



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

**TAMANHO DE AMOSTRA PARA O
TESTE-RETESTE NA DETERMINAÇÃO
DO COEFICIENTE DE CORRELAÇÃO
INTRACLASSE**

Tiago Bortoletto Lino dos Santos

Uberlândia-MG

2018

Tiago Bortoletto Lino dos Santos

**TAMANHO DE AMOSTRA PARA O
TESTE-RETESTE NA DETERMINAÇÃO
DO COEFICIENTE DE CORRELAÇÃO
INTRACLASSE**

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Rogério de Melo Costa Pinto

**Uberlândia-MG
2018**



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

Rogério de Melo Costa Pinto

Ednaldo Carvalho Guimarães

Marcelo Tavares

**Uberlândia-MG
2018**

RESUMO

O presente trabalho objetiva verificar qual é o tamanho mínimo de amostra necessário para a realização do teste-reteste. Para isso, um código computacional foi desenvolvido e implementado, com o auxílio do software estatístico R, para a realização do cálculo dos coeficientes de correlação intraclass (CCI) para dez tamanhos de amostras. Os dados utilizados foram obtidos pelo Banco de Itens Saúde Global do PROMIS (Patient-Reported Outcomes Measurement Information System), que é composto por 10 itens que avaliam aspectos de saúde física, mental e social, e cuja validação para língua portuguesa e coleta de respostas foi realizada em Zumpano et al. (2015). Como resultado, os coeficientes de correlação intraclass obtidos entre o teste e o reteste, por tamanho de amostra, apresentaram valores acima de 0,90 e baixa variabilidade, o que confirma a confiabilidade do teste. Além disso, ao comparar os resultados para os diferentes tamanhos de amostra, conclui-se que o tamanho mínimo de amostra necessário para a realização do teste-reteste é $n = 20$, pois a partir desse tamanho de amostra as médias dos coeficientes de correlação intraclass não apresentaram diferença significativa.

Palavras-chave: Confiabilidade, instrumento de medida, coeficiente de correlação intraclass.

ABSTRACT

The present study aims to verify which minimum sample size is required to perform the test-retest. For this, a computational code was developed and implemented, with the aid of statistical software R, for the intraclass correlation coefficient (ICC) realization for ten sample sizes. The data were obtained by PROMIS (Global Health Items Bank), which is composed of 10 items that assess aspects of physical, mental and social health, and whose validation for Portuguese language and data collection was performed in Zumpano et al. (2015). As a result, the intraclass correlation coefficient obtained between the test-retest, by sample size, presented values above 0.90 and low variability which confirms the reliability of the test. In addition, when comparing the results for the different sample sizes, it is concluded that the sample size required for the test-retest is $n = 20$, because from this sample size the mean of the intraclass correlation coefficients did not present a significant difference.

Keywords: Reliability, measurement instrument, intraclass correlation coefficient.

SUMÁRIO

Lista de Figuras	I
Lista de Tabelas	III
1 Introdução	1
2 Objetivos	5
2.1 Objetivo Geral	5
2.2 Objetivos Específicos	5
3 Material e Método	7
3.1 Coeficiente de correlação intraclasse	8
4 Resultados e Discussão	11
5 Conclusão	17
Referências Bibliográficas	19
Apêndice A Código do R	21

LISTA DE FIGURAS

4.1	Gráfico das médias das correlações entre teste e reteste.	12
4.2	Gráfico das médias das correlações do teste e reteste separadamente.	14

LISTA DE TABELAS

4.1	Cálculo das médias das correlações por tamanho de amostra.	11
4.2	Cálculo das variâncias das correlações por tamanho de amostra.	13
4.3	Análise de variância (ANOVA) referente ao teste.	13
4.4	Teste de Tukey para comparação das amostras referente ao teste.	13
4.5	Análise de variância (ANOVA) referente ao reteste.	13
4.6	Teste de Tukey para comparação das amostras referente ao reteste.	14
4.7	Análise de variância (ANOVA) dos coeficientes calculados entre o teste e reteste.	15
4.8	Teste de Tukey para comparação dos coeficientes calculados entre o teste e reteste.	15

1. INTRODUÇÃO

Medida é uma síntese da qualidade e da quantidade. Medir é determinar, tendo por base uma escala fixa, um padrão de unidade, e uma grandeza. Para medir, avaliar ou quantificar informações, o pesquisador precisa se atentar para os critérios de exigências de medidas provenientes de testes, instrumentos de coleta de dados, e técnicas de aferição, para que se possa aceitá-los como geradores de boas avaliações [10].

De acordo com Martins (2005), toda medida deve reunir dois requisitos essenciais: confiabilidade (ou fidedignidade) e validade. Medidas confiáveis são replicáveis e consistentes, isto é, geram os mesmos resultados em sucessivas medições. Já medidas válidas são representações precisas da característica que se pretende medir.

A confiabilidade de um instrumento para coleta de dados, teste, técnica de aferição, se trata da sua coerência, determinada através da constância dos resultados. Em outras palavras, a confiabilidade de uma medida é a confiança que a mesma inspira [10]. Segundo Cozby (2003), a confiabilidade de um instrumento de medição se refere ao grau em que sua repetida aplicação, ao mesmo sujeito ou objeto, produz resultados iguais.

Por exemplo, ao se medir de forma constante a temperatura de uma sala climatizada, o termômetro que apresentar resultados diferentes em cada medição deve ser considerado não confiável, pois, sabemos que nessas condições, não há motivo para mudanças de temperatura. Se ocorrerem resultados alterados o instrumento de medidas não terá a característica de fidedignidade e seus resultados não serão confiáveis.

De maneira ampla, uma medida fidedigna é consistente e precisa porque fornece uma medida estável da variável. Em outras palavras, confiabilidade refere-se à consistência ou estabilidade de uma medida. Um exemplo corriqueiro pode nos ajudar a compreender ainda mais este conceito: diz-se que se tem um relógio confiável quando o instrumento nos fornece o tempo preciso, raramente adiantado ou atrasado [10].

De acordo com Martins (2005), o desvio padrão (medida de dispersão em torno da média) pode ser um indicador do grau de confiabilidade de um instrumento de medidas, de modo que quanto menor o valor do desvio padrão maior será o grau de confiabilidade do instrumento de medidas. Além dessa maneira, a confiabilidade de um instrumento de medidas pode ser determinada mediante diversas técnicas, sendo a técnica do *teste reteste* um dos procedimentos mais conhecidos.

Na técnica do teste reteste, o instrumento de medidas é aplicado duas vezes a um mesmo grupo de pessoas ou objetos, depois de um período de tempo entre as aplicações. Se a correlação

entre os resultados das duas aplicações é fortemente positiva o instrumento pode ser considerado confiável.

O período de tempo entre as medições é um fator que se deve considerar quando esta técnica é aplicada. Períodos longos são suscetíveis às mudanças que podem comprometer a interpretação do coeficiente de confiabilidade (ou coeficiente de estabilidade) obtido. Um tempo longo demais favorece a aquisição de novas aprendizagens e pode provocar uma sub avaliação da estabilidade [10].

O conceito de sub avaliação da estabilidade poderá ser melhor explicado através de um exemplo: vamos supor que foi aplicado um questionário com a seguinte pergunta: o que você prefere como sobremesa? Com as seguintes alternativas: (1) Sorvete, (2) Torta de morango e (3) Não sei. Na primeira aplicação o entrevistado marcou a alternativa (3). Porém, as alternativas despertaram o entrevistado quanto às possibilidades de sobremesa. Depois de um longo tempo foi aplicado o reteste e ao deparar com a mesma questão a pessoa escolheu a alternativa (1). O pesquisador ao comparar as respostas pode ser induzido a afirmar que o instrumento de medidas não tem estabilidade, mas na verdade, trata-se de uma mudança real da pessoa.

Já se o período de tempo entre as medições é curto, os resultados podem ser contaminados pelo efeito memória. O intervalo curto entre a aplicação de um teste e reteste provoca um efeito conhecido como superavaliação da estabilidade. Esse efeito pode ser provocado pela lembrança das respostas que o indivíduo deu no primeiro teste e depois, simplesmente repete as respostas recordadas no reteste, ou seja, não são respostas espontâneas ou inteiramente pensadas [10].

Logo, o pesquisador estará mais seguro com uma interpretação dos efeitos da sub avaliação do que com a superavaliação, pois, no primeiro efeito o pesquisador concluirá que há necessidade de mais estudo sobre a medida em questão, enquanto o segundo efeito pode dar falsa segurança de estabilidade e se chegar a conclusões inválidas.

Conforme explicam Camines e Zeller (1979), para avaliar a confiabilidade pelo teste e reteste precisamos obter dois escores (medidas) de cada um de muitos indivíduos, ou objetos. Este critério de avaliação da confiabilidade só poderá ser aplicado quando o nível de mensuração da variável é intervalar. Não é comum se ter duas medidas de uma variável para os mesmos indivíduos ou objetos, fato que limita a aplicação deste critério.

Dessa forma, quando a variável sob análise apresentar nível intervalar de mensuração, pode-se calcular o coeficiente de correlação intraclasse, cujo conceito foi induzido por Fisher (1925). O coeficiente de correlação intraclasse (CCI) é umas das ferramentas estatísticas mais utilizadas para a mensuração da confiabilidade de medidas. O CCI é adequado para mensurar a homogeneidade de duas ou mais medidas e é interpretado como a proporção de variabilidade total atribuída ao objeto medido [15].

O coeficiente de correlação intraclasse corresponde a uma medida do grau de homogeneidade existente entre os elementos que pertencem a determinados grupos. Quando as unidades elementares dentro dos grupos são homogêneas, ou seja, quando são bastante similares de acordo com uma determinada característica, a correlação intraclasse é próxima de 1. Inversamente, se as unidades elementares nos grupos forem relativamente heterogêneas em relação a uma carac-

terística, o coeficiente de correlação intraclasse será próximo de zero positivamente ou no caso mais extremo negativo [9].

O CCI é utilizado em diversas áreas do conhecimento e teve suas primeiras implementações para medir a homogeneidade da herdabilidade [5]. Na área da saúde, por exemplo, existem estudos que utilizam o CCI para fazer a mensuração de homogeneidade genética [12], para medir a precisão de instrumentos de medida [6], para verificar a associação entre níveis plasmáticos de carotenóides e tipos diferentes de dieta [7], para medir a falta de concordância entre médicos na hora de avaliar os custos e a qualidade dos cuidados dos pacientes com doenças crônicas [17], entre outros.

2. OBJETIVOS

2.1 OBJETIVO GERAL

Com o presente trabalho objetiva-se verificar qual o tamanho mínimo de amostra necessário para a realização do teste-reteste.

2.2 OBJETIVOS ESPECÍFICOS

- Desenvolver e implementar um código computacional para a realização do cálculo do coeficiente de correlação intraclass (CCI) para cada tamanho de amostra, com o auxílio do software estatístico R.
- Comparar os resultados dos coeficientes de correlação intraclass obtidos para diferentes tamanhos de amostra.

3. MATERIAL E MÉTODO

No presente trabalho, foi utilizado o Banco de Itens Saúde Global do PROMIS (*Patient-Reported Outcomes Measurement Information System*), banco este cuja validação para língua portuguesa e coleta de respostas foi realizada em Zumpano et al. (2015). O banco de itens em questão é composto por 10 itens que avaliam aspectos de saúde física, mental e social, além de um item para classificação da saúde geral. Este instrumento é composto por duas escalas: Saúde Física Global e Saúde Mental Global, cada uma composta por quatro itens.

A escala Saúde Física Global inclui os itens de saúde física (Global 03), função física (Global 06), intensidade da dor (Global 07) e fadiga (Global 08). Já a escala Saúde Mental Global é constituída pelos itens de qualidade de vida (Global 02), saúde mental (Global 04), satisfação com atividades sociais e relacionamentos (Global 05), e problemas emocionais (Global 10). Dois itens, saúde geral (Global 01) e papéis sociais (Global 09), não são utilizados para a pontuação das escalas de Saúde Física Global e Saúde Mental Global.

Para a análise da confiabilidade teste-reteste, 100 participantes completaram o Banco de Itens Saúde Global após 14 dias da primeira aplicação. Este banco de itens possui como categorias de respostas uma escala do tipo Likert de 5 pontos, de modo que quanto maior a pontuação, melhor a saúde global do respondente. O banco de dados também é composto com a soma das pontuações dadas por cada indivíduo para todos os dez itens ou questões.

O banco de dados foi respondido por indivíduos com doenças crônicas ou saudáveis, com idade entre 18 e 60 anos, de ambos os sexos, diferentes raças e níveis de escolaridade, e com habilidade de escrita e leitura da língua portuguesa. Os participantes foram abordados, consecutivamente, no ambulatório de diversas especialidades médicas do Hospital de Clínicas da Universidade Federal de Uberlândia, Brasil.

A confiabilidade teste-reteste do banco de itens foi analisada por meio do cálculo do coeficiente de correlação intraclass (CCI) para dez tamanhos de amostra ($n = 10, 20, \dots, 100$). Para cada n , o CCI foi calculado 100 vezes para n amostras distintas escolhidas aleatoriamente do banco de dados, sem reposição. Dessa forma, o valor CCI utilizado para a comparação dos tamanhos de amostra foi obtido por meio da média dos coeficientes calculados em cada repetição. A variabilidade do CCI para cada n foi obtida de forma análoga.

3.1 COEFICIENTE DE CORRELAÇÃO INTRACLASSE

Para um grupo de medidas, a variância total dos dados (σ_T^2) é dada a partir da variância verdadeira (σ_t^2) e a variação do erro (σ_e^2). Da mesma forma, cada medida observada é composta pelo seu verdadeiro valor e o erro [14]. Na teoria, a verdadeira medida ou observação de um indivíduo reflete a média de um número infinito de medidas de um sujeito, enquanto que o erro é igual à diferença entre o verdadeiro valor da medida e a medida observada [4].

A origem do erro pode ser devida a variabilidade biológica, instrumentação, erro pelo sujeito e erro pelo examinador [19]. Se calcularmos a razão entre (σ_t^2) e (σ_T^2) das medidas observadas, onde (σ_T^2) é igual a soma de (σ_t^2) e (σ_e^2), é possível obter o seguinte coeficiente de confiabilidade:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}. \quad (3.1)$$

Quanto mais próxima essa razão é de 1,0, mais alta é a confiabilidade e menor é o erro (σ_e^2). Como o valor verdadeiro de cada observação não é conhecido, um índice de σ_t^2 é usado com base na variabilidade entre os indivíduos, ou seja, a variação devido a como os sujeitos diferem um dos outros. No contexto, confiabilidade (consistência relativa) é formalmente definida como [2]:

$$\text{Confiabilidade} = \frac{\text{Variabilidade entre os sujeitos}}{\text{Variabilidade entre os sujeitos} + \text{Erro}}. \quad (3.2)$$

O coeficiente de confiabilidade dado pela Equação 3.2 é mensurado pelo CCI. Embora a confiabilidade é conceituada com termos como reprodutibilidade, repetibilidade e concordância, ela é definida como acima. As estimativas das variâncias necessárias para o cálculo são obtidas por meio da análise de variância (ANOVA). Especificamente, os CCIs podem ser calculados a partir dos valores dos quadrados médios entre os indivíduos, considerando a ANOVA com um único fator (isto é, ANOVA com medidas repetidas).

O CCI é uma medida relativa de confiabilidade [3], em que a razão de variâncias é derivada da ANOVA, não apresenta unidade específica, e é conceitualmente mais semelhante ao R^2 da regressão [13] do que do coeficiente R de Pearson. O CCI pode teoricamente variar entre 0 e 1,0, onde um CCI de 0 indica a não confiabilidade, enquanto que um CCI de 1,0 indica perfeita confiabilidade.

Na prática, CCIs podem variar além do intervalo de 0 a 1,0 [8], embora com dados reais isso seja raro. A relativa natureza do CCI é refletida no fato de que a magnitude de um CCI depende na variabilidade entre os indivíduos. Isto é, se os sujeitos diferem pouco um dos outros, os valores CCI são pequenos, enquanto que se a diferença entre os sujeitos é alta, os CCIs podem apresentar valores grandes.

Segundo Streiner e Norman (1995), um teste não qualificado literalmente não apresenta

confiabilidade; e o coeficiente tem significado apenas quando aplicado em específicas populações. Além disso, é intuitivo que pequenas diferenças entre indivíduos são mais difíceis de serem detectadas do que grandes diferenças, e o CCI se reflete nisso [16].

Erros são tipicamente classificados em dois tipos: erro sistemático (exemplo, viés) e erro aleatório (imprecisão) [1], de modo que o erro total reflete em ambos os erros. O erro aleatório se refere a fontes de erro que são devido a mudança de fatores como sorte, estado de alerta, atenção do examinador e variabilidade normal biológica. Ambos os erros aumentam e diminuem os resultados dos testes com repetição. Assim, podemos expandir Equação 1 da seguinte maneira:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{es}^2 + \sigma_{ea}^2}, \quad (3.3)$$

onde σ_{es} é o erro sistemático e o σ_{ea} indica o erro aleatório.

O CCI pode ser calculado da seguinte forma:

$$ICC = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}, \quad (3.4)$$

de modo que σ_s^2 é a variância estimada entre os sujeitos ou indivíduos.

Interpretar o CCI é bastante simples, de modo que representa a proporção da variação de um conjunto de medidas, atribuíveis ao σ_t^2 . Um CCI de 0,95 significa que cerca de 95% da variação das medidas observadas são devido ao σ_t^2 . Já o restante da variação ($1 - \text{CCI} = 5\%$) é atribuída ao erro [18].

4. RESULTADOS E DISCUSSÃO

As médias dos coeficientes de correlação intraclassa por tamanho de amostra estão representadas na Tabela 4.1. Normalmente, valores de CCI acima de 0,75 são considerados adequados [11]. Dessa forma pode-se dizer que, de modo geral, as correlações se apresentaram satisfatórias.

No caso dos coeficientes obtidos para o teste e reteste separadamente, os resultados indicaram homogeneidade dentro dos grupos, ou seja, os grupos apresentaram bastante similaridade de acordo com determinadas características. Os resultados também apresentaram altos valores para os coeficientes de correlação intraclassa calculados entre o teste e o reteste, sendo todos os valores próximos de 1,0, o que indica uma boa confiabilidade do teste (Figura 4.1).

Tabela 4.1: *Cálculo das médias das correlações por tamanho de amostra.*

n	Teste	Reteste	Entre teste e reteste
10	0,7030	0,7729	0,9466
20	0,7422	0,7974	0,9473
30	0,7541	0,8100	0,9500
40	0,7552	0,8083	0,9521
50	0,7625	0,8162	0,9531
60	0,7703	0,8194	0,9527
70	0,7666	0,8176	0,9527
80	0,7625	0,8144	0,9526
90	0,7659	0,8164	0,9536
100	0,7685	0,8190	0,9530

Já na Tabela 4.2 estão representadas as variâncias dos coeficientes de correlação intraclassa por tamanho de amostra. Em geral, as correlações apresentaram baixa variabilidade, de modo que quanto maior o tamanho da amostra, menor a variação do CCI.

De acordo com o coeficiente de variação, as correlações obtidas apenas para o teste apresentaram maior variabilidade ($CV = 10,62\%$), seguido das correlações obtidas apenas para o reteste ($CV = 6,87\%$). As correlações entre o teste e reteste apresentaram menor variabilidade quando comparada com as demais ($CV = 1,57\%$).

A análise de variância (ANOVA) foi realizada com o intuito de verificar se existe diferença significativa entre as médias dos coeficientes de correlação intraclassa por tamanho de amostra. De acordo com a Tabela 4.3, em relação as correlações obtidas apenas para o teste, pode-se afirmar que pelo menos uma das médias das correlações é diferente das demais.

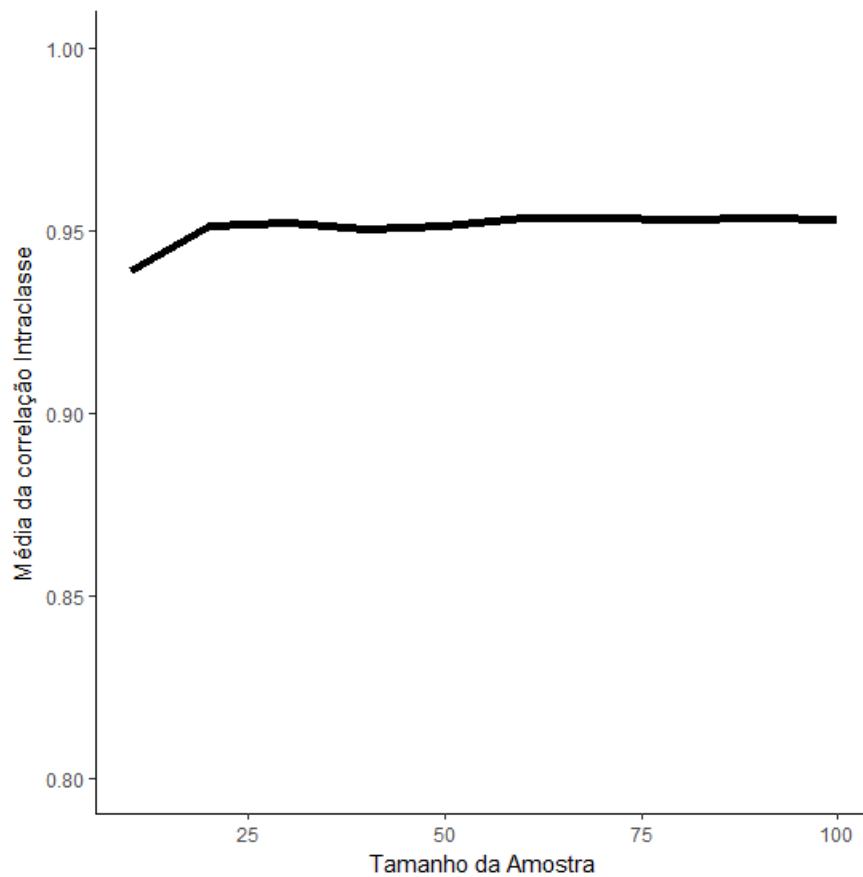


Figura 4.1: Gráfico das médias das correlações entre teste e reteste.

De acordo com o teste de Tukey (Tabela 4.4), tem-se que apenas a média da correlação da amostra de tamanho $n = 10$ apresentou diferença significativa em relação as correlações dos demais tamanhos de amostra. Dessa forma, pode-se dizer que a partir de uma amostra de tamanho $n = 20$ já é possível obter um resultado satisfatório para o CCI.

Quando se analisa os resultados da ANOVA para o reteste separadamente, pode-se concluir que existe diferença significativa entre as médias dos coeficientes de correlação intraclassa por tamanho de amostra (Tabela 4.5).

O teste de Tukey referente apenas ao reteste está representado na Tabela 4.6. Por meio dos resultados obtidos, é possível ver que a média da correlação da amostra $n = 10$ não apresenta diferença significativa em relação a média da correlação da amostra $n = 20$, mas se diferenciou significativamente dos demais tamanhos de amostra.

Além disso, verifica-se que a partir da amostra $n = 20$, as médias dos coeficientes de correlação intraclassa não apresentaram diferença significativa. Logo, uma amostra de tamanho $n = 20$ já pode ser considerada suficiente para obter um resultado de CCI adequado.

A Figura 4.2 ilustra as médias dos coeficientes de correlação intraclassa do teste e reteste separadamente. Por meio do gráfico, observa-se que as correlações referentes ao reteste apresentam valores maiores em comparação com as correlações obtidas para o teste. Além disso, é possível ver que os valores das correlações, tanto para o teste quanto para o reteste, vão crescendo a medida que o tamanho de amostra aumenta, e em determinado ponto apresentam certa estabilidade.

Tabela 4.2: *Cálculo das variâncias das correlações por tamanho de amostra.*

n	Teste	Reteste	Entre teste e reteste
10	0,0381	0,0171	0,0013
20	0,0100	0,0078	0,0004
30	0,0082	0,0025	0,0002
40	0,0028	0,0013	0,0001
50	0,0026	0,0009	0,0001
60	0,0011	0,0007	0,0000
70	0,0009	0,0003	0,0000
80	0,0005	0,0003	0,0000
90	0,0002	0,0001	0,0000
100	0,0000	0,0000	0,0000

Tabela 4.3: *Análise de variância (ANOVA) referente ao teste.*

	GL	SQ	QM	Fc	Pr>Fc
Tratamento	9	0,3651	0,0406	6,3071	<0,0001
Resíduo	990	6,3684	0,0064		
Total	999	6,7336			

Tabela 4.4: *Teste de Tukey para comparação das amostras referente ao teste.*

Grupos	Tratamentos	Médias
a	60	0,7703
a	100	0,7685
a	70	0,7666
a	90	0,7659
a	80	0,7625
a	50	0,7625
a	40	0,7552
a	30	0,7541
a	20	0,7422
b	10	0,7030

Tabela 4.5: *Análise de variância (ANOVA) referente ao reteste.*

	GL	SQ	QM	Fc	Pr>Fc
Tratamento	9	0,1841	0,0205	6,6121	<0,0001
Resíduo	990	3,0622	0,0031		
Total	999	3,2463			

Tabela 4.6: *Teste de Tukey para comparação das amostras referente ao reteste.*

Grupos	Tratamentos	Médias
a	60	0,8194
a	100	0,8182
a	70	0,8176
a	90	0,8164
a	50	0,8162
a	80	0,8144
a	30	0,8100
a	40	0,8083
ab	20	0,7974
b	10	0,7729

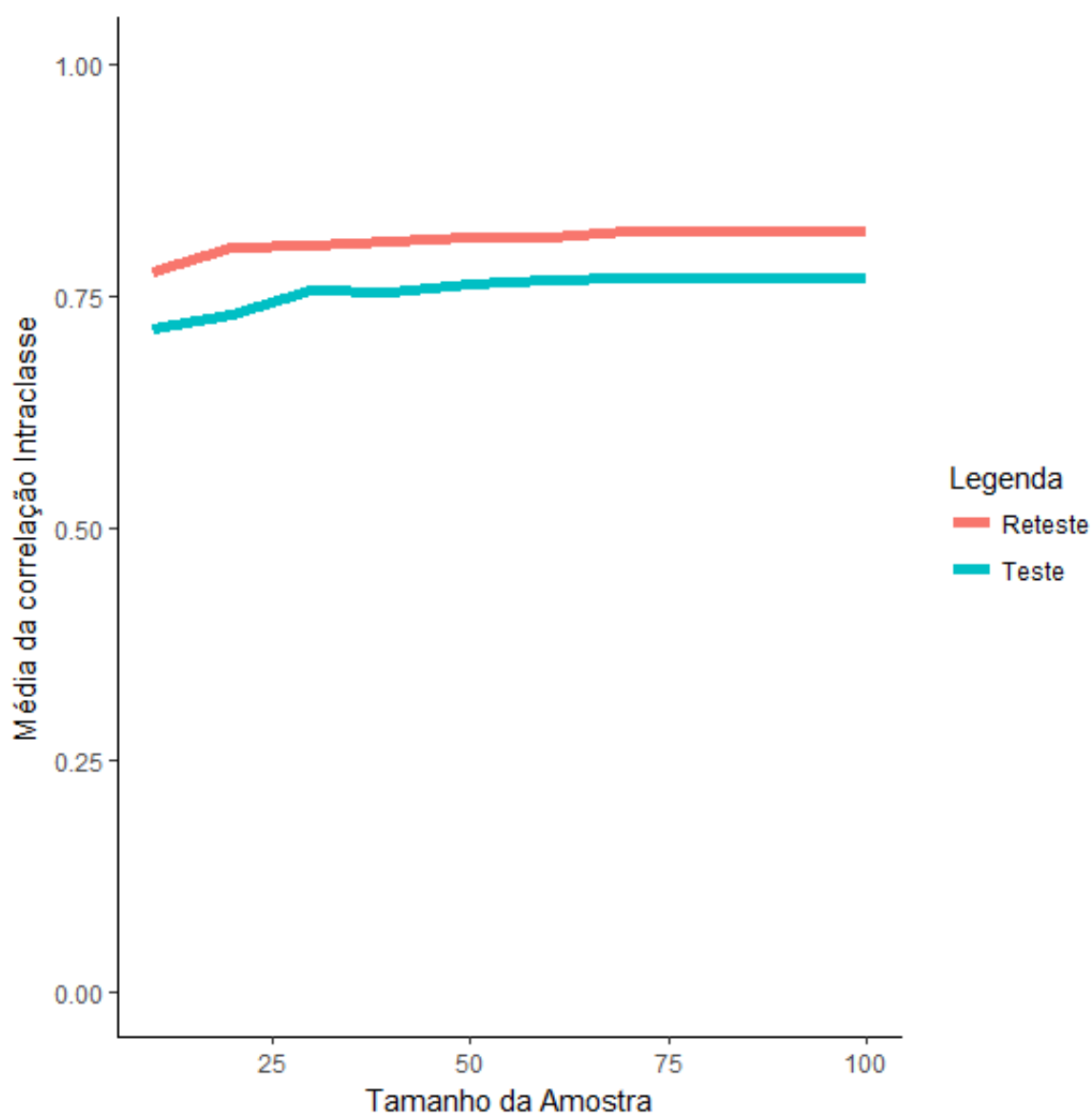


Figura 4.2: Gráfico das médias das correlações do teste e reteste separadamente.

Os resultados da ANOVA para as médias dos coeficientes de correlação intraclass entre o teste e o reteste também indicam a presença de diferença significativa entre as correlações para diferentes tamanhos de amostra.

Tabela 4.7: *Análise de variância (ANOVA) dos coeficientes calculados entre o teste e reteste.*

	GL	SQ	QM	Fc	Pr>Fc
Tratamento	9	0,0058	0,0006	2,8889	0,0022
Resíduo	990	0,2208	0,0002		
Total	999	0,2266			

Para os valores do CCI entre o teste e o reteste, os resultados do teste de Tukey, apresentados na Tabela 4.8, mostram que apenas as médias das correlações dos tamanhos de amostra $n = 10$ e $n = 90$ apresentaram diferença significativa entre si. O restante dos grupos não se diferenciam significativamente.

Tabela 4.8: *Teste de Tukey para comparação dos coeficientes calculados entre o teste e reteste.*

Grupos	Tratamentos	Médias
a	90	0,9536
ab	50	0,9531
ab	100	0,9530
ab	70	0,9527
ab	60	0,9527
ab	80	0,9526
ab	40	0,9521
ab	30	0,9500
ab	20	0,9473
b	10	0,9466

5. CONCLUSÃO

Os coeficientes de correlação intraclassa obtidos entre o teste e o reteste, por tamanho de amostra, para este questionário específico, resultaram em valores maiores que 0,90 e baixa variabilidade, o que confirma a confiabilidade do teste.

Além disso, por meio dos resultados obtidos pode-se concluir que o tamanho mínimo de amostra necessário para a realização do teste-reteste é $n = 20$, pois a partir desse tamanho de amostra as médias dos coeficientes de correlação intraclassa não apresentaram diferença significativa. Logo, uma amostra de tamanho $n = 20$ já pode ser considerada suficiente para obter um resultado de CCI adequado.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Atkinson, D. B. e Nevill, A. M.: *Statistical methods for assessing measurement error (reliability) in variables relevant to Sports Medicine*. Sports Med, 1998.
- [2] Baumgartner, T. A.: *Estimating reliability when all test trials are administered on the same day*. Res. Q., (40):222–225, 1969.
- [3] Chinn, S. e Burney, P. G.: *On measuring repeatability of data from self-administered questionnaires*. Int. J. Epidemiol, (16):121–127, 1987.
- [4] Feldt, L. S. e McKee, M. E.: *Estimation of the reliability of skill tests*. Res. Q., 1(29):279–293, 1958.
- [5] Fisher, R. A.: *Statistical methods for research workers*. Hafner Press, New York, 1ª ed., 1925.
- [6] Hade, E. M., Strickland, P. A. O. e et al.: *Intraclass correlation estimates for cancer screening outcomes: Estimates and applications in the design of group-randomized cancer screening studies*. Journal of the National Cancer Institute Monographs, 1(40):97–103, 2010.
- [7] Hofer, T. P., Hayward, R. A., Greenfield, S., Wagner, E. H., Kaplan, S. H. e Manning, W. G.: *The unreliability of individual physician “report cards” for assessing the costs and quality of care of a chronic disease*. American Medical Association, 22(281):2098–2105, 1999.
- [8] Lahey, M. A., Downey, R. G. e Saal, F. E.: *Intraclass correlations: there’s more than meets the eye*. Psychol. Bull., (93):586–595, 1983.
- [9] Lila, M. F. e Freitas, M. P. S.: *Uma Medida de Homogeneidade da Migração nos Municípios Brasileiros*. 2006.
- [10] Martins, G. A.: *Sobre Confiabilidade e Validade*. RBGN, 20(8):1–12, 2006.
- [11] Nunnally, J. C. e Bernstein, I. H.: *Psychometric theory*. McGraw-Hill, New York, 1ª ed., 1994.
- [12] Robert, C., Foulley, J. L. e Ducroq, V.: *Genetic variation of traits measured in several environments. I. Estimation and testing of homogeneous genetic and intra-class correlations between environments*. Genet. Sel. Evol., (27):111–123, 1995.

- [13] Rousson, V., Gasser, T. e Seifert, B.: *Assessing intrarater, interrater, and test-retest reliability of continuous measurements*. Stat. Med., (21):3431–3446, 2002.
- [14] Safrit, M. J. E.: *Reliability Theory*. American Alliance for Health, Physical Education and Recreation, Washington, 1ª ed., 1976.
- [15] Shrout, P. E. e Fleiss, J. L.: *Intraclass correlations: Uses in assessing reliability*. Psychological Bulletin, 2(86):420–428, 1979.
- [16] Streiner, D. L. e Norman, G. R.: *Measurement Scales: A practical guide to their development and use*. Oxford University Press, Oxford, 1ª ed., 1995.
- [17] Tiao, G. C. e Tan, W. Y.: *Bayesian analysis of random-effect models in the analysis of variance - I: Posterior distribution of variance components*. Biometrika, (52):37–53, 1965.
- [18] Traub, R. E. e Rowley, G. L.: *Understanding reliability*. Educ. Meas. Issues Pract., (10):37–45, 1991.
- [19] Weir, J. P.: *Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM*. Journal of Strength and Conditioning Research, 1(19):231–240, 2005.

A. CÓDIGO DO R

Segue os comandos implementados no software estatístico R para os cálculos dos coeficientes de correlação intraclasse por tamanho de amostra:

```
library("dplyr")
library("psych")
library("tidyr")
library("ExpDes.pt")
library("irr")
library("ggplot2")

require("dplyr")
require("psych")
require("tidyr")
require("ExpDes.pt")
require("irr")
require("ggplot2")

dados_teste=data.frame(read.csv(file.choose(),sep=";"),grupo="TESTE")
dados_reteste=data.frame(read.csv(file.choose(),sep=";"),grupo="RETESTE")

matriz_final=data.frame()

for(i in seq(from=10,to=100,by=10)){

tam_amostra=i

for(j in 1:100){

repeticao=j

amostra_teste=dados_teste[sample(nrow(dados_teste),i,replace=FALSE),]
amostra_teste=amostra_teste[order(as.numeric(row.names(amostra_teste))),]
amostra_reteste=dados_reteste[c(dados_reteste$"IDENTIFICACAO"%in%as.character(amostra_
```

```

teste$"IDENTIFICACAO"))],]

#####
# CÁLCULO DO CCI
#####

icc_t=icc(amostra_teste[,2:11],unit="average")$value
icc_r=icc(amostra_reteste[,2:11],unit="average")$value

matriz_aux=data.frame(tam_amostra,repeticao,icc_t,icc_r)
matriz_final=rbind(matriz_final,matriz_aux)

}}

#####
# CÁLCULO DAS MÉDIAS DOS COEFICIENTES POR TAMANHO DE AMOSTRA
#####

x=gather(matriz_final,key="tipo",value="vl_teste",c("icc_t","icc_r"))

matriz_final%>%group_by(tam_amostra)%>%summarise(media_teste=mean(icc_t),media_
reteste=mean(icc_r))
matriz_final%>%group_by(tam_amostra)%>%summarise(var_teste=var(icc_t),
var_reteste=var(icc_r))
dic(matriz_final$tam_amostra,matriz_final$icc_t)
dic(matriz_final$tam_amostra,matriz_final$icc_r)

df=x%>%group_by(tam_amostra,tipo)%>%summarise(media=mean(vl_teste,na.rm=TRUE))
ggplot(df,aes(x=tam_amostra,y=media,color=tipo))+geom_line(size=2)+theme_
classic()+ylab("Média da correlação Intraclass")+
xlab("Tamanho da Amostra")+ylim(c(0,1))+
scale_colour_discrete("Legenda",breaks=c("icc_r","icc_t"),
labels=c("Reteste","Teste"))

#####
# CÁLCULO DO CCI ENTRE O TESTE E O RETESTE
#####

```

```
matriz_final_t=data.frame()

for(i in seq(from=10,to=100,by=10)){

tam_amostra=i

for(j in 1:100){

repeticao=j

amostra_teste=dados_teste[sample(nrow(dados_teste),i,replace=FALSE),]
amostra_teste=amostra_teste[order(as.numeric(row.names(amostra_teste))),]
amostra_reteste=dados_reteste[c(dados_reteste$"IDENTIFICACAO"%in%as.character(amostra_
teste$"IDENTIFICACAO")),]

totais=data.frame(total_teste=amostra_teste[,12],total_reteste=amostra_reteste[,12])
icc_total=icc(totais[,c(1,2)],unit="average")$value

matriz_aux=data.frame(tam_amostra,repeticao,icc_total)
matriz_final_t=rbind(matriz_final_t,matriz_aux)

}}

matriz_final_t=matriz_final_t[complete.cases(matriz_final_t),]
y=gather(matriz_final_t,key="tipo",value="vl_teste",c("icc_total"))
matriz_final_t%>%group_by(tam_amostra)%>%summarise(media=mean(icc_total))
matriz_final_t%>%group_by(tam_amostra)%>%summarise(var_teste=var(icc_total))
dic(matriz_final_t$tam_amostra,matriz_final_t$icc_total)

dft=y%>%group_by(tam_amostra)%>%summarise(media=mean(vl_teste,na.rm=TRUE))
ggplot(dft,aes(x=tam_amostra,y=media))+geom_line(size=2)+theme_classic()+
ylab("Média da correlação Intraclass")+xlab("Tamanho da Amostra")+
ylim(c(0.8,1))+guides(fill=FALSE)
```