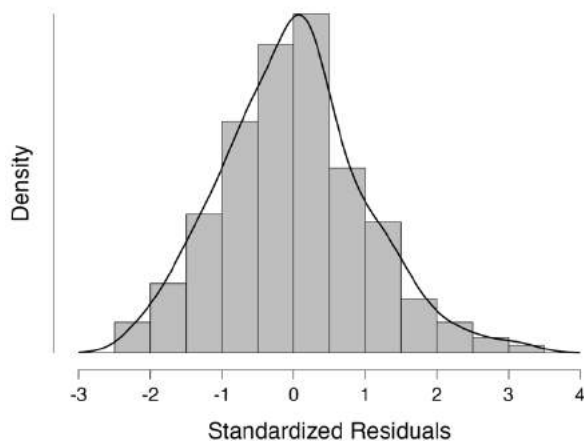
Cook maior que 1 e os gráficos residuais com os valores preditos, histograma com resíduos padronizados e Q-Q dos resíduos padronizados.

Iniciaremos nossa análise pela tabela de diagnósticos de casos, já que essa informa sobre os casos extremos e influentes do modelo que podem afetar outros pressupostos. Conforme tabela abaixo, não foi identificado nenhum caso extremo com distância de Cook maior que 1.

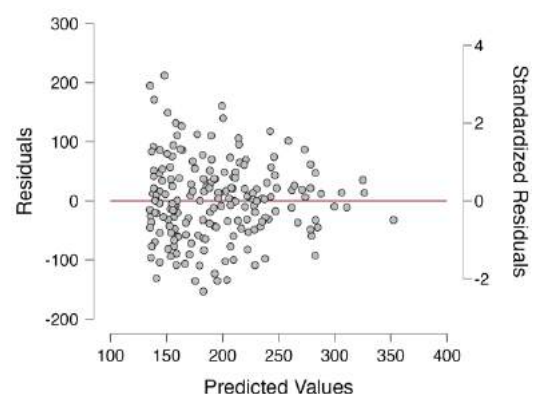
#### Casewise Diagnostics

Case Number	Std. Residual	sales	Predicted Value	Residual	Cook's Distance
.	.	.	.	.	.

Em seguida, avaliaremos os gráficos residuais. O histograma dos resíduos padronizados e o gráfico Q-Q dos resíduos padronizados apresentam de forma diferente indícios sobre o mesmo pressuposto, a normalidade dos resíduos. Com base nestes podemos verificar um pequeno desvio nos extremos, contudo o mesmo não é considerado suficiente para causar problemas na estimação do modelo.



O último gráfico a ser analisado é referente aos resíduos vs. valores preditos. Esse gráfico não deve apresentar qualquer formato ou padrão de distribuição para que não haja desvios na normalidade, linearidade ou homogeneidade das variâncias dos resíduos. Com base na figura abaixo, nota-se que não há desvios de linearidade ou normalidade, contudo há um desvio na



homogeneidade da variância, formando uma distribuição triangular, como pode ser observado ao lado.

Desvios na homogeneidade dos resíduos podem ser corrigidos por meio de transformações dos dados ou métodos robustos de estimação, como as correções *sandwich*. As últimas não estão disponíveis para o JASP. Por ora, seguiremos com a análise do modelo conforme formulado.

A primeira tabela a ser analisada será o sumário do modelo (*Modelo summary*). Nela podemos verificar que o modelo proposto ( $H_1$ ) explica 33,5% da variância das vendas ( $R^2 = 0,335$ ;  $R^2_{ajustado} = 0,331$ ) e que não foram identificadas autocorrelações no modelo, ou seja, os resíduos são independentes.

#### Model Summary - sales

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	RMSE	Durbin-Watson		
					Autocorrelation	Statistic	p
H <sub>0</sub>	0.000	0.000	0.000	80.699	0.086	1.807	0.171
H <sub>1</sub>	0.578	0.335	0.331	65.991	-0.044	2.032	0.823

Por fim, verificaremos a tabela dos coeficientes. Novamente, os resultados que nos interessam são referentes ao modelo  $H_1$ . Especificamente, podemos verificar que a variável anúncios explica significativamente a variável vendas e, com base no coeficiente padronizado (*Standardized*) verificamos que isso está associado a um efeito médio ( $\beta = 0,578$ ;  $p < 0,001$ ). Cabe ressaltar que o *Intercept* (intercepto) representa a estimativa da variável dependente quando todas as variáveis independentes não têm efeito sobre ela. Portanto, é importante considerar os valores médios das variáveis independentes ao interpretar o valor do intercepto.

#### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI	
							Lower	Upper
H <sub>0</sub>	(Intercept)	193.200	5.706		33.857	< .001	181.947	204.453
H <sub>1</sub>	(Intercept)	134.140	7.537		17.799	< .001	119.278	149.002
	adverts	0.096	0.010	0.578	9.979	< .001	0.077	0.115

## RELATO DOS RESULTADOS

Foi realizada uma análise de regressão simples considerando a venda (em milhares de libras) como variável dependente e os gastos em publicidade (em milhares de libras) como variável independente. Esse modelo apresentou aderência aos pressupostos da análise, com exceção de apresentar heteroscedasticidade. Os gastos com publicidade



explicaram 33,5% da variância das vendas ( $R^2$  ajustado = 0,33). De forma que um aumento nos gastos com publicidade é diretamente proporcional a um aumento nas vendas ( $\beta = 0,58$ ;  $p < 0,001$ ), compreende-se ainda que as vendas previstas irão corresponder a 134,14 + (0,096 \* cada milhar de libras adicionalmente gasto em publicidade).

## REGRESSÃO LINEAR MÚLTIPLA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a análise de Regressão Linear Múltipla e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Regressão Linear Múltipla no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

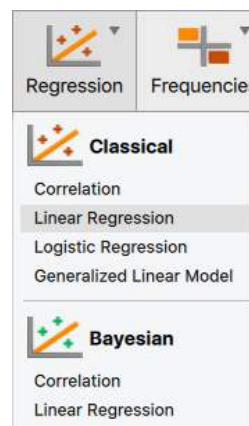
A regressão linear múltipla é, assim como a regressão linear simples, uma equação para estimar a condicional de uma variável  $y$  dados os valores de outras variáveis  $x_n$ . A principal diferença entre a regressão linear simples e múltipla diz respeito a quantidade de variáveis independentes (ou preditoras). Na regressão linear múltipla pode haver mais de uma variável do tipo escalar, ordinal ou categórica. Os mesmos pressupostos da regressão linear simples valem para a regressão linear múltipla, contudo ela apresenta um pressuposto adicional: ausência de (multi)colinearidade entre as variáveis independentes (ou preditoras).

### ROTEIRO DE ANÁLISE: REGRESSÃO LINEAR MÚLTIPLA

Para essa análise será utilizado o mesmo banco de dados do capítulo anterior, disponível na Biblioteca de Dados do JASP e denominado *Album Sales* que mensurou fatores que poderiam afetar a venda de 200 álbuns diferentes. Nesse banco quatro variáveis estão disponíveis: *adverts* (anúncios), o valor (em milhares de libras) gasto na promoção do álbum antes do lançamento; *sales* (vendas), as vendas (em milhares de cópias) de cada álbum na semana após o lançamento; *airplay*, quantas vezes as músicas do álbum foram tocadas em uma importante estação de rádio nacional na semana anterior

ao lançamento; e *attract* (atração), quão atraente as pessoas acharam a imagem da banda (variando de 1 a 10).

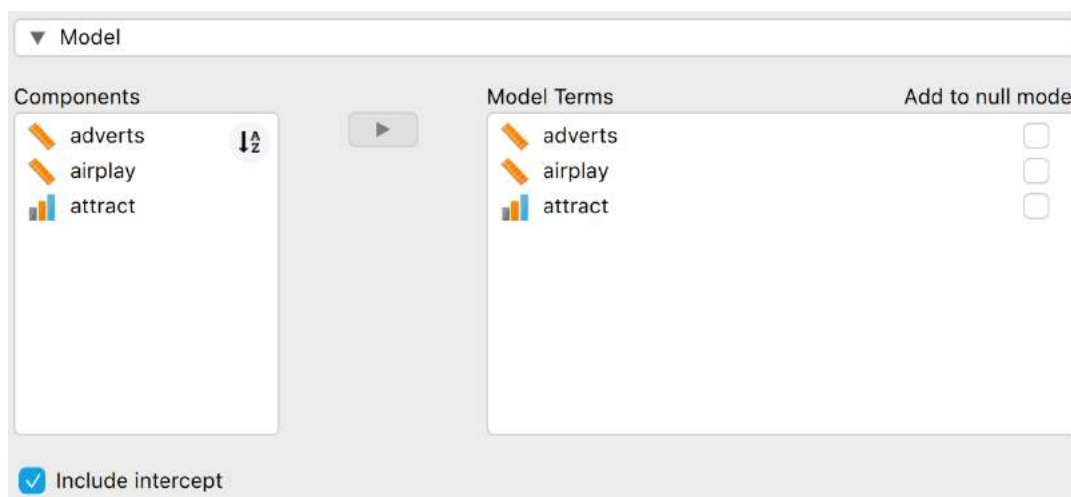
Para realizar a análise, deve-se selecionar a opção *Regression* (regressão) no menu superior do JASP e em seguida selecionar *Linear Regression* (regressão linear) no grupo *Classical* (clássico). Na tela de configuração da análise, múltiplas opções estão disponíveis. Primeiro iremos inserir nossa variável dependente (*sales* ou vendas) no quadro denominado *Dependent Variable* (variável dependente) e como variáveis independentes serão utilizadas *adverts*, *airplay* e *attract* que serão inseridas no quadro *Covariates* (covariáveis). Neste caso não utilizaremos as outras opções, que podem ser úteis na regressão linear múltipla, a saber: o quadro *Factors* (fatores) permite a inserção de variáveis categóricas no modelo; o quadro *WLS Weights* (pesos dos quadrados mínimos ponderados) permite determinar pesos para a ponderação dos cálculos dos quadrados mínimos; as opções *Methods* (métodos) permitem determinar a melhor forma de estimar o modelo, são elas apresentadas abaixo.



- *Enter* (inserir): todas as variáveis são inseridas no modelo simultaneamente;
- *Backward*: todas as variáveis são inseridas no modelo simultaneamente e, em seguida, novos modelos são estimados por meio da remoção de uma variável a cada passo com base no critério selecionado no menu *Method Specification* (especificação do método);
- *Forward*: novos modelos são estimados por meio da inclusão de uma variável a cada passo com base no critério selecionado no menu *Method Specification* (especificação do método);
- *Stepwise* (passo a passo): novos modelos são estimados por meio da inclusão de uma variável a cada passo com base no critério selecionado no menu *Method Specification* (especificação do método), contudo a cada etapa um preditor de menor relevância pode ser removido.

Para essa análise será utilizado o método *Enter* (inserir). No menu *Model* (modelo) algumas opções importantes se apresentam para modificar o modelo estimado. O quadro *Components* (componentes) apresenta todas as variáveis incluídas no modelo, o quadro *Model Terms* (termos do modelo) apresenta quais variáveis serão inseridas no modelo. Caso selecione duas ou mais variáveis ao mesmo tempo no quadro *Components* e as arraste ao quadro *Model Terms*, você criará interações (ou moderações) que são

representadas por “variável  $x_1$  \* variável  $x_2$ ”. Além disso, selecionando as caixas das variáveis abaixo do texto *Add to null model* (adicionar ao modelo nulo) no quadro *Model Terms* você pode incluir variáveis ao modelo nulo de forma que o modelo proposto seja comparado a esse modelo alterado. Nenhuma alteração ao modelo será realizada para essa análise.



No menu *Statistics* (estatísticas) as mesmas opções da regressão linear simples serão marcadas, a saber: *Estimates*, *confidence intervals*, *model fit*, *Durbin-Watson*, diagnósticos caso a caso por distância de Cook maior que 1. Adicionalmente será selecionado a opção de correlações parciais e semi-parciais que serão adicionadas aos resultados (*Part and partial correlations*) e os diagnósticos de colinearidade que permitirão testar o pressuposto de colinearidade da regressão linear múltipla. O menu *Method Specification* (especificação do método) é utilizado em caso do uso de outros métodos com exceção do inserir, tendo como opções o critério de inserção e remoção por meio do valor  $p$  (significância estatística) ou por meio do valor  $F$  (ANOVA).

Na tela dos resultados, inicialmente deve-se atentar aos testes de pressupostos, os quais foram tratados no capítulo de regressão linear simples. Para isso verificaremos a coluna *Collinearity Statistics* (estatísticas de colinearidade) na tabela *Coefficients* (coeficientes). A estatística de *Tolerance* (tolerância) deve ser acima de 0,1 e a estatística *VIF* deve ser abaixo de 10 para não indicar colinearidade, dessa forma não foram identificados problemas de colinearidade entre as variáveis.

O modelo explicou 66,5% da variância das vendas (variável *sales*),  $R^2 = 0,665$ ;  $R^2$  ajustado = 0,66. Ainda, por meio da tabela *Coefficients* (coeficientes) é possível notar que todas as variáveis auxiliam a explicar significativamente as vendas dos discos ( $p < 0,001$ ),

	<b>Tolerance</b>	<b>VIF</b>
<b>adverts</b>	0.986	1.015
<b>airplay</b>	0.959	1.043
<b>attract</b>	0.963	1.038

sendo que o tempo de reprodução de músicas do disco em rádios (*airplay*) é a variável que melhor explica a quantidade de vendas ( $\beta = 0,512$ ), conforme pode ser notado na tabela abaixo.

#### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI	
							Lower	Upper
H <sub>0</sub>	(Intercept)	193.200	5.706		33.857	< .001	181.947	204.453
H <sub>1</sub>	(Intercept)	-26.613	17.350		-1.534	0.127	-60.830	7.604
	adverts	0.085	0.007	0.511	12.261	< .001	0.071	0.099
	airplay	3.367	0.278	0.512	12.123	< .001	2.820	3.915
	attract	11.086	2.438	0.192	4.548	< .001	6.279	15.894

## RELATO DOS RESULTADOS

Foi realizada uma análise de regressão linear múltipla tendo as vendas (em milhares de libras) como variável dependente e os gastos com anúncios (em milhares de libras), o tempo de reprodução em rádio e a atratividade da imagem da banda como variáveis independentes. Esse modelo apresentou significância estatística com relação ao modelo nulo e explicou 66,5% da variância das vendas ( $R^2$  ajustado = 0,66). Todas as variáveis independentes auxiliaram a prever o aumento das vendas (isto é, são diretamente associadas às vendas de discos). Verifica-se que o tempo de reprodução em rádio ( $\beta = 0,51$ ;  $p < 0,001$ ) e o gasto em publicidade ( $\beta = 0,51$ ;  $p < 0,001$ ) explicaram em magnitude similar e proporcionalmente superior o número de vendas de discos do que a atratividade da imagem da banca ( $\beta = 0,19$ ;  $p < 0,001$ ).

## REGRESSÃO LOGÍSTICA BINÁRIA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a análise de Regressão Logística Binária e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Regressão Logística Binária no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A regressão logística é um modelo estatístico que permite estimar a chance da ocorrência de um determinado desfecho categórico ( $y$ ) em função de um ou mais preditores ( $x_n$ ), que podem ser contínuos ou categóricos. Quando a variável dependente apresenta apenas dois níveis ou classes, a regressão é chamada de binária. Quando há mais níveis ou classes, é chamada de multinomial.

É possível entender a regressão logística como um complemento da regressão linear aplicada a variáveis dependentes categóricas a partir de uma função de ligação, uma generalização do teste do qui-quadrado ( $\chi^2$ ) ou, de maneira geral, um caso particular da família dos modelos lineares generalizados (GLM), que implementa uma ligação *logit*. Há muitas pesquisas em que modelos logísticos são utilizados. Em áreas de saúde, eles tendem a ser os modelos de primeira escolha para verificar condições clínicas e de agravo à saúde. Em psicologia, por exemplo, ele é útil para descrever e investigar os possíveis preditores de condições diagnósticas bem definidas, tais como TDAH e outros transtornos psiquiátricos. Em epidemiologia, eles são fundamentais em estudos do tipo caso-control.

Essas características fazem com que os resultados da regressão logística informem sobre chances (*Odds*) e Razão de chances (*Odds Ratio*) e não sobre probabilidades

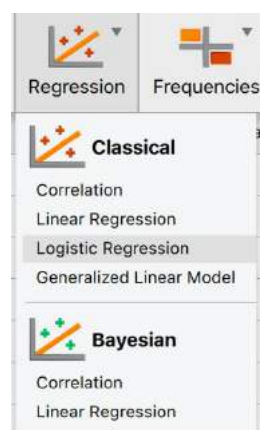
(Riscos ou *Risks*) diretamente. Existem muitas formas de demonstrar esta diferença. Entretanto, a forma mais simples é por sua estrutura matemática.

Compreendendo isto, a regressão logística binária apresentará pressupostos diferentes da regressão linear. Conforme já mencionado, a variável dependente deverá ser categórica e possuir apenas dois níveis (binária). Ainda, a relação entre as variáveis independentes (contínuas ou categorias) e o *logit* da variável dependente deve ser linear. Por fim, outros pressupostos são comuns a regressão linear, sendo eles: independência das observações (ou resíduos), ausência de (multi)colinearidade entre as variáveis independentes e ausência de casos extremos (*outliers*).

## ROTEIRO DE ANÁLISE: REGRESSÃO LOGÍSTICA BINÁRIA

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Titanic* que fornece informações sobre 1.313 passageiros do Titanic. Nesse banco cinco variáveis estão disponíveis: *Name*, nome do passageiro; *PClass*, classe do passageiro; *Age*, idade do passageiro; *Sex*, sexo do passageiro (se é homem ou mulher); e *Survived* (sobreviveu), se o passageiro sobreviveu (1) ou morreu (0) no naufrágio. Por meio desse banco de dados estaremos interessados em avaliar, especificamente, em que medida a idade do passageiro e a classe de sua cabine predizem a chance de sobrevivência do passageiro.

Para realizar a análise, deve-se selecionar a opção *Regression* (regressão) no menu superior do JASP e em seguida selecionar *Logistic Regression* (regressão logística) no grupo *Classical* (clássico). Na tela de configuração da análise, múltiplas opções estão disponíveis, algumas similares às apresentadas nos capítulos sobre regressão linear simples e múltipla. Primeiro iremos inserir nossa variável dependente (*Survived*) no quadro denominado *Dependent Variable* (variável dependente) e como variáveis independentes serão utilizadas *Age* (idade) que será inserida no quadro *Covariates* (covariáveis) e *PClass* (classe do passageiro) no quadro *Factors* (fatores), pois a última é uma variável categórica. O método (*Methods*) utilizado será o padrão, inserir (*Enter*).



Assim como apresentado no capítulo de regressão logística, o menu *Model* (modelo) permite criar interações (ou moderações), incluir variáveis ao modelo nulo e

solicitar a apresentação dos interceptos. Para essa análise não será feita nenhuma modificação neste menu. O menu *Statistics* (estatísticas) apresentam diversas opções diferentes da regressão linear e serão apresentados abaixo:

The screenshot shows the 'Statistics' menu in JASP, which is organized into several sections:

- Descriptives:** Includes an unchecked checkbox for 'Factor descriptives'.
- Coefficients:** Includes a checked checkbox for 'Estimates', a text input for 'From 5000 bootstraps', and several unchecked checkboxes for 'Standardized coefficients', 'Odds ratios', 'Confidence intervals' (with an interval of 95.0%), 'Odds ratio scale', 'Robust standard errors', 'Vovk-Sellke maximum p-ratio', and 'Multicollinearity Diagnostics'.
- Residuals:** Includes an unchecked checkbox for 'Casewise diagnostics' and three radio button options: 'Standard residual > 3', 'Cook's distance > 1', and 'All'.
- Performance Diagnostics:** Includes an unchecked checkbox for 'Confusion matrix'.
- Performance Metrics:** Includes several unchecked checkboxes for 'Accuracy', 'AUC', 'Sensitivity / Recall', 'Specificity', 'Precision', 'F-measure', 'Brier score', and 'H-measure'.

- *Descriptives* (descritivos): a opção *Factor descriptives* (descritivos dos fatores) apresenta a quantidade de participantes para cada variável categórica incluída como antecedente no modelo;
- *Coefficients* (coeficientes):
  - *Estimates* (estimativas): apresenta os coeficientes da regressão logística binária, pode ser solicitado que seu calculo seja baseado em reamostragem (bootstraps), seja apresentada versão padronizada (*Standardized coefficients*), a Razão de Chance (*Odds Ratio*) que é um tamanho de efeito para a regressão logística, os

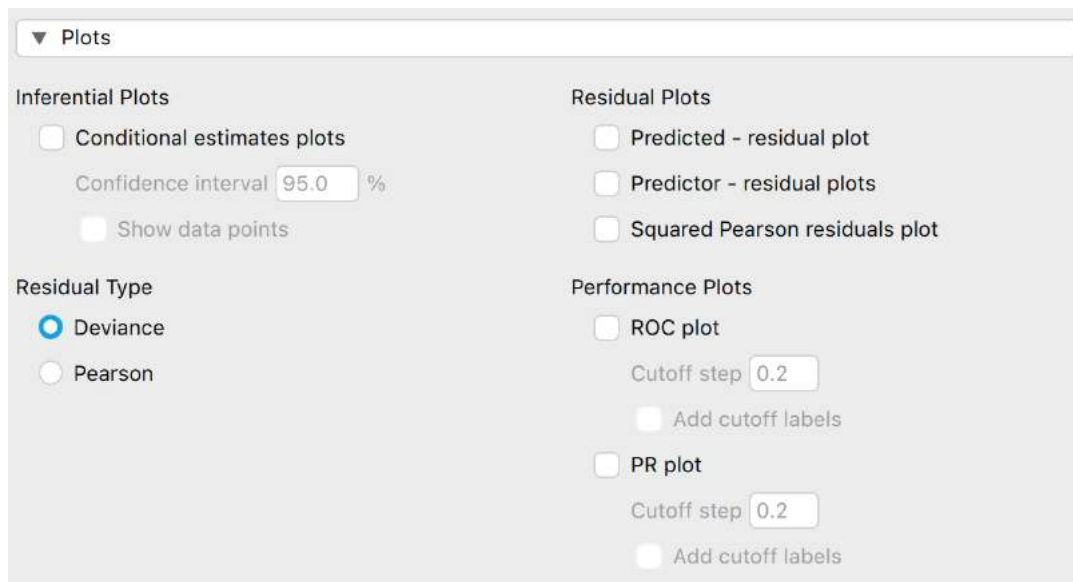


- intervalos de confiança (*Confidence Intervals*) e seja solicitado que o cálculo envolva erros padrão robustos (*robust standard errors*);
- *Multicollinearity diagnostics* (diagnósticos de multicolinearidade): apresenta os diagnósticos com base no *Tolerance* (tolerância) e VIF;
  - *Residuals* (resíduos): diagnósticos caso a caso (*casewise diagnostics*) permite que identifiquemos casos extremos (*outliers*) e influentes;
  - *Performance diagnostics* (diagnósticos de performance): permite solicitar a matriz de confusão (*Confusion matrix*) que apresenta em quantos casos o valor observado foi predito pelo modelo corretamente ou incorretamente, sendo assim um diagnóstico da adequação do poder de predição do modelo, a partir dessa matriz algumas métricas de performance podem ser calculadas;
  - *Performance metrics* (métricas de performance):
    - *Accuracy* (acurácia): a proporção de casos corretamente preditos pelo modelo;
    - *AUC* (*Area Under the Curve* ou área abaixo da curva): indica o poder preditivo do modelo, considerando  $AUC = 1,0$  como predição perfeita

e  $AUC = 0,5$  como ausência de poder preditivo (pois é igual ou inferior a chance aleatória);

- *Sensitivity / Recall* (sensibilidade): indica a proporção de verdadeiros positivos (1-1) identificados na matriz;
- *Specificity* (especificidade): indica a proporção de verdadeiros negativos (0-0) identificados na matriz;
- *Precision* (precisão): indica a proporção de verdadeiros positivos (1-1) em todos os positivos (0-1 e 1-1);

Por fim, o menu *Plots* (gráficos) pode ser utilizado para obter diversos gráficos diagnósticos e representativos do modelo. As opções serão melhor apresentadas abaixo.



- *Inferential Plots* (gráficos inferenciais): apresenta os gráficos inferenciais para cada variável independente, similar a linha de regressão para variáveis contínuas e a gráficos de caixa (boxplots) para variáveis categóricas;
- *Residual Plots* (gráficos residuais): permitem a análise dos pressupostos de linearidade do *logit* quando os resíduos se distribuem de forma similar a uma linha horizontal, com desvios a esse pressuposto sendo representados por padrões ou agrupamentos;
  - *Predicted – residuals plot*: gráfico residual dos valores preditos;
  - *Predictors – residuals plot*: gráfico residual dos preditores (variáveis independentes);
  - *Squared Pearson residuals* (resíduos do quadrado de Pearson): permite avaliar a dispersão dos resíduos, uma grande dispersão

indica que há maior variabilidade nos dados do que a mensurada pelo modelo;

- *Performance Plots* (gráficos de performance):
  - *ROC plot* (gráfico ROC): gráfico com curva indicativa do poder de predição do modelo, a área abaixo da curva gera a estimativa AUC.

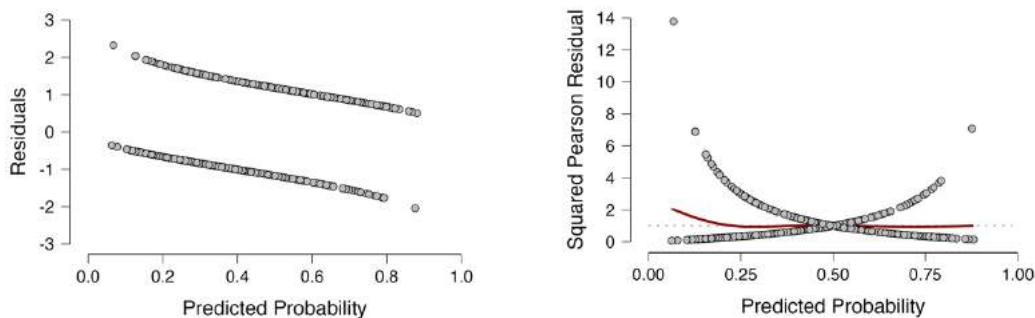
Dadas as opções, nessa análise solicitaremos as estimativas, as Razões de Chance, os intervalos de confiança com a escala de razões de chance, diagnósticos de multicolinearidade, diagnósticos caso a caso, a matriz de confusão e as métricas de performance de acurácia, AUC, sensibilidade, especificidade e precisão. Na tela dos resultados, iniciaremos pela análise dos pressupostos da regressão logística.

A tabela de diagnósticos de multicolinearidade **Multicollinearity Diagnostics** (*Multicollinearity Diagnostics*) apresentou tolerâncias maiores que 0,1 e VIF menores que 10, portanto não foi identificada multicolinearidade entre as variáveis. A

	<b>Tolerance</b>	<b>VIF</b>
Age	0.705	1.418
PClass	0.705	1.418

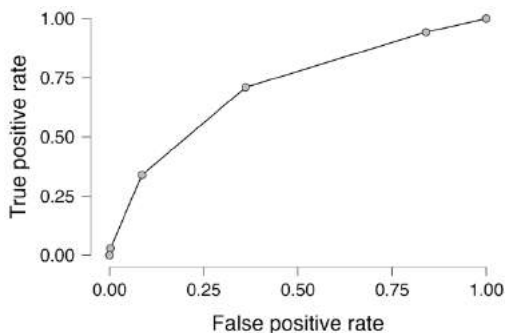
A tabela de diagnósticos caso a caso (*Casewise Diagnostics*) também não apresentou nenhum caso, de forma que não foi identificado nenhum caso extremo.

Com relação aos gráficos residuais, verifica-se que o gráfico de resíduos dos valores preditos se assemelha a uma linha horizontal, não havendo desvio da linearidade do *logit* (figura abaixo a esquerda). O gráfico de resíduos quadrados de Pearson indicaram variância acima do predito pelo modelo, contudo não aparenta ser suficiente para justificar problemas com tal pressuposto (figura abaixo a direita).



Por fim, o gráfico da curva ROC sugere que o poder preditivo do modelo é superior a chance aleatória (diagonal), conforme gráfico abaixo. Isso pode ser confirmado por meio

do índice AUC na tabela *Performance metrics* (métricas de performance), o qual foi superior a 0,5 (AUC = 0,72).



Por fim, analisaremos os resultados sobre o modelo de regressão logística binária. O modelo proposto apresentou uma melhora significativa em relação ao modelo nulo, conforme pode ser verificado na tabela abaixo por meio do teste do qui-quadrado significativo e da redução nos índices de AIC e BIC. Por fim, vários cálculos

diferentes para o  $R^2$  estão disponíveis na regressão logística, os quais apresentam resultados ligeiramente diferentes. Cabe ao pesquisador avaliar qual índice é mais adequado para seus dados e análise.

#### Model Summary - Survived

	Model Deviance	AIC	BIC	df	X <sup>2</sup>	p	McFadden R <sup>2</sup>	Nagelkerke R <sup>2</sup>	Tjur R <sup>2</sup>	Cox & Snell R <sup>2</sup>
H <sub>0</sub>	1025.573	1027.573	1032.201	755						
H <sub>1</sub>	909.917	917.917	936.429	752	115.656	< .001	0.113	0.191	0.148	0.142

Buscando compreender como as variáveis independentes impactam e auxiliam individualmente a explicar o modelo é importante analisar os resultados apresentados na tabela *Coefficients* (coeficientes).

#### Coefficients

	Estimate	Standard Error	Odds Ratio	z	Wald Test			95% Confidence interval (odds ratio scale)	
					Wald Statistic	df	p	Lower bound	Upper bound
(Intercept)	2.034	0.307	7.642	6.627	43.918	1	< .001	4.188	13.946
Age	-0.039	0.007	0.962	-5.882	34.602	1	< .001	0.950	0.975
PClass (2nd)	-1.146	0.220	0.318	-5.202	27.062	1	< .001	0.206	0.490
PClass (3rd)	-2.232	0.229	0.107	-9.745	94.962	1	< .001	0.069	0.168

Note. Survived level '1' coded as class 1.

Os resultados dessa tabela apontam que a idade (*Age*) e a segunda e terceira classe (*PClass*, considerando a primeira classe como valor de referência) foram significativos em explicar a sobrevivência dos passageiros ( $p < 0,001$ ). Ademais, as estimativas (*Estimate*) indicam que essa relação é negativa, ou seja, uma idade mais avançada e classes mais elevadas estão relacionadas a menor chance de sobrevivência. Ainda, considerando o tamanho de efeito (*Odds Ratio* ou razão de chance), é importante saber que: valores abaixo de 1 indicam relação negativa, enquanto valores acima de 1

indicam relação positiva com a variável dependente; ainda pode-se ler que pessoas com mais idade possuíam 0,96 (96,2%) mais chances de morrer no naufrágio, por exemplo.

Por fim, é importante avaliar a performance do modelo. Este apresentou forte especificidade (0,79) e poder preditivo (AUC = 0,72), além de acurácia (0,67) e precisão (0,63) medianos e uma baixa sensibilidade (0,51). Conforme tabela e matriz de confusão abaixo apresentadas.

<b>Performance metrics</b>		<b>Confusion matrix</b>			
	<b>Value</b>		<b>Predicted</b>		
		<b>Observed</b>	<b>0</b>	<b>1</b>	<b>% Correct</b>
Accuracy	0.672	0	350	93	79.007
AUC	0.721	1	155	158	50.479
Sensitivity	0.505	Overall % Correct			67.196
Specificity	0.790	<i>Note.</i> The cut-off value is set to 0.5			
Precision	0.629				

## RELATO DOS RESULTADOS

Foi realizada uma regressão logística binária para avaliar se a idade dos passageiros e a classe em que estavam prediria a chance de sobrevivência, ou não, dos mesmos ao desastre do Titanic. Os resultados apontam que o modelo foi significativo em prever a sobrevivência dos passageiros ( $\chi^2 [752] = 115,66$ ;  $p < 0,001$ ;  $R^2$  Nagelkerke = 0,19). Dessa forma, a chance de sobrevivência dos passageiros aumenta 96,2% quão mais jovens eles são ( $b = -0,04$ ;  $p < 0,001$ ;  $OR = 0,96$ ) e também aumenta 31,8% se forem da primeira classe em relação a segunda classe ( $b = -1,15$ ;  $p < 0,001$ ;  $OR = 0,32$ ) e apenas 10,7% se forem da primeira classe em relação a terceira classe ( $b = -2,32$ ;  $p < 0,001$ ;  $OR = 0,11$ ). Entende-se, portanto, que passageiros mais jovens e da primeira classe sobreviveram mais, contudo destaca-se também que passageiros da terceira classe tiveram maior chance de sobrevivência do que os passageiros da segunda classe. Esse modelo apresentou forte especificidade (0,79) e poder preditivo (AUC = 0,72), além de acurácia (0,67) e precisão (0,63) medianos e uma baixa sensibilidade (0,51).

## ANÁLISE DE VARIÂNCIA ONE-WAY

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Análise de Variância e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Análise de Variância no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A Análise de Variância (comumente conhecida como ANOVA) é um teste paramétrico que compara médias de três ou mais grupos. É equivalente ao Teste  $t$ , utilizado para dois grupos. O teste inferencial da ANOVA avalia a probabilidade de que as diferenças entre grupos sejam devido a um erro amostral. A rejeição da hipótese nula (não há diferenças entre os grupos) em favor da hipótese alternativa indica que essas diferenças se devem ao grupo, ou seja, o efeito do grupo identificado é estatisticamente significativo sobre os resultados dos participantes.

Pode-se dizer que a ANOVA procura diferenças entre médias de grupos. Quando as médias são diferentes, existe um alto grau de variação entre as condições. Se não existirem diferenças entre as médias dos grupos, não existe variação. As variações entre e dentro dos grupos surgem como efeito de tratamentos ou experimentos, diferenças individuais e erro experimental. Para analisar essas variações a ANOVA compara a variância entre grupos e a variação intragrupo ou dentro do grupo.

A referida análise apresenta múltiplas possibilidades de aplicações nos mais diversos campos de estudo. Dessa forma, a técnica estatística se mostra como robusta ferramenta vastamente utilizada no contexto acadêmico e profissional. A ANOVA, como qualquer análise estatística, baseia-se em alguns pressupostos matemáticos que são necessários para que os resultados sejam confiáveis. Os pressupostos estatísticos da ANOVA são: independência das observações (ou resíduos), a variável dependente possuir

distribuição normal e homogeneidade das variâncias (ou homoscedasticidade) entre os grupos.

## ROTEIRO DE ANÁLISE: PRESSUPOSTOS

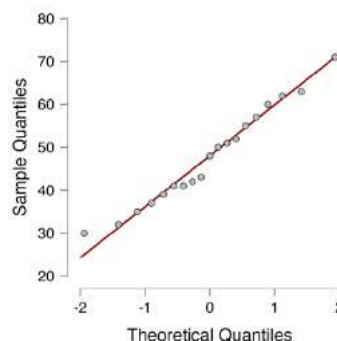
Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Pain Threshold* que mensurou a tolerância a dor de 19 participantes que tinham cores de cabelo diferentes. Nesse banco três variáveis estão disponíveis: *V1* (número de identificação do participante), *Hair Color* (cor de cabelo do participante), *Pain Tolerance* (nível de tolerância a dor do participante).

A ANOVA é um teste estatístico univariado, ou seja, que possui variável dependente única, a qual também deve ser do tipo escalar. Já as variáveis independentes (*fixed factors* ou fatores fixos, no JASP) devem ser nominais ou intervalares. No caso da ANOVA One-Way haverá apenas uma variável independente.

Inicialmente precisaremos realizar a análise dos pressupostos citados. O primeiro pressuposto a ser avaliado é a distribuição normal dos dados, que pode ser avaliado por diferentes formas no JASP, como: o teste de Shapiro-Wilk, assimetria, curtose e gráfico de probabilidade normal por quantis Q-Q. Para isso, acesse a análise Descritivos (*Descriptives*) no menu superior, inclua a variável dependente (*Pain Tolerance*) ao quadro de Variáveis (*Variables*) e selecione no menu Estatísticas (*Statistics*) as opções: *Skewness* (assimetria), *Kurtosis* (curtose) e *Shapiro-Wilk Test* (teste de Shapiro-Wilk). Ainda, para gerar o gráfico Q-Q, selecione no menu Gráficos Básicos (*Basic plots*) o *Q-Q plots* (Gráfico Q-Q). Na tela de resultados será gerada uma Tabela e um Gráfico, conforme abaixo:

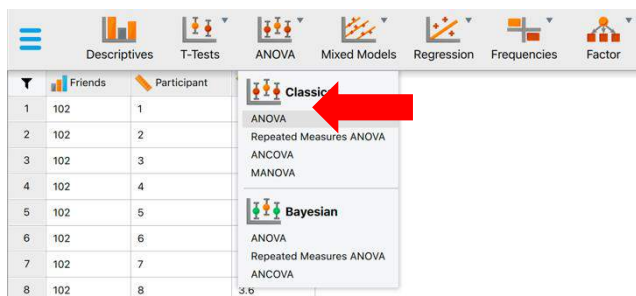
### Descriptive Statistics

	<b>Pain Tolerance</b>
Skewness	0.293
Std. Error of Skewness	0.524
Kurtosis	-0.735
Std. Error of Kurtosis	1.014
Shapiro-Wilk	0.971
P-value of Shapiro-Wilk	0.800



Considerando o teste de Shapiro-Wilk, pode-se compreender que a variável Pain Tolerance apresenta distribuição normal, pois o valor- $p$  é maior que 0,05, ou seja, não foi encontrada diferença significativa entre a distribuição dos dados dessa variável e a distribuição normal. Já os testes de assimetria e curtose vão indicar desvio da normalidade quão mais distante forem de 0, alguns autores sugerem que valores acima de 1 ou 2 já indicariam algum desvio da normalidade. De toda forma, a variável Pain Tolerance apresentou distribuição muito próxima a normal. Esse resultado é corroborado no Gráfico Q-Q, o qual apresenta apenas pequenos desvios dos dados (pontos) da reta (distribuição normal).

O segundo pressuposto a ser avaliado é a homogeneidade das variâncias, o qual pode ser analisado por meio do teste de Levene. Para isso é necessário estimar o modelo da ANOVA. No menu superior clique em



ANOVA e no grupo *Classical* (Clássico) clique novamente em ANOVA, conforme imagem ao lado. Na tela de análise você irá inserir as variáveis conforme indicado a seguir:

- Variável dependente (*Dependent Variable*): Pain Tolerance;
- Fatores fixos (*Fixed Factors* ou variável independente): Hair Color.

Em seguida desça abra o menu *Assumption Checks* (Checagem de Pressupostos) e selecione os *Homogeneity tests* (testes de homogeneidade). Uma Tabela será gerada na página de resultados constando do teste de Levene. Esse teste avalia se há diferença significativa na variância da variável dependente (Pain Tolerance) para cada grupo (variável independente, Hair Color), dessa forma um valor- $p$  menor que 0,05 (neste caso  $p = 0,69$ ) indica não haver diferença significativa na variância entre os grupos, ou seja, as variâncias dos dados são homogêneas.

#### Test for Equality of Variances (Levene's)

F	df1	df2	p
0.493	3.000	15.000	0.692

Caso identifique em seus dados a quebra do pressuposto de homogeneidade dos dados, o JASP possui duas correções que permitem o uso da ANOVA nestes casos. Para isso selecione, na tela de análise, no menu de Checagem dos Pressupostos, a correção



de homogeneidade (*Homogeneity correction*) desejada Brown-Forsythe ou Welch. Livros e artigos estão disponíveis sobre os diferentes tipos de correções existentes.

O último pressuposto da ANOVA, relativo à independência das observações (ou resíduos) não pode ser testado por meio do JASP até a presente versão. Contudo, a dependência das observações ocorrerá quando alguns casos (ou participantes) apresentarem variação similar por motivo não considerado ou controlado, comumente isso ocorre quando existem agrupamentos comuns na amostra, como parte dos participantes fazerem parte de uma mesma turma, formação ou grupo. Ademais, quando o mesmo participante é exposto às duas condições de agrupamento, os resultados apresentarão uma variação comum relacionada ao participante. Neste, e em outros casos, métodos robustos para tal desvio estão disponíveis, como a ANOVA de Medidas Repetidas (*Repeated Measures ANOVA*) ou os modelos mistos (*Mixed Models*).

## ROTEIRO DE ANÁLISE: ANOVA

Verificados os pressupostos da análise, é possível prosseguirmos para a ANOVA. Na tela de análise, selecione, abaixo dos quadros de variáveis, as estatísticas descritivas (*Descriptive statistics*) e estimativas de tamanhos de efeito (*estimates of effect size*). Selecione também o eta quadrado ( $\eta^2$ ) e o eta quadrado parcial (partial  $\eta^2$ ).

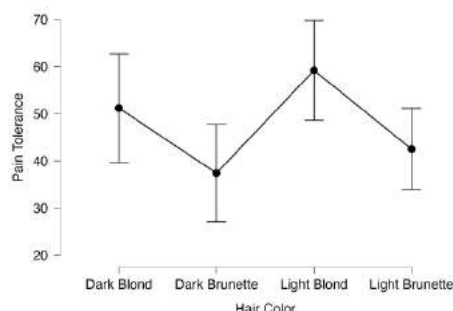
### ANOVA - Pain Tolerance

Cases	Sum of Squares	df	Mean Square	F	p	$\eta^2$	$\eta^2_p$
Hair Color	1360.726	3	453.575	6.791	0.004	0.576	0.576
Residuals	1001.800	15	66.787				

*Note.* Type III Sum of Squares

O resultado da ANOVA é representado pelo teste  $F$ , o qual indica haverem diferenças significativas entre as médias dos grupos quando o valor- $p$  for menor que 0,05. A diferença entre as médias dos grupos também pode ser verificada por meio de um gráfico de intervalos, para isso inclua a variável independente (Hair Color) no quadro *Horizontal Axis* (Eixo Horizontal) e marque a opção *Display error bars* (Mostre barras de erro) no menu *Descriptives Plots* (Gráficos Descritivos).

O gráfico ao lado será gerado. Neste gráfico os pontos representam a média dos grupos e as barras representam o limite inferior e superior dos intervalos de confiança (neste caso em 95%). A diferença entre



as médias dos grupos será significativa sempre que a média de um grupo não estiver dentro do intervalo de confiança do outro. Dessa forma, o gráfico de intervalos confirma a existência de diferenças entre os grupos de cor de cabelo (Hair Color) para a variável de tolerância a dor (Pain Tolerance).

Ainda com relação a Tabela, algumas estimativas do tamanho de efeito são apresentadas. O eta quadrado ( $\eta^2$ ) indica o tamanho do efeito do modelo testado, ou também o quanto da variável dependente (neste caso Pain Tolerance) é explicado pelas variáveis independentes (neste caso Hair Color). Um tamanho de efeito maior sugere que esse efeito seria melhor identificado na população ou considerando variáveis intervenientes, já um efeito menor pode não ocorrer na população. Alguns autores sugerem pontos de corte para considerar o eta quadrado pequeno ( $\eta^2 < 0,01$ ), médio ( $0,02 < \eta^2 < 0,06$ ) ou grande ( $\eta^2 > 0,14$ ).

O eta quadrado parcial (partial  $\eta^2$ ) é uma medida de tamanho de efeito que permite identificar o tamanho de efeito de uma variável independente na variável dependente. Dessa forma, na ANOVA One-Way o eta quadrado parcial será sempre igual ao eta quadrado (partial  $\eta^2 = \eta^2$ ). Devido a forma de seu calculo para remover a variância explicada por outras variáveis, o eta quadrado parcial não pode ser utilizado como medida de variância explicada da variável dependente.

Por fim, sendo identificadas diferenças significativas entre os grupos, é importante compreender quais grupos apresentam tais diferenças. Para isso pode-se utilizar de: testes por contraste, que podem ser configurados no menu *Contrasts* (Contrastes) da tela de análise e devem ser utilizados sempre que não houver sentido na comparação entre todos os grupos; ou testes Post Hoc, menu *Post Hoc tests* que permitem comparar todos os grupos.

No menu *Post Hoc tests*, mova a variável independente (Hair Color) para o bloco da direita e, em seguida, selecione as opções de interesse. As opções estão apresentadas abaixo:

- *Type* (Tipo): permite selecionar o tipo de teste post hoc a ser realizado;
  - *Standard* (Padrão): pré-selecionado automaticamente, esse teste é o mais usual e permite solicitar que seja executado a partir de técnica de Bootstrap (reamostragem) e que apresente o respectivo tamanho de efeito (*d* de Cohen);
  - *Games-Howell*, *Dunnett* e *Dunn*: são testes post hoc apropriados para quando há quebra do pressuposto de homogeneidade das

variâncias, diferenças, vantagens e desvantagens de cada teste devem ser consultados.

- *Correction* (Correção): nos testes post hoc várias estatísticas inferenciais serão calculadas com o mesmo conjunto de dados, para evitar que isso aumente a taxa de Erro Tipo I uma correção ao valor-*p* é realizada.
  - As diversas opções possuem vantagens e desvantagens, cabendo ao pesquisador avaliar qual será mais adequada ou interessante para os dados, contudo as correções mais comuns são de Tukey e Bonferroni.
- *Display* (Exibir):
  - *Confidence intervals* (Intervalos de confiança): permite incluir os intervalos de confiança para a diferença média entre os grupos e os tamanhos de efeito;
  - *Flag significant comparisons* (marcar comparações significativas): adiciona asteriscos (\*) para comparações significativas ao final da Tabela.

Para o caso da presente análise foi selecionado um teste post hoc com correção de Tukey, tendo sido solicitado os tamanhos de efeito das comparações e que comparações significativas fossem marcadas. A Tabela abaixo representa os resultados do teste.

#### Post Hoc Comparisons - Hair Color

		Mean Difference	SE	t	Cohen's d	<i>p</i> <sub>Tukey</sub>
Dark Blond	Dark Brunette	13.800	5.169	2.670	1.689	0.074
	Light Blond	-8.000	5.169	-1.548	-0.979	0.436
	Light Brunette	8.700	5.482	1.587	1.065	0.415
Dark Brunette	Light Blond	-21.800	5.169	-4.218	-2.668	0.004**
	Light Brunette	-5.100	5.482	-0.930	-0.624	0.789
Light Blond	Light Brunette	16.700	5.482	3.046	2.043	0.037*

*Note.* P-value adjusted for comparing a family of 4

\*  $p < .05$ , \*\*  $p < .01$

Na Tabela nota-se que foram identificadas diferenças significativas ( $p < 0,05$ ) apenas entre os grupos *Dark Brunette* e *Light Blond* e os grupos *Light Blond* e *Light Brunette*. O sentido (positivo ou negativo) dos valores do teste *t* e do tamanho de efeito (*d* de Cohen) apenas indicam qual grupo apresentou média maior. Valores positivos indicam

que o grupo da primeira coluna apresentou média maior, enquanto valores negativos indicam que o grupo da segunda coluna apresentou média maior. Essa comparação também deve ser qualificada pelos resultados descritivos da média e desvio padrão de cada grupo, conforme apresentado na Tabela ao lado.

## RELATO DOS RESULTADOS

Foram identificadas diferenças significativas na tolerância a dor entre pessoas com cores de cabelo diferente ( $F[15; 3] = 6,79; p = 0,004$ ) e essa diferença apresentou tamanho de efeito grande ( $\eta^2 = 0,58$ ). Pessoas com cor de cabelo loiro claro ( $M = 59,20; DP = 8,53$ ) apresentaram maior tolerância média a dor em relação a pessoas com cores de cabelo castanho claro ( $M = 42,50; DP = 5,45; t = 3,05; p = 0,04$ ) e pessoas com cores de cabelo escuro ( $M = 37,40; DP = 8,33; t = 4,22; p = 0,004$ ).

## ANCOVA: ANÁLISE DE COVARIÂNCIA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Análise de Covariância e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Análise de Covariância no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A Análise de Covariância (ANCOVA) é uma extensão da ANOVA que avalia os efeitos principais e as interações entre variáveis independentes, após a remoção estatística dos efeitos de uma covariável sobre a variável dependente. Dessa forma, essa técnica está fundamentada na análise de variância e na regressão. Em síntese, uma ANCOVA simples indica se os escores dos grupos diferem em uma variável dependente enquanto se mantêm fixos os efeitos (sobre a variável dependente) de uma outra variável denominada covariável. Uma covariável é uma variável que apresenta um relacionamento linear com a variável dependente. Dessa forma, a ideia é eliminar estatisticamente os efeitos de uma variável, a covariável sobre a variável dependente.

Em síntese, a ANCOVA iguala os grupos em relação a covariável para que se possa mensurar apenas os efeitos da variável de grupo sobre a variável dependente. Um dos requisitos para a ANCOVA é que a variável dependente esteja associada significativamente a covariável. Dessa forma, a ANCOVA apresenta pressupostos similares a ANOVA, sendo estes: normalidade da distribuição dos dados; homogeneidade da variância (ou homoscedasticidade) entre os grupos; homogeneidade das retas de

regressão; linearidade da relação entre variável dependente e covariáveis para cada nível das variáveis independentes.

## ROTEIRO DE ANÁLISE: PRESSUPOSTOS

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Viagra* que mensurou a libido de 30 participantes após o uso de doses diferentes de viagra. Nesse banco três variáveis estão disponíveis: *dose* (dose de viagra que cada participante tomou), *libido* (nível da libido do participante após uso do medicamento) e *partnerLibido* (nível da libido do/a parceiro/a do participante).

Assim como na ANOVA, a ANCOVA vai possuir uma variável dependente que deve ser do tipo escalar (a variável *libido*). As variáveis independentes (*fixed factors* ou fatores fixos no JASP) devem ser nominais ou intervalares, já que são as variáveis de agrupamento (neste caso, a variável *dose*). Por fim, temos também as covariáveis (*covariates*), aquelas sobre as quais desejamos controlar os efeitos na variável dependente (aqui, a variável *partnerLibido*).



Para a análise dos pressupostos, iniciaremos pela análise dos pressupostos específicos a ANCOVA. Esses pressupostos podem ser analisados de forma simples no JASP. Para verificar a linearidade da relação entre variável dependente e covariáveis para cada nível das variáveis independentes, selecione no menu superior do JASP o menu *Regression* (Regressão) e, em seguida, *Correlation* (Correlação) dentro do grupo *Classical* (Clássico). Conforme imagem ao lado.

Na tela de configuração da análise inclua as variáveis *libido* e *partnerLibido* no quadro *Variables* (Variáveis) e a variável *dose* no quadro *Partial out* (não se esqueça de alterar a variável dependente e covariável para tipo escalar!). A partir disso será gerada uma Tabela de correlação de Pearson entre as variáveis dependente e covariável condicionada nos grupos da variável categórica. Podemos afirmar que o pressuposto de linearidade foi aceito caso seja identificada uma relação significativa ( $p <$

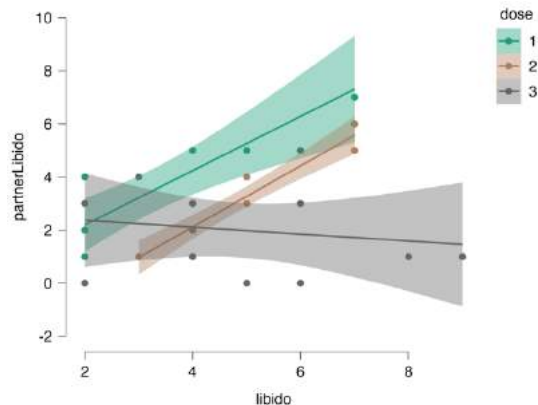
0,05) entre as variáveis. No caso de nossa análise foi identificada uma correlação mediana e significativa, o que permite prosseguirmos com a análise.

### Pearson's Partial Correlations

	Pearson's r	p
libido - partnerLibido	0.411	0.027

*Note.* Conditioned on variables: dose.

Em seguida, iremos testar a homogeneidade das retas de regressão. Esse pressuposto exige que não exista processos de interação (ou moderação) entre as variáveis de interesse. Para verificar esse pressuposto, selecione no menu superior do JASP a opção *Descriptives* (Descritivos) e na tela de análise inclua *libido* e *partnerLibido* no quadro *Variables* (Variáveis) e *dose* no quadro *Split* (Dividir). Em seguida clique no menu *Customizable Plots* (Gráficos customizáveis), selecione *Scatter plots* (Gráfico de dispersão), e marque *Add regression line* (Adicionar linha de regressão) do tipo linear e um gráfico similar ao que se encontra ao lado será gerado.

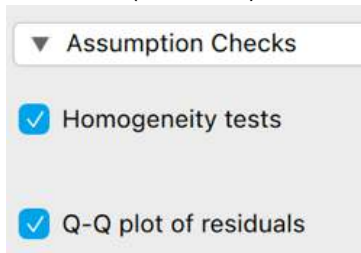


Para que o pressuposto de homogeneidade das retas de regressão seja aceito, todas as retas devem ser paralelas. Neste caso, verificamos que a reta correspondente a dose 3 não está paralela às outras e, portanto, leva a uma quebra do pressuposto, impedindo a continuidade da análise. Neste caso, entretanto, uma possível solução seria filtrar o grupo que tomou a dose 3. Faremos isso utilizando uma das opções de filtro dos dados indicada no capítulo de Estatísticas Descritivas.

Os outros pressupostos da ANCOVA são similares aos da ANOVA e foram tratados no capítulo referente a essa. A normalidade da distribuição dos dados da variável dependente pode ser testada conforme apresentado no capítulo da ANOVA, porém, outra estratégia está disponível e pode ser utilizada de forma mais simples durante o processo de estimação do modelo da ANCOVA (também disponível para a ANOVA). Dessa forma, para avaliar conjuntamente os pressupostos da normalidade dos dados e da homogeneidade das

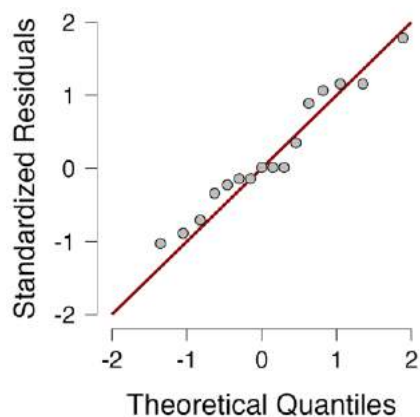


variâncias, selecione ANOVA no menu superior do JASP e selecione ANCOVA no grupo *Classical* (Clássico), conforme indicado ao lado.



Na tela de análise arraste a variável *libido* para o quadro *Dependent Variable* (Variável Dependente), a variável *dose* para o quadro *Fixed Factors* (Fatores fixos) e a variável *partnerLibido* para o quadro *Covariates* (Covariáveis). Em seguida, no menu *Assumption Checks* (Checagem de pressupostos) marque *Homogeneity tests* (Testes de Homogeneidade) e *Q-Q plot of residuals* (Gráfico Q-Q dos resíduos), conforme ao lado.

Isso irá gerar uma Tabela com os resultados do teste de Levene que avalia a homogeneidade da variância, sendo desejado que o valor- $p$  seja menor que 0,05 para que



o pressuposto não seja violado. Será gerado também um gráfico Q-Q que permite a avaliação da normalidade dos resíduos, para esse gráfico é esperado que os pontos estejam o mais próximo possível da reta, indicando assim a normalidade dos dados.

#### **Test for Equality of Variances (Levene's)**

<b>F</b>	<b>df1</b>	<b>df2</b>	<b>p</b>
5.130	1.000	15.000	0.039

Com base nos resultados apresentados acima, identifica-se que não houve desvio significativo da normalidade no gráfico Q-Q, contudo há quebra do pressuposto de homogeneidade das variâncias no teste de Levene ( $p = 0,039$ ). Ao contrário da ANOVA em que estão disponíveis correções para esse tipo de quebra de pressuposto, o JASP não fornece correções para quebra do pressuposto de homogeneidade das variâncias na ANCOVA. Portanto, soluções para esse tipo de problema devem ser buscadas por meio da transformação dos dados. Apesar da limitação na quebra desse pressuposto, iremos prosseguir com a ANCOVA de forma a apresentar o roteiro de análise e forma de interpretação.

## **ROTEIRO DE ANÁLISE: ANCOVA**

A partir da análise construída na última etapa, de avaliação dos pressupostos, selecione no grupo *Display* (Mostrar) as opções: *Descriptive statistics* (Estatísticas



descritivas) e *Estimates of effect size* (estimativas de tamanho de efeito), marque também o eta quadrado ( $\eta^2$ ) e o eta quadrado parcial (partial  $\eta^2$ ).

A Tabela com os resultados principais da ANCOVA serão apresentados de forma similar a abaixo. Com base nos resultados podemos verificar, com base no valor- $p$  significativo ( $p < 0,05$ ) que a libido do parceiro (*partnerLibido*) impacta na libido do participante, assim como a dose de viagra. Para além disso, verifica-se que o tamanho de efeito da libido do parceiro na libido do participante é alto ( $\eta^2 = 0,61$ ) e maior do que o efeito da dose de viagra ( $\eta^2 = 0,27$ ). Contudo, verifica-se que ao retirar o efeito da outra variável, essa diferença não é tão grande (considerando o  $\eta^2$  parcial).

#### ANCOVA - libido

Cases	Sum of Squares	df	Mean Square	F	p	$\eta^2$	$\eta^2_p$
partnerLibido	33.928	1	33.928	73.053	< .001	0.610	0.839
dose	15.226	1	15.226	32.783	< .001	0.274	0.701
Residuals	6.502	14	0.464				

*Note.* Type III Sum of Squares

Tendo identificado um efeito relevante no modelo geral, é importante prosseguir com os testes Post Hoc para compreender melhor tais resultados e verificar as diferenças em comparações dois a dois. Para isso, abra o menu *Post Hoc Tests* (testes Post Hoc) e arraste a variável *dose* para o quadro à direita. As outras opções são iguais às apresentadas na ANOVA, dessa forma verifique neste capítulo informações mais detalhadas sobre testes Post Hoc. Para essa análise iremos manter as opções que já vem automaticamente marcadas (*Type* [tipo]: *Standard* [padrão]; *Correction* [correção]: Tukey), marque também a opção *Effect size* (tamanho de efeito). A partir disso será gerada uma Tabela similar a apresentada abaixo, no caso de nossa análise só teremos uma comparação entre grupos devido a termos filtrado um dos grupos dos dados. Contudo, o mais comum é que existam diversos grupos sendo comparados e, portanto, diversas linhas na Tabela.

#### Post Hoc Comparisons - dose

	Mean Difference	SE	t	Cohen's d	$p_{tukey}$
1 2	-1.903	0.332	-5.726	-2.793	< .001

Com base nos resultados podemos verificar que há diferenças significativas ( $p < 0,001$ ) entre o grupo que tomou a dose 1 e a dose 2. A partir do sentido (positivo ou

negativo) do teste  $t$  e também do  $d$  de Cohen, é possível verificar que o grupo que tomou a dose 2 apresentou maior libido do que o grupo que tomou a dose 1.

Isso também pode ser verificado com mais detalhes na tabela *Descriptives* (descritivos). Dessa forma é possível verificar que o grupo que tomou a dose 2 apresentou média maior que o grupo que tomou a dose 1.

<b>Descriptives - libido</b>				
<b>dose</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>	<b>SE</b>
1	9	3.222	1.787	0.596
2	8	4.875	1.458	0.515

## RELATO DOS RESULTADOS

Foram identificadas diferenças significativas na libido dos participantes com base na dose de viagra utilizada ( $F[1; 14] = 32,78; p < 0,001$ ), mesmo quando controlado o efeito da libido do/a parceiro/a ( $F[1; 14] = 73,05; p < 0,001$ ) e essa diferença apresentou tamanho de efeito grande ( $\eta^2 = 0,27$ ). Participantes que tomaram a dose 2 ( $M = 4,88; DP = 1,46$ ) apresentaram maior libido em relação aos participantes que tomaram a dose 1 ( $M = 3,22; DP = 1,79; t = 5,73; p < 0,001$ ).

## MANOVA: ANÁLISE MULTIVARIADA DE VARIÂNCIA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Análise Multivariada de Variância e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Análise Multivariada de Variância no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

A Análise Multivariada de Variância é uma extensão da ANOVA que permite trabalhar com mais de uma variável dependente simultaneamente. Alguns pesquisadores, quando possuem mais de uma variável dependente métrica realizam várias ANOVAs separadamente. O grande problema neste caso é que o Erro Tipo I se torna inflacionado. A exemplo, se você tiver cinco variáveis dependentes, o Erro Tipo I do conjunto será  $1 - 0,95^5$ .

Para além da correção deste erro, a MANOVA possui outras vantagens, como a possibilidade de avaliar o relacionamento entre todas as variáveis dependentes e independentes. ANOVAs separadas não permitem avaliar o relacionamento entre todas as variáveis, o que não permite informações sobre as correlações entre as variáveis. A MANOVA avalia se grupos se diferem ao longo de uma combinação de dimensões.

Tendo em vista essas questões, é de fácil compreensão a relevância da técnica estatística e suas vantagens em relação à execução de ANOVAs em modelos com mais de uma variável dependente. Dessa forma, a análise a ser executada se apresenta como importante ferramenta utilizada nos mais diversos contextos de pesquisa e análise de dados.

A MANOVA possui os seguintes pressupostos: normalidade multivariada, as variáveis resposta devem ter distribuição multivariada normal para cada grupo das variáveis independentes; independência das observações (ou resíduos); homogeneidade

das variâncias para cada grupo das variáveis independentes; e a ausência de *outliers* (casos extremos) multivariados.

## ROTEIRO DE ANÁLISE: MANOVA

A Biblioteca de Dados do JASP não possui exemplos para MANOVA, portanto usaremos o banco de dados *OCD* disponibilizado por Andy Field como material suplementar de seu livro “Discovering statistics using IBM SPSS Statistics”. Esse banco de dados mensurou a frequência de pensamentos e comportamentos compulsivos em pessoas com transtorno obsessivo-compulsivo após três tipos de tratamentos psicológicos (nenhum tratamento, terapia comportamental e terapia cognitivo-comportamental), a variável *Group* (grupo) indica qual tipo de tratamento o participante passou, a variável *Actions* (ações) indica a frequência de comportamentos compulsivos e a variável *Thoughts* (pensamentos) indica a frequência de pensamentos compulsivos.



Para realizar a MANOVA, iremos selecionar ANOVA no menu superior do JASP e, em seguida, selecionar MANOVA no grupo clássico (*Classical*), conforme imagem ao lado. Vamos arrastar a variável independente, *Group*, para o quadro *Fixed Factors* (fatores fixos) e as variáveis dependentes, *Actions* e *Thoughts*, para o quadro de mesmo nome (*Dependent Variables*).

O menu *Model* (modelo) permite incluímos ou retirarmos variáveis do modelo, assim como estabelecer interações entre as

variáveis independentes. Contudo, nosso modelo possui apenas uma variável independente, portanto não iremos utilizar esse menu.

O menu *Additional Options* (opções adicionais) permite tomarmos decisões importantes sobre o cálculo da MANOVA e os resultados a serem apresentados. Vejamos cada uma das opções abaixo:

▼ Additional Options

<b>Test</b>	<b>Assumption Checks</b>
<input checked="" type="checkbox"/> Pillai	<input type="checkbox"/> Homogeneity of covariance matrices
<input type="checkbox"/> Wilks	<input type="checkbox"/> Multivariate normality
<input type="checkbox"/> Hotelling-Lawley	
<input type="checkbox"/> Roy	
<b>Display</b>	
<input type="checkbox"/> ANOVA tables	
<input type="checkbox"/> Vovk-Sellke maximum p-ratio	

O grupo *Test* (teste) permite selecionar qual forma de cálculo será utilizada para a MANOVA. As opções são:

- Traço de Pillai: é um teste mais robusto quando há desvios no pressuposto de homogeneidade das variâncias e os grupos são iguais ou quando os grupos são pequenos;
- Lambda de Wilk: tende a ser o teste mais recomendado, em geral, quando não há nenhum desvio nos pressupostos da análise;
- Traço de Hotelling-Lawley: é um teste mais robusto quando há desvios no pressuposto de homogeneidade das variâncias e os grupos são iguais;
- Maior Raiz de Roy: o mais recomendado quando há apenas duas variáveis dependentes.

O grupo *Display* (mostrar) adiciona duas opções aos resultados da MANOVA: as Tabelas de ANOVA (*ANOVA tables*), que apresenta os resultados para cada variável dependente individualmente (sendo similar à ANOVA); e *Vovk-Sellke maximum p-ratio*.

Por fim, o grupo *Assumption Checks* (checagem de pressupostos) permite avaliar a normalidade multivariada, por meio do teste multivariado de Shapiro-Wilk, e a homogeneidade das variâncias, por meio do teste Box M. Cabe ressaltar que ambos os

testes devem apresentar resultados não significativos (em  $p > 0,05$ ) para que não haja desvios dos pressupostos. Esses serão os primeiros resultados que iremos analisar.

Conforme Tabelas abaixo, pode-se notar que não houve desvio da normalidade multivariada segundo o teste de Shapiro-Wilk ( $p = 0,522$ ), assim como não houve desvio na homogeneidade das variâncias segundo o teste de Box M ( $p = 0,18$ ).

<b>Box's M-test for Homogeneity of Covariance Matrices</b>			<b>Shapiro-Wilk Test for Multivariate Normality</b>	
$\chi^2$	df	p	Shapiro-Wilk	p
8.893	6	0.180	0.969	0.522

Considerando a ausência de desvios nos pressupostos da análise e a presença de apenas duas variáveis dependentes no modelo, iremos prosseguir a análise com o teste da Maior Raiz de Roy. Também iremos solicitar as tabelas de ANOVA para facilitar a interpretação dos nossos resultados.

Aqui cabe destacar duas limitações do JASP para análise de MANOVA. Até a versão atual não é possível solicitar estatísticas de tamanho de efeito para a MANOVA ou ANOVA e também não é possível solicitar testes Post Hoc ou de Contraste para as ANOVAs. De forma que, quando necessários, testes Post Hoc devem ser solicitados por meio da execução da análise de ANOVA, o que gera um inflacionamento do teste de significância estatística (valor- $p$ ).

Os resultados da MANOVA, apresentados na Tabela abaixo, indicam que o tipo de tratamento psicológico é significativo em explicar diferenças nos comportamentos e pensamentos compulsivos.

<b>MANOVA: Roy Test</b>						
Cases	df	Approx. F	Largest Root	Num df	Den df	p
(Intercept)	1	745.230	57.325	2	26.000	< .001
Group	2	4.520	0.335	2	27.000	0.020
Residuals	27					

Para compreender melhor esses resultados iremos analisar também as Tabelas de ANOVAs individuais. A primeira ANOVA para comportamentos compulsivos indicou que os participantes não apresentaram diferenças significativas nos diferentes grupos de tratamentos.

<b>ANOVA: Actions</b>					
Cases	Sum of Squares	df	Mean Square	F	p
(Intercept)	616.533	1	616.533	326.400	< .001

**ANOVA: Actions**

Cases	Sum of Squares	df	Mean Square	F	p
Group	10.467	2	5.233	2.771	0.080
Residuals	51.000	27	1.889		

A segunda ANOVA para os pensamentos compulsivos indicou que os participantes também não apresentaram diferenças significativas nos diferentes grupos de tratamentos. Dessa forma, os resultados sugerem que as diferenças identificadas na MANOVA decorrem unicamente da interação entre os comportamentos e os pensamentos compulsivos.

**ANOVA: Thoughts**

Cases	Sum of Squares	df	Mean Square	F	p
(Intercept)	6336.533	1	6336.533	1402.348	< .001
Group	19.467	2	9.733	2.154	0.136
Residuals	122.000	27	4.519		

**RELATO DOS RESULTADOS**

Foi realizada uma MANOVA para compreender se o uso de diferentes tratamentos psicológicos afeta os comportamentos e pensamentos compulsivos de pessoas com transtorno obsessivo compulsivo. Devido a haver apenas uma combinação linear entre as variáveis dependentes optou-se por realizar a análise por meio da Maior Raiz de Roy, ainda não foi verificada quebra nos pressupostos da análise. Os resultados da MANOVA apontaram que o tipo de tratamento psicológico gera diferenças significativas nos comportamentos e pensamentos compulsivos ( $F [2; 27] = 4,52; p = 0,02$ ). Contudo, ao analisar as ANOVAs, não foi identificado efeito significativo do tipo de tratamento psicológico nos comportamentos compulsivos ( $F [2; 27] = 2,77; p = 0,08$ ) ou nos pensamentos compulsivos ( $F [2; 27] = 2,15; p = 0,14$ ) de forma individual.

## ANÁLISE FATORIAL EXPLORATÓRIA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Análise Fatorial Exploratória e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Análise Fatorial Exploratória no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

Certos conceitos das ciências sociais e comportamentais não são bem definidos e há muitas discussões a respeito do real significado de termos como inteligência, personalidade, grupo social, equipes, personalidade extrovertida e outros. Um conceito utilizado para explicar um fenômeno pode ser considerado um construto, o qual apoia a construção de medidas de mensuração do fenômeno. Um construto pode ser mensurado por uma variável observada ou por várias e quando é mensurado por mais de uma variável estas compõem uma variável latente, que explicaria o construto de interesse.

A exemplo, na aplicação de um teste de inteligência ou de personalidade, há muitos itens, os quais são denominados como variáveis observadas. As variáveis observadas dizem respeito aos dados empíricos obtidos pelo pesquisador. Esses itens poderão compor um ou mais componentes ou fatores, os quais passam a ser denominados de variáveis latentes. As variáveis latentes são aquelas que se referem ao construto e que resumem os achados provenientes das variáveis observadas. Nesse sentido, variáveis latentes não são observadas diretamente.

Para construir variáveis latentes é necessário construir medidas ou buscar medidas existentes que relacionem um conjunto de indicadores, comportamentos, atitudes, crenças com valores observáveis e associados ao que se pretende mensurar (por exemplo, preconceito racial). A Análise Fatorial é um conjunto de técnicas estatísticas que pretende



avaliar a adequação dos itens ou das variáveis observadas em compor e representar adequadamente uma, ou mais, variáveis latentes (fatores).

Por isso, a Análise Fatorial é considerada uma análise estatística multivariada que é utilizada com frequência por psicometristas para construção e buscar evidências de validade de instrumentos ou por outros pesquisadores para verificar se tais evidências conformam-se a amostra alcançada. Ao empregar esta análise o interesse encontra-se na correlação ou covariância de um grupo de variáveis observadas em explicar um construto não mensurado diretamente que é a variável latente.

Mais precisamente, um fator é um construto, uma entidade hipotética, uma variável não observada diretamente, que se supõe estar subjacente às respostas das pessoas a testes, escalas, itens ou a um conjunto de indicadores de um determinado fenômeno. Como construtos, os fatores apenas explicam a variância de variáveis observadas, tal como se revelam pelas correlações ou covariâncias entre as variáveis analisadas, ou seja, a única realidade científica que os fatores possuem provém das correlações entre respostas a testes ou ao comportamento de múltiplas variáveis pesquisadas. Para que um fator seja detectado numa matriz de respostas a um teste ou numa matriz de dados econômicos é preciso que existam correlações entre as variáveis observadas que compõem essas matrizes. No caso dos testes psicológicos, as respostas das pessoas aos itens devem estar correlacionadas entre si. No caso do conjunto de variáveis econômicas, esses indicadores teriam que estar correlacionados para que se possa supor fator(es) subjacente(s) a eles.

A análise fatorial faz reduções de dados inicialmente mais complexos a um conjunto de tamanho manuseável para que o pesquisador possa interpretar mais facilmente os resultados. Além disto, ela irá gerar uma matriz cargas fatoriais. A carga fatorial é um coeficiente, positivo ou negativo, que em sua versão padronizada (e mais comum) varia entre -1 e 1 e expressa o quanto uma variável observada está carregada de um fator. Em outras palavras, quanto maior for a carga de um item em um fator, mais a variável representa esse fator (variável latente).

Além das diversas técnicas envolvidas na Análise Fatorial, essa se divide também em: Análise Fatorial Exploratória (AFE) que auxiliará na exploração inicial de um conjunto de variáveis observadas para determinar sua relevância na explicação do(s) construto(s) proposto(s) teoricamente; e Análise Fatorial Confirmatória (AFC) que a partir de um modelo pré-determinado com os fatores (variáveis latentes) e sua composição (variáveis observadas), verifica a adequação do modelo testado com base nos dados coletados. A

AFC pode confirmar, ou não, os achados da AFE, assim como múltiplas AFEs podem apresentar diferentes resultados a depender das técnicas utilizadas e amostra coletada.

Em síntese, a Análise Fatorial Exploratória é um método a ser utilizado para determinar o número de fatores existentes em um conjunto de dados, para determinar quais variáveis observadas pertencem a quais fatores, e em que extensão as variáveis observadas estão saturadas no fator subjacente. Devido a sua complexidade, esse método de análise apresenta pressupostos mais complexos.

## PRESSUPOSTOS DA ANÁLISE FATORIAL EXPLORATÓRIA

**Tamanho da amostra.** O tamanho da amostra é uma questão complexa para a Análise Fatorial e que não pode ser solucionada facilmente por meio de análises de poder. O número adequado de participantes irá depender diretamente do nível de correlação das variáveis observadas e da quantidade de fatores, contudo essas informações geralmente são desconhecidas ao pesquisador na etapa inicial de coleta.

Para solucionar esta questão uma possibilidade é utilizar de estudos piloto e técnicas de simulação de dados, contudo o primeiro gera maior custo operacional à pesquisa e o segundo envolve técnicas complexas de análise de dados. Portanto, diversos autores propuseram normas simples que auxiliam a guiar tais decisões, algumas são:

- No mínimo 100 participantes;
- No mínimo 300 participantes;
- 5 a 10 participantes para cada item (Hair et al., 2009; Pasquali, 2010).

Além disso, no caso de amostras muito grandes é importante que sejam testadas subamostras, esse procedimento permite dividir aleatoriamente os dados em  $n$  amostras iguais e testar tais subamostras utilizando as mesmas técnicas para comparar seus resultados, o que é denominado validação cruzada.

**Dados ausentes.** Dados ausentes prejudicam as análises estatísticas por reduzir o tamanho da amostra por isso devem ser evitados. Algumas soluções estatísticas estão disponíveis atualmente (como a *Multiple Imputation* [Imputação Múltipla]), entretanto elas podem gerar correlações “ilusórias” ou falsas. Portanto, deve-se cuidar dos dados ausentes verificando se há alguma tendência ou relação entre as ausências.

**Normalidade.** A Análise Fatorial é relativamente robusta a desvios da normalidade, contudo é importante que haja normalidade multivariada para a realização da análise.

Quando identificado ou suspeita de desvios de normalidade alguns métodos de estimação serão mais recomendados por serem mais robustos.

**Linearidade.** A Análise Fatorial baseia-se no pressuposto de que há relações lineares entre as variáveis. Contudo, no caso de variáveis que apresentem relações não lineares pode-se optar pelo uso da matriz de Correlações Policóricas no lugar da matriz de Correlações de Pearson ou pela transformação das variáveis que não apresentam relação linear. A matriz de Correlações Policóricas é uma medida de associação não paramétrica que é adequada para dados ordinais ou categóricos. A matriz de Correlações de Pearson, por outro lado, é baseada na suposição de que as variáveis seguem uma distribuição normal e linear.

**Multicolinearidade e singularidade.** Algum nível de multicolinearidade entre as variáveis é desejável para que exista relação entre as mesmas, entretanto correlações altas (acima de 0,70 ou 0,80) sugerem multicolinearidade e podem gerar problemas de singularidade.

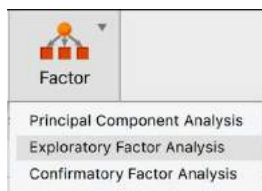
**Outliers nos casos e nas variáveis.** Casos extremos (outliers) nas variáveis impactam na matriz de correlações por diminuir ou ampliar a magnitude dos coeficientes de correlações. Variáveis com outliers influentes tendem a ser excluídas dos fatores por terem suas cargas fatoriais prejudicadas.

**Fatorabilidade da matriz de correlação/covariância.** Previamente a Análise Fatorial é importante identificar se a matriz de correlação/covariância é passível de ser fatorada. Isto pode ser averiguado por meio de testes como o teste de esfericidade de Bartlett e o índice de Kaiser-Meyer-Olkin (KMO).

## ROTEIRO DE ANÁLISE: ANÁLISE FATORIAL EXPLORATÓRIA

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *G Factor* que possui os dados das notas escolares e escores de discriminação sensorial de 23 alunos. Esse banco possui 17 variáveis, contudo iremos focar nas variáveis que iniciam em *Residuals*, ou seja: *Residuals Pitch* (discriminação sensorial auditiva), *Residuals Light* (discriminação sensorial visual), *Residuals Weight*

(discriminação sensorial de força), *Residuals Classics* (estudos clássicos), *Residuals French* (Francês), *Residuals English* (Inglês) e *Residuals Mathematics* (matemática).



Para iniciar as análises, selecione a opção *Factor* (Fator) no menu superior e *Exploratory Factor Analysis* (Análise Fatorial Exploratória). Na tela de configuração da análise transfira as variáveis observadas, acima mencionadas, para o quadro *Variables* (Variáveis). Outras opções vêm automaticamente selecionadas, mas podem (e devem) ser alteradas de acordo com a sua necessidade de análise. Trataremos cada opção individualmente abaixo:

 A screenshot of the JASP configuration window for Factor analysis. The window is divided into several sections:
 

- Number of Factors based on:** Includes radio buttons for 'Parallel analysis' (selected), 'Based on PC', and 'Based on FA'. There is a 'Seed' input field with the value '1234'.
- Eigenvalues:** Includes a radio button for 'Eigenvalues' and an input field for 'Eigenvalues above' with the value '1'.
- Manual:** Includes a radio button for 'Manual' and an input field for 'Number of factors' with the value '1'.
- Factoring method:** A dropdown menu set to 'Minimum residual'.
- Rotation:** Includes radio buttons for 'Orthogonal' and 'Oblique' (selected). There are dropdown menus for rotation type, currently set to 'none' and 'promax'.
- Base analysis on:** Includes radio buttons for 'Correlation matrix' (selected), 'Covariance matrix', and 'Polychoric/tetrachoric correlation matrix'.

- *Number of Factors based on* (Número de fatores baseado em): as opções desse grupo auxiliarão na determinação do número de fatores a serem extraídos pela análise;
  - *Parallel Analysis* (Análise Paralela): técnica robusta que gera conjuntos de dados aleatórios baseados nos dados inseridos e calcula os autovalores para os dados aleatórios, o número de fatores a serem extraídos será igual ao número de autovalores calculados para os dados reais que sejam maiores que a média dos autovalores

- calculados aleatoriamente, pode se basear nos Componentes Principais ou na Análise Fatorial;
- *Eigenvalues* (autovalores): os fatores com autovalor maior que 1 (ou outro definido manualmente) serão mantidos;
  - Manual: o usuário determina quantos fatores serão extraídos.
  - *Factoring Method* (Método de Fatoração ou Método de Estimação): determina qual método será utilizado para estimar as cargas fatoriais e outras matrizes resultantes da Análise Fatorial;
    - Diversas opções estão disponíveis e devem ser estudadas e analisadas cuidadosamente para determinar qual a melhor para seus dados;
    - *Principal Axis Factoring* (Fatoração por Eixos Principais): método de estimação mais comum no SPSS, tem sido substituído por opções mais robustas;
    - *Minimum Residual* (Mínimo Residual): busca reduzir os resíduos gerados pela fatoração dos dados, método padrão no JASP;
    - *Maximum Likelihood* (Máxima Verossimilhança): método que apresenta melhores resultados quando estes apresentam distribuição normal;
    - *Ordinary/Weighted/Generalized Least Squares* (Quadrados Mínimos Ordinários/Ponderados/Generalizados): métodos baseados na regressão, são preferidos em casos de variáveis ordinais.
  - *Rotation* (Rotação): processo de transformação dos resultados que simplifica a interpretação dos fatores sempre que há mais de um, pode ser
    - *Orthogonal* (Ortogonal): rotações que tornam os fatores ortogonais, isto é, não permite que haja correlação entre eles. Esse tipo de rotação gera resultados mais fáceis de serem interpretados e devem

ser utilizados sempre que os fatores apresentarem correlações baixas;

- *Oblique* (Oblíqua): rotações que permitem a correlação entre os fatores, é preferível nas ciências sociais devido a alta correlação entre os construtos estudados;
- Ambos os tipos de correlações possuem diversas técnicas para cálculo que devem ser avaliadas com base nas necessidades do pesquisador e compreendidas para embasar sua escolha.
- *Base analysis on* (Basear as análises em):
  - *Correlation matrix* (matriz de correlações): matriz padrão, os resultados serão baseados numa matriz de correlações de Pearson entre as variáveis;
  - *Covariance matrix* (matriz de covariância): os resultados serão baseados na matriz de covariância entre as variáveis;
  - *Polychoric/tetrachoric correlation matrix* (matriz de correlação policórica/tetracórica): deve ser utilizada quando os dados não são

escalares (por exemplo, escalas Likert de até 6 pontos) ou apresentam desvios de normalidade multivariada.

Há também o menu de opções de saída (*Output Options*) que permite solicitarmos outros resultados relevantes para a Análise Fatorial, conforme apresentado abaixo:

The screenshot shows the 'Output Options' dialog box in SPSS. It is divided into several sections:

- Display loadings above:** A vertical slider with a white knob and a blue bar, currently set to 0.4.
- Order factor loadings by:** Two radio buttons: 'Factor size' (selected) and 'Variables'.
- Tables:** A list of checkboxes: 'Structure matrix', 'Factor correlations', 'Additional fit indices', 'Residual matrix', and 'Parallel analysis'. Under 'Parallel analysis', there are two sub-options: 'Based on PC' and 'Based on FA'.
- Plots:** A list of checkboxes: 'Path diagram', 'Scree plot', and 'Parallel analysis results'.
- Assumption checks:** A list of checkboxes: 'KMO test', 'Bartlett's test', and 'Mardia's test', all of which are unchecked.
- Missing Values:** A list of radio buttons: 'Exclude cases pairwise' (selected) and 'Exclude cases listwise'.

- *Display loadings above* (Mostrar cargas acima): arraste o ponto branco na barra para aumentar ou reduzir o tamanho mínimo que a carga fatorial de uma variável deve possuir para ser apresentada na Tabela dos resultados. Por padrão vem selecionado em 0,40, contudo alguns autores sugerem que cargas acima de 0,30 sejam mantidas. Ainda, o ideal é solicitar que todas as cargas fatoriais sejam apresentadas (arraste para 0) de forma que seja

possível comparar diferenças pequenas que seriam ocultadas por esse filtro inicial.

- *Order factor loadings by* (Ordenar as cargas fatoriais por): determina o método de ordenação das variáveis na tabela de resultados que apresenta as cargas fatoriais;
  - *Factor size* (tamanho do fator): as variáveis são ordenadas por tamanho decrescente das cargas fatoriais;
  - *Variables* (variáveis): ordem de inclusão das variáveis na análise;
- *Assumption checks* (Checagem de pressupostos):
  - *KMO test*: índice que avalia a qualidade da matriz dos dados para fatoração, é considerado adequado acima de 0,5 e ótimo acima de 0,8;
  - *Bartlett's test*: teste que testa a hipótese de que a matriz correlacional não é uma matriz identidade, isto é, não é uma matriz onde todos os elementos da diagonal são iguais a 1 e o restante dos coeficientes da matriz iguais a zero, se o valor- $p$  do teste de Bartlett for menor que o nível de significância escolhido (geralmente  $p < 0,05$ ), isso indica que há evidências suficientes para rejeitar a hipótese nula de que a matriz de correlação é igual à matriz identidade. Isso significa que as variáveis estão correlacionadas e que a análise fatorial é apropriada, este é um teste sensível a desvios;
  - *Mardia's test*: teste que avalia a normalidade multivariada dos dados, quando o valor- $p$  é significativo indica desvio de normalidade multivariada e quebra do pressuposto, portanto deve ser não significativo ( $p > 0,05$ ).
- *Tables* (Tabelas):
  - *Structure matrix* (matriz de estrutura): gera uma tabela que apresenta a estrutura fatorial proposta pela análise, ou seja, como as variáveis observadas se organizam dentro dos fatores, essa tabela já é apresentada em conjunto com outras colunas na Tabela *Factor Loadings* (Cargas Fatoriais);
  - *Factor correlations* (Correlações inter-fatoriais): gera tabela que apresenta a correlação entre os fatores gerados, é importante para



determinar o tipo de rotação a ser utilizada e quando há mais de um fator;

- *Additional fit indices* (Índices de ajuste adicionais): gera tabela que apresenta os índices de ajuste dos dados ao modelo e que indicam o quanto a estrutura está adequada. Esses índices são comuns a Análise Fatorial Confirmatória e se configuram como Análise Fatorial Semi-Confirmatória quando apresentados na AFE;
- *Residual matrix* (Matriz residual): gera tabela que apresenta a matriz dos resíduos (erro) da análise;
- *Parallel Analysis* (Análise Paralela): gera tabela com os resultados da Análise Paralela, importante para justificar a determinação dos números de fatores a serem extraídos.
- *Plots* (gráficos):
  - *Path diagram* (diagrama de trajetória): gera gráfico que apresenta as trajetórias entre as variáveis observadas e variáveis latentes (fatores);
  - *Scree plot* (gráfico Scree): gera gráfico que apresenta os valores de autovalores para cada possível número de fatores extraídos, permite determinar o número de fatores a serem extraídos e também apresenta de forma visual os resultados da Análise Paralela desde que selecionada a caixa *Parallel Analysis results* (resultados da Análise Paralela).
- *Missing values* (valores ausentes):
  - *Exclude cases pairwise* (Excluir casos pairwise): opção padrão, exclui os casos ausentes apenas das variáveis relevantes para a análise, reduzindo a quantidade de casos perdidos;
  - *Exclude cases listwise* (Excluir casos listwise): opção que exclui todos os casos ausentes independente da variável considerada, de forma que a análise se dará apenas com dados completos, é a mais recomendada.

Como visto, a Análise Fatorial (e a AFE) é um conjunto de técnicas estatísticas que exige a tomada de diversas decisões. Para esse exercício realizaremos a análise utilizando: a Análise Paralela baseada na FA (Análise Fatorial) para determinação do número de fatores a serem extraídos; método de estimação do Mínimo Residual; rotação

Oblíqua, promax (padrão do JASP), contudo verifique a tela de resultados e verá que apenas um fator foi extraído, portanto essa configuração não é necessária; e análise baseada na matriz de correlações. Ainda solicitaremos que todas as cargas fatoriais sejam apresentadas (factor loadings above 0), os testes KMO, Bartlett e Mardia, a apresentação das tabelas com os índices de ajuste adicionais e resultados da Análise Paralela e o *Scree plot* da Análise Paralela. Esse banco não conta com dados ausentes, portanto nenhuma alteração é necessária nessas configurações.

Na tela de resultados, primeiro avaliaremos os testes de pressupostos. O teste de Bartlett foi significativo ( $\chi^2 [21] = 86,74; p < 0,001$ ), indicando que a matriz de correlação dos dados é passível de ser fatorada, assim como o teste KMO (0,67), o qual deve ser avaliado por meio da linha *Overall MSA* na Tabela ao lado. O teste de Mardia permite avaliar a normalidade multivariada dos dados e, conforme, resultados apresentados na Tabela abaixo não se verificou desvios na normalidade multivariada dos dados para *skewness* (assimetria) e *kurtosis* (curtose).

#### **Kaiser-Meyer-Olkin test**

	<b>MSA</b>
Overall MSA	0.670
Residuals Pitch	0.565
Residuals Light	0.646
Residuals Weight	0.229
Residuals Classics	0.757
Residuals French	0.620
Residuals English	0.863
Residuals Mathematics	0.739

#### **Mardia's Test of Multivariate Normality**

	<b>Value</b>	<b>Statistic</b>	<b>df</b>	<b>p</b>
Skewness	25.648	98.317	84	0.136
Small Sample Skewness	25.648	114.726	84	0.015
Kurtosis	59.219	-0.808		0.419

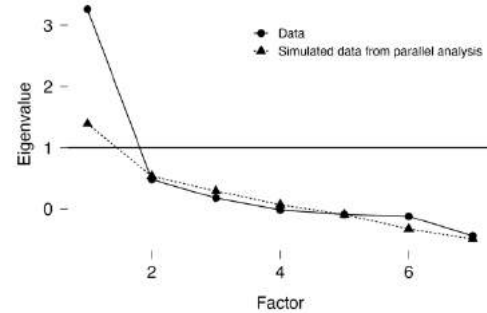
*Note.* The statistic for skewness is assumed to be  $\chi^2$  distributed and the statistic for kurtosis standard normal.

Em seguida, devemos avaliar os resultados da Análise Paralela. Este indicou que apenas um fator deveria ser extraído.

### Parallel Analysis

	Real data factor eigenvalues	Simulated data mean eigenvalues
Factor 1*	3.264	1.393
Factor 2	0.483	0.536
Factor 3	0.178	0.293
Factor 4	-0.015	0.069
Factor 5*	-0.087	-0.090
Factor 6*	-0.119	-0.327
Factor 7*	-0.439	-0.488

Note. '\*' = Factor should be retained.  
Results from FA-based parallel analysis.



A partir desses resultados iremos avaliar os resultados da análise fatorial, isto é, a tabela de cargas fatoriais. Essa tabela apresenta cada um dos itens com sua carga fatorial em cada fator e informações sobre a comunalidade (*uniqueness*) dos itens. Autores sugerem que cargas fatoriais acima de 0,30 devem ser mantidas desde que não haja carga cruzada (carga fatorial com diferença menor que 0,1 em dois fatores). Além disso, as comunalidades indicam o índice de regressão da variável, considerando cada a

variável indicada como dependente e as outras como independentes, valores extremos (muito próximos a 0 ou 1) de comunalidade indicam problemas. A comunalidade indica a variância comum entre a variável analisada e as demais.

### Factor Loadings

	Factor 1 Uniqueness	
Residuals Classics	0.965	0.069
Residuals French	0.906	0.180
Residuals Mathematics	0.851	0.276
Residuals English	0.729	0.469
Residuals Pitch	0.431	0.814
Residuals Light	0.218	0.952
Residuals Weight	0.157	0.975

Note. Applied rotation method is promax.

Outra Tabela relevante é sobre as características dos fatores (*Factor Characteristics*), apresentada abaixo. A coluna *SumSq.Loadings* indica os autovalores de cada fator; a coluna *Proportion var.* indica a variância proporcional explicada por cada fator (se multiplicado o valor por 100 você obterá o valor em porcentagem); e a coluna *Cumulative* apresenta a variância explicada cumulativa (somando sempre a linha atual com a linha abaixo), que neste caso será igual

a variância explicada. Os resultados indicam que o fator explica 46,6% da variância dos itens.

<b>Factor Characteristics</b>			
	<b>SumSq.</b>	<b>Loadings</b>	<b>Proportion var. Cumulative</b>
Factor 1	3.264	0.466	0.466

Por fim, os dados sobre a Análise Fatorial Semi-Confirmatória (AFSC), apresentados na Tabela *Additional Fit Indices*, indicam o ajuste dos dados ao modelo. Alguns valores de corte são sugeridos para tais índices: considera-se o RMSEA (Root Mean Square Error of Approximation) adequado quando abaixo de 0,08 ou intervalo de confiança superior abaixo de 0,10; o TLI (Tucker-Lewis Index) e o CFI (Comparative Fit Index) é considerado adequado quando acima de 0,90. Nesse caso identificou-se que o modelo é adequado considerando o TLI e CFI, porém apresenta nível de resíduo alto com base no RMSEA.

<b>Additional fit indices</b>					
<b>RMSEA</b>	<b>RMSEA 90% confidence</b>	<b>SRMR</b>	<b>TLI</b>	<b>CFI</b>	<b>BIC</b>
0.099	0 - 0.247	0.092	0.908	0.942	-26.074

## RELATO DOS RESULTADOS

Foi realizada uma Análise Fatorial Exploratória a partir dos dados residuais de cinco variáveis observadas buscando averiguar a presença de um fator comum entre estes. Para tal foi utilizado o método de estimação do Mínimo Residual a partir de uma matriz de correlações de Pearson. A quantidade de fatores a serem retidos foi estipulada por meio da técnica de Análise Paralela e a adequação da matriz para o procedimento foi avaliada por meio do índice KMO. Por fim, foi realizada a Análise Fatorial Semi-Confirmatória que permite verificar o ajuste do modelo testado aos dados.

O índice KMO (0,67) sugere uma matriz passível de ser fatorada e a Análise Paralela indicou a retenção de apenas um fator que apresentou autovalor superior à média dos autovalores dos dados simulados. Seguindo essa proposta não foi necessário utilizar técnicas de rotação para simplificar a estrutura fatorial. Em geral, os itens apresentaram cargas fatoriais elevadas, com exceção das variáveis relacionadas a discriminação sensorial visual (0,22) e discriminação sensorial de força (0,16) que apresentaram carga

fatorial inferior ao limiar sugerido (0,30). O fator extraído explicou 46,6% da variância dos itens.

Com relação aos índices de ajuste, verificou-se que índices comparativos do modelo, como o CFI e o TLI apresentaram ajustes adequados. Contudo, índices residuais, como o RMSEA e o SRMR, apresentaram valores altos. Avalia-se, contudo, que tais índices foram impactados negativamente pela manutenção dos itens relacionados a discriminação sensorial visual e discriminação sensorial de força que apresentaram cargas fatoriais baixas.

## ANÁLISE FATORIAL CONFIRMATÓRIA

### OBJETIVOS DO CAPÍTULO

- Apresentar teoricamente a Análise Fatorial Confirmatória e seus pressupostos;
- Descrever as rotinas de comandos para a realização da Análise Fatorial Confirmatória no JASP;
- Apresentar sugestões de interpretação e relato dos resultados encontrados.

### APRESENTAÇÃO TEÓRICA

Como parte da família de Análises Fatoriais, recomenda-se que leia inicialmente a apresentação teórica do capítulo sobre Análise Fatorial Exploratória (AFE) para melhor compreender a Análise Fatorial Confirmatória (AFC).

O papel da análise fatorial confirmatória é determinar se o modelo (isto é, os fatores e os itens que os compõem, assim como suas relações) é capaz de explicar a variação e covariação entre um conjunto de medidas observáveis. O modelo da AFE é tido como um modelo saturado/irrestrito, isso porque todas as dimensões latentes explicam a variação em todos os itens. Já na AFC, ou Modelo Restrito, é imposto algumas restrições ao modelo, por exemplo, não haver cargas cruzadas de um fator com itens de outro fator.

As técnicas que envolvem a AFC compõem estatísticas multivariadas que são utilizadas para estimar a estrutura de uma medida ou instrumento e verificar a qualidade dessa estrutura em representar os dados coletados, ou identificados por meio das variáveis observadas. Dessa forma, essas técnicas exigem que seja determinada uma estrutura a ser testada.

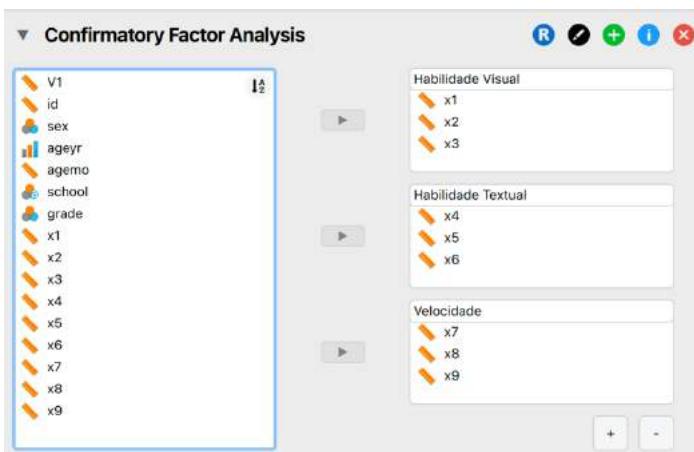
No caso de validações transculturais de instrumentos é importante sempre realizar a AFE e a AFC, conforme recomendado pela literatura. Utilizando as duas técnicas conjuntamente será possível que o pesquisador verifique como se estrutura o instrumento naquele contexto por meio da AFE, enquanto testa o ajuste da estrutura identificada em

novos dados. Por fim, os pressupostos da AFC são similares aos pressupostos da AFE, portanto cabe revisitar o capítulo anterior para maiores detalhes.

## ROTEIRO DE ANÁLISE: ANÁLISE FATORIAL CONFIRMATÓRIA

Para essa análise será utilizado o banco de dados disponível na Biblioteca de Dados do JASP denominado *Mental Ability* que possui os resultados de testes de habilidade mental de 301 crianças do 8º e 9º ano de duas escolas diferentes. Esse banco possui 16 variáveis, contudo iremos focar nas variáveis que compõem três fatores: Habilidade Visual, composto pelas variáveis  $x_1$ ,  $x_2$  e  $x_3$ ; Habilidade Textual, composto pelas variáveis  $x_4$ ,  $x_5$  e  $x_6$ ; e Velocidade, composto pelas variáveis  $x_7$ ,  $x_8$  e  $x_9$ . Testaremos se os dados se ajustam esse modelo proposto e verificaremos as cargas fatoriais identificadas em tal modelo.

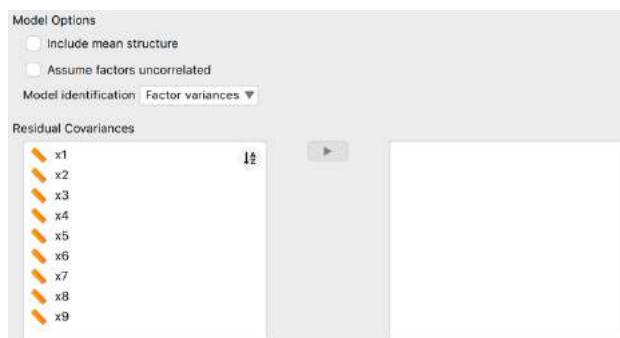
Para iniciar as análises, selecione a opção *Factor* (Fator) no menu superior e *Confirmatory Factor Analysis* (Análise Fatorial Exploratória). Na tela de configuração da análise, no quadro direito escreva o nome do primeiro fator e transfira as variáveis que correspondem a ele; em seguida clique no ícone de + e repita o procedimento para os outros fatores. Ao final você terá uma tela similar a apresentada ao lado. Outras opções



vêm automaticamente selecionadas, mas podem (e devem) ser alteradas de acordo com a sua necessidade de análise. Trataremos cada opção individualmente abaixo.

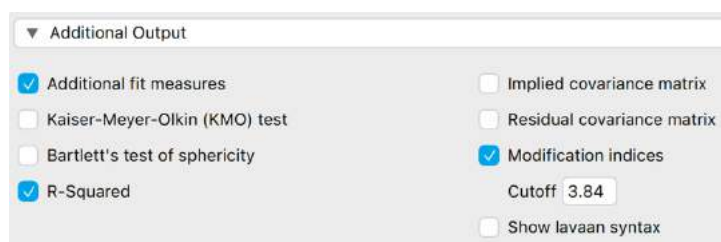
No menu *Second-Order Factor* é possível indicar se será estimado modelo hierárquico, no qual há um ou mais fatores mais amplos que apresentam alta correlação ou variância compartilhada e explicariam os fatores subjacentes. No caso do banco de dados que estamos utilizando não há proposição da existência de um fator de segunda-ordem, contudo seria possível considerar um fator geral de habilidade mental a partir das três habilidades identificadas como fatores de primeira ordem. Para determinar fatores de segunda ordem basta enviar os nomes dos fatores que formariam o fator de segunda

ordem (no quadro à esquerda) para o quadro a direita. Essa análise não será realizada aqui.



O menu *Model Options* (Opções do Modelo) permite que outras informações sejam solicitadas. Ao abri-lo as seguintes opções serão apresentadas:

- *Include mean structure* (incluir estrutura média): inclui as médias (ou interceptos) dos indicadores, ou das variáveis latentes no caso da CFA Multigrupo;
- *Assume factors uncorrelated* (assumir fatores não correlacionados): define a correlação entre os fatores para zero;
- *Model Identification* (identificação do modelo): permite selecionar um método para identificação e mensuração do modelo;
  - *Factor variances* (variância do fator): opção padrão no JASP e mais comumente utilizada, fixa a variância dos fatores para 1, de forma que todas as variáveis observadas apresentarão variância variável;
  - *Marker variable* (variável marcadora): a variância do fator será fixada para o valor de sua primeira variável observada que, por sua vez, terá variância fixada em 1;
  - *Effects coding* (codificação de efeitos): a média das cargas fatoriais das variáveis observadas é fixada em 1, saiba mais sobre formas de interpretar a codificação de efeitos em Veen et al. (2018).
- *Residual covariances* (covariância residual): permite que duas variáveis observadas de fatores distintos possam covariar, de forma a reduzir o resíduo (ou erro) devido a variância não explicada pelos fatores, para permitir a covariância residual arraste as variáveis identificadas do quadro esquerdo ao direito em pares.



O menu *Additional Output* permite solicitar que outros resultados sejam apresentados pela análise ou que outros



cálculos estatísticos sejam realizados para analisar o modelo.

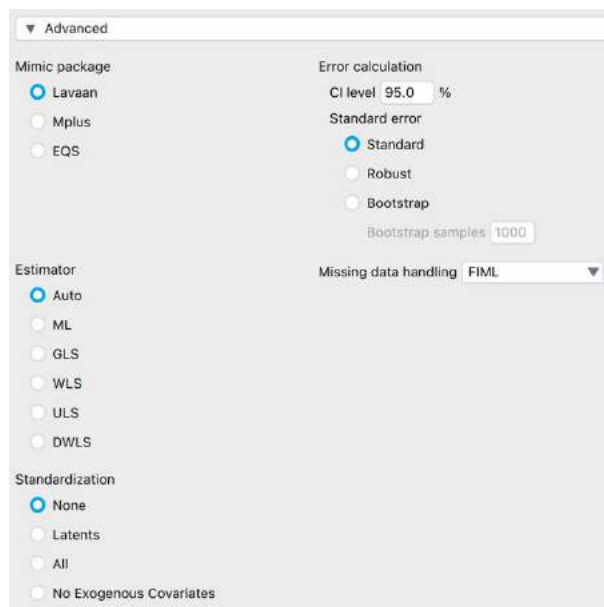
- *Additional fit measures* (medidas de ajuste adicionais): ao selecionar essa opção serão geradas três tabelas com diversos índices de ajuste do modelo;
  - *Fit índices* (índices de ajuste): tabela que apresenta os índices Comparative Fit Index (CFI) e Tucker-Lewis Index (TLI) que são considerados adequados quando superiores a 0,90, dentre outros;
  - *Information criteria* (critério de informação): tabela que apresenta índices úteis para a comparação de diferentes modelos, como o Log-likelihood e seus graus de liberdade (*Number of free parameters*), *Akaike Information Criteria* (AIC) e *Bayesian Information Criteria* (BIC), considera-se um modelo melhor aquele que apresenta esses índices como inferiores;
  - *Other fit measures* (outras medidas de ajuste): tabela que apresenta diversas outras métricas de ajuste, dentre os quais o *Root Mean Square Error of Approximation* (RMSEA), que deve ser inferior a 0,08 ou possuir intervalo de confiança superior menor que 0,1, e o *Standardized Root Mean Square Residual* (SRMR).
- *Kaiser-Meyer-Olkin (KMO) test e Bartlett's test of sphericity* (teste de esfericidade de Bartlett): gera tabela apresentando os resultados dos testes de KMO e Bartlett, mais comumente utilizados na AFE, porém também úteis para verificar a adequação da matriz de correlações/covariância previamente a análise;
- *R-squared* (R quadrado): apresenta a variância explicada ou  $R^2$  por cada variável observada;
- *Implied covariance matrix* (matriz de covariância implícita): retorna essa matriz em formato de tabela;
- *Residual covariance matrix* (matriz de covariância residual): apresenta uma matriz com as covariâncias residuais entre as variáveis observadas após estimação do modelo;
- *Modification indices* (índices de modificação): exhibe os índices de modificação do modelo a partir de um valor de corte (por padrão 3,84, contudo alguns autores recomendam 20 ou 50). O índice de modificação mostra o quanto o valor do qui-quadrado mudaria e o ajuste do modelo

melhoraria a partir da inclusão de determinada relação no modelo, o EPC mostra a mudança esperada no parâmetro;

- *Show lavaan syntax* (mostrar sintaxe do lavaan): exhibe o código necessário para executar esse modelo no lavaan, R.

O menu *Multigroup CFA* (AFC Multigrupo) permite estimar o modelo proposto para diferentes grupos a partir de uma variável de agrupamento (*grouping variable*), assim como compará-los para verificar a invariabilidade da medida (*invariable testing*). As opções de invariabilidade da medida incluem: *configural*, a mais básica, na qual os grupos apresentam a mesma estrutura fatorial; *metric* (métrica), na qual as cargas fatoriais são restritas a serem iguais entre os grupos; *scalar* (escalar), na qual as cargas fatoriais e as médias são restritas a serem iguais entre os grupos; *strict* (estricto): na qual as cargas fatoriais, médias, variâncias residuais e covariâncias residuais são restritas a serem iguais entre os grupos.

O menu *Plots* (gráficos) permite solicitar o gráfico do modelo (*model plot*) e o gráfico de desajuste (*misfit plot*). O gráfico de desajuste apresenta de forma visual a matriz de covariância residual padronizada do modelo. O gráfico do modelo apresenta o gráfico de trajetórias com as variáveis observadas e latentes, assim como suas relações e estimativas para o modelo estimado.



Por fim, o menu *Advanced* (avançado) apresenta configurações importantes para determinar a estimação do modelo proposto. Por meio das seguintes opções:

- *Mimic package* (emular pacotes): permite emular os resultados que seriam apresentados utilizando os softwares mais

comuns, por padrão é selecionado o Lavaan, pacote do R para AFC;

- *Estimator* (estimador): a opção padrão *Auto* utiliza ML para variáveis unicamente escalares e WLS para quando há variáveis ordinais;
  - *ML: Maximum Likelihood* (máxima verossimilhança);
  - *GLS: Generalized Least Squares* (quadrados mínimos generalizados);
  - *WLS: Weighted Least Squares* (quadrados mínimos ponderados);
  - *ULS: Unweighted Least Squares* (quadrados mínimos não ponderados);
  - *DWLS: Diagonally Weighted Least Squares* (quadrados mínimos ponderados em diagonal).
- *Standardization* (padronização): determina qual forma de padronização é apresentada nos resultados, por padrão nenhuma (*None*);
- *Error calculation* (cálculo de erro): permite determinar o intervalo de confiança e forma de cálculo do erro padrão (*standard* ou padrão, *robust* ou robusto e *Bootstrap*, por reamostragem);
- *Missing data handling* (tratamento para dados ausentes): permite alterar a forma de tratamento aos dados ausentes, por padrão utiliza FIML (*Full Information Maximum Likelihood*).

Apresentadas as opções existentes no JASP para a AFC, solicitaremos para este banco de dados um modelo identificado por meio das variâncias dos fatores, tabelas dos índices de ajuste adicionais e de índices de modificação, o gráfico do modelo com apresentação dos parâmetros estimados padronizados, por meio do pacote Lavaan, com estimador automático (ML), padronização em todas as variáveis e erro padrão robusto.

Os primeiros resultados apresentados na tela de Resultados são relativos ao ajuste do modelo. Quatro tabelas são apresentadas, porém iremos focar aqui em apenas alguns dados de cada tabela. O teste do qui-quadrado permite avaliar a adequação dos dados ao modelo proposto. O *Baseline model* (modelo de linha de base) corresponde a um modelo nulo, sem nenhuma relação definida entre as variáveis, enquanto o *Factor model* (modelo fatorial) corresponde ao modelo proposto pela análise. Considera-se que o modelo fatorial apresentou uma melhora com relação ao modelo nulo quando o teste do qui-quadrado é significativo. Para além da significância estatística, o valor do qui-quadrado dividido pelos graus de liberdade

<b>Chi-square test</b>			
<b>Model</b>	<b>X<sup>2</sup></b>	<b>df</b>	<b>p</b>
Baseline model	918.852	36	
Factor model	85.306	24	< .001

( $\chi^2 / df$ ) pode ser utilizado e tem sido considerado melhor. Segundo Kline (2015) uma razão menor que 3 é adequada, já Hair et al. (2009) indicam que seria adequado quando menor que 5. Nesse caso temos a razão de qui-quadrado por graus de liberdade igual a 3,55.

Ainda buscando avaliar o ajuste do modelo, temos: CFI = 0,93; TLI = 0,90; RMSEA = 0,09, IC 90% [0,07; 0,11]; SRMR = 0,06. Focaremos nestes índices por serem os principais e não analisaremos os índices apresentados na tabela de critérios de informação, pois não há nenhum outro modelo a ser comparado. Brown (2006) sugere que o CFI e o TLI acima de 0,90 são considerados adequados, o RMSEA seria adequado abaixo de 0,06 ou com intervalo de confiança superior menor que 0,10 e o SRMR seria adequado quando inferior a 0,05.

A segunda parte dos resultados refere-se as estimativas dos parâmetros calculados. Iremos focar aqui nas cargas fatoriais (*factor loadings*). Nessa tabela são apresentadas as cargas fatoriais (*estimate*) de cada item em seu respectivo fator, assim como seu erro padrão (*std. Error*), significância estatística (*p*), intervalo de confiança da carga fatorial em 95% e a carga fatorial padronizada (*std. Est. (all)*).

#### Factor loadings

Factor	Indicator	Symbol	Estimate	Std. Error	z-value	p	95% Confidence Interval		Std. Est. (all)
							Lower	Upper	
Habilidade Visual	x1	$\lambda_{11}$	0.900	0.100	8.973	< .001	0.703	1.096	0.772
	x2	$\lambda_{12}$	0.498	0.088	5.681	< .001	0.326	0.670	0.424
	x3	$\lambda_{13}$	0.656	0.080	8.151	< .001	0.498	0.814	0.581
Habilidade Textual	x4	$\lambda_{21}$	0.990	0.061	16.150	< .001	0.870	1.110	0.852
	x5	$\lambda_{22}$	1.102	0.055	20.146	< .001	0.994	1.209	0.855
	x6	$\lambda_{23}$	0.917	0.058	15.767	< .001	0.803	1.031	0.838
Velocidade	x7	$\lambda_{31}$	0.619	0.086	7.193	< .001	0.451	0.788	0.570
	x8	$\lambda_{32}$	0.731	0.093	7.875	< .001	0.549	0.913	0.723
	x9	$\lambda_{33}$	0.670	0.099	6.761	< .001	0.476	0.864	0.665

Devido a forma de identificação do modelo selecionada por nós, a variância dos fatores será sempre 1, não havendo necessidade de analisar tais informações. A tabela de

covariância dos fatores (*factor covariances*) nos permite verificar a correlação entre os fatores estimados.

### Factor Covariances

		Estimate	Std. Error	z-value	p	95% Confidence Interval		Std. Est. (all)
						Lower	Upper	
Habilidade Visual	↔ Habilidade Textual	0.459	0.073	6.258	< .001	0.315	0.602	0.459
Habilidade Visual	↔ Velocidade	0.471	0.119	3.954	< .001	0.237	0.704	0.471
Habilidade Textual	↔ Velocidade	0.283	0.085	3.311	< .001	0.115	0.451	0.283

Por fim, os índices de modificação são apresentados em duas tabelas. A primeira, cargas cruzadas (*cross-loadings*) refere-se a variáveis observadas que apresentam alta variância explicada por outro fator. A segunda, covariâncias residuais (*residual covariances*) refere-se à correlação entre duas variáveis observadas que não é explicada por um fator. Esses índices auxiliam a que realizemos mudanças no modelo para que o mesmo se torne matematicamente melhor.

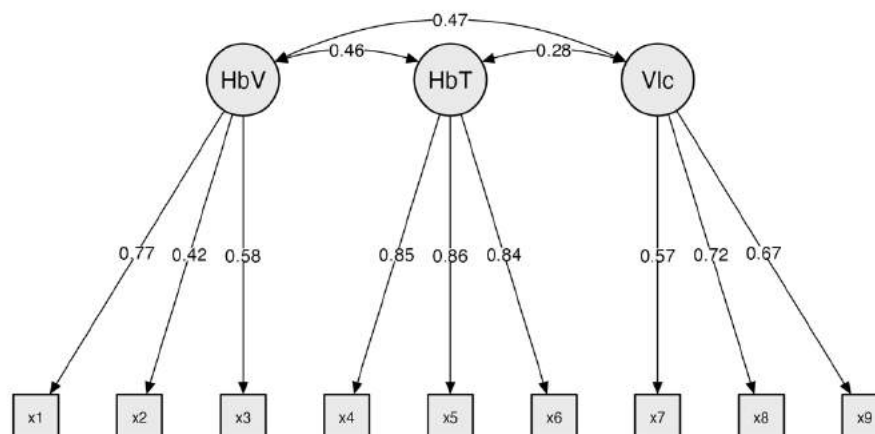
### Cross-loadings

	Mod. Ind.	EPC
Habilidade Visual → x9	36.411	0.519
Habilidade Visual → x7	18.631	-0.380
Habilidade Textual → x3	9.151	-0.269
Habilidade Textual → x1	8.903	0.347
Habilidade Visual → x5	7.441	-0.189
Habilidade Textual → x9	4.796	0.137
Habilidade Visual → x8	4.295	-0.189

### Residual covariances

	Mod. Ind.	EPC
x7 ↔ x8	34.145	0.536
x8 ↔ x9	14.946	-0.423
x2 ↔ x7	8.918	-0.183
x2 ↔ x3	8.532	0.218
x3 ↔ x5	7.858	-0.130
x1 ↔ x9	7.335	0.138
x4 ↔ x6	6.221	-0.235
x4 ↔ x7	5.920	0.098
x1 ↔ x7	5.420	-0.129
x7 ↔ x9	5.183	-0.187
x3 ↔ x9	4.126	0.102

O gráfico do modelo com os parâmetros estimados padronizados torna os resultados visualmente atrativos e pode substituir algumas tabelas.



## RELATO DOS RESULTADOS

Foi realizada uma Análise Fatorial Confirmatória buscando averiguar a adequação dos itens (9) aos três fatores propostos pelo modelo (habilidade visual, habilidade textual e velocidade). Para tal, utilizou-se o estimador de Máxima Verossimilhança e foram avaliados os índices de ajuste do modelo. O modelo fatorial apresentou melhora significativa com relação ao modelo nulo ( $\chi^2 [24] = 85,31$ ;  $p < 0,001$ ;  $\chi^2 / df = 3,55$ ), os índices de ajuste comparativos sugeriram a adequação do modelo (CFI = 0,93; TLI = 0,90) e os índices de ajuste residuais foram limiares (RMSEA = 0,09; 90% IC [0,07; 0,11]; SRMR = 0,06).

Os itens apresentaram cargas fatoriais adequadas em seus respectivos fatores, contudo os resultados identificados nos índices de ajuste residuais podem ser compreendidos por meio dos índices de modificação elevados (acima de 20) que sugerem carga cruzada do item 9 com o fator de Habilidade Visual ( $\Delta\chi^2 = 36,41$ ) e a variância residual compartilhada entre os itens 7 e 8 no fator Velocidade ( $\Delta\chi^2 = 34,15$ ). Por fim, os fatores apresentaram correlações medianas, com exceção do fator Habilidade Textual e Velocidade ( $r = 0,28$ ) que apresentaram correlação fraca.

## REFERÊNCIAS

- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*. Bookman.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford.
- Pasquali, L. (2010). *Instrumentação psicológica: Fundamentos e práticas*. Artmed.
- Veen, D., Little, T. D., & van de Schoot, R. (2018). *Effects coding as unbiased alternative to scale scores* [Project Page]. OSF. <https://osf.io/3z47e/>

